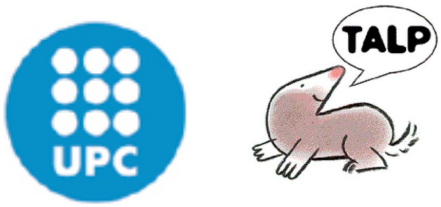
Ph.D. Dissertation

NEW STATISTICAL AND SYNTACTIC MODELS FOR MACHINE TRANSLATION

Maxim Khalilov

Thesis advisor: **J. Adrián Rodríguez Fonollosa**



TALP Research Center, Speech Processing Group

Department of Signal Theory and Communications

Universitat Politècnica de Catalunya

Barcelona, October 2009

We can only see a short distance ahead, but we can see plenty there that needs to be done.

Alan Turing

Abstract

Significant improvements have been achieved in machine translation (MT) over the past few years, mostly motivated by the appearance of statistical machine translation (SMT) technology, which is currently considered the best way to perform MT of natural languages.

The main goal of this thesis is to enhance the classical SMT models, introducing syntactical knowledge in the pre-translation step by reordering the source side of the corpus. To a great extent, our interest is in the value of syntax in reordering for languages with high word order disparity. A secondary objective consists of determining the potential of different language model (LM) enhancement techniques in order to improve the performance and efficiency of SMT systems.

We start with a comprehensive study of the SMT state-of-the-art, describing the fundamental models underlying the translation process, along with a brief description of the main methods of automatic evaluation of translation quality. We emphasize phrase-based and N-gram-based SMT, analyzing the major differences between these two approaches.

Subsequently, we concentrate on language modeling methods that have not received much attention in the SMT community. We report on experiments in applying N-gram-based SMT system adaptation to a speech transcription task, describe a positive impact of accurate cut-off threshold selection both on the model size and LM noisiness, and finally present a continuous-space LM, estimated in the form of an artificial neural network.

Moreover, we propose a novel syntax-based approach to handle the fundamental problem of word ordering for SMT exploiting syntactic representations of source and target texts. The idea of augmenting SMT by using a syntax-based reordering step prior to translation, proposed in recent years, has been quite successful in improving translation quality, especially for translation between languages with high word order disparity.

We provide the reader with a thorough study of the state-of-the-art reordering techniques and introduce a new classification of reordering algorithms for SMT. We then propose a

new non-deterministic reordering strategy based on a syntactically augmented alignment of source and target texts and automatically extracted hierarchical reordering patterns. In the next step, we couple the novel reordering module with decoding in a deterministic way; our goal in this is to effectively tackle both global and local reordering dependencies. Finally, we propose a novel translation units blending scheme, combining bilingual tuples extracted from the parallel corpora with monotone and reordered source parts.

The experiments are carried out on N-gram- and phrase-based SMT systems. We contrast the obtained results with the ones produced by the state-of-the-art reordering algorithms and demonstrate our methods' improvements over alternative distortion models.

The major conclusion to be drawn from the thesis is that syntactic information is useful in handling global reordering, and it achieves better MT performance than the standard phrase-based and N-gram-based model.

Resum

Durant els últims anys s'han aconseguit millores significatives en traducció automàtica (TA), motivades en gran part per l'aparició de la tecnologia basada en traducció automàtica estadística (TAE), la qual es considera actualment la millor manera de traduir automàticament llenguatges naturals.

L'objectiu principal d'aquesta tesi és la millora del models clàssics de TAE mitjançant la introducció de coneixement sintàctic en l'etapa de pre-traducció a través d'un reordenament en la banda d'origen del corpus. En gran mesura, el nostre interès rau en el valor de la sintaxi en el reordenament per a llengües amb una alta disparitat en l'ordre de les paraules. Un segon objectiu consisteix a determinar el potencial de diverses tècniques de millora del model del llenguatge (ML) per tal de millorar el funcionament i el rendiment dels sistemes de TAE.

Comencem amb un estudi exhaustiu de l'estat de la qüestió en TAE, i descrivim els models fonamentals subjacents en el procés de traducció, així com una breu descripció dels mètodes principals d'avaluació de TA. Fem èmfasi en la TAE basada en sintagmes i N-grames, tot analitzant les diferències principals entre aquestes dues propostes.

Tot seguit, ens concentrem en mètodes de modelització del llenguatge que no han estat objecte de gaire atenció en la comunitat de TAE. Així, presentem els experiments sobre adaptació de sistemes de TAE basats en N-grames a la tasca de transcripció de la parla, descrivim un impacte positiu de la selección d'un llindar límit tant per a la mida del model com per al soroll del ML, i finalment presentem un ML d'espai continu en forma de xarxes neuronals artificials.

A més a més, proposem una nova aproximació basada en la sintaxi per tractar el problema fonamental d'ordenació de paraules per a la TAE tot explotant representacions sintàctiques de textos d'origen i de destí. La idea d'augmentar la TAE utilitzant un pas de reordenament basat en sintaxi previ a la traducció, tal com s'ha proposat els últims anys,

ha resultat força vàlida a l'hora de millorar la qualitat de la traducció, especialment en traduir entre llengües amb un alt grau de disparitat en l'ordre de les paraules.

Oferim al lector un estudi detallat de l'estat de la qüestió en tècniques de reordenament i introduïm una nova classificació d'algorismes de reordenament per a la TAE. Proposem, aleshores, una nova estratègia de reordenament no determinista basada en un alineament augmentat sintàcticament dels textos d'origen i de destí i patrons de reordenament jeràrquic extrets automàticament. En el següent pas combinem el nou mòdul de reordenament amb la descodificació de forma determinista, tot perseguint l'objectiu de fer front amb eficàcia a les dependències de reordenament global i local. Finalment, proposem un nou esquema de mescla d'unitats de traducció, combinant tuples bilingües extretes dels corpus paral·lels amb les bandes d'origen monòtones i reordenades.

Els experiments es duen a terme en sistemes de TAE basats en N-grames i sintagmes. Es contrasten els resultats obtinguts amb els que s'han mostrat mitjançant els actuals algorismes de reordenament, i es demostren millores en els models de distorsió alternatius.

La conclusió principal que s'extreu de la tesi és que la informació sintàctica és útil per tractar el reordenament global i la TE assoleix un millor funcionament en base al model estàndard basat en sintagmes i en N-grames.

Acknowledgments

First, I would like to thank my thesis advisor José Adrián Rodríguez Fonollosa for his professional support and friendly advice and for giving me the freedom to pursue my ideas. You have encouraged and guided my endeavors during every step of the way toward this dissertation, and it would not be possible without your kindly assistance. I learned a lot from you, and I consider myself very lucky to have had the opportunity to do my Ph.D. under your supervision. You have been so patient with my disorganization and I am excited to continue our collaboration.

The next person I would like to thank is José Mariño for his patience and almost mysterious ability to arouse interest in many things that seem routine and boring at a first glance, but which turn out to be simple and transparent after your comprehensive explanation.

My deepest gratitude goes to my internship supervisor Mark Dras from the Centre for Language Technology at Macquarie University. Thank you for making several insightful comments on many aspects of my research work and for your constant support that has lasted for almost two years after my internship in your marvelous country has ended.

My four years in Barcelona turned out to be an exciting adventure and an unforgettable experience, mostly thanks to the daily support from my colleagues and friends. I was lucky to be part of a great group of scientifically ambitious, outstanding researchers at the machine translation group at the TALP Research Center. Thank you for your help and for all the great moments we spent together, including the sleepless nights before evaluation deadlines. Thank you Adrià, Josep María, Marta Ruiz, Rafael, Adolfo and Carlos. I owe a lot to all of you, and maybe especially to Marta, Adrià and Josep María, who always had a minute for a brief consultation.

My gratitude definitively includes all my colleagues at the TALP Research Center and

the Department of Signal Theory and Communications: Climent, Coralí, Enric, Henrik, Ignasi, Israel, Gloria, Jesús, Jordi, Katya, Marta Casar, Martí, both Martins, Mateu, Mireia, Nem, Yesika, Sisco, and many, many others. I learned most of what I know about natural language processing from the present and past students, postdoctoral researchers, and visitors who have been my colleagues in the lab.

I am extremely grateful to Tatyana and Dani for their comprehensive introduction to the mysterious world of the Spanish language and to all the other individuals who helped me with Spanish and Catalan languages. Furthermore, I want to thank Mariella for her endless cheerfulness and for always having a captivating story to share with me.

I am indebted to many external collaborators during my Ph.D. study, especially to the members of the Department of Information Systems and Computation at the Technical University of Valencia, namely Francisco Zamora Martínez, María José Castro-Bleda and Salvador España-Boquera for their great collaboration, important comments and providing me with a neural network language model for re-ranking experiments.

I also want to thank the all the members of the Centre for Language Technology at Macquarie University for the wonderful three months I spent there during my internship. I learned many interesting things there and was lucky to work with you.

I also feel very privileged in having incredible friends who are always there for me. Andrey, Mariella, Ania, Tanya, Edu, Oleg, Daniella, Giovanna, Martin, Celia and other people whose kindness helped me to feel at home in the amazing city of Barcelona – you not only helped me much in research by expressing your support, but also buoyed my spirits, especially during difficult times. Thank you for all the great parties and adventures we have spent together and for keeping me sane.

Needless to say, I also thank the Spanish Ministry of Education and Science (MEC) for making this thesis possible by awarding me with a FPU grant.

Last, but not least, I would like to extend my heartfelt thanks to my parents, Eleonora and Firudin, and to Oksana for their love, support and encouragement. I could not have done this without you. Thank you so much.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This Ph.D. thesis is focused on statistical machine translation (SMT), which is a specific approach to machine translation (MT). MT is a field of computational linguistics that investigates the translation of texts from one human language to another, while SMT, in contrast to many automatic rule-based translation systems, is a translation paradigm based on statistical learning techniques.

Our world is currently in a period of globalization, which implies increasing interaction and the intertwining of different language communities. Information globalization extends to all corners of the world, and although English is becoming a universal second language, users in general still feel more comfortable in their own native language. Consequently, multi-linguality should be seen as a strategic issue for all companies aiming to play an important role in the future information society.

Kaija Poysti's[1] statement that "you can always buy in your own language, but you must sell in your customer's language" has become more and more relevant these days. A modern conception of social communications must include engaging customers, including commercial companies and users, in any information in its textual representation regardless of geography and cultural expectations.

Another important aspect is the socio-political importance of translation in communities where more than one language is generally spoken, as these communities often experience a high need for routine translation. MT is particularly attractive for the European Union (EU) since it already experiences high demands in terms of translation; as of January 1, 2007, there are 23 official EU working languages, and the EU spends more than EUR

[1] Ex-CEO of Trantex, one of the largest localization and translation agencies in Europe

1,000,000,000 on translation costs each year. In addition, the EU has sufficient funding potential to support scientific research in MT.

The present-day information society can be easily characterized by a broad accessibility to a great number of information resources from all over the world that are presented in various languages. Given the lack of quick and quality translation, one confronts a language barrier, further hampering information exchange in a multilingual context.

It would be utopian to believe that at the current state of the information society, MT could completely substitute for human-based translation. However, it can be very useful for translation tasks in which quality may be less important than usability. Moreover, these imperfect MT systems can and are being used by millions of users to translate web-pages and routine everyday documents for which translation quality is not crucial. In these cases, the main goal is to give to the user an idea of the content. An on-going paradigm shift in global network services involves an increase in the demand for on-line real-time multilanguage communication, which will have a great impact on the future user and MT technology. Considering the translation industry, the majority of the work done by professional interpreters involves routine and non-literary translations that are not of great cultural value, and this type of work tends to be the most appropriate for the application of MT.

The imperfection of automatic translation, which is one of the most difficult tasks for natural language processing (NLP), is explained by the high complexity of human languages. Apart from that, there is no single perfect translation of a source string, as there are many factors that influence translation tasks in addition to a great number of poorly formalizable (or not formalizable at all) dependencies [Jur00].

The semantic concept of human languages implies additional hidden challenges that the MT research community faces. In many cases, a word in the source language does not mean exactly the same as its closest counterpart in the target language, i.e., the semantic "spots" for almost each word in the target and source languages do not quite coincide. Word homonymy and polysemy are additional problems that complicate the issue of MT.

Furthermore, a professional interpreter makes decisions about translation based not only on the subject of the phrase, but also by involving additional in-domain knowledge that may be contained in the preceding context. Generally speaking, any professional interpreter can state that a good translation is not just an interpretation of words and expressions as they are, but rather a transfer of thoughts, concepts, images and a human vision of reality,

which are highly influenced by personal and cultural experience.

This chapter will go into detail on MT history, as well as outline the major and secondary goals and objectives of this Ph.D. dissertation. The contributions of this dissertation and its organization are also provided.

1.1 Machine Translation: A Historical Overview

MT can be traced back to the 19th century, when the text carved on the Rosetta Stone was translated using a statistical approach. The Ancient Egyptian language was an enigma for a long while, until the French scholar Jean-François Champollion decrypted the signs present on the stone in three languages, two of which were lost (Egyptian hieroglyphics, demotic script, and well-known Greek), but each communicating the same message [Par99]. These findings served as a starting point for a new approach to translation based on statistical models and volumes of aligned bilingual texts.

The foundation of modern MT was laid in the early 1950s when Warren Weaver published his outstanding paper [Wea55]. This work drew on Shannon's information theory [Sha48] and was inspired by successes in code breaking during the Second World War [Sha51]. It posited that the decryption of codes and automatic translation share the same universal principles.

Enthusiasm faded a decade after, when two reports were published. First, the so-called "Bar-Hillel report" [BH60] posited that MT research set excessively ambitious and utterly unrealistic goals and that a high-quality MT is not achievable without a complete understanding of the translated text. Second, "the ALPAC report" [Pie66] was published in 1966 and stated that actual progress was in fact very poor, with ten years of research yielding results incommensurable with the funds spent.

From the 1950s to the early 1980s, research in the field of MT was substantially restricted, apart from some projects like Russian-to-English translation that were specifically motivated by the US and Soviet governments (one example is the Systran system developed for Russian-to-English translation with a limited vocabulary in 1970 [Tom70]). Gradually, MT research activities shifted to Canada, where a MÉTÉO English-French Translation System [Lan05] was developed at the Université de Montréal to translate weather forecasts. MT research activities also moved to Western Europe, where the English-French version of Systran was introduced by the Commission of the European Communities for helping

Europe with its heavy translation burden.

A number of mainframe translation systems for European and Asian pairs of languages appeared in the 1980s ("Logos" for German-to-English, English-to-French, and English-to-Vietnamese translations, "Metal" for German-to-English translations, and several MT systems for Japanese-to-English and vice versa translations). From the mid-1980s, less expensive MT systems began to appear due to reductions in microcomputers prices, the wide availability of text processing tools and a resurrected interest in MT technology from commercial organizations. Nearly all MT research activities at this time were focused on the exploration of methods for linguistic analysis aimed at elaborating an MT system based on traditional rule-based transfer and interlingua approaches (see subsection 1.2). Attention during this period mainly emphasized MT with human assistance, which is a modern technology of translation memory based on the ALPNET developments of that time [Hut86].

MT got a new boost in the 1990s when the first SMT systems were developed. Its appearance was the result of the tremendous progress made in computer technology and software engineering in the previous few years, as MT began to be used in personal computers and workstations as opposed to mainframes. About that time, IBM began developing one of the first full-scale SMT solutions. Unlike previous approaches to MT, SMT performed translations generated using statistical models based on data derived from the analysis of bilingual text corpora (that is, a collection of texts and their reference translations). Recent research activities have been substantially stimulated by the growing availability of parallel corpora, thereby allowing valuable information to be extracted for a given language pair.

Currently, the general trend in MT is toward the generalization of morphologic, syntactic and semantic abstractions that operate within a statistical translation system and complement fundamental models. The driving force behind this on-going paradigm shift is that the statistical state-of-the-art approach is limited to fully bilingual lexical examples that are extracted from parallel training data, despite the existence of many purely statistical and hybrid translation systems with greater powers of generalization.

MT is in high demand, and it is the subject of worldwide research and development. In 2004, the Translation Automation User Society was established; this society made an important contribution with respect to instilling a positive mindset toward MT among the Internet user society. It is also becoming a commercial software product, and it is at the stage of a new business model elaboration that aims to provide increasing interaction among

producers, sellers and customers. From today's point of view, the future of MT is mostly related to corpus-based techniques, while user communities have great expectations for MT.

1.2 MT approaches

There are several methodologies for MT classification, and the most popular one is based on the level of linguistic analysis it performs. The MT pyramid suggested in [Vau68] specifies processing method comparable to that used by human translators and is presented in Figure 1.1. Three major levels are determined: *direct* translation, a *transfer* approach, and *interlingual* MT. The analysis becomes more and more complicated as one climbs up toward the top of the pyramid.

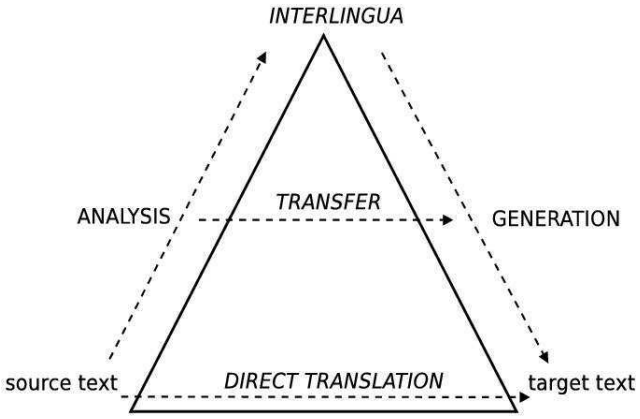


Figure 1.1: *Machine Translation pyramid.*

Pyramid shows comparative depths of intermediary representation.

- The *direct* translation approach represents translation without performing any linguistic analysis at all. This is the simplest approach to MT, as it involves performing "dictionary-style" word-to-word translation. Presently, this approach is not in use.
- A *transfer-based MT* system realizes a deeper level of intermediary representation, usually on the morphological (i.e., parts-of-speech (POS) analysis or lemmatization) level or the syntactic level. The main idea of this approach is to establish a collection of transfer rules (either automatically learned or human-made) that define a correspondence between the structure of the source language and that of the target

language. It normally consists of two steps, namely, analyses of the source language text and target language string generation.

- The *interlingual MT* approach is found on the top of the pyramid and is based more on the structural similarities of languages than on an assumed identical mapping of meanings. According to this approach, translation is considered as a mapping between "semantic spaces" of a particular word in the source and target languages, and the choice of the correct translation hypothesis is conditioned by the "on-site" semantic meaning of the word from the target language.

  An interlingual MT system has access to the complete semantic scheme of the source sentence represented in the form of neutral (that is, abstract) language, which is used as a "bridge" between the source and target languages. An example of such a language can be found in [Dav01], and is known as *UniversalWords*. It was originally proposed as a tool for informational unification over communication through the Internet. In [Hir93], a universal language is established in accordance with different perspectives (i.e., sentence concept, sentence structure, and the intention and perspective of the speaker).

  Despite the fact that this approach intuitively seems to provide the best translation quality, it suffers from many difficulties. Insurmountable problems have sometimes been faced by researchers of interlingual MT, such as a lack of semantic language-dependent analytical resources; a lack of tools and resources required for semantic and morphological synthesis from the artificial abstract language to the natural one; arbitrarily deep syntactic word representation; divergence of information contained in the target and source training corpora that can lead to a shortage of information required for the generation step; and the impossibility of combining parsing and generation steps.

There is an ongoing debate regarding whether semantic information is contained in the syntax of a language [Rap02]. This claim erases the border between the pure interlingual and transfer-based MT paradigms.

Considering system design criteria, MT systems can be decomposed as shown in Figure 1.2.

Rule-based systems are accepted as a classical approach to MT. Translation systems based on this approach use the set of linguistic rules (normally set by human experts)

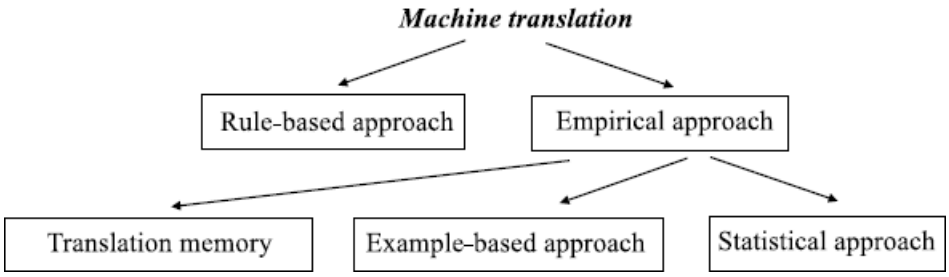


Figure 1.2: *Decomposition of machine translation approaches according to design criteria.*

specifically describing the translation process [Str98, Cha02]. The weak point of these systems is that they require a great deal of linguistic knowledge, which is expensive in terms of both time and money.

An empirical approach (EA) to MT, appeared at the beginning of the 1990s and is a new paradigm to challenge and enrich the established rule-based MT. The appearance and popularization of these data-driven methods is possible due to a big leap in computing technology and availability of textual data in huge quantities. This approach is based on the parallel corpora (pure data) and the reusing of examples of already-existing translations to generate a final translation.

There are three major classes of EA systems: *example-based MT (EMT)*, *translation memory*, and *statistical MT*.

EMT and *translation memory* both deal with finding and matching examples that contribute to a translation on the basis of their similarity with the input sequence, thereby translating the source string on the basis of recognizing bits that have been previously translated and then sticking them together. They differ in that in translation memory, the example extraction stage is carried out by humans during the post-editing step [Pla05, Ger02], while EMT provides it automatically and is sometimes considered to be an extension of the latter technology [Car98, Zha01]. The essence of the SMT method is to generate translations using statistical models in which parameters are estimated on the basis of bilingual text corpora. In this way, it fundamentally differs from EMT since the ranking between fragments is done with probabilities rather than matching measures.

The area of written language MT is the significant part of academic research. Another aspect of SMT applications is closely related to the automatic speech recognition *(ASR)* field, namely speech translation. In contrast to written language MT, speech transcription is characterized by special peculiarities typical of recognition tasks, like noise, non-grammatical input, spontaneous speech phenomena, and so on. There is much interest in exploring new techniques in SMT and ASR integration, as indicated in [Kay92, Kni98, Vid97a]. The ultimate goal of the SMT/ASR integration approach is to develop a speaker-independent real-time translation system that is likely domain-oriented.

1.2.1 A statistical approach to machine translation

The modern SMT originates from work carried out by IBM in the early 1990s [Bro90a, Bro93]. This research was inspired by experiments in the field of speech recognition, while the ideas underlying SMT came from information theory. A statistical approach to MT is based on the principle of translating a source sentence into a sentence in the target language using statistical information drawn from the parallel training corpus. The problem can be reformulated as the selection of the most probable translation from a set of target sentence hypotheses. A detailed description of the SMT paradigm can be found in chapter 2.

Initially operating at the word level (that is, word-to-word translation), MT ignored the context in which a word was used and could not tackle situations in which the fertility[2] is less than 1, typical for translations from morphology poor (like English) to morphology rich languages (like Spanish).

Later, this approach was deployed using algorithms, that attempted to overcome this disadvantage by learning the translation of a sequence of source- and target-side words, though not necessarily a linguistically motivated set (see [Zen02] for example). In the next step, the final translation is generated by a composition of partial translations that are subject to a certain reordering algorithm.

The present-day popularity of the SMT approach both among the MT community and the speech recognition scientific community is mainly explained by:

- The constantly growing availability of parallel aligned corpora, along with monolingual texts, which are necessary for high-quality target language modeling;

[2]The ratio of source translating sequence length to target translating sequence length. In other words, fertility tells how many foreign words each native word produces.

- The good performance of existent SMT systems, which have proven to be competitive with or have even outperformed rule-based MT systems in various evaluation campaigns (see Appendix C);

This technology has some clear advantages over traditional MT methods:

- Once constructed, a SMT system can be developed for a new pair of languages given parallel data with minimal human effort (at least in theory);
- SMT can deal with lexical word ambiguity involving context or other informational sources;
- SMT methods are more robust to non-grammatical input or grammatical faults typical of live speech;
- SMT algorithms can theoretically address idiomatic expressions that occur in the training corpus;
- Ignoring the syntax of a sentence, can generate more natural translations based on human-made examples from the training data.

Along with obvious benefits, this approach must still address one challenge: a lack of access to the structure of the sentence. SMT does not explicitly deal with syntax, as it is classically not involved in the translation process or in the word-reordering step, and there are no conditioning translations for syntactically related words. The syntax of the sentence is crucial to model many translation phenomena, such as systematic differences between the word order of the languages following distinct word order schemes or the place of the modifiers for nouns. For this reason, a number of possible ways of introducing syntactic information incorporation within a SMT system have been presented over the last years.

While there is endless debate over whether SMT is better than orthodox approaches to MT, it is obvious that most global IT companies have clearly created a situation favorable for SMT. Even the well-known Google Translate service provided by Google Inc. has adopted a purely statistical approach to MT, and it recently augmented its system with a target language structure expressed in the form of a parse tree [Zol08]. More details about the syntax-based and statistical translation approach hybridization are provided in chapter 4.

By the late 2000s, SMT has become an area of major interest and considerable funding from state and private institutions. Several large-scale multinational research activities,

projects, and initiatives have been recently founded that have among their main goals improvements to the SMT approach to speech and text translation (TC-STAR[3], LC-STAR[4], MataHari[5], AVIVAVOZ[6], etc.). These recent developments firmly establish SMT technology both as a subject of legitimate research and as a useful application of technology.

Another important factor stimulating state-of-the-art progress in MT is the open source software and tools that make feasible it to perform SMT experiments in a fast and simple way, thereby providing technology access to the average user. Along with the growing availability of language-dependent linguistic resources and tools (e.g., taggers, parsers, and lemmatizers), specialized components of SMT software can be found, including the GIZA++ word alignment toolkit[7] [Och03b], the open source Moses toolkit[8] [Koe07a], the MARIE decoder[9][Cre05a], and so on.

1.3 Thesis outline and research contributions

This dissertation involves research performed between 2005 and 2009 at the Center of Speech and Language Applications and Technology (TALP) at the Universitat Politècnica de Catalunya (UPC).

The thesis consists of **six** chapters:

Chapter 2 outlines the current state-of-the-art SMT technology. First, it introduces the mathematical frameworks of early word-based SMT systems and presents phrase-based translation as a natural evolution of the original approaches, as well as the TALP-UPC N-gram-based system as an alternative approach. It also details word alignment algorithms, the introduction of additional feature functions and weight optimization procedures. It concludes with a brief description of the mechanisms for MT performance evaluation. This chapter serves as the foundation on which the remaining chapters build.

Chapter 3 introduces a detailed description of the research, focused on language model improvement. The first part of the chapter describes LM adaptation experiments for verbatim translation tasks, while the second part presents a continuous-space LM driven by

[3] http://www.tc-star.org
[4] http://www.lc-star.com
[5] http://www.dcs.qmul.ac.uk/~christof/html/projects.html
[6] http://www.avivavoz.es
[7] http://code.google.com/p/giza-pp/
[8] http://www.statmt.org/moses/
[9] gps-tsc.upc.es/veu/soft/soft/marie/

a neural network along with its incorporation into N-gram-based SMT. Both approaches demonstrate a significant improvement in translation accuracy when applied to a corresponding translation task. The last part of the chapter introduces an accurate LM pruning with an optimal cut-off threshold set to the minimal count of n-grams of order n included in the LM.

Chapter 4 focuses on a detailed study of state-of-the-art reordering frameworks and sets the stage for the following chapter. We first review previous work that serves as inspiration for our approach. More particularly, we briefly review state-of-the-art reordering frameworks and syntax-based translation systems, which to a certain extent motivate the reordering approach presented as a main contribution of the Ph.D. dissertation. We then identify some limitations of the existing literature and describe how the word-reordering problem is treated therein.

Chapter 5 presents research on *Syntax-Based Reordering (SBR)*, which is the main contribution of this thesis. It discusses the novel distortion model, which is based on a set of syntax-based reordering rules derived via a syntactically augmented alignment of source and target texts. We apply a word-reordering scheme that captures local and global word distortion dependencies. We describe the algorithms in detail, and analyze the scheme theoretically and empirically by testing it on various translation tasks.

Chapter 6 contains conclusions drawn from the dissertation and highlights the contributions of the thesis. It also suggests possible extensions of the research conducted for this Ph.D. dissertation.

At the end of the Ph.D. dissertation are four appendices. **Appendix A** provides an overview of the corpora used in the experiments. Meanwhile, the Ph.D. thesis is completed in the context of two research projects detailed in **Appendix B**. Third, **Appendix C** describes the participation of the international evaluation campaigns in which the TALP-UPC took part between 2004 and 2009. Finally, **Appendix D** references a list of publications by the author related to the Ph.D. research.

In this Ph.D. dissertation, we make the following contributions to the SMT field:

- We introduce a novel approach to word reordering problem for SMT, which is called *Syntax-Based Reordering*, similar in spirit to the modern hybrid statistical translation systems augmented with syntax. This technique drives local and long-range word reorderings by automatically extracted permutation patterns operating with source

language constituents and augmenting them with non-isomorphic sub-tree transfers. The algorithm is applied in the step prior to translation, reformulating translation task from the *plain source-to-target* to the *reordered source-to-target translation*, which makes a mutual word order closer to monotonic and leads to a simplification of the translation task. Furthermore, the statistical model is enhanced with a source input word lattice, which is used by decoder during taking decision about final translation.

- We propose to estimate a *continuous-space LM (CC LM)* presented in form of a neural network, which is then used in a SMT system. We report our experiments for a smaller translation task with a limited amount of training material, which is the most applicable field for CC LM. An N-gram-based SMT system enhanced with CC LM applied during the n-best list rescoring step shows a statistically significant improvement in translation accuracy.

- We study how *LM task adaptation* influences overall translation system performance and can be used in real-life applications, such as in translation of automatic speech recognizer output. Although the approach of linear LM interpolation is not new in itself, we show how the N-gram-based SMT system can benefit from the interpolated task-adapted LM and report experiments on verbatim translation tasks.

- We address the problem of *optimal threshold-based LM pruning* by isolating its impact on translation quality and model size for several LM pruning decisions. We show how accurate and rational selection of the number of n-grams can positively influence not only the model size and system processing speed but also the translation accuracy.

- We also report extensive contributions to the TALP-UPC research group's effort in more than 10 *international evaluation campaigns* in which the TALP-UPC system was regularly ranked quite high. System construction for an evaluation competition normally implies that different combinations of statistical methods and model parameters have been tried with the aim of maximizing translation quality for a given translation task(s). This leads to important design decisions that are taken in each system configuration and allows for advancing the state-of-the-art of the technology.

This research addresses several important questions that the research community is facing regarding the impact and value of models addressing the translation process by mainly concentrating on the distortion model and target-side LM.

On the theoretical side, this dissertation focuses on some aspects of the hybrid distortion model formalized through a context-free grammar representation. On the practical side, we show via small- and large-scale experiments that the proposed algorithms improve the accuracy of state-of-the-art SMT systems.

Note that the research developed in this Ph.D. has been published in a number of publications, which are referenced in the appropriate chapters and in Appendix D.

Chapter 2

State of the art

This chapter provides a brief overview of modern approaches to SMT and gives a review of the models addressing the translation process. It also summarizes the most significant works from the rapidly growing field of SMT that are relevant to this Ph.D. research.

We first cover the mathematical foundations of SMT in §2.1; we describe the source-channel approach followed by the first SMT system since its introduction by the IBM research group in the early 1990s. Afterward, we present a set of feature functions that augment the posterior probability calculation and their combination with the maximum entropy approach.

Introduced almost 20 years ago, the word-based IBM approach to SMT has lost its popularity nowadays as a primary translation engine, but it serves as a basis for the modern SMT framework, as it establishes relationships between words in source and target languages. This procedure is called statistical word alignment and is outlined in §2.2.

§2.3 describes the current approaches to SMT as a natural evolution of this initial word-based paradigm. Modern SMT systems operate with sequences of consecutive words by considering them to be translation units rather than dealing with them as isolated words. We describe a classical phrase-based SMT (§2.3.1), briefly outline factored translation model (§2.3.2), hierarchical (§2.3.3), and syntax-based (§2.4) approaches, and then concentrate on a description of the N-gram-based (tuple-based) system (§2.3.4).

§2.5 concludes this chapter with a presentation of the most popular metrics used in the automatic and human evaluation of translation quality.

2.1 Mathematical background

In SMT we are given a source language string $F = f_1^J = f_1 \dots f_j \dots f_J$ (traditionally referred to as "*French*"), which is to be translated into a target language sentence $E = e_1^I = e_1 \dots e_i \dots e_I$ ("*English*"), where I and J represent the number of words of the sentences in target and source languages. The translation problem is defined as *arg max* operation, as described by the following equation:

$$\hat{E} = \arg\max_E \{ p(E|F) \} \tag{2.1}$$

Hence, the translation problem is formulated as choosing the translation hypothesis with the highest probability among the set of target sentences.

Modern SMT is based on the so-called *noisy-channel* (or *source-channel*) approach [Bro90b] according to which the equation 2.1 is decomposed using Bayes' rule, as follows:

$$\hat{E} = \arg\max_E \left\{ \frac{p(F|E) \cdot p(E)}{p(F)} \right\} \tag{2.2}$$

$$= \arg\max_E \{ p(F|E) \cdot p(E) \} \tag{2.3}$$

This decomposition into two knowledge sources allows for an independent modeling of the *target language model* $P(E)$ and *bilingual translation model* $P(F|E)$. The source string probability $P(F)$ is usually omitted within the bounds of the *arg max* operation because it does not affect the choice of the translation hypothesis.

The translation model (TM) establishes linkages between the source and target strings, taking into account word-to-word links; a detailed description of word alignment model can be found in §2.2. The set of model probabilities is automatically estimated from an aligned parallel bilingual corpus.

The target-side language model (LM) assigns probabilities to target word sequences and is built on a monolingual dataset. Statistical language modeling has been successfully used for many NLP applications, including speech recognition, part-of-speech tagging, syntactic parsing and information retrieval [Roa07, Son99, Che99]. Typically, n-gram models are used in SMT; however, alternative models exist, and more details can be found in chapter 3.

Training materials for TM and LM can vary; in real translation systems, monolingual complements are frequently used for language modeling.

The overall architecture of the statistical translation system in the noisy channel approach is depicted in figure 2.1.

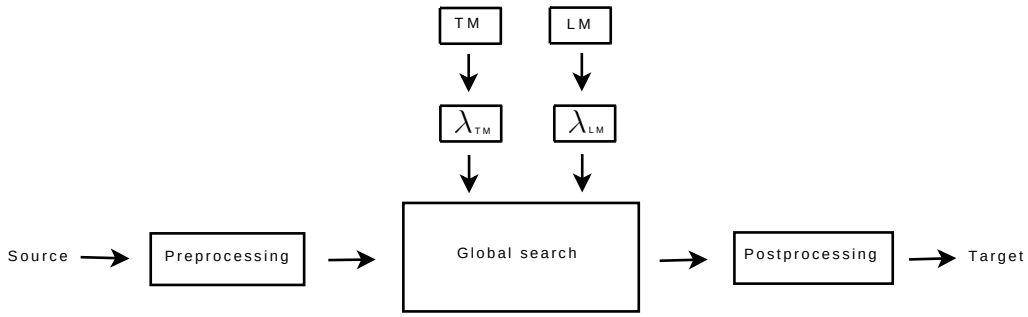


Figure 2.1: *Noisy channel approach.*

The classical noisy channel approach has been supplemented with an alternative *maximum entropy framework* [Ber96a], which was proposed in [Pap98] for a natural language understanding problem and successfully applied to the SMT task, as shown in [Och02b]. According to this approach, the posterior probability $P(e|f)$ is directly modeled as a log-linear combination of the set of feature functions and is considered a generalization of the source-channel paradigm. Mathematically, it is expressed as follows:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{ p(e_1^I|f_1^J) \} \tag{2.4}$$

$$= \arg\max_{e_1^I} \left\{ \frac{exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^{I'}_1} exp(\sum_{m=1}^{M} \lambda_m h_m(e'^{I'}_1, f_1^J))} \right\} \tag{2.5}$$

$$= \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \tag{2.6}$$

where the feature functions h_m refer to the system models and the set λ_m refers to the scaling factors corresponding to these models. The model weight coefficients are trained according to the maximum class posterior criterion [Och02b] or with respect to the translation quality measured in the form of error criterion [Och03a]. The maximum entropy approach is a generalization of the noisy channel approach that considers only two feature functions (i.e.,

the translation and language models) and equal scaling factors. The schematic of the maximum entropy approach is displayed in figure 2.2.

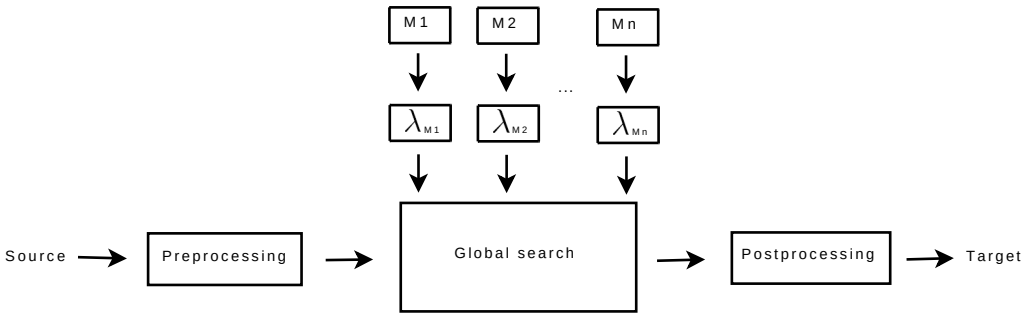


Figure 2.2: *Maximum entropy approach.*

2.2 Word alignment

Typically, the number of words and the order of counterpart appearances in translated sentences are different. Hence, the first challenge in statistical translation is to establish the correspondence between the words of the target sentence and the words of the source sentence, i.e., to model the string translation probability $P(f|e)$. In [Bro93] this modeling problem is addressed through a hidden variable a aimed at accounting for all possible pairwise alignment links between two sentences:

$$P(f|e) = \sum_{a} P(f, a|e) = P(J|e) \sum_{j=1}^{J} P(a_j|f_1^{j-1}, a_1^{j-1}, e, J) \cdot P(f_j|f_1^{j-1}, a_1^{j}, e, J) \qquad (2.7)$$

where J is the length of the source sentence f, f_j stands for the word of the sentence f in position j, and a_j refers to the hidden alignment of word s_j describing the position in the target sentence where the word that aligns to s_j is placed. Note that alignment a_j can take on a zero value, i.e., $a_j = 0$ with the artificial NULL word to account for the source word that is not aligned to any target word.

Word alignment using mechanisms similar to Hidden Markov Models (HMM) specifies the modified word order when a sentence is translated into another language and, given a sentence and its translation, specifies links at the word level. Figure 2.3 shows a visualization

of an alignment for a parallel English-Spanish sentence.

English:	*This*	*is*	*an*	*important*	*matter*	.
Spanish:		Es	un	asunto	importante	.

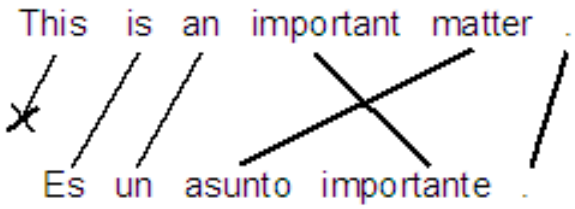


Figure 2.3: *Word alignment example.*

The three fundamental models developed to calculate the probability in equation 2.7 are decomposed as follows:

- *Fertility model.* This accounts for the probability that a target word e_i generates ϕ_i words in the source sentence, i.e., the suggested number of source words that are generated by each target word.
- *Lexicon model.* This accounts for the probability of producing a source word f_j given a target word e_i, i.e., strict dependencies are suggested between source and target words.
- *Distortion model.* This model tries to explain the phenomenon of placing a source word in position j given that the target word is placed in position i in the target sentence, i.e., the reordering of the set of the source words is suggested that best complies with the target language. This is also used with inverted dependencies and is known as the alignment model.

Different combinations of these models are known as *"IBM machine translation models"*:

- IBM1 - assigns a uniform distribution to the alignment probability (*lexical probabilities only*);
- IBM2 - introduces a zero-order dependency with position in the source (*lexicon plus absolute position*);

- Homogeneous HMM - IBM2 modification, which introduces first-order dependencies in alignment probabilities [Vog96, Dag94] (*lexicon plus relative position*);
- IBM3 - describes the choice of a fertility ϕ_i, which depends only on e_i (*plus fertility*);
- IBM4 - models relative movements conditioning the linking decision on previous linking decisions (*inverted relative position alignment*);
- IBM5 - limits the waste of a probability mass on impossible situations (*non-deficient version of IBM4*).

The generative process underlying IBM models calculation is illustrated in Figure 2.4.

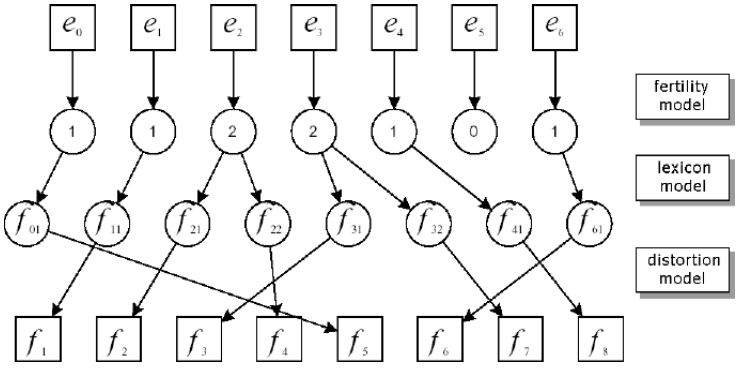


Figure 2.4: *IBM models: translation.*

Estimation algorithms for some of the IBM models, including fertility parameters (IBM3, IBM4, and IBM5), usually have problems with respect to poor local optima. The algorithm proposed in [Bro93], which mitigates this challenge, is an instance of an unsupervised learning technique, namely the *Expectation-Maximization* (EM) algorithm [Dem77]. It is used to increase the likelihood of the parallel data given the model. Iterative training is initiated with a simple model, and later on the parameters of the simple model are used to initialize the training procedure of a more complex model.

More detailed information about IBM 1-5 Models can be found in [Och02a], and a systematic performance comparison is presented in [Och03c].

The manual annotation of word alignments is an expensive and frustrating task. The present-day popularity of the statistical approach to MT to a certain extent can be traced

back to 1999, when a freely-available tool called GIZA implemented IBM models to generate Viterbi alignment[10] [AO99]. This tool was released as a part of the EGYPT toolkit. In 2001 and 2003, an improved extension of the program appeared, which was called GIZA++[11] [Och03d].

In the majority of currently available translation systems, this piece of software coupled with the mkcls[12] tool, which allows for statistical word clustering for better generalization [Och99], is used for training statistical TMs from bitext [Mar06b, Koe07b, Ima05].

GIZA++ implements a word alignment algorithm in an unsupervised fashion. However, recent research efforts on SMT systems seem to be shifting toward supervised and semi-supervised alignment models [CB04, Fra06, Lam08, Moo06, Wu06].

One of the problems with the IBM models is that they generate both one-to-many and many-to-one source-to-target alignments. Several heuristic symmetrization algorithms have been recently proposed as supplements to these two alignments in order to decrease the effect of incorrectly aligned multi-word units. A many-to-many alignment is usually obtained using the following algorithm:

- An **intersection** of alignments;
- A **union** of alignments;
- Using **refined** symmetrization method, as described in [Och00a];
- Employing **grow-diag-final-and** heuristic [Koe05b].

The latter is a widely accepted and most popular strategy based on the step-by-step extension of the intersection of original alignments, which consequently involves neighboring left, right, top, or bottom elements (**grow**); the diagonally neighboring alignment points (**grow-diag**); the non-neighboring one-direction aligned words (**grow-diag-final**); and the alignment points between two unaligned words (**grow-diag-final-and**).

The data sparseness problem is crucial for word alignment, as for many other SMT tasks. In [dG06], this issue is addressed with linguistic classifications done before the alignment step. Base forms (that is, lemmas), stems, and reduced verb morphology are used instead of preface word forms, and a positive impact on translation system performance is reported.

[10] We use the term "Viterbi alignment" to denote the most probable alignment given the estimated IBM models using a Viterbi search rather than the true Viterbi alignment.

[11] code.google.com/p/giza-pp/

[12] http://www.fjoch.com/mkcls.html

There are two main measures to evaluate the quality of word alignment. The most widespread criteria is **AER (alignment error rate)**, as proposed in [Och00b]. Given a manual gold standard alignment with the criterion of sure and possible links, *recall, precision, and AER measures* are defined. However, in [Fra06], it is shown that the AER measure does not always correlate with MT accuracy, but it does with the *F-measure value* because of its capacity to penalize precision and recall components. Recently, new measures of word alignment quality have appeared [Aya06], showing quite promising results.

Since this Ph.D. thesis is not directly related to statistical word alignment evaluation, further description of the alignment evaluation metrics falls out of the scope of the thesis.

2.3 Current SMT approaches

In this section, we briefly outline the most popular modern approaches to SMT:

- A **phrase-based** model operating with sequences of phrases instead of single words, as a coherent and natural evolution of the IBM translation models (§2.3.1);
- A **factored translation** model taking into account additional information such as morphology, lemmas, and so on during translating (§2.3.2);
- A **hierarchical** approach to SMT that intends to address the most significant problems of phrase-based translation (§2.3.3);
- An alternative to the phrase-based SMT N**-gram-based** approach to SMT stemming from a finite-state perspective (§2.3.4);
- A **syntax-based** MT system that uses source, target, or both-side parse trees (§2.4).

2.3.1 Phrase-based translation

In human-made translation, it is common to translate contiguous sequences of words as a single unit. For example, a Spanish expression *"casa blanca"* is translated into English as *"white house"*. IBM translation models do not take into account local context but rather translate the words *"casa"* and *"blanca"* separately and then place them in a monotone order, without intervening words. This strategy intuitively can be improved using the original contiguous sequence *"casa blanca"* as a fundamental translation element. The appearance of an approach in which models address units longer than simple words has

caused a groundbreaking improvement in SMT, and this data-driven paradigm has proven to be a possible practical means to the longstanding goal of cheap MT.

Modeling word sequences (**phrases**[13] or **bilingual phrases**) rather than single words in both the alignment and lexicon models makes sense with respect to the frequency-based nature of natural language. Moreover, this treatment makes it possible to exploit the ability to handle collocational relations within the sentence.

Phrases are extracted from a bilingual aligned corpus using the following two fundamental constraints:

1. The words in the source and target parts are consecutive,
2. They are consistent with the word alignment matrix.

The phrase extraction procedure is illustrated in Figure 2.8. The original bilingual sentence is:

Spanish:	*quisieramos*	*lograr*	*traducciones perfectas*
English:	we would like	to achieve	perfect translations

This approach was first presented in [Och04] and was named the *alignment template approach*. The translation process consists in grouping source words into phrases; source phrases are mapped onto target phrases and are allowed to be generatively inserted as lexically motivated by word context. Thus, translation is performed in a monotone phrase order, allows for word classes and includes internal word alignment. Example of translation procedures can be found in Figure 2.5.

A simplified version of the alignment template approach is the so-called phrase-based SMT ([Zen02]). The simplification consists in (1) handling words ignoring word classes, (2) ignoring internal alignment information, and (3) assuming one-to-one phrase alignments. This can be formally expressed as shown in equation 2.8:

$$P(f_1^J|e_1^I) = \alpha(e_1^I) \cdot \sum_B P(\tilde{f}_k|\tilde{e_k}) \tag{2.8}$$

where the hidden variable B is the segmentation of the sentence pair in K bilingual phrases $(\tilde{f}_k|\tilde{e_k})$, and $\alpha(e_1^I)$ assumes equal probability for all admissible segmentations.

[13] Hereafter, we use the term "phrase" to refer to any consecutive sequence of words that do not necessarily coincide with their linguistic analogues.

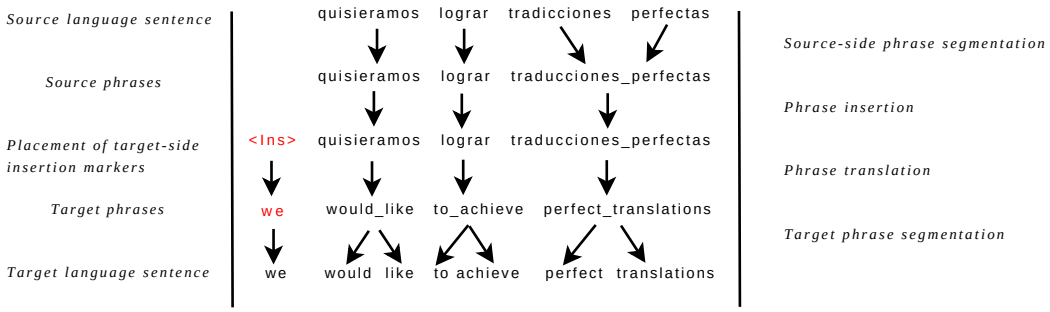


Figure 2.5: *Alignment template model with monotone phrase order.*

The phrase translation probabilities are commonly estimated by relative frequency over all bilingual phrases in the corpus for both translation directions:

$$P(f|e) = \frac{N(f,e)}{N(e)} \tag{2.9}$$

$$P(e|f) = \frac{N(f,e)}{N(f)} \tag{2.10}$$

where $N(f,e)$ refers to the number of times the phrase f is translated by e. In addition, $N(f)$ and $N(e)$ refer to the number of times the source or target phrase, respectively, appears in the training corpus.

According to this approach, phrase-based translation is considered a three step algorithm:

1. The source sequence of words is segmented in phrases;
2. Each phrase is translated into the target language using the translation table;
3. The target phrases are reordered to follow the natural order of the target language.

There are alternative methods of phrase extraction. In [Koe03], the authors look for methods to efficiently build phrase translation probability tables. They also demonstrate that the approach based on word alignment outperforms a syntax-based phrase generation

method [Yam01], and a joint-probability model learning phrase translation and alignment probabilities simultaneously from a set of parallel sentences [Mar02].

Another important question raised in [Koe03] is *"How long do phrases have to be to achieve high performance¿'* The best results were obtained when limiting the length to a surprisingly low level (3 in the case of European Parliament[14] German-English translation task). Learning longer phrases does not yield much improvement and can even lead to worse results. The size of phrase tables rises almost linearly with the training corpus size (Figure 2.6).

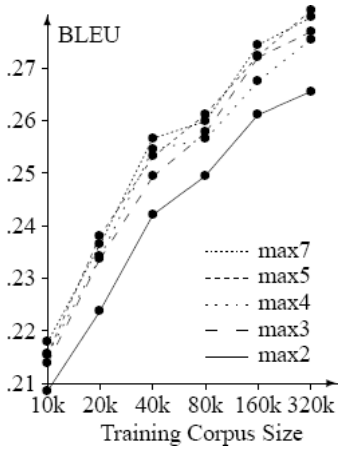


Figure 2.6: *Different limits for maximum phrase length. Source: [Koe03].*

2.3.2 Factored translation model

A noteworthy extension of phrase-based SMT is the **factored translation model**, which behaves like phrase-based models while being able to take into account morphology, lemmas, stems, and other informative sources at the word level in the source and target languages during decoding [Koe07b, Koe07c].

The tight combination of morphological, syntactic, or semantic information has demonstrated to be valuable by integrating it in before or after processing steps. In a factored translation, a word is represented as a vector of factor elements characterizing different levels of annotation (see Figure 2.7).

The mapping of foreign words onto target language words is broken up into two steps:

[14] See Appendix A.

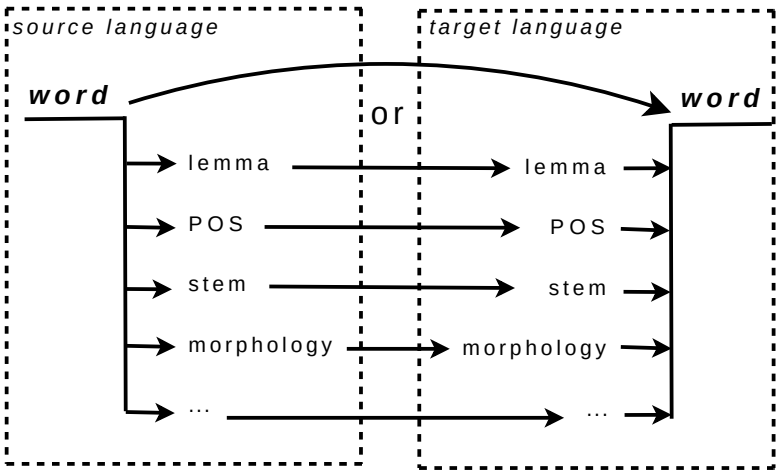


Figure 2.7: *Factored translation model: input/output representation and translation process.*

- *Translation step*: source language factors are mapped onto the target language factors (phrasal level);
- *Generation step*: target language factors are mapped onto the target language factors and words (word level).

Then, the surface forms are generated on the target side from the translated vector elements.

Note that in modern systems, alternative decoding paths are allowed only for unknown words that the system is not able to translate. In other words, the model allows both surface and morphgen (model that do morphological analysis and decomposition during the translation process) TMs, preferring a surface model for known words and using a morphgen model to act as back-off.

The Moses toolkit[15] is a widely-known implementation of the factored-based model.

2.3.3 Hierarchical translation

One more recent enhancement of the classical phrase-based approach to SMT is a **hierarchically structured model** that defines transduction rules, which are interpretable

[15] www.statmt.org/moses/

as components of a bilingual (synchronous) formal grammar [Wu97, Chi05, Chi07, Igl09]. These models intend to address the principal limitations of phrase-based SMT, i.e., the sparse data effect and distance-dependence of the distortion models.

In [Wu97], the twin concept of bilingual language modeling and simultaneous parallel parsing was proposed. It involves context-free inversion transduction grammar formalisms, which learn a grammar and simultaneously generate two trees extracted from a parallel text without any syntactic annotations. The system models order shifts between languages and balance needed flexibility against complexity constraints.

A hierarchical phrase-based system called *Hiero* was presented in [Chi05] and enhanced with a *cube pruning* method used to efficiently apply LMs at the search step as described in [Chi07]. A hierarchical structure is applied to capture translations with scope larger than a few consecutive words. Thus, *Hiero* removes the limitation on contiguous phrases and allow phrases to include indexed placeholders, thus turning the phrase-based SMT into a parallel parsing problem over a grammar with one non-terminal symbol.

Formally, each rule from the generalized rule hierarchy[16] can be expressed as follows:

$$N \longrightarrow f_1 \dots f_m / e_1 \dots e_n \tag{2.11}$$

and can be extended by another existing rule:

$$M \longrightarrow f_i \dots f_u / e_j \dots e_v \tag{2.12}$$

where $1 \leq i < u \leq m$, $1 \leq j < v \leq n$ and the right-hand side of the rule constitutes a phrase pair under the word alignment, to obtain a new rule:

$$N \longrightarrow f_1 \dots f_{i-1} M_k f_{u+1} \dots f_m / e_1 \dots e_{j-1} M_k e_{v+1} \dots e_n \tag{2.13}$$

where k is an index for the non-terminal M that indicates a one-to-one correspondence between the new M tokens on the two sides. Note that adjacent non-terminals are prohibited due to "spurious ambiguity" and over-generation issues [Chi07].

Lately, a hierarchical phrase-based translation has emerged as one of the dominant current approaches to SMT due to an efficient combination of phrase-based translation advantages and strengths of the hierarchical architecture that underlie any natural language.

[16] Formally, equations 2.11, 2.12 and 2.13 describe a generalization of the Chiang's approach with multiple non-terminals. Classically, *Hiero* employs only a single non-terminal [Chi07].

2.3.4 N-gram-based translation

In conjunction with the phrase-based approach, the N**-gram-based** approach also appeared [Cas02, Cas04]. It has roots in the finite-state transducers paradigm primarily proposed for speech translation [Vid97b, Kni98, dG02] and extends the log-linear modeling framework, as shown in [Mar06b]. A system that follows this approach deals with bilingual n-grams, which are the so-called **tuples**. Tuples are extracted from a word-to-word alignment and are composed of one or more words from the source language and zero or more words from the target one.

Here, the translation procedure is regarded as a stochastic process maximizing the joint probability $p(f, e)$, which is approximated at the sentence level as described by the following equation:

$$\hat{e}_1^I = \arg\max_{e_1^I} \{ p(e_1^I, f_1^J) \} \tag{2.14}$$

$$= \arg\max_{e_1^I} \left\{ \prod_{n=1}^{N} p((f,e)_n | (f,e)_{n-x+1}, \cdots, (f,e)_{n-1}) \right\} \tag{2.15}$$

where x is the length of the context.

The tuples induce a unique segmentation of the pair of sentences, as shown below. This way the context used in the TM is bilingual, and it not only takes the target sentence into account, but both languages are linked together by means of tuples. The TM can be seen here as an n-gram LM of an imaginary language composed of two-language units.

The main difference between phrase-based and N-gram-based approaches lies in their distinct representations of bilingual units, which are the components of the TM. While the regular phrase-based SMT considers context only for phrase reordering but not for translation, the N-gram-based approach conditions translation decisions on previous translation decisions.

The N-gram-based SMT has proven competitive with state-of-the-art systems within recent evaluation campaigns [Kha08, Lam07b].

Tuples extraction. The size and content of tuples vocabulary strongly rely on the particular set of word-to-word alignments. In [Ban00], tuples are extracted from one-to-one

alignments, while in [Cas04] a one-to-many word alignment set is used.

In contrast, in all the experiments with N-gram-based SMT, we adopt an approach according to which the *union* of the source-to-target and target-to-source alignment sets (that is, many-to-many) is used for tuple extraction. This approach has proven efficient, as shown in [Mar06b], where positive results are reported for some specific translation tasks.

Tuples are extracted from the parallel corpus following a set of three rules:

- Given a certain word-to-word alignment, a monotonic segmentation of each bilingual sentence pair is produced;
- No word in a tuple is aligned to words outside of it;
- No smaller tuples can be extracted without violating the previous constraints.

Note that phrase-based SMT does not consider the last rule; hence, tuples are formally defined as minimal length bilingual phrases that provide monotonic segmentation of the corpus.

Figure 2.8 shows an example of tuple extraction from a bilingual sentence pair and contrasts it with the phrases extraction procedure (see §2.3.1).

Tuples:
1. NULL : we
2. quisieramos : would like
3. lograr : to achieve
4. traducciones perfectas : perfect translations

Phrases:
1. quisieramos : would like
2. quisieramos : we would like
3. lograr : to achieve
4. traducciones : translations
5. perfectas : perfect
6. quisieramos lograr : would like to achieve
7. quisieramos lograr : we would like to achieve
8. traducciones perfectas : perfect translations
9. lograr traducciones perfectas : to achieve perfect translations
10. quisieramos lograr traducciones perfectas : would like to achieve perfect translations
11. quisieramos lograr traducciones perfectas : we would like to achieve perfect translations

Figure 2.8: *Bilingual phrase and tuples extraction.*

There are two issues regarding tuple extraction procedure:

- *Embedded words.* The embedded words problem arises for the tuples that include more than one word. A certain number of one-word units are left out of the tuple vocabulary because they are embedded in longer units. Considering an example from Figure 2.8, the words "perfect" and "translation" are not provided with translation probabilities if they are not encountered in one-word tuples in the training corpus.

  The solution of this problem was proposed in [dG02]. The tuple n-gram model is enhanced by incorporating single-gram translation probabilities for all embedded words detected during the tuple extraction step. These unigram translation probabilities are computed from the *intersection* of source-to-target and target-to-source alignments, in contrast to the first-step tuples extraction in which *union* symmetrization is used.

- *Source-side NULLs.* One of the hardest decisions during tuples segmentation is how to handle units with NULL on the source side. Some words linked to NULL end up producing tuples with NULL source side (for example, *NULL#we* tuple in Figure 2.8). However, no NULL input is expected to appear in tuples, and this tuple is not allowed. This problem is solved by modifying word alignments before the tuple extraction step; any target word that is linked to NULL is attached either to the preceding or following word.

  A challenging issue is a binomial choice between two attachment directions. Three tuple segmentation strategies are proposed and contrasted in [dG06]. The main conclusion drawn from the results reported for English-to-Spanish, Spanish-to-English, and Arabic-to-English translation tasks is that the best performance is demonstrated by the system that regards the tuple segmentation problem around source-NULLs as a monolingual decision relying on statistics conditioned on associated target-side POS tags that are consistent with human intuition.

The tuples-based approach is basically considered monotonous in the sense that the local context is modeled in a sequential order of tuples during training. Therefore, it is more appropriate for pairs of languages with relatively similar word order schemes. Currently, many research efforts are being made toward adapting the N-gram-based approach to language pairs with different word orders [Cj09, Cre05c]. These recent developments will be overviewed in chapter 4.

Modeling. Like many other outstanding SMT systems [Dóf8, Mat06, Che08], the TALP-UPC N-gram-based translation system follows a log-linear approach to inform the decoder with probabilistic information based on a set of feature functions. Apart from the TM presented above, the feature models typically taken into consideration are:

- **A target-side LM**, typically represented in the form of an n-gram model, or a continuous-space LM trained in the form of a neural network (see chapter 3). The model accounts for the target language statistical dependencies and favors those partial translation hypotheses that are more likely to correctly constitute structured target sentences over those that are not.

  A clear-cut distinction of an N-gram-based system from a phrase-based SMT is its unique representation of the LM; for a phrase-based system, it is an integrated part, whereas in an N-gram-based SMT, LM is used as an additional feature since the target language is modeled *inside* the bilingual n-gram model and is considered a way of improving translation accuracy.

- **A tagged target-side LM**, normally implemented as an n-gram model of word classes associated with target-side words and act as a method for reducing data sparseness. The word classes can be statistically or linguistically motivated (POS) [Kha07, Cj06a].

- **A word bonus (or word penalty) model**, which is implemented in order to compensate for the system's preference for short output sentences. Technically, the bonus depends on the total number of words in the partial translation hypothesis and is determined as follows:

$$P_{WP}(t_k) = exp(\text{number of words in } \mathrm{t_k}) \tag{2.16}$$

- **Forward and backward lexical models**, provide lexicon translation probabilities for each tuple based on the word-to-word IBM 1 probabilities. These models estimate lexical weights according to the formula below:

$$P_{Lex}((\tilde{f},\tilde{e})_k) = \frac{1}{(I_k+1)^{J_k}} \prod_{j=1}^{J_k} \sum_{i=0}^{I_k} P_{IBM1}(f_j|e_i) \tag{2.17}$$

where f_j and e_i are the j-th and i-th words in the source and target parts of the tuple $(\tilde{f}, \tilde{e})_k$; J_k and I_k are the corresponding total numbers of words on either side of the tuple. Giza++ word-to-word source-to-target and target-to-source alignments are used in the calculation of forward and backward lexical models, respectively. Note that the lexicon models give probabilities to tuples of different source and target length and actually constitute complementary TMs.

Translation scheme and modeling issues. A typical training scheme for an N-gram-based SMT system is illustrated in Figure 2.9.

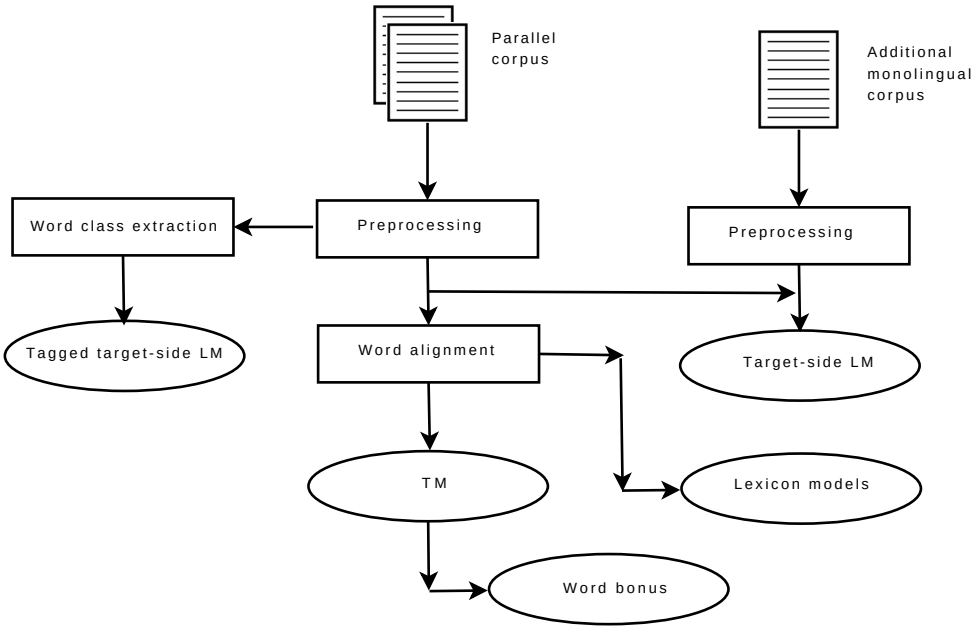


Figure 2.9: *Feature models estimation scheme. Data flow diagram.*

Normally, parallel data is provided as already aligned on the sentence level, i.e., only one-to-one links (one source sentence aligned to one target sentence) are found in the corpus. However, in some situations, additional sentence or paragraph alignment is needed, which can be done by splitting the training sentences by full stops on both sides of the bilingual text when the number of stops is equal.

The preprocessing step consists in tokenization[17] and the application of language-dependent morphology reduction algorithms, especially relevant for morphologically rich languages like Spanish or Arabic.

Morphology reduction algorithms can include:

- Contraction separation. For example, *del* → *de el, al* → *a el* can be used when translating to or from the English version *of the* and *to the*, respectively;
- Postfix or suffix splitting. For example, prepositions, conjunctions, articles, and future marker separation may be useful for Arabic-to-English translation;
- Verbal forms splitting. For example, Spanish verbs, such as *mándaselo* with gloss *send to him it*, can be split into three single words *manda/send*, *se/him*, and *lo/it*);

Equivalence word class extraction can be done statistically in a language-independent way according to the algorithm outlined in [Och99] or linguistically using POS taggers for the target side of the training corpus.

In the following step, word alignment is performed by estimating IBM translation models and finding the Viterbi alignment in accordance with them. This process is carried out using the GIZA++ toolkit or BIA tool [Lam07a].

LM, tagged LM, and TM estimation are done with the SRI LM toolkit [Sto02]. Parameters and configuration of monolingual LM models for words and word classes, such as the n-gram order (history length), smoothing technique and pruning strategy, are empirically adjusted to minimize perplexity value measured on the preselected test dataset.

Since the tuples bilingual model is implemented as a standard n-gram model dealing with "bi-language", widely-known language modeling problems and challenges are still relevant for this model. TM parameters can hardly be estimated as a function of perplexity computed on a reference. Instead, the parameters are adjusted to maximize automatically measured system performance (see §2.5). A comprehensive study and comparative analysis of different smoothing techniques for bilingual n-gram models can be found in [dG06].

[17] Simple normalization strategies tending to reduce vocabulary size without information loss (i.e., which can be reversed if required). For example, separating punctuation marks, classifying numerical expressions into a single token, and so on.

2.3.5 Decoding and optimization

Decoding In general, SMT can be interpreted simplistically to be seen as a two-fold problem consisting in modeling and search. While modeling is the subject of this Ph.D. dissertation, the search part of the global problem is addressed by an in-house decoder called MARIE[18] [Cre05b, Cre08a]. The tool implements a beam-search algorithm with reordering capabilities (i.e., an input permutation graph) and allows for three different pruning strategies:

- **Histogram pruning**, which limits partial-translation hypotheses to the n-best ranked instances;
- **Threshold pruning**, which discards the partial-translations with the score below a certain threshold value;
- **Hypothesis recombination**, which is a risk-free strategy according to which partial-translation hypotheses are seen by a decoder as identical in case if they coincide exactly in both the present tuple and tuple n-gram history.

Scale factors optimization. Given a set of m feature model parameters $\boldsymbol{\lambda} = \lambda_1 \ldots \lambda_m$ and a corpus of K aligned sentence pairs $(f_1, e_1), (f_2, e_2), \ldots, (f_K, e_K)$ that do not overlap either the training set or the test dataset, let $F(\boldsymbol{\lambda})$ be a real-valued scalar function that characterizes particular aspects of a training procedure to optimize the parameters $\boldsymbol{\lambda}$. The optimization problem is then formally stated as:

$$\hat{\lambda} = \arg\max_{\boldsymbol{\lambda}} F(\boldsymbol{\lambda}) \tag{2.18}$$

The parameters $\boldsymbol{\lambda}$ are estimated to minimize the translation error, using a *minimum error rate (MER)* iterative strategy as described in [Och03a]. Theoretically, the estimation criteria should be entropy maximization in the development set, but in practice, it has been proven to be too complex computationally. Instead, the BLEU metric, or a weighted combination of translation quality metrics applied to one or more reference translations of the development test, are used.

The optimization procedure is performed by a tool implementing the downhill simplex [Nel65] or the SPSA [Lam06] algorithms. The method uses a geometrical figure called

[18] http://gps-tsc.upc.es/veu/soft/soft/marie/

"simplex" consisting in N dimensions of $N+1$ points and all their interconnecting line segments, polygonal faces, and so on. The starting point is a set of $N+1$ points in parameter space, which defines an initial simplex. At each step, the downhill simplex algorithm performs geometrical operations (that is, reflections, contractions, and expansions) until a local minimum is reached. In our case, it adjusts the log-linear weights so as to maximize an objective function. The optimization scheme is depicted in Figure 2.10.

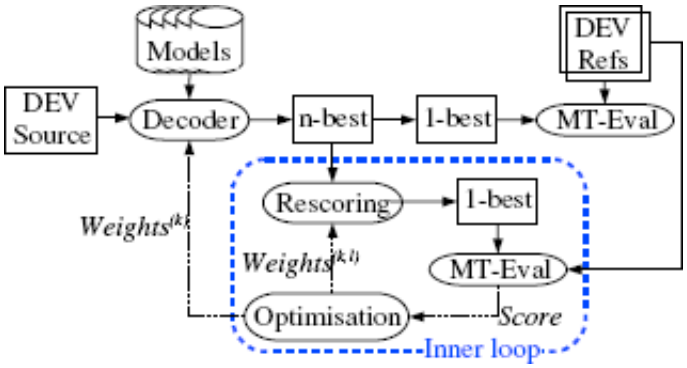


Figure 2.10: *Optimization scheme. Flow diagram.*

Translation quality is numerically estimated over the list of n best translations of the development dataset. Inside the inner loop, the n-best list is first generated by the decoder. Then, the optimization algorithm is used to minimize the translation error while rescoring the n-best list and fine-tuning the values of the set of scale factors. Once the coefficients are optimized, a new decoding cycle is launched and an updated n-best list is produced [Ber06].

The algorithm does not guarantee that the global minimum is found. Usually a multi-start method for global optimization is used in order to decrease the probability that the process is trapped in local minima. A rescoring process employs the same models used in the overall search (i.e., decoding). The algorithm converges when no improvement in translation quality is observed or when a maximum number of translations is achieved. More details on the optimization procedure can be found in http://www.statmt.org/jhuws/.

2.4 Syntax-based translation

In the next lines, we give a brief overview of the translation systems that model the translation process using parse tree structures since their underlying principles are directly connected to the main contribution of this Ph.D. thesis. Since a detailed presentation of several state-of-the-art reordering techniques based on syntax can be found in chapter 4, their description will be omitted in this section.

The main motivation of syntax-based TMs is to overcome widely-known problems inherent in SMT, such as:

- The data sparseness problem, which is even more serious when the source, target, or both languages are highly inflected and rich in morphology.
- Long-distance reordering, since the statistical distortion model is only based on the movement distance that causes a computational resource limitation [Och04].
- Dependencies difficult to capture with purely statistical methods (clause restructuring or word order scheme transformation (SOV↔VOS), for example).

Linguistic syntax is a potential solution to many of these problems, as it accurately models many systematic differences between source and target languages. That is why the challenge of incorporating syntactic information in a statistical framework has been of increasing interest to many researchers in the past several years. However, a concept that seems elegant and promising at first glance has not achieved state-of-the-art results until recently, when the scientific community's perseverance was awarded with the creation of well-performing hybrid MT systems combining natural language syntax and machine learning methods. One of these systems is presented in [Mar06a], which has obtained state-of-the-art results in Arabic-to-English and Chinese-to-English large-size data tasks. Another very promising algorithm integrating syntax into hierarchical SMT is implemented in the SAMT toolkit [Ven06] (see below).

Modern syntax-based systems operating with parse trees perform translation in two separate steps, namely parsing and decoding. They can be classified in accordance with the way syntax is used, whether syntax is used (1) on the source side only (**tree-to-string**), (2) the target side only (**string-to-tree**), (3) or on both sides (**tree-to-tree**).

Tree-to-string and string-to-tree approaches. Tree-to-string translation systems use source-side language. The source-side dependency parsers induce target-language dependency structures, as shown in [Men05] and [Lin04b]. As an alternative, several systems make use of target-side syntax [Mar06a, Gal06, Cha03].

In [Yam01], a set of operations on each node of the parse tree is defined, and leaf nodes are translated. Extra words are inserted at each node, while reordering is modeled by permutations of children nodes. This approach is criticized for its high computational costs and poor performance when dealing with non-grammatical input typical of spoken language.

The hierarchical TM introduced in §2.3.3 does not have any linguistic motivation, and it is induced from a parallel text without linguistic annotations. Improvements are achieved through context-free grammar (CFG). The open-source toolkit SAMT[19] [Zol06] is an implementation of the MT system, which provides further evolution of this approach. While Chiang's model operates with only two non-terminals (that is, a substantial phrase category and "a glue marker") in [Zol06, Ven06], a significant improvement in terms of translation quality has been reported if complete or partial syntactic categories (which are derived from the target-side parse tree) are assigned to the phrases. SAMT was also implemented for the *MapReduce* model [Ven09].

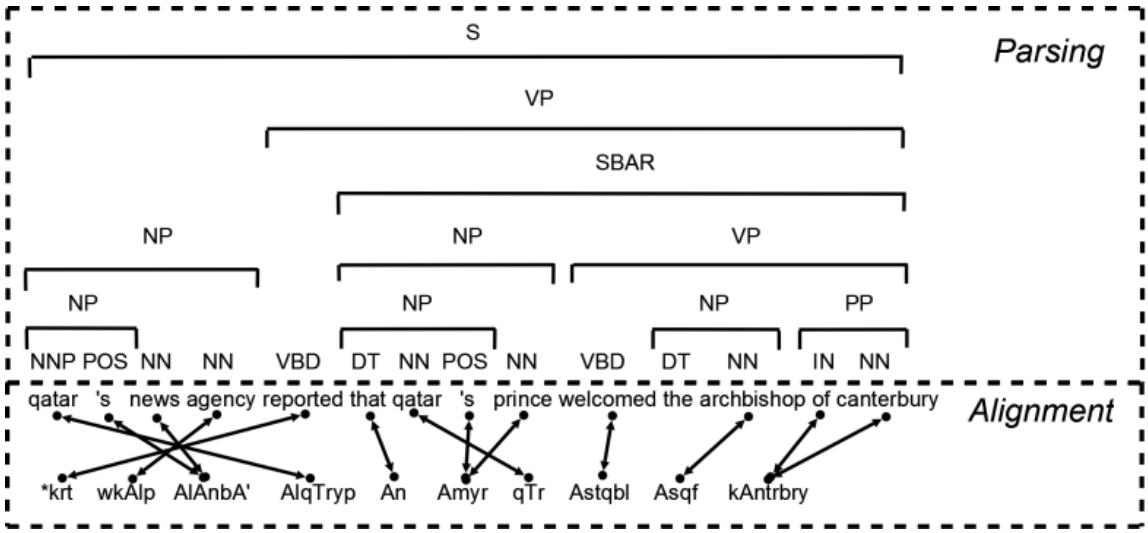
Figure 2.11 shows an example of the rule extraction procedure driven by the target-side parse tree structure with underlying word-to-word alignment. The set of initial translation rules is augmented with more complex rules constructed using a Combinatory Categorial Grammar (CCG) [Ste00].

Tree-to-tree approach. A number of researchers have proposed models in which the translation process involves syntactic representations of both the source and target languages. The formal description of many of these models is based on tree transducers, which describe operations on tree fragments rather than on strings.

There has been a good deal of research on tree-to-tree *theoretical methods*. Among these studies is [Mel04] in which some important theoretical results are discussed. A comprehensive theoretical framework for generalized synchronous parsing and translation using multitexts[20] grammars is proposed. In [J.E03], non-isomorphic tree-to-tree mappings in the context of the synchronous tree-substitution grammar formalism are discussed, although no

[19] www.cs.cmu.edu/~zollmann/samt

[20] Parallel texts between an arbitrary number of languages.



Initial rules:
NN → agency, wkAlp
DT → that, An
...
NP → qatar 's news agency, wkAlp AlAnbA' AlqTryp
NP → that qatar 's prince, An Amyr qTr

CCG rules:
DT/NP → qatar 's prince, Amyr qTr
VBD+NP → reported that qatar 's prince, *krt An Amyr qTr
NP+PP → the archbishop of canterbury, Asqf kAntrbry

Figure 2.11: *SAMT: example of rule extraction.*

results are reported. [Gra04] present a tree transducer able to compute transformations on trees.

The complete syntax-driven SMT system based on a two-side sub-tree transfer is described in [Ima05]. The authors propose a probabilistic non-isomorphic tree mapping model

based on a context-free breakdown of the source and target parse trees; they extract alignment templates that incorporate the constraints of the parse trees; and then apply syntax-based decoding.

The idea of using synchronous grammars in probabilistic form for MT has been recently presented by the Computer Science Group of the Harvard University. In [Shi07] a conceptual basis for thinking of MT in terms of synchronous tree-adjoining grammar is provided. In [Nes06] is shown that a probabilistic synchronous variation of tree-adjoining grammar can be useful for SMT.

In [Wu97], a formalism called inversion transduction grammar is described and used to parse source and target languages synchronously with an inversion ordering rule. According to the approach proposed by Wu, CFG rules simultaneously generate both non-terminal symbols[21] and terminal symbols. A natural trade-off between translation accuracy and computational complexity is found.

In [Cow08], a statistical tree-to-tree mapping algorithm based on *aligned extended projection*, which is a parse-tree syntactical structure intending to model NLP phenomena and used as a target-language structure when mapping from source parse tree to target language, is performed.

Hybrid MT approaches. Some hybrid MT systems utilize methods from the phrase-based approach, for example, by improving phrasal coverage [DeN07].

A tree-to-tree *Treelet* approach projects a target dependency tree onto target-language translation using word alignment [Qui05]. Once the projection is complete, dependency treelets are extracted, and a tree-based ordering model can be trained.

[Mar06a] present a hybrid Chinese-to-English tree transducer implementation, that uses target-language syntax to augment source-target phrase pairs. The probabilities of these phrases are estimated as frequency counts and are employed in the context of a log-linear framework.

A synchronous tree substitution grammar trained from the source and target language parse trees serves as the engine of the *Tectogrammatical* English-to-Czech translation system [Boj07]. The adopted chart parsing algorithm is described in [O.07]. A probabilistic model assigns a conditional probability to each phrase, informing the decoder how probable is the rule applied to a given pair of slot types.

[21] Non-terminal symbols are shared by both languages.

2.5 Evaluation

Research on MT has revealed many difficulties, among which performance evaluation is one of the most challenging and subjective tasks. Human evaluation of MT output remains the most reliable method to assess translation quality, despite its costliness and time consumption. To overcome some of the drawbacks of MT evaluation, many automatic evaluation metrics that are much faster and cheaper than human-made evaluation have been recently introduced. However, so far, the scientific community has not accepted unified criteria for MT evaluation, as has been the case in the ASR field[22].

This section briefly describes common evaluation metrics used for the assessment of translation quality and the error analysis technique of MT output, which makes possible the identification of the most prominent source of errors in a given system.

Automatic evaluation Usually, the task of automatic translation quality evaluation is performed by producing some kind of similarity measure between the translation hypothesis and a set of human-made reference translations, which represent the expected solution of the system.

The commonly accepted criteria that defines the quality of the evaluation metric is its level of correlation with human evaluation. Another reason to have automatic measure of MT quality is that human scoring can be subjective and may vary depending on the human and his/her particular point of view with respect the correct translation.

As shown in [Eck05], the above-mentioned automatic measures have attained good correlation results at the system level, while the degree of correlation achieved at the sentence level, crucial for an accurate error analysis, is much lower.

Nowadays, there are several automatic measures widely used. Selection of robust and consequent evaluation criteria is of crucial importance, as MT systems are normally trained to optimize an automatic evaluation measure.

BLEU metric. The BLEU (Bilingual Evaluation Understudy) metric has been proposed by [Pap02]. It calculates the geometric mean of the precision of n-grams ($n \in 1, \ldots, k$, where k refers to a BLEU order) between a hypothesis and a set of reference translations with an exponential brevity penalty factor to compensate for inadequately short translations (i.e., shorter than references), as shown in equation 2.19:

[22] Word error rate - WER.

$$BLEU_k = exp\left(\frac{\sum_{i=1}^{k} bleu_i}{k} + LP\right) \tag{2.19}$$

where $bleu_i$ and LP (length penalty) are a cumulative count, updated sentence by sentence, and refer to the whole evaluation corpus (test and reference datasets). $bleu_i$ is computed by dividing the matching n-grams by the number of n-grams in the test dataset, as shown in equation 2.20:

$$bleu_i = log\left(\frac{Nmatched_i}{Ntest_i}\right) \tag{2.20}$$

BLEU implies a very high score with a short output, so long as all its n-grams are present as a reference. Roughly speaking, one BLEU point represents a minor but appreciable difference in the recovery of n-grams.

BLEU is mostly a precision metric, taking recall into account in a very simple way by considering only the measure for sentence length; this component of BLEU acts like a cheating detector. The introduction of the length penalty component is motivated by the idea that if a candidate receives a high score then it must match the reference in length, in word choice and in word order. If the candidate and reference translation are of approximately the same length, a translation must produce the same words in roughly the same order as the references in order to obtain a high precision score.

The formula for LP calculation is as follows:

$$LP = min\left\{0.1 - \frac{shortest\_ref\_length}{Ntest_1}\right\} \tag{2.21}$$

$Nmatched_i$, $Ntest_i$ and $shortest\_ref\_length$ scores are cumulative counts and are calculated as shown below:

$$Nmatched_i = \sum_{n=1}^{N} \sum_{ngr \in S} \min\left\{N(test_n, ngr), \max_r \{N(ref_{n,r}, ngr)\}\right\} \tag{2.22}$$

where S is the set of n-grams of size i in sentence $test_n$. $N(sent, ngr)$ is the number of occurrences of n-gram ngr in sentence $sent$. N is the number of sentences that must be evaluated. $test_i$ is the i-th sentence of the test set. R is the number of different references

for each test sentence, and $ref_{n,r}$ is the r-th reference of the n-th test sentence.

$$Ntest_i = \sum_{n=1}^{N} length(test_n) - i + 1 \tag{2.23}$$

$$shortest\_ref\_length = \sum_{n=1}^{N} \min_r \{length(ref_{n,r})\} \tag{2.24}$$

Note that the final score is not computed by accumulating a given sentence score, but instead matching counts are estimated on a sentence-by-sentence basis.

Despite being a useful characteristic and *de facto* standard in MT evaluation, BLEU can often unnecessarily penalize syntactically valid but slightly altered translations with low n-gram matches. It is specifically designed to perform the evaluation on a corpus level and can perform badly if used over isolated sentences.

METEOR. The METEOR (Metric for the Evaluation of Translation with Explicit ORdering) score is a metric for the evaluation of MT output, which is calculated as an averaged mean of precision and benefited recall by considering stems and synonym matching.

It was introduced in [Ban05] and [Lav08], which showed that METEOR produces adequate correlation with human judgment at the sentence or segment level, thereby distinguishing it from the BLEU metric in that BLEU seeks correlation at the corpus level.

The evaluation algorithm consists of two steps:

1. First, a set of mappings between unigrams from the translation output and the reference translation is iteratively created. Every unigram in the translation output must map to zero or one unigram in the reference translation and vice versa. In any alignment, a unigram in one sentence cannot map to more than one unigram in another sentence.

   Three information sources are used in this step: "exact" words (preface forms), stems, and WordNet synonyms [Mil91], with the default ordering of application as appears in the text.

2. Once the alignment is produced, the sequence of matched unigrams between the two strings is divided into the fewest possible number of "chunks" such that the matched unigrams in each chunk are mutually adjacent and in identical word order. Then, the METEOR score for this pairing is computed as follows:

$$METEOR = (1 - penalty) \cdot F_{mean} \tag{2.25}$$

where

$$penalty = \gamma \cdot frag^{\beta} \tag{2.26}$$

γ scales penalty value ($0 \leq \gamma \leq 1$); $frag = ch/m$ is a fragmentation fraction calculated as a ratio between the number of chunks ch and the number of matches m. The value of β determines the functional relation between fragmentation and penalty.

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \tag{2.27}$$

where *precision* $P = m/t$, *recall* $R = m/r$. m refers to the number of mapped unigrams found between two strings, t refers to the total number of unigrams in the translation, and r refers to the number of unigrams in the reference.

The free parameters in the metric, namely α, β and γ, are highly language dependent, as shown in [Lav07], and must be fine-tuned to achieve maximum correlation with human evaluations. In the framework of this dissertation, the parameters are set to the following values: $\alpha = 0.9$, $\beta = 3$, $\gamma = 0.5$[23]

The metric evaluates a translation by computing a score based on explicit word-to-word matches between the translation and a reference translation. If more than one reference translation is available, the given translation is scored against each reference independently, and the best score is reported.

The main problems with METEOR are its limited area of application (at present, it can be used for English, Spanish, Czech, French, and German) and its evaluation speed (METEOR performs much slower than BLEU).

Other automatic evaluation metrics. There are many other evaluation metrics that have been recently proposed and claim a strong correlation with human intuition. In the next few lines, we briefly outline the most important measures:

[23] The parameters are optimized for English. Strictly speaking, due to high need for linguistic tools, only the English version of METEOR is fully supported at the moment.

- **NIST.** The NIST scoring system, developed by the National Institute of Standards and Technology, is a sensitive metric of MT quality. It is based on the BLEU score but weights n-grams in order to mark less informative n-grams with higher weights [Dod02]. It is again based on n-gram precision, but it employs the *arithmetic average* of n-gram counts rather than a geometric average. Consequently, the n-grams are weighted according to their information contribution, as opposed to just counting them, as in BLEU.

  The idea behind this metric is to offer a higher evaluation if a system provides an adequate translation of a difficult segment (that is, it obtains an n-gram match that is rare), but to offer a lower evaluation an n-gram match that is easy. The NIST metric has the same weak points as BLEU.

- **The WER or mWER.** The word error rate or multireference word error rate was introduced in [McC04] as a standard speech recognition evaluation measure and is calculated as the minimum word-level Levenshtein distance between a translation system output and a reference translation. The WER is the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference target sentence. In case of multiple references, a whole set of reference translations is used. For each translation hypothesis, the edit distance to the most similar sentence is calculated, so that the final measure for a given corpus is based on the cumulative WER for each sentence.

- **The PER and mPER.** The position-independent word error rate or multireference position-independent word error rate is a variation of (m)WER metric, alleviating the effect of a possibly different word order between an acceptable translation hypothesis and reference translation(s) [Til97].

- **ORANGE and IQMT.** Finally, there are several works devoted to designing a uniform metric that considers information at distinct linguistic levels and permits combining metric scores into a single measure of MT quality. In this case, automatic evaluation is considered the application of similarity metrics between a set of candidate translations and a set of reference translations.

  The Oracle RANking for Gisting Evaluation (ORANGE) [Lin04a] is defined as the ratio between the average rank of the reference translations within the merged list of k

best-produced translations and reference translations and the size of the list. However, ORANGE does not allow the simultaneous consideration of different metrics.

The Inside Qarla Machine Translation (IQMT) evaluation framework is presented in [Gim06]. This tool follows a "divide and conquer" strategy so that one can define a set of metrics and then combine them into a single measure of MT quality in a robust and elegant manner, avoiding scaling problems and metric weightings.

Human evaluation Human-driven methods of MT evaluation require a certain degree of human intervention in order to obtain a quality score. Manual evaluation is performed by human judges, which are instructed to estimate the performance of a system based on a sample of its output. For the most part and thanks to their linguistic competence, bilingual language users are able to perform an intuitive evaluation on the quality of MT system output and can be considered the reference for a number of language processing tasks.

However, there is also considerable variation across their ratings due to many aspects, such as task- and domain-dependence, the evaluator level of conscious linguistic knowledge, some personal evaluator preferences (such as whether the evaluator weighs content or grammar highest) or dynamic learning from evaluation (e.g., evaluating different types of errors would allow one to distinguish among systems that perform more or less the same). Because of these reasons, the automatic evaluation is sometimes referred to as an *objective* evaluation, while the human kind appears to be more *subjective*.

The three most popular strategies of human evaluation are described below:

1. *Adequacy-fluency.*

   Accuracy and fluency is a widespread means of manual evaluation. Usually, these measures of generated translation are evaluated according to a 1 to 5 quality scale. *Fluency* indicates how natural the hypothesis sounds to a native speaker of the target language. *Adequacy* is assessed after the fluency judgment is done, and the evaluator is presented with a certain reference translation and must judge how much of the information from the original translation is expressed in the translation by selecting one of the proposed grades.

2. *Ranking.*

A simplified approach that has been gaining popularity in the last evaluation campaigns is the *ranking* of sentences. Annotators have to rank up to five sentences from best to worst relative to the other choices, with ties usually allowed. As stated in [CB07, CB08], this approach yields both greater inter- and intra-annotator agreement.

3. *Post-editing.*

   Another trend is to manually *post-edit* the references with information from the test hypothesis translations so that differences between a translation and reference account only for errors; in this case, the final score is not influenced by the effects of synonymy. The human targeted reference is obtained by editing the output with two main constraints; that is, the resultant references must preserve the meaning, and it must be fluent.

   In this case, we refer to the measures as their human-targeted variants, such as HBLEU, HMETEOR, or HTER, as in [Sno05]. These measures are rooted the paradigm of semi-automatic evaluation metrics for interactive MT, which advocates for counting metrics like KSR (key-stroke ratio), the number of key-strokes required to produce the single reference translation, or MAR (mouse-action ratio), which measures the number of mouse pointer movements [Civ06].

   Unfortunately, despite its promising potential, post-editing evaluation techniques are costly and cannot be used to evaluate minor system improvements.

The manual evaluation of MT output is extremely time-consuming and expensive to perform, and so comprehensive comparisons of multiple translation systems are rare. However, mostly thanks to international evaluation campaigns aiming to set up a fair framework for objectively comparing different MT systems, human evaluation metrics are also used in order to compare different systems. An overview of the most relevant MT evaluation campaigns can be found in Appendix C.

Chapter 3

Improved Language Modeling for SMT

As an important part of any MT system, language modeling has not received much specialized attention within the SMT community, which has preferred to focus on more specialized translation models, decoding algorithms, and training techniques. In contrast, in other fields of NLP, particularly in ASR, there exists a large body of research that addresses the specific problems of statistical language modeling. To a great extent, this discrepancy is a consequence of noisy experimental results and inconsistencies between a LM configuration and a translation system performance. However, recent progress in the availability of training data has made the application of such monolingual techniques quite promising since typically the greater amount of data is used to estimate the parameters of the LM, the better the LM performance is.

In this chapter, we first cover major works in the area of language modeling (§3.1) including probabilistic models and language modeling using syntactic structure. We then present three methods of LM enhancement. In §3.2, we report on the experiments regarding standard SMT system adaptation to the verbatim translation tasks by general and specific target-side LMs interpolation. In §3.3, we present our experiments on threshold-based LM pruning for LMs of different history length and describe the impact of an accurate cut-off threshold selection for both the model size and LM noisiness.

§3.4 reports a different language modeling technique based on a continuous-space LM representation trained in artificial neural networks (NN). For the most part, this approach aims to improve translation quality for tasks that lack translation data. The scores produced

by the continuous-space LM are assigned to each sentence from a pre-calculated list of n best-translation hypotheses. In the next step, the list is rescored selecting the best translation by taking into account the score generated by the continuous-space LM.

3.1 Related work

Techniques for language modeling can be classically decomposed in three main approaches.

Statistical language modeling

The first class of approaches includes statistical corpus-based probabilistic models, which is a powerful and simple method for language modeling. Traditionally, statistical LMs have been designed to assign probabilities to strings of words (or tokens, which may include punctuation) according to the so-called n-gram models. These models assign high probabilities to frequent sequences of words by considering the history of $n-1$ preceding words in an utterance. As such, they have comprised a *de facto* standard for language modeling in the state-of-the-art SMT systems [Zen02, Mar06b, Che05, Ven06]. The idea behind the n-gram model consists in dividing sentences into fragments that small enough to be frequent (and thus appear in the corpus) but are large enough to contain some language information. The probability of each fragment is then calculated. A sentence that contains many frequent fragments is placed in good order and should have a high probability.

One related issue is that n-grams that do not occur in the corpus will be assigned a zero probability and will void an entire sentence's probability. The most prominent technique to avoid this loss of generality is through the use of a smoothing algorithm, which redistributes probability from observed events to unobserved ones (*back-off*). There are a lot of smoothing methods proposed over the past few years that follow different strategies with the common idea of taking some of the probability "mass" (normally, a very small but positive value) from the known n-grams and redistributing them to the unseen ones. An excellent discussion of smoothing techniques for n-gram LMs may be found in [Che99].

However, one obvious disadvantage of the n-gram model is that it can not capture the long-distance dependencies in data. Various alternative algorithms have been recently proposed.

The skip LM [Ros94, Mar99] gets influence of words further away without increasing dimensionality by skipping some words in the word history in a probabilistic way. It allows

to condition word probability on the context different from the previous $n-1$ words and estimate long-range dependencies probabilistically. A widely-known SRI LM toolkit [Sto02] includes skip LM in a package.

The trigger language model, which was described in [Lau93, Ros94, SM07], is designed to model the fact that content words are more likely to be used repeatedly within a single conversation than to occur evenly spread throughout all speech. In other words, a trigger feature indicates the number of times in a conversation that a certain word is seen preceded by a previous instance of this word.

Syntax-based language models

The second type of LMs includes syntax-based LMs incorporating syntactic structure with the standard n-gram model. To a certain extent a bunch of publications on the syntax-based language modeling, appeared in recent years, is explained by the recent progress in statistical parsing.

[Cha03] demonstrates how LMs might be improved by adding syntactic structure by rescoring a tree-to-string translation forest with a lexicalized parser comprising synchronous context-free grammars. Other parsers operating in a left-to-right manner attempt to build the syntactic structure incrementally while traversing the sentence from left to right [Che98, Xu03] or make use of semantic dependencies in a maximum entropy model for accurate language modeling [Wu99].

The model proposed in [Sar07] tends to reduce the number of out-of-vocabulary words preserving the predictive power of the whole words for the task of SMT. It also allows incorporation of additional available semantic, syntactic and linguistic information about the morphemes and words into the language model.

Factored language models

Another approach to language modeling includes factored language representation, which treat each word is not only a token, but a vector of factors that represent a variety of additional information sources (lemmas, part-of-speech tags, etc.). According to the factored LM, presented in [Bil03, Kir05], each word is dependent not only on a single stream of temporally preceding words, but also on its other factors given the factorized history. This representation is considered as an extension of the n-gram word-based language modeling

to tightly integrate additional information, in order to exploit sparse training data more effectively.

An alternative to the aforementioned factored model is a LM extension with lexical linguistic representations ("supertags"), associated with at least one lexical item. "Supertags" employment assigned for each word according to Lexicalized Tree-Adjoining Grammar and Combinatory Categorial Grammar is described in [Has07], while the usefulness of their integration into the target-side LM and the target side of the translation model is shown in [Has08].

3.2 Language modeling for verbatim translation tasks

The challenge of adapting LM to a specific task has been largely reinvented in the last twenty years, like many other computational problems. Language modeling is a key component of any SMT system; however, monolingual training material available for some specific translation tasks often is not enough to estimate a well-performed LM. The quality and quantity of the monolingual data mostly determines the quality of the LM and, indirectly, of the translation output.

At the same time, there are huge amounts of monolingual data available from popular domains (news domain, for example). Many systematic distinctions between datasets from different data domains can be considered, namely average sentence length, vocabulary, and so on.

These data can be efficiently used for language modeling in SMT tasks. An extended target-side LM is obtained by considering additional information from alternative monolingual sources. This LM is actually computed according to one of the algorithms that implements an interpolation of independently computed in-domain and out-of-domain LMs. In the case of linear or log-linear interpolation, the weights of the combination are adjusted so that translation system performance is maximized with respect to a given test dataset.

The idea of LM adaptation for SMT has been around for years. For example, in [Hil05], information retrieval methods are used to form an adapted training corpus by selecting sentences similar to the test set. In [Xu07], a domain-specific phrase-based system is constructed exploiting a combination of feature weights to discriminate multiple domains. In [Sch08], several accurately selected target-side LMs are interpolated in a linear way. [Fos07] presented various LM combinations and analyze in detail the challenges to LM

domain adaptation.

One of the main motivations beyond the research work presented in this section is the opportunity to expand existing and optimized algorithms of LM *domain adaptation* to the wider problem of *task adaptation*. Here, we try to specialize the translation system trained on the "pre-edited" data to a different task of verbatim speech translation by optimizing the target LM, which is one of the features contributing to the N-gram-based SMT system.

3.2.1 Task description

LM task adaptation experiments have been conducted within the framework of the second open evaluation campaign[24] organized by the European TC-STAR project[25]. Two translation directions are considered: Spanish-to-English and English-to-Spanish.

The Spanish and English languages exhibit slightly different characteristics with respect to word order, for example, which may affect the role of the target LM.

Three input conditions were proposed to the participants:

- An EPPS (European Parliament Plenary Session transcription) plain text run (**FTE** - Final Text Edition);
- A **verbatim** dataset;
- The output of an automatic speech recognition system (**ASR output**)[26].

The **FTE** of an EPPS is a main run of the evaluation. It is a manually corrected clean transcription of European Parliament Plenary Sessions that slightly differs from the verbatim ones. Some sentences are rewritten. The text data do not include transcriptions of spontaneous speech phenomena. In this context the FTE run can be described as a "formal"-style translation of the original speeches given by politicians, which is not a verbatim transcription or "literal" translation but rather a text that aims for easy readability.

Verbatim transcription includes spontaneous speech phenomena (e.g., hesitations, false-starts, half-words, and corrections). The text data is case sensitive with punctuation marks.

[24] http://www.elda.org/en/proj/tcstar-wp4/tcs-run2.htm
[25] Technology and Corpora for Speech to Speech Translation
[26] See Appendix B for details about the project's framework, and see Appendix C for further details about participation in the international evaluation campaigns.

AST output is the "raw" run that is taken directly from the automatic speech recognizer. The text is also case sensitive, and punctuation marks are provided. The data are automatically segmented at syntactic or semantic breaks.

The difference between FTE, verbatim, and ASR texts is illustrated in the example below:

FTE: I am starting to know what Frank Sinatra must have felt like
Verbatim: I'm I'm I'm starting to know what Frank Sinatra must have felt like
ASR outuput: and i'm times and starting to know what frank sinatra must have felt like

In this study, we deal with the verbatim run.

3.2.2 Corpora

For this study, we had two task specific training corpora at our disposal: a Spanish-English parallel FTE corpus containing 1.3 M of running words and a smaller EPPS verbatim monolingual corpora with 70 K words for Spanish and 73 K for English. Both corpora belong to the same domain (i.e., plenary session transcription), but they are characterized by different edit levels. For a detailed description of the EPPS (EuroParl) FTE and EPPS verbatim corpora, refer to §A.1 and §A.2 in Appendix A, respectively.

For both tasks, only the first 500 sentences from the development datasets (*Dev 5K*) were used to find the best combination of weights when interpolating FTE and verbatim LMs; the same sentences were used to fine-tune the system models. Testing was conducted over the complete test sets.

3.2.3 LM task-dependent interpolation

Two main problems need to be solved to adapt a SMT system to a particular domain or task. The first one is how to build task-specific SMT systems when training; the second is how to perform task adaptation during decoding. For the first problem, we adapt a LM mixing procedure interpolating the task-dependent LMs with optimized weights. The second problem is solved in a straight-forward way by incorporating the interpolated LM into a set of feature models used by the decoder to replace the initial LM.

For each target language, we build two task specific LMs: $P_{LM}(LM \in 1,2)$, where P_1 refers to the EPPS FTE LM, P_2 to the EPPS verbatim LM. Following the maximum-entropy

LM optimization, we adjust the weight coefficients λ_1 and λ_2 ($\lambda_2 = 1 - \lambda_1$) to obtain an adapted LM:

$$P(w) = \lambda_1 \cdot P_1^w + \lambda_2 \cdot P_2^w \tag{3.1}$$

where P_1^w and P_2^w are probabilities assigned to the word sequence w by the LM estimated on FTE and verbatim data, respectively.

The mixing procedure was conducting using the linear interpolation algorithm and was implemented with the *n-gram* tool from the SRI language modeling toolkit [Sto02], which allows to read the second model for interpolation purposes and adjust weight coefficients when interpolating the principal model.

A straightforward way to optimize the weight coefficients of interpolated LMs is to use perplexity measured on a pre-selected development corpus, which is a popular and easy measure of LM quality. Perplexity can be intuitively interpreted as the geometric mean of the branching factor of language [Jel97]. A language with perplexity x has roughly the same difficulty as another language in which every word can be followed by x different words with equal probabilities.

However, it is not absolutely clear if perplexity is a good criterion to predict improvements when the LM is used in a SMT system. The LM's lack of direct influence at the word sequence production, aggregation and scoring steps makes it difficult to precisely correlate the perplexity of the LM with the metric of translation quality. Without a direct link between a LM's perplexity and the BLEU score of a translation system that uses it, guiding the construction of a LM for translation can be difficult.

Therefore, we optimize the coefficients directly as a function of the BLEU score measured on the output of the translation system, providing a more robust fine-tuning of the system.

However, we report perplexity results for information purposes. Even if perplexity does not always yield a high correlation with SMT systems performance, it is still an indicator of LM complexity as it gives an estimate of the average branching factor in a LM [Den05].

An adequate algorithm for such a task is the single-parameter *Expectation-Maximization (EM)* algorithm [Dem77], which mitigates this challenge by considering it an instance of an unsupervised learning technique. The EM algorithm was applied to optimize the BLEU score by iteratively adjusting the weights for FTE and verbatim LMs within the optimization loop. The EM iterative step is repeated until the improvement in BLEU is below 0.01%.

For clarity's sake, the EM algorithm was used on each iteration only to select the values

of λ_{FTELM} and $\lambda_{verbatim_LM}$. Furthermore, the FTE and verbatim LMs are interpolated using these coefficients, and the adapted LM, among other feature models under consideration, is passed to the simplex optimization module to implement the standard single-loop optimization algorithm, as described in [Cj09].

The flow diagram illustrating the optimization procedure can be found in figure 3.1.

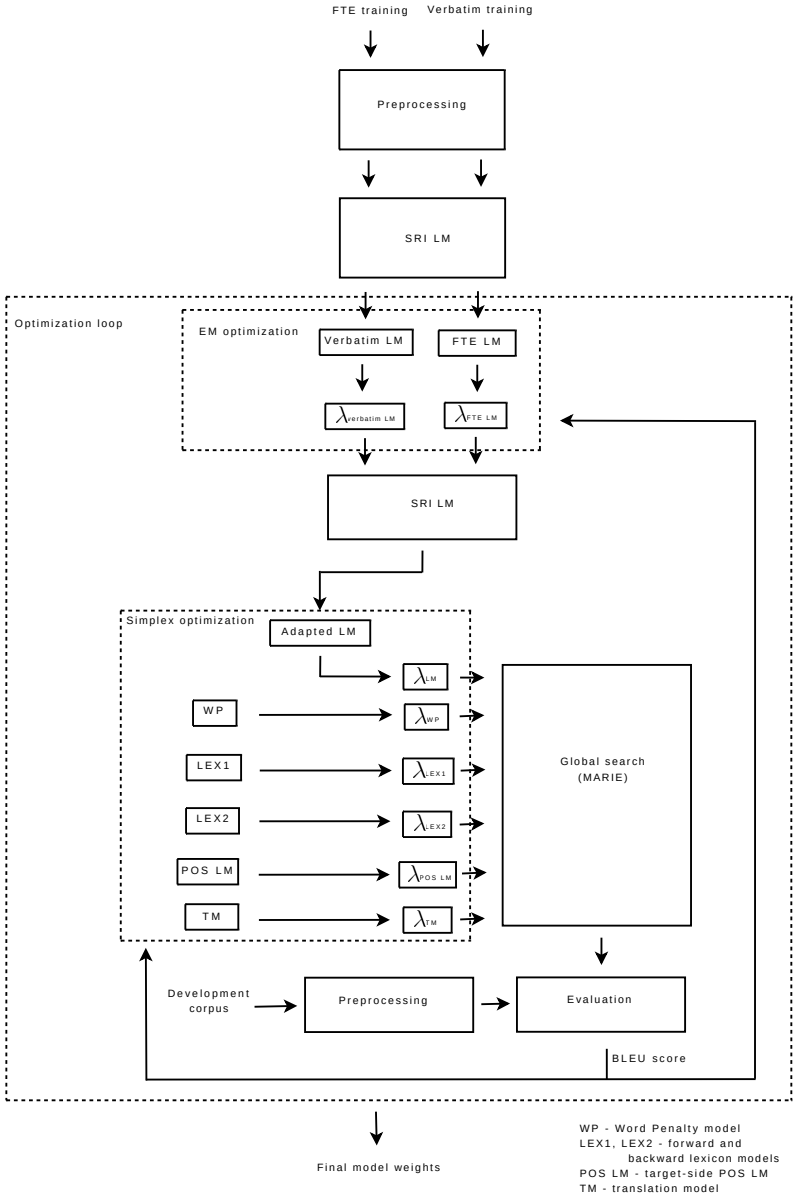


Figure 3.1: *SMT system with adapted LM. Optimization procedure.*

3.2.4 System description

Results are reported for verbatim N-gram-based translation systems trained on the FTE parallel training corpus using an adapted target-side FTE-verbatim LM in the log-linear feature combination. A summary of the most important system characteristics for each translation direction is presented in Table 3.1.

	Verbatim run	
	English-to-Spanish	Spanish-to-English
Word alignment	GIZA++	
Symmetrization	Union	
NULL-source tuples	IBM1 model	
Embedded words	Yes	
TM	5-gram, Kneser-Ney discounting	
LM	4-gram, Kneser-Ney discounting	
Other features	WP, LEX1, LEX2, POS LM	
POS LM	5-gram, Good-Turing discounting	
Decoding	Monotonic, beam=50 (10 during optimization)	
Pruning	Histogram, tnb=30	Histogram, tnb=20
Reordering	No	ABC
Optimization criteria	BLEU	

Table 3.1: *Verbatim system parameters.*

Notes to table:
ABC - word reordering based on alignment block classification, as described in [Cj08b],
tnb - the number of most frequent tuples for each source-side instance which are kept when decoding.

POS TNT tagger [Bra00] and Freeling tagger [Car04] were used for English and Spanish corpora tagging, respectively. All BLEU scores are case sensitive with punctuation marks considered.

Notice that target-side POS LMs were not interpolated since word class data is not informative enough to adapt a SMT system to a particular task of verbatim translation. More details can be found in §C.1.1 and [Mn06].

3.2.5 Experiments and results

We consider the evolution of system performance as a function of the weight combination corresponding to the FTE and verbatim target-side LMs within the adapted LM. The baseline (BL) is the FTE system without the use of the verbatim LM ($\lambda_{verbatimLM} = 0$). We also consider the use of only verbatim LM ($\lambda_{FTELM} = 0$).

Development results. The BLEU score obtained for the development sets as a result of the simplex optimization procedure for some FTE-verbatim LM interpolation points are presented in Tables 3.2 and 3.3. Best scores and corresponding system configurations are placed in cells filled with grey. Graphical representations of the obtained results are provided in Figures 3.3 and 3.2.

FTE	VBT	BLEU Dev
1.0 (BL)	0.0	44.05
0.9	0.1	44.39
0.8	0.2	44.01
0.5	0.5	43.90
0.2	0.8	43.52
0.1	0.9	43.17
0.0	1.0	42.03

Table 3.2: *Results for English-to-Spanish system (Dev).*

FTE	VBT	BLEU Dev
1.0 (BL)	0.0	50.53
0.9	0.1	50.66
0.8	0.2	50.84
0.5	0.5	50.76
0.2	0.8	50.92
0.1	0.9	50.99
0.0	1.0	50.97

Table 3.3: *Results for Spanish-to-English system (Dev).*

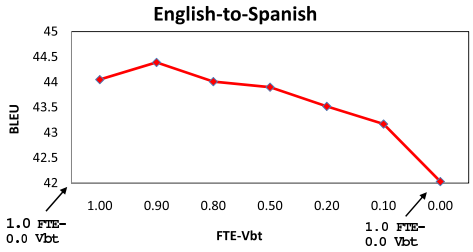


Figure 3.2: *Results for English-to-Spanish system (Dev).*

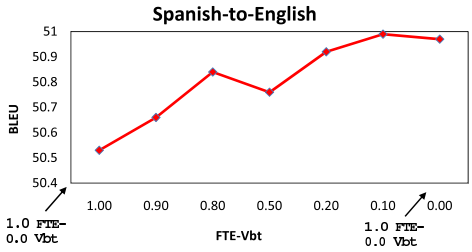


Figure 3.3: *Results for Spanish-to-English system (Dev).*

The performance of the translation system varies for the different LM configurations. The highest final BLEU score was given by the system corresponding to the "0.1 FTE-0.9 VBT" configuration for Spanish-to-English translation task and to the "0.9 FTE-0.1 VBT" for English-to-Spanish task. LM adaptation allowed a gain of about 0.5 BLEU points with respect to the baseline FTE system for the Spanish-to-English task and a gain of about 0.35 BLEU points for English-to-Spanish translation.

Test results and examples. We also investigated the BLEU scores obtained on the official test data of the 2nd TC-STAR open evaluation campaign. The results are summarized in Tables 3.4 and 3.5 and are depicted in Figures 3.4 and 3.5.

FTE	VBT	BLEU Test
1.0 (BL)	0.0	44.19
0.9	0.1	44.46
0.8	0.2	44.48
0.5	0.5	44.75
0.2	0.8	44.06
0.1	0.9	43.62
0.0	1.0	42.71

Table 3.4: *Results for English-to-Spanish system (Test).*

FTE	VBT	BLEU Test
1.0 (BL)	0.0	52.24
0.9	0.1	52.49
0.8	0.2	52.83
0.5	0.5	52.81
0.2	0.8	52.87
0.1	0.9	52.72
0.0	1.0	52.63

Table 3.5: *Results for Spanish-to-English system (Test).*

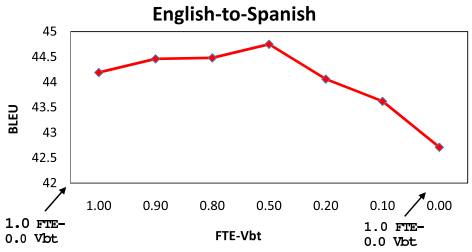


Figure 3.4: *Results for English-to-Spanish system (Test).*

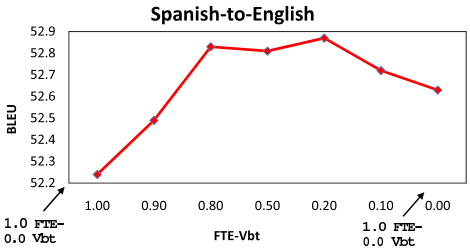


Figure 3.5: *Results for Spanish-to-English system (Test).*

It can be seen that despite the absolute maxima achieved on the test and the variance of the development corpora, the general trend of dependence and the shape of the graph *BLEU-interpolation point* are consistent with results obtained on the development corpus.

In contrast to frequently observed situations, when the improvements on test data are smaller than those on the development data, the N-gram-based system with adapted LM outperforms the BL by ≈0.6 BLEU points.

Regarding the Spanish-to-English task, the "0.0 FTE - 1.0 VBT" system again outperforms the system using FTE LM only by about 0.4 BLEU point. The optimal points maximizing the BLEU score on the test corpus are observed near the equal-weighted configuration.

Table 3.6 shows perplexity measured on a concatenation of both reference translations of the test corpus.

Configuration		Perplexity	
FTE	VBT	English-Spanish	Spanish-English
1.0 (BL)	0.0	63.33	72.44
0.9	0.1	61.27	67.63
0.8	0.2	63.21	67.51
0.5	0.5	75.43	72.65
0.2	0.8	111.51	90.67
0.1	0.9	148.01	106.64
0.0	1.0	150.11	138.74

Table 3.6: *Perplexity results for English-to-Spanish and Spanish-to-English systems.*

A comparison of LMs done on the perplexity basis allows for contrasting LMs with

respect to their usefulness for particular corpora. Minimal perplexity corresponds to the lowest uncertainty of language events as described by a given LM.

As expected, the system weighting the verbatim LM more than the FTE one shows poor perplexity for FTE data for both tasks. However, this system's configuration achieves surprisingly good performance for the Spanish-to-English track. In this case, translation is driven for the most part by the bilingual TM, while a low-vocabulary and high-perplexity specific-target LM is assigned with a low weight as a result of simplex optimization.

For the English-to-Spanish translation, the scores are more correlated with perplexity values. This is probably due to optimization procedure imperfections, which cannot exclude the negative factor of verbatim but only target-side LM. The best improvements were obtained for the points close to the FTE configuration that correlate with the minimal perplexity criteria.

In addition to perplexity and BLEU scores, we give an example of typical translations done by the FTE and adapted N-gram-based systems. Examples show how the adapted LM can have an influence on the translation decision and correct some errors caused by noisy input. Apart from simple input filtration (e.g., hesitations, repeated words, and other spontaneous speech effects removal), the adapted target-side LM manages to improve the fluency of the translation, as shown in the Spanish-to-English example.

In case of English-to-Spanish translation, adapted LM eliminated the hesitation ("*eh*") and produced the more fluent translation of the English phrase "*I have to say*", that is "*tengo que decir que*".

3.2.6 Conclusions

In this section of the chapter, we presented a brief study of the possible ways to improve verbatim texts translation by means of applying modified LM to the N-gram-based translation system. The baseline translation system represents a UPC-TALP submission to the TC-STAR 2006 Evaluation Campaign (verbatim run). We have discussed the task adaptation issue in the context of SMT systems and have proposed a statistical method to build task-specific SMT systems through task-dependent language modeling.

Here, task adaptation during the translation of test sets is implemented to solve monolingual text classification problems. We have demonstrated that the proposed technique provides better generalization for both considered target languages and corresponding translation tasks, namely English-to-Spanish and Spanish-to-English translation directions.

Spanish-to-English

Spa.: *no , no , señor Mote , no la he recibido .*
Gloss: no , no , Mister Mote , no it have received .
Eng. ref.: ' no , Mister Mote , I have not received it . '
FTE transl.: no , no , Mr Mote , not I have received .
Adapted LM transl.: no , Mr Mote , I have not received it .

A generated clause with wrong word order "*not I have received*" was corrected by the adapted LM to the more fluent phrase "*I have not received*".

English-to-Spanish

Eng.: *uh I have to say that I disagree fundamentally with that approach .*
Gloss: eh tengo que decir que discrepo totalmente con ese enfoque .
Sp. ref.: ' tengo que decir que discrepo totalmente con ese planteamiento . '
FTE transl.: eh he de decir que , fundamentalmente , discrepo con este enfoque .
Adapted LM transl.: tengo que decir que , fundamentalmente , discrepo con este enfoque .

The fact that the adapted LM can also be applied to other NLP tasks raises the prospect of using the presented algorithm for speech recognition and information retrieval.

As an alternative approach, an additional verbatim LM can be also included in the set of functions combined in a log-linear way.

3.3 Threshold-based target-side LM pruning

A well-known fact is that LMs can reach a tremendous size and sometimes cause a memory overflow problem. Since a LM is an integrated component or additional feature of a translation system, it significantly affects the system performance. A LM pruning strategy is definitely needed, as it reveals an *efficiency-performance* trade-off, which generally causes performance degradation for smaller models. However, carefully determined pruning strategies can significantly accelerate the translation process, save disk space, and even increase translation quality by means of reducing system noise.

While possible ways of TM threshold pruning were shown in [dG06], this part of the Ph.D. research is dedicated to a target-side LM pruning strategy based on rational threshold selection.

3.3.1 Target-side LM pruning experiments

The commonly-used strategy of *counts cut-off (threshold pruning)* [Goo00b] implies that all n-grams occurring less than a certain number of times are discounted to zero. According to this pruning model, long n-grams occurring less than a certain predetermined number of times in the training material are considered as important as all n-grams that do not occur at all.

A set of threshold values is defined for each n-gram order; in a "complete" system, the threshold would be 1 for all the n-grams. Large counts (i.e., numbers of occurrence) are taken to be reliable, and thus, they are not subject to any discounting. Therefore, this strategy is efficiently used for large n values, and it has as a consequence less costly models with improved performance. The unigram threshold is permanently set to 1, as we do not intend to reduce n-gram vocabulary.

The issue of LM size reduction was considered in many recent works (for example, [Gao02, Goo00a]). However, the interaction between LM pruning strategy and smoothing technique, which is normally applied to avoid zero probability estimates for unseen data, has received a lot less attention. The only study, to our knowledge, is [Sii07], where the results are given in terms of perplexity and cross-entropy, and are not directly related to SMT. It is shown in [Sii07] that n-grams seen only a few times can be discarded without significantly degrading the LM and we expected a similar behavior in relation to final translation scores.

The LM pruning was performed using the SRI language modeling toolkit [Sto02], which

enables users to set a minimal count of n-grams included in the LM for each n. The experiments were conducted on the EPPS Spanish-English corpus (FTE run). Corpus statistics can be found in §A.1 (Appendix A).

Major features of the translation system used in experiments are presented in Table 3.7. The BLEU score is case insensitive and includes punctuation marks.

	FTE run	
	English-to-Spanish	Spanish-to-English
Word alignment	GIZA++	
Symmetrization	Union	
NULL-source tuples	IBM1 model	
Embedded words	No	
TM	4-gram, Kneser-Ney discounting	
LM	3-gram, 4-gram, 5-gram unmodified Kneser-Ney discounting	
Other features	WP, LEX1, LEX2	
Decoding	Monotonic, beam=50 (10 during decoding)	
Pruning	Histogram, tnb=30	Histogram, tnb=20
Reordering	No	ABC
Optimization criteria	BLEU	

Table 3.7: *LM pruning experiments. N-gram-based SMT system parameters.*

The experimental results both for the development and official 2006 test datasets are shown in Table 3.8. Best scores and corresponding pruning configurations are placed in cells filled with grey.

Experiments on the unpruned high-ordered models have not been performed due to a lack of memory resources and decoder limitations. Consequently, the minimally pruned system configuration includes threshold 2 for 4-gram and 5-gram LMs and threshold 1 for low-order n-grams.

The LM configuration that provides the best trade-off between BLEU score and model size is the 4-gram model with thresholds set to 2 for 4- and 3-grams and 1 for unigrams and bigrams (that is, the **4-2211** system configuration).

3.3.2 Discussion and conclusions

SMT decoding can be considered a computationally intensive process in which model size is a crucial factor that has a significant influence on decoding time. The threshold setting considerably reduces the model size, while the BLEU score remains constant or even

N-gram order	Pruning threshold				BLUE		Model size, millions				
	2	3	4	5	Dev	Test	1	2	3	4	5
Spanish-to-English											
3	1	1	-	-	65.28	56.41	0.11	2.29	9.43	-	-
	1	2	-	-	65.18	56.84	0.11	2.29	2.95	-	-
	2	2	-	-	65.20	56.38	0.11	2.29	2.95	-	-
4	1	1	2	-	65.50	56.66	0.11	2.29	9.43	3.74	-
	1	2	2	-	65.59	56.81	0.11	2.29	2.75	3.74	-
	2	2	2	-	65.55	56.74	0.11	0.99	2.75	3.74	-
5	1	1	2	2	65.32	56.85	0.11	2.29	9.43	3.45	3.40
	1	2	2	2	65.10	56.82	0.11	2.29	2.75	3.45	3.40
	2	2	2	2	65.18	56.42	0.11	0.99	2.75	3.45	3.40
English-to-Spanish											
3	1	1	-	-	55.94	49.63	0.14	2.52	9.52	-	-
	1	2	-	-	55.38	50.23	0.14	2.52	2.93	-	-
	2	2	-	-	55.64	49.79	0.14	1.03	2.93	-	-
4	1	1	2	-	55.79	49.71	0.14	2.52	9.52	3.86	-
	1	2	2	-	56.07	50.13	0.14	2.52	2.68	3.86	-
	2	2	2	-	55.90	49.57	0.14	1.03	2.68	3.86	-
5	1	1	2	2	55.84	50.07	0.14	2.52	9.52	3.52	3.67
	1	2	2	2	55.54	49.69	0.14	2.52	2.68	3.52	3.67
	2	2	2	2	55.49	49.93	0.14	1.03	2.68	3.52	3.67

Table 3.8: *LM pruning experiments. Model sizes and BLEU scores on the development and test data.*

increases.

Just as in the case of TM pruning described in [dG06], setting thresholds to 2 produces a very important model size reduction, whereas translation performance remains stable or even increases for both tasks. This means that about 7 million of the 3-grams turn out to be more useless than useful as part of the N-gram-based MT system. It allows for the reduction of the model size without losing translation quality.

In the following experiments, the LM pruning strategy of setting the threshold 1 to unigrams and bigrams and 2 to all the higher-order n-grams (if any) will be adopted.

Another conclusion that can be drawn from this study is that the increase in LM history length does not directly improve the system performance, as no correlation between

LM order and final BLEU score is observed. It can possibly be explained by the specific character of the N-gram-based system and the particular role played by the target-side LM as an additional feature. The results presented in this section correlate with the research findings from [Goo00a, Sii07] concerning effective LM pruning strategy and demonstrate the usefulness of the threshold LM pruning when applied to SMT.

An interesting area to work on in the future is to include a threshold-based pruned LM into the set of feature functions of the phrase-based SMT system.

3.4 Neural network language modeling

In the third part[27] of this chapter, we step aside from traditional language modeling techniques and introduce a continuous-space LM based on a neural network to exploit its ability to learn distributed representations in order to reduce the impact of the curse of dimensionality. We also show that this can be used to improve an N-gram-based SMT system using an example of a small Italian-to-English translation task.

3.4.1 Motivation and computational issues

Problem discussion

Regardless of the approach that a SMT system follows, it typically takes as its basis a maximum entropy approach in which the target language sentence is seen as distorted by the channel conditioned by a set of feature functions in the foreign language. This combination normally includes a target-side LM, which informs a translation decoder and provides it with an idea of the correctness of a given sentence and, in our case, of the fluency of the translation hypothesis.

The approach presented in this section can be considered a coherent and natural evolution of the probabilistic LMs. We propose to use a continuous-space LM that deals better with the smoothing challenge and thereby provides better generalizations to unknown n-grams.

While the use of a continuous-space representation of a language has been successfully applied in recent NN approaches to language modeling [Xu04, Ben03, Cas03] and speech recognition [Sch07a], the neural network language model (NN LM) application in the state-of-the-art SMT systems is not so popular and can be traced back to the works done in LIMSI[28], in which the NN LM was applied both to train a target-side LM [Sch06] in the form of a fully-connected multilayer perceptron, as well as to smooth the probabilities involved in the bilingual tuple TM [Sch07b].

[27] This work was done in collaboration with the Department of Information Systems and Computation at the Technical University of Valencia with F. Zamora-Martínez, M.J. Castro-Bleda and S. España-Boquera

[28] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

Motivation

In this work, we address the challenge of MT by using an Italian-English parallel corpus with a limited amount of training material (§A.3). This translation task is characterized by an extremely limited amount of training data (about 150 K of tokens in the training corpus), a similar but slightly different word order and distinct inflectional characteristics.

The heavy tailed structure of any modern natural language results in the fact that one is likely to encounter new n-grams that were never witnessed during training. The n-gram LM is often criticized because they lack any explicit representation of dependencies longer than $n-1$ tokens, while the effective range of dependency is significantly longer than this, although long range correlations drop exponentially with distance for any Markov model. We address the problem of LM smoothing in a continuous domain using a connectionist LM trained in a neural network.

A major difference between classical n-gram LM and the approach we are following to train a NN LM lies in the distinct mechanism used to implement a smoothing process. Unmodified Kneser-Ney discounting, which is the smoothing algorithm used for the n-gram models in the framework of the study, is an extension of absolute discounting. The idea behind this smoothing algorithm is to optimize the LM taking into account the fact that the lower-order model is significant only when count is small or zero in the higher-order model [Jam00, Che99]. In contrast to interpolated smoothing models[29] (for example, Jelinek-Mercer or interpolated Chen-Goodman models described in [Che99]), Kneser-Ney back-off model does not use information from lower-order models in determining the probability of n-grams with non-zero counts, however it does in determining the probability of n-grams with zero counts.

Within a NN LM, posterior probabilities are interpolated for any possible context of length $n-1$ rather than backing-off to shorter contexts. We expect a better performance from the NN LM in comparison with the n-gram LM with unmodified Kneser-Ney discounting since the former allows to capture all possible $n-1$-gram combinations seen in the training corpus instead of the $n-1$-grams preceding the word under consideration.

Unfortunately, a computational problem arises here. In general, a NN LM has a complexity of $O(NH)$, where N is the size of the word vocabulary, and H is the size of the hidden layer of the NN (which is, in practice, much smaller than the vocabulary). This complexity quickly overwhelms modern computational resources for even average-size vocabulary tasks

[29] See experimental section §3.4.4 for more details.

and thus noticeably limits the area of NN LM application.

However, Zipf's law [Zip49] states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Consequently, the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, and so on. This observation opens the way to limit the input and output of the NN to the m most frequent words from the vocabulary; that is, a shortlist can be created without significant loss of generality. Therefore, the NN LM is believed to be most appropriate for tasks with limited resources.

In our experiments, we mostly concentrate on the issue of shortlist and n-gram order selection by searching for an optimal trade-off between NN LM training time and SMT system performance.

3.4.2 Neural network language models

Model architecture

A NN LM is a statistical LM that follows the same equation as n-grams estimates the LM probability for a sequence of words of length $|W|$ in the following way:

$$p(w_1 \dots w_{|W|}) \approx \prod_{i=1}^{|W|} p(w_i | w_{i-n+1} \dots w_{i-1}) \tag{3.2}$$

and where the probabilities that appear in that expression are estimated within a NN. The model naturally fits under the probabilistic interpretation of the outputs of NNs. If a NN is trained as a classifier, the outputs associated to each class are estimations of the posterior probabilities of the defined classes. The demonstration of this assertion can be found in a number of places, for example, [Bis95] and [Ben03].

The training set for a LM is a sequence $w_1 w_2 \dots w_{|W|}$ of words from a vocabulary Ω. In order to train a NN to predict the next word given a history of length $n-1$, each input word must be encoded. A natural representation is a local encoding following a "1-of-$|\Omega|$" scheme. The problem for this encoding regarding tasks with large vocabularies (as is often the case) is the huge size of the resulting NN. We have solved this problem following [Ben03] by developing a *distributed representation* for each word.

A general definition of a distributed representation of a word is a vector of features that

characterizes the meaning of the word and are not mutually exclusive [Lar06]. In the case of language modeling without additional word markers, it is transformed into a set of indices that unambiguously characterizes the words in a vocabulary.

The idea behind NN LM is to project these word indices onto a continuous space by using a probability estimator to smooth this space. Since the resulting probability functions are smooth functions of the word representation, a better generalization to the unknown n-grams can be expected.

Figure 3.6 illustrates the architecture of the feed-forward NN used to estimate the NN LM. The input is composed of words $w_{i-n+1}, \ldots, w_{i-1}$ of equation 3.2 (for example, the input words are w_{i-3}, w_{i-2}, and w_{i-1} for a 4-gram).

Each word is represented using local encoding. P is the projection layer of the input words formed by $P_{i-n+1}, \ldots, P_{i-1}$ subsets of projection units. The subset of projection units P_j represents the distributed encoding of input words w_j. The weights of this projection layer were linked; that is, the weights from each local encoding of input word w_j to the corresponding subset of projection units P_j are the same for all input words j.

H denotes the hidden layer, while the output layer O has $|\Omega|$ units, one for each word of the vocabulary. Trained as a classifier, this NN predicts the posterior probability of each word of the vocabulary given the word history of word w_i, i.e., $p(w_i|w_{i-n+1} \ldots w_{i-1})$.

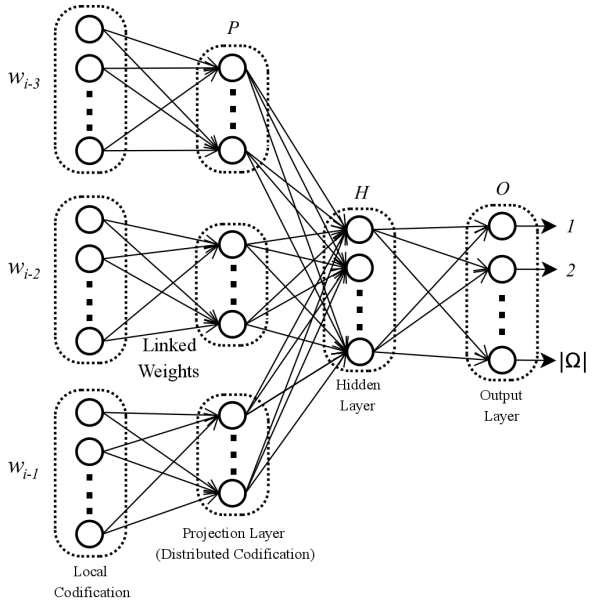


Figure 3.6: *Architecture of the continuous-space NN LM.*

NN LM		NN Topology	# Weights
Vocabulary	n-gram	Input–Projection–Hidden–Output	
k=5 2,148	3-gram	2×2,148–2×32–64–2,148	2×68,768 + 143,788
	4-gram	3×2,148–3×32–64–2,148	3×68,768 + 145,828
k=3 3,093	3-gram	2×3,093–2×32–64–3,093	2×99,008 + 205,205
	4-gram	3×3,093–3×32–64–3,093	3×99,008 + 207,253

Table 3.9: *Selected NN LM configurations: size and configuration.*

In order to achieve an adequate configuration in terms of topology and parameters for each translation task undertaken by the NN LM, exhaustive scanning using a fine-tuning set was performed. The activation function for the hidden layers was the *hyperbolic tangent* function, and the *softmax* function was chosen for the output units. Best configurations used a projection layer of 32 units for each word.

To illustrate the huge sizes of the NNs used, Table 3.9 shows the topology and number of weights of the selected NN LMs to be 2,148 words (words with less than k=5 occurrences were discarded from the BTEC corpus) and 3,093 (corresponding to k=3), respectively. The third columns shows the topology of the used NNs, including the number of input, projection, hidden and output units, and the last column shows the number of weights. First, the weights are replicated $n-1$ times at the projection layer, and second, the weights are shown at the hidden and output layers.

Rescoring

A NN LM model is integrated in the N-gram-based SMT system within a discriminative rescoring/reranking framework, which incorporates complex feature functions by using the entire translation hypothesis to generate a score.

During the first step, the MARIE decoder produces a list of k candidate translations[30] based on the weights vector trained over the m basic features (excluding orthodox n-gram LM in order not to obscure the NN LM effect). Then, the statistical scores of each generated translation candidate are rescored using information provided by the NN LM that presumably should add information not included during decoding to better distinguish between

[30] In all NN LM experiments k was set to 1,000.

higher and lower quality translations. During this step, a rescoring vector is trained over $k+1$ features and provides different, better choices for the single-best translation hypothesis.

An alternative way of incorporating NN LM into a SMT system is to use the continuous-space LM directly during decoding. We decided not to pursue this strategy since this would result in a dramatic increase of decoding time.

3.4.3 Experimental setup

The experiment results were obtained using the Italian-English 2006 BTEC corpus (see §A.3), which is a collection of spoken dialogue data. Along with regular sentences, like "Questo traghetto si sta dirigendo verso un'isola" (*"This ferry is heading for an island."*), it contains many colloquial or simple expressions, such as "Hm! non mi sento bene." (*"Hm! I am not feeling well."*).

The Italian part of the bilingual corpus was preprocessed. This step included tagging, lemmatization, and separation of contractions as described in [Cre06a].

3.4.4 Baseline system

The *baseline* system characteristics are summarized in Table 3.10. To provide reasonable comparison with NN LM experiments, we consider the use of the regular LM for rescoring in the same way as the NN LM (integrating the n-gram LM on the rescoring step).

Alternatively, we consider the inclusion of the regular LM as a feature in the set of functions combined in a log-linear way during decoding (*Dec*). Results shown by the (*Dec*) system correspond to the performance of a standard N-gram-based SMT.

Unmodified Kneser-Ney discounting was chosen to compute a smoothed n-gram LM since it has demonstrated the best results in terms of perplexity and the final translation score (BLEU) measured on the concatenation of the reference translations (development dataset). We compared the original Kneser-Ney discounting with Good-Turing and Chen-Goodman (uninterpolated and interpolated versions) discounting algorithms. Application of the unmodified Kneser-Ney technique demonstrated significant improvement in perplexity ($\approx 12\%$) and translation quality according to the BLEU score ($\approx 3.9\%$) in comparison with alternative smoothing algorithms. Interpolation of higher- and lower-order n-grams has no positive effect on the MT scores.

Automatic evaluation conditions were case-sensitive and included punctuation marks.

	Italian-to-English
Word alignment	GIZA++
Symmetrization	Union
NULL-source tuples	Entropy of POS distribution
Embedded words	No
TM	4-gram, Kneser-Ney discounting
LM	4-gram, unmodified Kneser-Ney discounting (during rescoring)
Other features	WP, LEX1, LEX2
Decoding	Monotonic, beam=50 (10 during optimization)
Pruning	Histogram, tnb=30
Reordering	Input reordering graph
Optimization criteria	100BLEU+4NIST

Table 3.10: *Baseline system parameters. NN LM experiments.*

Note to table:
Input reordering graph - word harmonization algorithm as described in [Cre06b]. More details can be found in chapter 4.

3.4.5 Continuous-space LM experiments

We considered two key parameters of the continuous-space NN LM:

- *A shortlist size* defines a word frequency threshold t that implies that all of the words occurring less than t times in the training corpus are discarded;
- *An n-gram order* limits a word history to n preceding words. 3- and 4-gram configurations were tested.

When re-estimating the weight coefficients for the new log-linear model with the NN LM, different starting points were tried, and the best set of weights resulted from the 100 BLEU + 4 NIST criteria. Table 3.11 and Table 3.12 show the BLEU, NIST and METEOR scores when the NN LMs were integrated as a part of the combined SMT system for the development and the test sets.

As can be observed, considerable improvements were obtained using a NN LM. The best system configuration is highlighted in both aforementioned tables.

		BLEU	NIST	METEOR
Baseline		29.09	6.19	69.22
Dec		29.22	6.37	69.26
NN LM k=5	3-gram	30.02	6.31	69.44
	4-gram	30.07	6.17	69.19
NN LM k=3	3-gram	30.54	6.44	69.61
	4-gram	30.01	6.10	69.45

Table 3.11: *Evaluation scores on the development dataset.*

		BLEU	NIST	METEOR
Baseline		24.79	5.80	63.91
Dec		24.93	5.83	64.01
NN LM k=5	3-gram	25.17	5.86	63.70
	4-gram	25.07	5.79	63.99
NN LM k=3	3-gram	25.23	6.02	64.10
	4-gram	25.29	5.81	63.63

Table 3.12: *Evaluation scores on the test dataset.*

For the development dataset, the BLEU score for the NN LM experiments is higher than the one for the baseline system for all NN LM systems. Concerning the METEOR score, again all of the scores produced by the NN LM systems are slightly higher than the reference one.

Our previous experience shows that for small translation tasks with a lack of training material, poor correlation of development and test results is frequent, although this has not been the case in these experiments. Besides, the improvements on the test data are usually smaller than those on the development data that was used to fine-tune the parameters.

The BLEU and METEOR scores calculated for the test dataset are improved when the NN LM is applied in comparison with the baseline level, while the METEOR values generated by the NN LM configurations vary around the score produced by the system integrated with a conventional n-gram LM.

As stated in §3.2, the output sentences from a SMT system are built by aggregating word sequences that have a high-scoring combination of probabilities provided by TM and a

set of feature models, including LM. Therefore, this is not a clear breakdown of the impact of the LM perplexity on the assembled translation. However, perplexity is a measure of a LM's predictive power, which can be used to compare how well a LM can predict the next word in a previously-unseen piece of text.

Table 3.13 represents perplexity values for stand-alone LMs measured on the merged set of translation references of the test corpus.

Language Model	Parameters	Perplexity
Conventional 4-gram	-	103.52
NN LM 3-gram, k=5	281,316	88.54
NN LM 3-gram, k=3	403,221	101.08
NN LM 4-gram, k=5	352,132	91.04
NN LM 4-gram, k=3	504,277	100.71

Table 3.13: *Perplexity results for different language models.*

3.4.6 Discussion and example

The architecture of a SMT system implies that the smaller the amount of available training data is, the worse is the performance of a translation system. Obviously, new or specially-adapted methods to use limited information in more efficient way are needed. The technique presented in this section improves the performance of a SMT system by incorporating the NN LM when only a small amount of training material is available.

Considering the development and test data translation scores, the 3-gram k=3 NN LM system allows a gain up to 0.3 BLEU points for the test set over the system that includes conventional n-gram LM as a feature in the decoder (*DEC*); and a gain of about 0.4 BLEU points for the test dataset over the system that use the regular LM for rescoring (*baseline*). This difference is statistically significant for a 95% confidence interval and 1,000 re-samples using the bootstrap re-sampling method, as described in [Koe04].

Considering the NIST score, the baseline test results are exceeded for both 3-gram systems. Concerning the METEOR score, only the 3-gram, k=3 system provides a better LM generalization.

Perplexity values characterizing the 3-gram and 4-gram NN LMs with a word frequency threshold set to 3 are comparable with the results shown by the conventional n-gram model. A very important reduction (10-14 %) is observed with an increase in the size of the shortlist

that includes the k most frequent words. However, as was seen in previous experiments, it does not influence the translation results to the same extent.

The correlation of automatic and subjective human evaluation metrics (fluency and adequacy) is one of the main topics in the area of MT evaluation. As was reported in [Pau06] for small BTEC translation tasks, fluency correlates best with BLEU, while adequacy correlates best with METEOR. The NIST metric has only a moderate correlation to both subjective human evaluation metrics. Our work demonstrates the potential for the application of NN LMs to SMT systems to improve translation fluency, while adequacy remains the same. The positive impact of higher n-gram is not clear, and this is possibly due to the relatively short sentences provided within the BTEC corpus. Another possible issue is that higher n-gram order only slightly decreases translation quality, yet at the same time, it introduces more noisy translation hypotheses.

An example of a typical sentence from the BTEC corpus is shown in Figure 3.7. The Italian expression *"Oggi abbiamo a scelta"* is translated by the baseline system as *"Today we have selection at"*, whereas three of four NN LM systems provide a more fluent translation *"Today we have to choose from"*.

Source	Oggi abbiamo a scelta insalata ai frutti di mare insalata di patate e insalata mista.
References	Today we have a choice of seafood salad potato salad and wild vegetables salad. We are serving seafood salad potato salad and wild vegetables salad today. As for today's salad you can enjoy seafood potato and wild vegetables. For salad we have seafood potato and wild vegetables today. Today's selections are the seafood salad potato salad and wild vegetables salad. For today we have the seafood salad potato salad and wild vegetables salad. For today you can choose to have the seafood salad the potato salad or the wild vegetables salad.
Baseline	Today *we have selection at* the seafood salad potato salad and mixed salad.
3-**gram** k=**5**	Today *we have to choose from* the seafood salad potato salad and mixed salad.
4-**gram** k=**5**	Today we have selection at the seafood salad potato salad and mixed salad.
3-**gram** k=**3**	Today *we have to choose from* the seafood salad potato salad and mixed salad.
4-**gram** k=**3**	Today *we have to choose from* the seafood salad potato salad and mixed salad.

Figure 3.7: *An example of translation.*

In this section, we show the robustness of the NN LM, even for a highly limited training

corpus. The in-domain NN LM provides a significantly better generalization of the target language, better smoothed SMT output and enhanced improvement in the automatically-evaluated translation scores.

A main disadvantage of the continuous-space LM is that it has a very high computational cost. While traditional n-gram LMs can be trained in few minutes using the SRI language modeling toolkit, it can take several days to estimate a continuous-space LM for a large-vocabulary task. A possible solution to this problem can be the application of fast-training techniques (lattice regrouping and the utilization of specialized NN libraries with an ability of parallel calculation). However, at the moment, low-vocabulary tasks with a lack of training data appear to be the most appropriate domains for NN LM application.

Related publications:

- M. Khalilov **Target language modeling improvement techniques for statistical machine translation**. Proceedings of the Doctoral Consortium at the 8th EUROLAN Summer School, pp. 39-45, Iasi (Romania), July-August 2007.
- M. Khalilov and J.A.R. Fonollosa **Language modeling for verbatim translation task**. Proceedings. of the IV Jornadas en Tecnología del Habla - the IV Biennial Workshop on Speech Technology, pp. 83-87, Zaragoza (Spain), November, 2006.
- M. Khalilov, J.A.R. Fonollosa, F. Zamora-Martínez, M.J. Castro-Bleda and S. España-Boquera **Arabic-English translation improvement by target-side neural network language modeling**. Proceedings of HLT&NLP within the Arabic World International Workshop at LREC'08, Marrakech (Morocco), May 2008.
- M. Khalilov, J.A.R. Fonollosa, F. Zamora-Martínez, M.J. Castro-Bleda and S. España-Boquera **Neural Network Language Models for Translation with Limited Data**. Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, pp. 445-451, Dayton, Ohio (USA), November 2008.

Chapter 4

Word reordering problem

One of the most challenging problems facing MT is how to place the translated words in such order that they fit the target language. Some languages, like English or Spanish, have relatively restrictive word orders and follow more monotone mutual word order than, say, Chinese and English. Others, for example Slavic or Baltic languages, allow more flexibility in word order, which, in many cases, serves to define the relationship between the actions and the entities.

Divergent glottogony of languages is the main reason for the word ordering problem, especially when dealing with Asiatic and European languages. Study of the origins of the languages can help to understand different ways in which languages arrange the constituents of their sentences relative to each other, but it offers limited help in finding a solution for the reordering problem.

The phrase and to a greater extent the tuple internal reordering, in conjunction with distance-based distortion, implying restrictions in the search space of translation units, has provided SMT with improved robustness in *local reordering*, when the words to be reordered are adjacent or close within the sentence.

Spanish-to-English

Spa.:	*un*	***programa***	*espicífico*	*y*	*local*
Gloss:	a	**program**	specific	and	local

Eng. ref.: ' a specific and local **program** '

Figure 4.1: *Example of short-distance (local) reordering.*

Figure 4.1 demonstrates a typical difference in Spanish and English word order: the

Spanish adjective usually comes after the noun, unlike its English counterpart. So, in order to allow for monotone translation, the original Spanish phrase should be transformed to '*un espicífico y local programa*', which is not a natural word sequence in Spanish.

The problem is especially important if the distance between words that should be reordered is high (*long-distance or global reordering*); in this case, the reordering decision is very difficult to make based on statistical information, due to the dramatic expansion of the search space with the increased number of words involved in the search process[31].

Long-distance reordering is exemplified by Figure 4.2 for Arabic-English[32] language pair, where *'AElnt/announced*', which is a typical verbal structure in Arabic, should move to the right to obtain the English word order, as seen in the gloss.

Arabic-to-English

Ar.:	***AElnt***	*Ajhzp*	*AlAElAm*	*l*	*bEvp*	*fy*	*syrAlywn*	*An*	*...*
Gloss:	**announced**	press	release	by	mission	in	sierra leone	that	...

Eng. ref.: ' a press release by the mission to sierra leone **announced** that ... '

Figure 4.2: *Example of long-distance (global) reordering.*

A monotone SMT system often suffers from weakness in the distortion model, even if it is able to generate correct word-by-word translations. The problem here is that one syntactic and semantic unit in the source language might appear in a different position in the target language. There are no powerful mechanisms incorporated within a monotone SMT system to efficiently handle different word order if this word order disparity is not found within the limits of a multiword translation unit (*internal reordering*), as shown in Figure 4.3. The system implicitly memorizes each pair of source and target phrases in the training stage.

For the majority of translation tasks, however, word reordering disparity cannot be modeled with standard translation units, and additional extended techniques need to be employed.

This distortion-restricted rearrangement of translation units, conducted by the decoder or done prior to the translations, is called *external reordering*.

[31] Global/local reordering classification is extremely subjective and is mostly determined by the level of computer technology development, which limits the capability of the distance-based reordering model. At the moment, a regular SMT system defines the constraints of the distance-based distortion such that a sequence of five to seven words might be involved in the word reordering process.

[32] Hereafter, all the Arabic translations are provided in the Buckwalter transliteration [Buc94]

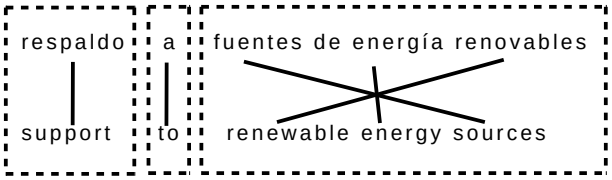


Figure 4.3: *Example of internal reordering.*

One of the fundamental processes underlying any natural language is a linguistic topology defined in terms of the finite verb (V), its subject (S) and its object (O). Different languages follow different word topology schemes. For example, English is mostly SVO, Spanish more often - SVO (although non-SVO orders are extremely common in native varieties of Spanish), while VSO is the most frequent order in Arabic, but SVO is almost equally common.

Topological disparity leads to particularly bad translation performed by monotone SMT systems and to the need for global reordering models with the capability to model long-range dependencies. Because a monotone translation approach often cannot deal with the necessary reordering to resolve such disparities, we used a parse tree structure in our own work. A detailed description of our parse tree approach will be given in chapter 5.

The word reordering problem has attracted a great deal of attention recently. There have been abundant publications on purely statistical techniques dealing with the word reordering challenge, as well as on approaches involving lexical information (context) or using additional information to reorder the target words in such a way that they fit the target language.

In this chapter, we will take a closer look at the state-of-the-art reordering models employed in phase-based and N-gram-based SMT, and we will describe the empirical classification of the most well-known distortion models for SMT.

4.1 State-of-the-art reordering approaches

In practice, a reordering model operates on a sentence level and is carried out based on word-reordering rules derived from the training corpus or motivated by the structural differences between the source and target languages.

Figure 4.4 presents the ways word reordering is approached in modern SMT systems. Due to the high complexity of many of the reordering methods, there are no clear criteria to determine the boundaries between categories; therefore, this classification is very subjective, and some of the considered algorithms cannot be unambiguously categorized.

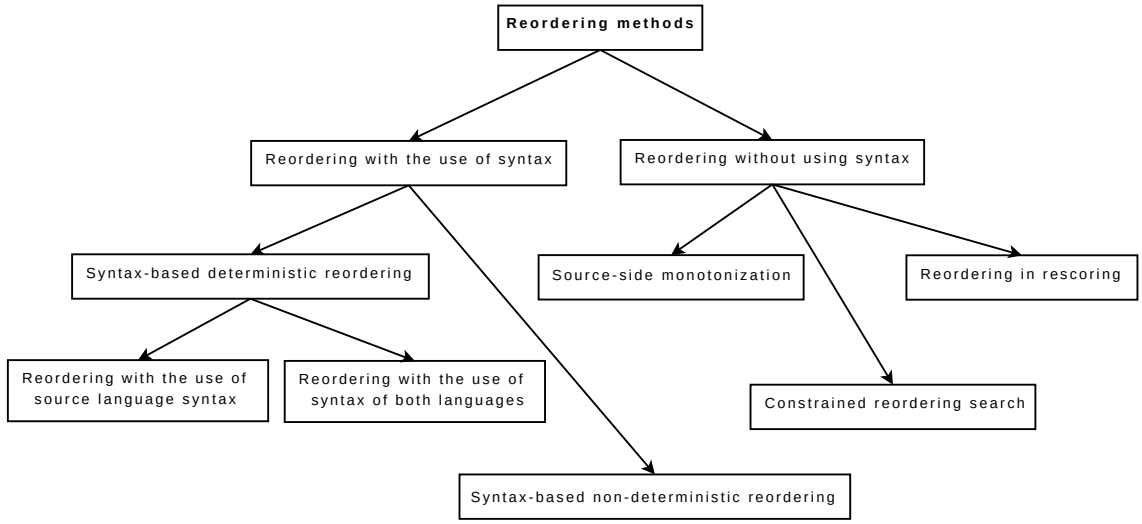


Figure 4.4: *Classification of state-of-the-art reordering algorithms.*

Alternatively, reordering models can be divided into two large groups depending on the reordering units with which the algorithm operates. Some of the methods (mostly dependent on the reordering of the phrase pairs from the alignment matrix) operate with bilingual units (phrases or tuples), others with words.

4.1.1 Statistical reordering methods

Constrained Reordering Search

In most cases, the generation of a translation hypothesis is computationally expensive. A brute force approach to reordering looks through all possible combinatorial permutations of the processed sets of source sentence words (construction of a fully reordered search graph), which results in a dramatic increase of decoding time with longer sentences. As shown in [Kni99], the search problem is NP-hard if arbitrary word reorderings are permitted. On the other hand, a polynomial-time search algorithm can be obtained if the reordering constraints are defined in an appropriate way.

Reordering constraints aim to limit the search space with minimal loss of generality during decoding and to introduce a balance between computational efficiency and translation quality. The possible ways to restrict word reorderings are briefly discussed below:

- The **IBM** constraint is intended to make the search feasible by introducing restrictions of the search space at the word level in the spirit of the IBM constraints [Ber96b, Til03]. A coverage vector is kept to mark the source positions that have already been placed. At each step, the selection of the next word to place is made among the first k yet-uncovered word positions.

- The **ITG** (Inverse Transduction Grammar) constraint, proposed in [Wu96, Wu97], allows a polynomial-time search algorithm. This constraint has been demonstrated to be useful for SMT, as shown in [Ben04]. The input sentence is interpreted as a sequence of word blocks. For each two adjacent blocks, a decision is taken either to invert the original order, or to leave it as is. A systematic comparison of the ITG and the IBM constraints can be found in [Zen03].

- A notable evolution of the ITG constraint is the **maximum entropy** model, transforming the reordering prediction into a classification problem, providing phrasal reordering depending on context, while its generalization capability is based on features automatically learned from a parallel text [Xio06].

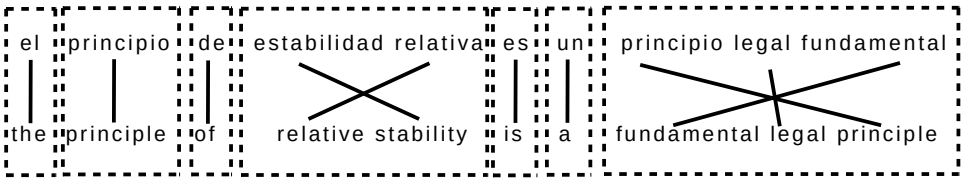
- The **local** constraint, presented in [Kan05], has proved to be very efficient for language pairs in which words are only shifted a few positions to the left or to the right. This constraint is a simplification of the IBM constraint allowing for local permutations only. Here, the next word must be contained within the first k words, starting from the first uncovered position.

- The **MaxJumps** constraint numerically limits the number of reorderings and the locality boundaries specified by two parameters [Cre05d]:

  - m - a maximum distance measured in words, that a source word, phrase, or tuple can be reordered (*a distortion limit*)
  - j - a maximum number of "jumps" within a sentence (*a reordering limit*)

  This algorithm requires the bilingual units (tuples, in particular) to have been extracted with using the *unfolding* technique described in detail in [Cre05d]. It allows

the generation of shorter tuples, increasing the system's reordering flexibility if a decoder is enhanced with reordering capabilities and, at the same time, alleviating the problem of embedded units (see §2.3.4). A comparative example of regular and unfolded tuples extraction is provided in Figure 4.5 for Spanish-to-English translation.



Regular tuples:

el#the
principio#principle
de#of
estabilidad relativa#relative stability
es#is
un#a
principio legal fundamental#
fundamental legal principle

Unfolded tuples:

el#the
principio#principle
de#of
estabilidad#stability
relativa#relative
es#is
un#a
principio#principle
legal#legal
findamental#fundamental

Figure 4.5: *Bilingual tuples extracted with regular and unfolded methods.*

The problem of word reordering has been approached since the origin of the modern era of MT: [Ber96b] introduced in their alignment models what they called distortion models, in an effort to include in their SMT system a solution for the reordering problem.

Word class-based reordering patterns were part of the alignment template system [Och04] and the classical phrase-based system [Koe03]; the RWTH template system [Ben04] follows a similar approach. While local reorderings within alignment templates are fixed in training, reordering of alignment templates tends to be ITG restricted for difficult translation tasks. Non-monotonic phrase alignments are penalized depending on the word distance between successively translated source phrases.

The main criticism of this approach is that it shows poor performance for pairs of languages with very distinct word order; it is explicitly appropriate primarily for local reorderings [Cre08a]. Another clear disadvantage of this technique is that the orientation of the phrase alignments and the contexts of the source and target phrases are not taken into account.

More recently, several attempts have been made to introduce global reordering into an SMT system. In [Nag06], the reordering is predicted based on the current phrase pair and the previous one, through introducing four types of reordering patterns, namely MA (monotone adjacent), MG (monotone gap), RA (reverse adjacent), and RG (reverse gap).

The modern state-of-the-art translation system Moses, along with a distance-based distortion model [Koe03], implements the so-called *lexicalized reordering* [Til04, Til05, Koe05b]. According to this model, in the first step the boundaries of bilingual phrase clusters of words are established. Later on, a probabilistic and lexicalized MSD (Monotone-Swap-Discontinuous) block-oriented model is learned from training data. During decoding, translation is viewed as a monotone block sequence generation process with the possibility of swapping a pair of neighbor blocks. Again, this approach is context-dependent, and reordering patterns are related to a set of particular phrases, which means it can be subject to the data sparseness problem.

Summing up the ideas presented in the constrained reordering approaches, all the algorithms can be divided into two classes: *content-independent* reordering models, e.g. distance-based or flat reordering models, which learn nothing for use in reordering from parallel corpora, and *content-dependent* reordering models, such as lexicalized reordering models, which are totally dependent on bilingual phrases and are not enhanced with generalization capabilities.

A boosting approach to MT that has gained many adherents over the past few years is the concept of hierarchical phrase-based SMT [Chi05, Chi07, Wat06, Igl09] described in §2.3.3. Classically, systems following this approach do not use any kind of syntactic information except for a synchronous context-free grammar. The hierarchical orientation model efficiently captures long-distance dependencies, introducing some lexical evidence without fully lexicalizing the translation/reordering rules by converting subphrases to variables, which are then used in other levels of the model hierarchy.

Reordering in the rescoring stage

A less popular but nevertheless efficient approach to modeling word-reordering phenomena consists of identifying the n best list instances expressed in the correct target-side word order. In the first step, the list of n-best translations is generated with a SMT system. The rescoring algorithm is then applied to the list, in which each translation hypothesis is enriched with a score provided by an appropriate additional feature function.

An example of a system using a reordering model operating with a set of automatically extracted patterns can be found in [Che06]. The reordering rules take advantage of POS or plain words, while generalization is achieved by the system's capability to swap a pair of blocks consisting of several consecutive words.

The features used in rescoring can be equally syntactical. We classify this set of reordering algorithms as syntax-based non-deterministic methods, described in §4.1.2.

Source-side monotonization

In classical phrase-based translation, the input sentence s is translated into the output sentence t. The translation process is considered as a three step algorithm: (1) the source sequence of words is segmented into phrases, (2) each phrase is translated into the target language using a translation table, (3) the target phrases are reordered to fit the target language.

A rather popular class of reordering algorithms involves the monotonization of the source part of the parallel corpus prior to translation (deterministic approach). Thus, a "zero" step in which the input is rearranged, intended to simplify the translation task, is inserted before segmentation. Reordering is not integrated with the translation system and is placed outside the system in order to make the source sentence word order resemble that of the target language. The majority of the reordering models that incorporate this preprocessing step do use the syntax structures of the source and/or the target languages and are described in §4.1.2.

However, many modern translation systems follow more advanced approaches, using special structures such as word lattices or confusion networks containing the k best reordering hypotheses coded in a directed graph. Though founded on the same idea, reordering graphs and confusion networks neither follow the same objective nor can be used for the same goals. The former are used to couple reordering and SMT decoding, while the latter are aimed at coupling speech recognition and MT. In the process of decoding the confusion network, one word of each column is picked and used as the input word. Thus, input sentences of different lengths can be hypothesized using the special service word. However, reordering cannot be implemented using the confusion network approach without additional constraints.

Pre-translation reordering models can be defined and applied following different deterministic and non-deterministic criteria exploiting statistically learned rules as described

below:

- *Corpus monotonization.* Some of the approaches operating within the statistical framework use IBM alignments to reorder the input sentences and produce a new bilingual pair, composed of the reordered input sentence s' and the output sentence t, whose translation (decoding) is monotonous. An example illustrating the aforementioned reordering strategy can be found in Figure 4.6.

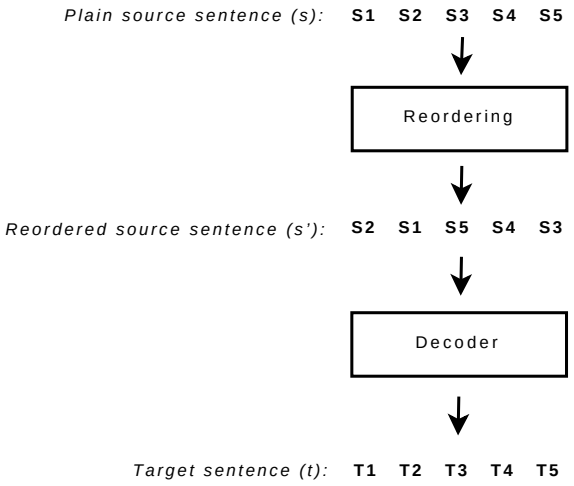


Figure 4.6: *Source-side monotonization prior to translation.*

A strength of this approach is that information in a reordering rule concerns a difference between source and target word order, thereby modeling the cross-language word order transfer. Consequently, better mutual word order is achieved and the translation task is simplified. Another advantage is that an accurate source-side shifting of words makes it possible to account for global phenomena as local [Zwa07].

The idea stems from the work described in [Nie04], where the use of combined morpho-syntactic information for improving translation quality in frameworks with scarce resources has been successfully tested. In [Pop06a], POS tag information is used to rewrite the input sentence between Spanish-English and German-English language pairs.

One more example of a word order monotonization strategy can be found in [Cj06b], where a technique called Statistical Machine Reordering (SMR) was presented. Here,

a monotone sequence of source words is translated into the reordered sequence using SMT techniques. In theory, this approach is intended to tackle long-range reordering. In practice, however, a number of long-distance dependencies are not considered due to high sparseness of data. In contrast to the work of Nießen and Ney, reordering here is treated as a purely statistical process, and no syntactic knowledge of the language is used. Generalization is achieved by using statistical word classes and POS tags. The impact of different types of word classes on the final translation score is analyzed in [Cj07a].

Notice that, within the corpus monotonization framework, word order harmonization can in principle be done equally well on the target language as on the source side of the bicorpus. The only reason to choose the latter way is the target-side LM, which is much more robust when estimated on a large amount of unreordered monolingual data.

- *Input reordering graph.*

  A coherent evolution of the previous approach is done in a non-deterministic fashion and consists of providing the decoder with multiple word orders compressed with the help of a *reordering (permutation) graph*, represented in the form of a *word lattice* [Cre08a, Koe07b], or *confusion networks* [Ber05, Ber07]. These representations differ in the methods used to code topological complexity. Structural constraints imply different space efficiency characteristics, but they nevertheless provide a strict permutation of the input words.

  To infer new possible reorderings, word graph structures are used instead of word sequences themselves as the input to the SMT system. They can be considered a generalization of the previous approach, allowing for a number of alternative reorderings of the source side of the training corpus. Consequently, the translation module has access to both the original word order and the reorderings predicted by a certain set of rewrite rules.

  Figure 4.7 depicts the idea of a multi-way input reordering graph, where multiple word order options are coded.

  Almost all of the approaches that can be found within the field of deterministic multiple reorderings use syntax in one form or another and will be described in section §4.1.2. However, some of the algorithms are purely statistical or are implemented

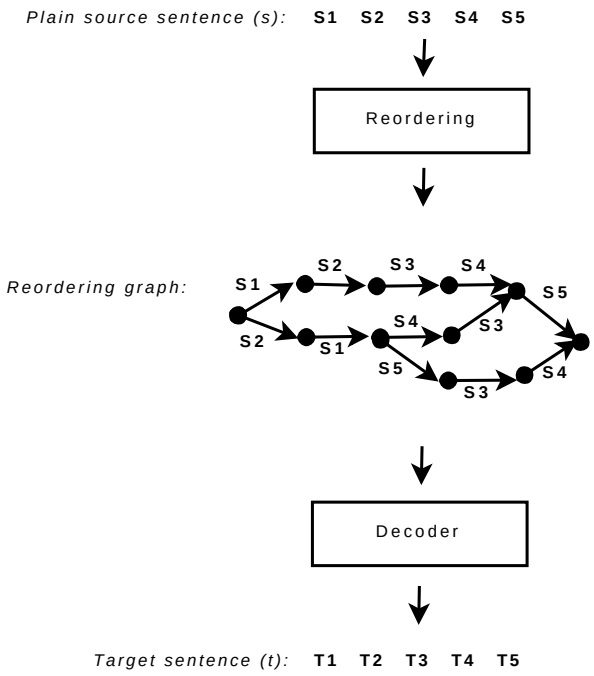


Figure 4.7: *Source input graph.*

with the use of morphological information, as in [Cre06b], where a linguistically motivated reordering model employs monotonic search graph extension. Another significant work is [Cj08a], where coupling of the SMR algorithm and the search space extension via generating a set of weighted reordering hypotheses has demonstrated a significant improvement in translation quality.

4.1.2 Reordering models based on syntax

Recently there has been growing optimism in the MT community about the use of syntax to improve SMT. The first class of research efforts has been concentrated on the development of complete syntax-based systems with the use of statistical models, or of hybrid SMT systems enhanced with syntactical models (briefly described in §2.4). Another class of algorithms includes the methods designed to address the crucial disadvantage of SMT, that is, little or no use of syntactical information in handling the word reordering problem.

The intuitive concept is that use of the syntactical information from the source, target, or both languages seems to be able to thoroughly handle long-range dependencies and accurately model many systematic differences between the word orders of languages [Bon94]. Traditional clump-based SMT models are not expected to handle all possible reorderings, but instead search for the correct reordering option within a limited distortion space. The distortion model provides only a control of the reordering combinations attempted by the decoder.

On the other hand, reordering error can be minimized by means of POS tags, or, more probably, by utilizing structural analysis of different languages. In this vein, a great impetus has been given to syntax-driven reordering by recent advances in the field of NLP, namely in natural language tree bank parsing [Cha00, Col99], which has facilitated incorporation of syntactic information into MT for the purpose of improved handling of reordering. Furthermore, exploiting the generalization power of synchronous context-free grammars proved to be helpful for statistical reordering and led to significant improvements, such as those shown in [Zha07a, Ven06].

Motivated by the lack of robustness in global reordering as performed by phrase-based SMT, syntax-driven systems are more relevant for translation between languages with very different word order structures. On the other hand, it is believed that short-range reorderings are modeled adequately by local phrases [Li07, Zha07b]. Shorter reorderings done on adjacent words that belong to different phrases can be also modeled with the help of syntactical structure, since it does not matter how many words are spanned by the swapping constituents [Elm08].

An empirical MT system including syntax-based reordering usually contains three modules: a syntactical analysis module, a reordering module, and an MT module, as depicted in §4.8. In the syntactical analysis step, constituent or dependency trees are generated for the target or source and target languages; the reordering module infers word or word-block permutations relying on the information provided by the parse tree(s); finally, the MT module handles translation with or without some sort of higher linguistic knowledge.

One potential solution is a pre-translation step of reordering the source sentence so that its word order resembles that of the target language. As in the statistical framework, the syntax-driven reordering models that can be found within the state-of-the-art can be divided into *deterministic* and *non-deterministic* approaches.

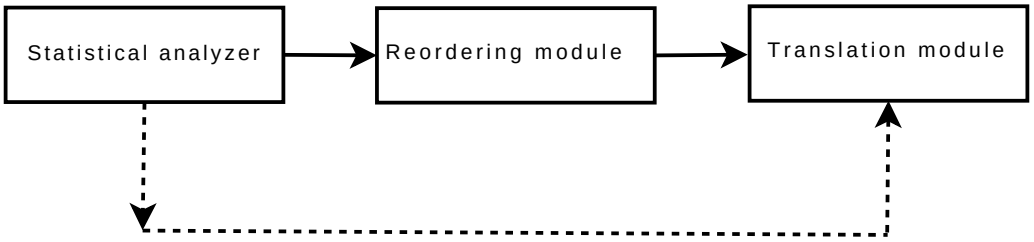


Figure 4.8: *Architecture of empirical MT systems exploiting syntax-based reordering model.*

Deterministic reordering models based on syntax

In recent years, a number of works modeling reordering in a deterministic way and employing syntactical information to mutually monotonize the source and target languages have begun to emerge. Here, the reordering process is performed in a separate modular element outside the SMT system. Translation is augmented by using syntactical algorithms that exploit manually written language-dependent rules or automatically extracted patterns driven by the syntactical structure of the languages.

In the next step, the rules are applied to the source part of the same training corpus, changing the structure of the source sentence so that it more closely matches the word order of the target language. This reordering simplifies the translation task by reducing the average length of bilingual units encountered when translating an unseen set.

Examples of clause restructuring performed with hand-crafted reordering rules for German-to-English and Chinese-to-English tasks were presented in [Col05] and [Wan07], respectively. In these works, the emphasis is placed on the distinction between German/Chinese and English clause structure, recombining the most prominent reordering candidates. In [Zwa07], the natural language tendency to minimize the distance between a head and its dependents derived from the dependency trees is exploited to automatically reorder source-side constituents.

Beyond the aforementioned works, other tree-to-string restructuring models were also introduced. For example, in [Ber96b], the authors present an approach for French where the phrases of the form NOUN1 DE NOUN2 were reordered prior to translation; in [Nie04], the input of a German-to-English SMT system is rearranged so that the verbs are combined with their associated particles, and also the question sentences are reordered.

In [Xia04], a set of automatically extracted POS rules is learned from a dependency-parsed parallel corpus prior to translation and then is applied to a French-to-English translation task. Rewrite patterns operate on context-free rule productions and are acquired automatically.

Non-deterministic reordering models based on syntax

A non-deterministic way to address reordering limitation with the use of syntax is to generate an input word lattice comprising different paths of syntactically motivated reordering, providing the decoder with alternative, possibly weighted, suggestions.

An example of a system performing reordering in this way can be found in [Cre07b], where syntactic structure on the source side is exploited to automatically learn rules, which are then used to reorder the input into a word lattice in an unweighted manner, slightly expanding the monotonic search space.

In [Zha07b], a similar strategy is proposed to address the word reordering problem through a source input graph. Here, the set of reordering hypotheses is defined using an intermediate syntax between POS and parse tree (chunks) as the basic reordering units.

A non-deterministic weighted approach was successfully applied in [Elm08], where automatically extracted syntactically motivated rewrite patterns are first combined in the weighted lattice of alternative translations, reflecting the structural structure of the languages, and then integrated in phrase-based SMT.

Very recent ideas to handle long-range reordering involve combining continuous (VVIMP VMFIN PPER → PPER VMFIN VVIMP) and discontinuous (VAFIN * VVPP → VAFIN VVPP *) POS reordering rules in a word lattice. This work is done in the spirit of hierarchical model framework [Nie09].

In [Zha07b, Dye08, Rot07, Cre08b], the authors take similar approaches to encode multiple source-side word segmentations in a lattice, using POS and chunk tags to generalize the reordering rules. These models mostly differ in the way the weights are assigned to different reordering patterns.

4.2 Research contribution in the field of syntax-based word reordering

This chapter has presented a comprehensive tutorial overview of word reordering algorithms for SMT. We aimed to cover a wide variety of modern reordering systems; therefore, some parts of the discussion have been left abstract.

It is well understood that word reordering is of crucial importance to SMT and is currently considered one of the most difficult problems in MT. Because of the differences in syntactic and informational structures across languages, grammatical or dependency relations may not always be preserved during the translation process. These changes consequently lead to the generation of systematic reordering errors by any purely statistical orthodox SMT system.

It is clear that, at the current stage of SMT technology evolution, the issue cannot be considered completely resolved by the use of state-of-the-art methods, mostly due to their limitations in handling long-distance reorderings. There are still many hurdles and open questions in the field of word reordering for SMT: some reordering algorithms involve no parsing and very little linguistics in the reordering process, which makes them weak in guiding long-distance movements. Other methods exploit syntactical information, which is certainly a potential solution to global reordering. However, the use of syntax is not a universal panacea for reordering: syntax-based models are typically criticized for frequently failing in the handling of non-syntactic phrase pairs (phrase pairs that are not subsumed by any syntax subtrees) [Li07, Zha07b].

In order to confront the reordering challenge, we propose a novel approach called *Syntax-based Reordering (SBR)*, described in detail in the next chapter. This reordering mechanism was initially presented in [Kha09] and is the primary contribution of this Ph.D. research to the field.

SBR can be classified as a *deterministic* approach with the use of both-side syntax, which is additionally combined with a state-of-the-art statistical *non-deterministic* method [Cre08a] to better handle local reordering dependencies. This approach is designed in the spirit of hybrid MT, integrating the syntax transfer approach and statistical methods to achieve better MT performance than the standard state-of-the-art models.

Our goal in developing SBR is twofold. The first motivation is to integrate syntactic information with the SMT approach to long-distance reordering, preserving the strength of

statistical techniques in local reordering.

Furthermore, we intend to enrich the reordering power of syntax-based reordering by incorporating extra reordering patterns that are beyond the scope of one-level tree transduction. Thereby, better-formed reordering rules with multi-level subtrees on the source side of the parallel corpus are incorporated into extended tree transduction, providing longer than regular (in most cases, binary) displacements.

A detailed description of the SBR approach, along with an analysis of its strengths and limitations, can be found in chapter 5.

Chapter 5

Syntax-based reordering

In this chapter, we develop an approach to handling the fundamental problem of word ordering for SMT[33]. We propose to alleviate the word order challenge including morpho-syntactical and statistical information in the context of a pre-translation reordering framework.

In particular, we suggest a word reordering technique which tackle:

1. the long-distance reordering problem in a deterministic way, by converting the source portion of the parallel corpus into an intermediate representation, in which source words are reordered to more closely match the target language;

2. short-range reorderings in a non-deterministic way using POS information and an input graph model, as described in the literature [Cre07a].

Our major interest is in the value of syntax in word reordering for SMT. For this purpose we examine the proposed approach from the theoretical and experimental points of view, analyzing its advantages and limitations in comparison with some of the state-of-the-art methods described in chapter 4.

This chapter describes the initial results of applying the syntax-based model to translation tasks with a great need for reordering (Chinese-to-English and Arabic-to-English).

We first investigate sparse training data scenarios, in which the translation and reordering models are trained on a sparse bilingual data. We then scale the method to a

[33]Much of this chapter is written with extremely useful suggestions by Mark Dras (http://www.ics.mq.edu.au/~madras/)

large training set and demonstrate that the improvement in terms of translation quality is maintained.

The chapter is organized as follows:

- First, in §5.1 we review the architecture and modeling of the proposed *syntax-based reordering* system and provide details about rule extraction, generalization, and application.
- In §5.2 we describe how to couple multiple word reorderings with a translation system, with the objective of handling local reordering dependencies through an input graph.
- §5.3 evaluates the contribution of our model to the performance of N-gram-based and phrase-based SMT systems, presenting the baseline systems and experimental setup along with the obtained results. In §5.3.5, we present a novel technique that demonstrates how a purely generalized (i.e., with no lexical information involved) syntax-based reordering system can help to improve the N-gram-based SMT.
- Finally, §5.4 is a summary of the chapter, providing analysis of resulting data and highlighting the main conclusions drawn.

5.1 Syntax-based reordering framework

This section introduces the *Syntax-Based Reordering (SBR)* approach. Like other preprocessing methods, it splits translation into two independent stages:

$$S \rightarrow S' \rightarrow T \tag{5.1}$$

where a sentence of the source language S is first reordered with respect to the word order of the target language, and then the reordered source sentence S' is monotonically translated into a target sentence T.

SBR deals with the $S \rightarrow S'$ part of the equation 5.1. Once the reordering of the training corpus is ready, it is realigned, and the monotonized alignment is used to extract information for the particular translation task $S' \rightarrow T$. The latter is thought of as a simplification of the original translation task $S \rightarrow T$ due to a shorter minimal length of bilingual units, which are more likely to be found when translating an unseen set.

Local and long-range word reorderings are driven by automatically extracted permutation patterns operating with source language constituents and underlain by non-isomorphic sub-tree transfer. The target-side parse tree use is optional, but it greatly affects system performance: it is considered as a filter constraining the reordering rules to the set of patterns covered by both the source- and target-side sub-trees. Apart from the reordering rules representing the order of child nodes, a set of additional rewrite rules based on a deep top-down sub-tree analysis is considered.

5.1.1 Motivation and sources of inspirations

Objectives

The clear objective of developing an MT system is to meet the growing demand for high-quality translation. As described in the previous chapter, one of the most common sources of errors for MT is the highly challenging problem of correct word order.

Our main goal is to create a model capable of placing the translated words in the natural order of the target language. The importance of reordering models as preprocessing in SMT can be found in many sources; however, in contrast with many distance-based methods, in our approach the emphasis is placed on the long-distance reorderings, which are difficult to capture with standard n-gram language models. Although statistical distortion models achieve the best results within a certain distortion limit and do not incorporate any linguistic analysis, it has been recognized that syntactic information can be efficiently used to capture global reorderings. This is, especially the case for translation tasks that deal with a pair of languages, one of which is a European language and the other an Asian or Semitic one [Wan07].

Approach

As described in §4.1, some translation systems employ deterministic reordering, where by word reordering is done as a preprocessing step aimed at transforming the order of the source sentence to make it closer to the target language. Another approach is non-deterministic, in which the decoder is provided with multiple reordering options.

Among the rationales behind the idea of a deterministic approach is that, in many cases, and especially when the sentences are long, the search space containing the permutations is too wide, even with constraints introduced to restrict the number of alternative reorderings.

In addition, we will show that a multi-step pruning is efficient enough to provide an accurate choice of a single-best reordering hypothesis. In this step, we aim to develop an extended tree transducer, which takes into account only the word order, without performing word/clump translation directly.

On the other hand, the deterministic approach is claimed to be problematic in a statistical framework in that it makes hard decisions about word order that cannot be undone during decoding [AO06]. The deterministic approach implies that reordering and decoding are two independent processes, whereas intuition tells us that the choice of reorderings is not independent of other translation factors; consequently, reordering mistakes cannot be corrected by the decoder.

These concerns convinced us to find an intermediate way to combine the clear advantages of the deterministic approach without losing the simplicity and compatibility proposed by the non-deterministic strategy. We propose to use a two-step integrated approach, in which we apply a deterministic syntax-based algorithm in the first stage and then construct a POS-based input graph of possible reordering permutations. Finally, we feed this lattice to the word lattice decoder, which provides a better final translation because it takes into account the dynamic relationship between word selection and the reordering option. A reordering system built in this way can be considered a compromise between two fundamental approaches and is expected to demonstrate robustness with respect to training errors.

Inspiration

Our work is heavily inspired by the approach proposed in [Ima05], in which a complete syntax-driven SMT system based on a two-sided sub-tree transfer is described.

In this approach the researchers constructed a probabilistic non-isomorphic tree mapping model based on a context-free breakdown of the source and target parse trees and used both Japanese and English parsers to limit the computational complexity of syntax-based SMT. They then extracted alignment templates that incorporated the constraints of the parse trees and applied syntax-based decoding. Three tree-mapping models were used to calculate the final score of the translation model, namely: source and target tree mapping models, and the tree-to-tree transfer mapping model.

One prominent advantage of the proposed method is that not only hierarchical syntactically motivated reorderings but also the monotonic phrases handled in conventional phrase-based SMT (the authors call them “flat“ phrases) can be directly applied to the

translation.

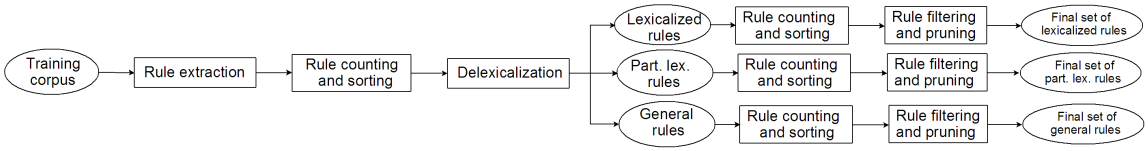
This approach succeeds in outperforming a phrase-based SMT baseline in a Japanese-to-English BTEC translation task by about 15% in terms of BLEU score and by about 8 % according to human evaluation. It is also shown that the exclusion of the "flat" phrases from the TM leads to a significant degradation in translation quality (up to 11 % in terms of BLEU score for the BTEC task).

We propose to use a similar non-isomorphic sub-tree mapping to extract reordering rules, but instead of involving the rules directly in the translation process, we use them to monotonize the source portion of the bilingual corpus. We expect an efficient combination of the probabilistic tree-to-tree mapping, explicitly involving syntactic information in the word reordering process, and the powerful techniques developed for purely statistical MT.

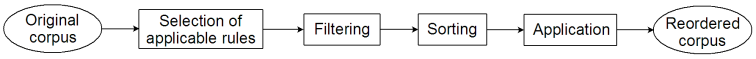
Another source of inspiration was the work presented in [Xia04], in which the authors propose a system for French-to-English translation, based on the principle of automatic rewrite pattern extraction using a parse tree and phrase alignments. Although the aforementioned approach and the SBR technique share the idea of sub-tree transfer, the former differs in many ways from the one that we used in the deterministic stage of our algorithm. Among other distinctions, we use a more complete lexical model underlying the sub-tree syntax transfer, which, apart from direct structural divergences, involves elementary reordering structures that cannot be captured within one-level production. We also propose a different generalization probabilistic model that makes use of lexical information based on three linguistic levels of generalization and that is not restricted to operations on syntactic tree structure nodes.

Training and testing steps

A block diagram of the training and testing processes of the SMR deterministic model can be found in Figure 5.1. During the training step, three sets of reordering rules are extracted from the initial fully lexicalized rules and then processed separately. During reordering of unseen data, a set of potentially applicable rules is first extracted from the space of reordering rules found in the training step, then this array of candidate patterns is filtered and sorted, and finally it is applied to the source-side parse tree. All the above-mentioned procedures are described in detail in the following sections.



(a) Training stage.



(b) Testing stage.

Figure 5.1: *Block diagram of the training and testing processes of the SBR deterministic model.*

5.1.2 Notation

SBR operates with source and target parse trees that represent the syntactic structure of a string in source and target languages.

This representation is usually formally defined as

$$G = \langle N, T, R, S \rangle \tag{5.2}$$

where N is a set of nonterminal symbols (corresponding to source-side phrase and part-of-speech tags); T is a set of source-side terminals (the lexicon); R is a set of production rules of the form $\eta \rightarrow \gamma$, with $\eta \in N$ and γ a sequence of terminal and nonterminal symbols; and $S \in N$ is the distinguished symbol.

The reordering rules then have the form

$$\eta_0@0 \ldots \eta_k@k \rightarrow \eta_{d_0}@d_0 \ldots \eta_{d_k}@d_k \,|\, \eta_0\langle\langle w_{0,0} ... w_{0,J_0}\rangle\rangle, \ldots, \eta_k\langle\langle w_{k,0} ... w_{k,J_k}\rangle\rangle \,|\, p \tag{5.3}$$

where $\eta_i \in N$ for all $0 \leq i \leq k$; $(d_0 \ldots d_k)$ is a permutation of $(0 \ldots k)$; $\eta_0\langle\langle w_{0,0} ... w_{0,J_0}\rangle\rangle, \ldots,$ $\eta_k\langle\langle w_{k,0} ... w_{k,J_k}\rangle\rangle$ is *a source-side lexicon*, where the non-terminal η_i spans the substring $w_{i,0} \ldots w_{i,J_i}$ of the length $J_i + 1$ and $i \in [0, k]$; and p is a probability associated with the rule[34].

[34]Notice that in the framework of this dissertation, the vertical bar ("|") is used to separate the transduction, lexical and probabilistic components of reordering rules.

5.1.3 Reordering rule extraction

Concept

The SBR system requires access to source and target language parse trees, along with the source-to-target and target-to-source word alignments intersection. We extracted a set of bilingual patterns allowing for reordering as described below:

(1) Align the monotone bilingual corpus with GIZA++ and find the intersection of the direct and inverse word alignments, resulting in the construction of the projection matrix P (see below);

(2) Parse the source and the target parts of the parallel corpus;

(3) Extract reordering patterns from the parallel non-isomorphic trees based on the word alignment intersection, considering POS and constituents equally and saving lexical information about all the elements of the pattern.

Step 2 is achieved using external tool (Stanford parser). Steps 1 and 3 need more detailed explanation, which can be found below.

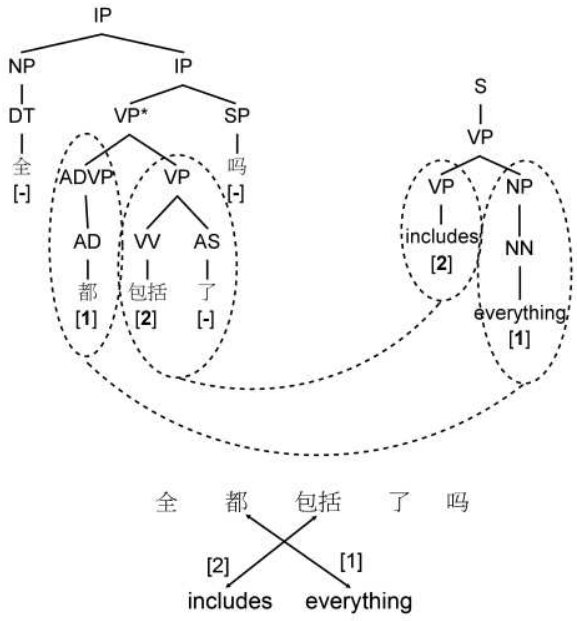


Figure 5.2: *Example of reordering rules extraction (Example 1).*

Figure 5.2 shows an example of the generation of two lexicalized rules for a Chinese-English bilingual phrase 5.4 (we use this phrase for further explanation and call it *Example 1*), where SfP refers to sentence-final particle:

Zh:	全	都	包括	了	吗
Gloss:	whole	all	include	SfP	
Eng. ref.:	' includes everything '				

(5.4)

In this example, "都/*all*" should move to the right of the Chinese block "包括 了/*include*" to get the English word order as seen in the gloss.

Notice that the first word "全/*whole*" is a Chinese redundant adverb, which is not translated into English (optionally, the second word "都/*all*" can be omitted during translation).

Reordering rules, which can be directly extracted from the presented structure, are:

ADVP@0 VP@1 → VP@1 ADVP@0 | ADVP@0 << 都 >> VP@1 << 包括 了 >>
AD@0 VP@1 → VP@1 AD@0 | AD@0 << 都 >> VP@1 << 包括 了 >>

The rules are equivalent, since constituents ADVP and AD are a unary chain (see below) for which reordering rules are extracted for each level in this chain.

An instance of a more complex syntax transfer can be found in Figure 5.3. It illustrates the algorithm's potential for capturing long-distance permutations for the following sentence 5.5 (*Example 2*), where BA is an indicator of a verbal construction, forming an SOV clause:

Zh:	我	把	他 的	电话号码	和	住址	给	你
Gloss:	I	BA	his	telephone	and	address	give	you
Eng. ref.:	' I give you his telephone number and address '							

(5.5)

The extracted rule reflects the divergence between Chinese and English languages in clause construction, which is out of the scope of the distance-based model. To generate an ideal translation into English, immediately after translating the first word "我/*I*", the decoder needs to move across six source words and translate the last sequence "给 你/*give you*" to get the correct word order. However, swapping NP and VP constituents in this case will help to generate the correct translation.

For this sentence, the only extracted rule is:

NP@0 VP@1 → VP@1 NP@0 | NP@0 << 他 的 电话号码 和 住址 >> VP@1 << 给 你 >>

This pattern implements an NP↔VP swap transformation, modeling the ordering of the VP in respect to the previous NP constituent and lexical filling of both constituents.

Projection matrix

Bilingual content can be represented in the form of words or sequences of words, depending on the syntactic role of the corresponding grammatical element (constituent or POS).

Given two parse trees and a word alignment intersection, a projection matrix P is defined as an $M \times N$ matrix such that M is the number of words in the target phrase; N is the number of words in the source phrase; and a cell (i, j) has a value based on the alignment intersection — this value is zero if word i and word j do not align, and it is a unique non-zero link number if they do.

For the tree in Figure 5.2,

$$P_1 = \begin{pmatrix} 0 & 0 & \mathbf{2} & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 \end{pmatrix}$$

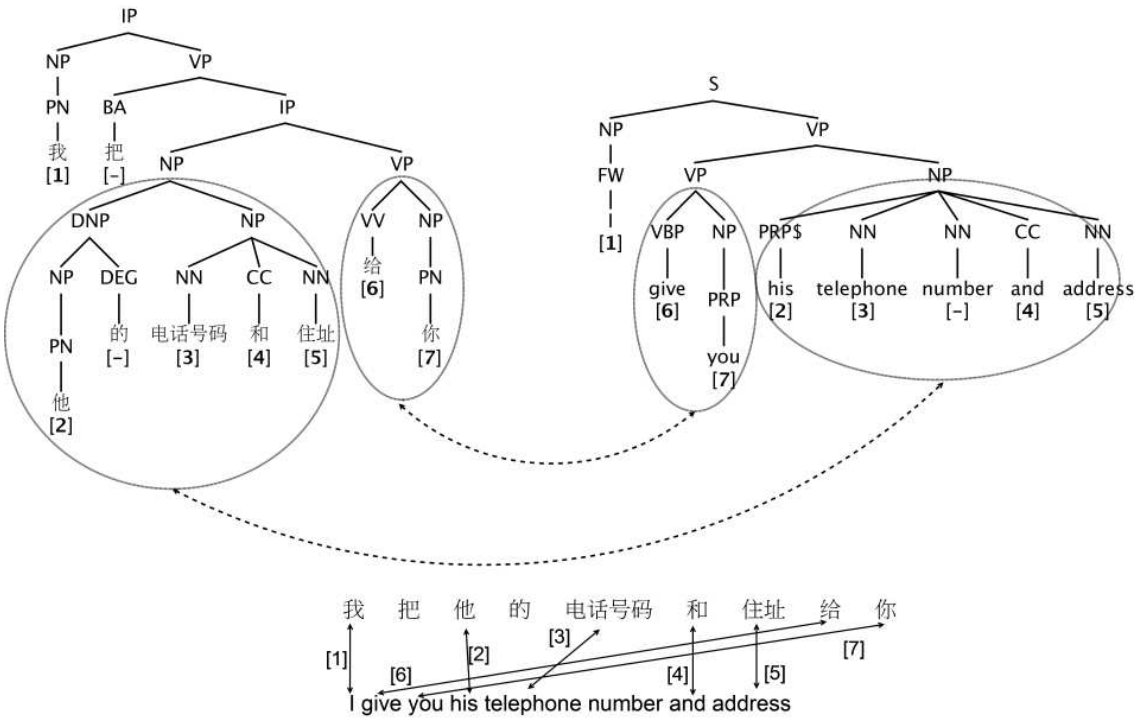


Figure 5.3: *Example of reordering rules extraction (Example 2).*

The schematic representation of the reordering block, describing the orientation within the matrix P_1, can be found in Figure 5.4.

	全	都	包括	了	吗
includes	0	0	2	0	0
everything	0	1	0	0	0

Figure 5.4: *Word reordering for the translation direction of Chinese into English (Example 1).*

For the tree, which can be found in Figure 5.3, the projection matrix is:

$$P_2 = \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{7} \\ 0 & 0 & \mathbf{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{5} & 0 & 0 \end{pmatrix}$$

The corresponding block orientation scheme is provided in Figure 5.5.

	我	把	他	的	电话号码	和	住址	给	你
I	1	0	0	0	0	0	0	0	0
give	0	0	0	0	0	0	0	6	0
you	0	0	0	0	0	0	0	0	7
his	0	0	2	0	0	0	0	0	0
telephone	0	0	0	0	3	0	0	0	0
number	0	0	0	0	0	0	0	0	0
and	0	0	0	0	0	4	0	0	0
address	0	0	0	0	0	0	5	0	0

Figure 5.5: *Word reordering for the translation direction of Chinese into English (Example 2).*

If a word that is aligned in only one direction appears in the branch that is considered a

candidate to be involved in a reordering pattern, it does not change the alignment projection matrix.

Alignment and sub-trees interaction

Each non-terminal from the source and target parse trees is assigned a string, which we call *alignment intersection ordering* (AIO). It carries information about elements from the alignment intersection that are contained in its child nodes, taking into account the order of their appearance in the tree. For example, the AIO string assigned to the source-side internal node VP* in Figure 5.2 is "1 2" (we define this string as AIO_{VP^*}) and to the target-side VP is "2 1" (AIO_{VP}). This information is used to indicate the source-side nodes that are to be reordered according to the target-language syntactical structure. Reordering patterns are extracted following the source and target-side AIOs as shown in Figure 5.2 (we call them "*main rules*").

If more than one non-zero element of the projection matrix is reachable through the child nodes, the AIO has a more complex structure, providing information about elements from the alignment intersection belonging to one or another child node. An example for the bilingual sentence 5.6 can be found in Figure 5.6.

Zh:	你	需要	什么	尽管		告诉	我
Gloss:	you	need	what	don't	hesitate	tell	me

Eng. ref.: ' don't hesitate to tell me what you need ' (5.6)

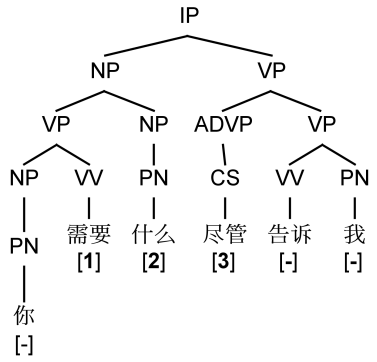


Figure 5.6: *Example of complex AIO structure.*

Here, the sub-tree IP is assigned with the AIO_{IP} = "(1 2) 3", meaning that it has two child nodes: the first contains the elements 1 and 2 from the alignment intersection, and

the second, element 3 (we call this subsequence "*closed*"). The reordering system considers nodes assigned with one or more children equally discerning the nodes with different-order alignment elements.

Unary chains

Given a unary chain of the form $X \rightarrow Y$, rules are extracted for each level in this chain. For example, in Figure 5.2, the unary chain is "$ADVP \rightarrow AD \rightarrow ...$", and the directly extracted reordering rules are equivalent since the node ADVP leads to the leaf through the node AD and does not have other edges. In Figure 5.3, the unary chains are NP → PN for the Chinese tree and NP → FW for the English tree.

The role of target-side parse tree

Conceptually speaking, the use of target-side parse tree is optional. Although reordering is performed on the source side only, the target-side tree is of great importance: the reordering rules can be extracted only if the words covered by the rule are entirely covered by a node both in the source and in the target trees. It allows for more accurate determination of the coverage of the rules and their limitation.

5.1.4 Non-isomorphic tree mapping

There are many nodes for which a comparison of AIOs indicates that a sub-tree transfer can be done, but segmentation of child nodes is not identical. This phenomenon stems from the well-known problem of non-isomorphism between source and target trees, which limits the transduction capability of synchronous systems. It was shown in [Gal04, Gra04] that one-level structures often lack expressive power and that many common translation patterns fall outside the scope of the child-reordering model.

In many recent syntax-based and hybrid MT systems only isomorphic trees have been essentially assumed [Yam01, Als00]. However, in fact many trees are not isomorphic. Figure 5.7 illustrates this situation. AIO strings assigned to the root nodes of the trees contain the same elements, but segmentation and order of appearance of elements do not coincide. These sub-trees cannot be directly used for pattern extraction; hence, more in-depth analysis is required.

To alleviate the problem of incomplete coverage, systems employing enhanced expressive

power and primarily based on synchronous grammars have been proposed [Shi90, Eis03]. Another attempt to move to more robust grammars that adapt to the parallel training corpus was done in [Hua06]. Here, the idea of extended domain of locality that spans multi-level sub-trees on the source side of the corpus, first presented in [Jos97], was used to implement an extended tree-to-string transducer.

We implement the strategy to address discrepancies between source and target parse trees through postorder source-side tree traversal with the aim of capturing multi-level sub-trees. In addition to using a set of elementary trees with enumerated "anchors" from alignment intersection, we also look into the corresponding target-side sub-trees spanning the same set of alignment elements.

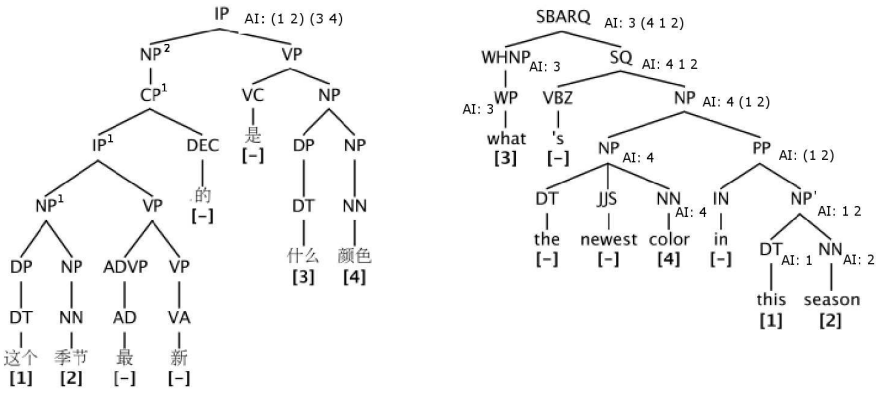


Figure 5.7: *Example of "secondary" rule extraction.*

Extracted rules:

NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << 什么 >> NP@2 << 颜色 >>
NP@0 DP@1 NN@2 → DP@1 NN@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << 什么 >> NN@2 << 颜色 >>
NP@0 DT@1 NP@2 → DT@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DT@1 << 什么 >> NP@2 << 颜色 >>
NP@0 DT@1 NN@2 → DT@1 NP@2 NN@0 | NP@0 << 这个 季节 >> DT@1 << 什么 >> NN@2 << 颜色 >>
CP@0 DP@1 NP@2 → DP@1 NP@2 CP@0 | CP@0 << 这个 季节 最 新 的 >> DP@1 << 什么 >> NP@2 << 颜色 >>
CP@0 DP@1 NN@2 → DP@1 NN@2 CP@0 | CP@0 << 这个 季节 最 新 的 >> DP@1 << 什么 >> NN@2 << 颜色 >>
CP@0 DT@1 NP@2 → DT@1 NP@2 CP@0 | CP@0 << 这个 季节 最 新 的 >> DT@1 << 什么 >> NP@2 << 颜色 >>
CP@0 DT@1 NN@2 → DT@1 NN@2 CP@0 | CP@0 << 这个 季节 最 新 的 >> DT@1 << 什么 >> NN@2 << 颜色 >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 最 新 的 >> DP@1 << 什么 >> NP@2 << 颜色 >>
NP@0 DP@1 NN@2 → DP@1 NN@2 NP@0 | NP@0 << 这个 季节 最 新 的 >> DP@1 << 什么 >> NN@2 << 颜色 >>
NP@0 DT@1 NP@2 → DT@1 NP@2 NP@0 | NP@0 << 这个 季节 最 新 的 >> DT@1 << 什么 >> NP@2 << 颜色 >>
NP@0 DT@1 NN@2 → DT@1 NN@2 NP@0 | NP@0 << 这个 季节 最 新 的 >> DT@1 << 什么 >> NN@2 << 颜色 >>
...

We adopt the following six-step algorithm for each parent node from the source-side

parse tree:

1. Find the AIO sequence for the source-side top-level element (in the example, the IP node is assigned "(1 2) (3 4)").
2. Look down through the target-side tree, finding AIOs for each node.
3. Find all target-side "closed" (see §5.1.3) subsequences for the source-side AIO found in step 1. In the example, it is the subsequence "(1 2)".
4. Find all target-side isolated nodes corresponding to the elements that were not covered in step 2. In the example, these elements are "3" and "4".
5. Extend the set of source-side nodes found in steps 2 and 3 with equivalent branches. Since the words that are not presented in the alignment intersection do not affect the projection matrix, "equivalence" means here that all the branches spanning the elements from the given instance are considered equally (for example, elements NP^1 are equivalent to the nodes IP^1, CP^1 and NP^2).
6. Place them in order corresponding to the target-side AIO and construct the final reordering patterns (*"secondary rules"*).

To illustrate the limitations incurred by a target-side parse tree, the potential reordering pattern, referring to the top node in the Chinese tree

NP@0 VP@1 → VP@1 NP@0 | NP@0 << 这个 季节 最 新 的 >> VP@1 << 是 什么 颜色 >>

is not allowed due to distinct source- and target-side tree coverage.

The strategy of "secondary" rule extraction specifically allows extending the set of initial rules extracted for the Example 1 with the following fair patterns:

ADVP@0 VV@1 AS@2 → VV@0 AS@2 ADVP@0 | ADVP@0 << 都 >> VV@1 << 包括 >> AS@2<< 了 >>
AD@0 VV@1 AS@2 → VV@0 AS@2 AD@0 | AD@0 << 都 >> VV@1 << 包括 >> AS@2<< 了 >>

It is worth noting that the latter pattern is a rewrite rule, employing only POS tags, where each raw word relates to a corresponding lexical category. This class of rules is a special case of the SBR framework modeling one-by-one word permutations, where each block that is subject to reordering consists of only one word.

For Example 2, the extended set includes 12 new "secondary rules" capturing word order regularities within four lower level product:

DNP@0 NP@1 VP@2 →
VP@2 DNP@0 NP@1 | DNP@0 << 他 的 >> NP@1 << 电话号码 和 住址 >> VP@2 << 给 你 >>

NP@0 DEG@1 NP@2 VP@3 → VP@3 NP@0 DEG@1 NP@2 |
NP@0 << 他 >> DEG@1 << 的 >> NP@2 << 电话号码 和 住址 >> VP@3 << 给 你 >>

...

PN@0 DEG@1 NN@1 CC@3 NN@4 VV@5 PN@6 → VV@5 NP@6 PN@0 DEG@1 NN@1 CC@3 NN@4 |
PN@0 << 他 >> DEG@1 << 的 >> NN@2 << 电话号码 >> CC@3 << 和 >> NN@4 << 住址 >>
VV@5 << 给 >> PN@6 << 你 >>

5.1.5 Rule organization

Once the list of fully lexicalized reordering patterns is extracted, the number of times each rule occurred is counted, and the set of initial rules is sorted. Then, all the rules are progressively processed, reducing the amount of lexical information. Initial rules are iteratively expanded such that each element of the pattern is generalized until all the lexical elements of the rule are represented in the form of fully unlexicalized categories. Hence, from each initial pattern with N lexical elements, $2^N - 2$ partially lexicalized rules and 1 general rule are generated. An example of the process of delexicalization can be found in Figure 5.8. Generalized elements of partially lexicalized rules are marked as "NON".

Initial rule:

NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << 什么 >> NP@2 << 颜色 >>

Partially lexicalized rules:

NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << NON >> NP@2 << NON >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << NON >> DP@1 << 什么 >> NP@2 << NON >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << NON >> DP@1 << NON >> NP@2 << 颜色 >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << NON >> DP@1 << 什么 >> NP@2 << 颜色 >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << NON >> NP@2 << 颜色 >>
NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0 | NP@0 << 这个 季节 >> DP@1 << 什么 >> NP@2 << NON >>

General rule:

NP@0 DP@1 NP@2 → DP@1 NP@2 NP@0

Figure 5.8: *Examples of lexical rules expansion.*

Thus, three types of rules are finally available:

1. *Fully lexicalized* (initial) rules
2. *Partially lexicalized* rules and
3. *Unlexicalize*d (general) rules

In the next step, these three sets of rules are processed separately: patterns are pruned and ambiguous (see below) rules are removed. We apply different pruning strategies to different groups of rules, introducing three independent thresholds: k_{lex} for fully lexicalized rules, k_{part} for partially lexicalized rules, and k_{gener} for general rules. All the rules from the corresponding set that appear fewer than k times are directly discarded. The probability of a pattern is estimated from its frequency in the training corpus, and only the most probable rule is stored.

In this version of the reordering system, only the single-best reordering hypothesis is used in other stages of the algorithm, so the rule output functioning as an input to the next rule can lead to situations reverting the change of word order that the previously applied rule made. Therefore, the rules that can be ambiguous when applied sequentially are exhaustively searched and pruned according to the higher probability principle[35].

For example, for the pair of patterns with the same lexicon (which is empty for a general rule), such as

$$\mathsf{NP@0\ VP@1} \rightarrow \mathsf{VP@1\ NP@0}\ p_1 \text{ and } \mathsf{VP@0\ NP@1} \rightarrow \mathsf{NP@1\ VP@0}\ p_2$$

which lead to a recurring contradiction,

$$\mathsf{NP\ VP} \rightarrow \mathsf{VP\ NP} \rightarrow \mathsf{NP\ VP}$$

the less probable rule is removed.

The same strategy is adapted to the overlapping rules leading to a contradiction in the set of applicable patterns, such as,

$$\mathsf{ADVP@0\ NP@1\ VP@2} \rightarrow \mathsf{VP@2\ NP@1}\ p_3 \text{ and } \mathsf{VP@0\ NP@1} \rightarrow \mathsf{NP@1\ VP@0}\ p_4$$

where the filtering decision is made depending on the values p_3 and p_4.

Finally, there are three resulting parameter tables analogous to the "r-table" as stated in [Yam01], consisting of POS- and constituent-based patterns allowing for reordering and monotone distortion.

[35] Computational impact is not an issue here since the brute-force search space is constrained to the set of rules applicable to a given sentence.

5.1.6 Source-side monotonization

Rule application is performed as a bottom-up parse tree traversal following two principles:

(1) The longest possible rule is applied; that is, among a set of nested rules, the rule with the longest left-side covering is selected. For example, in the case of the appearance of an NN JJ RB sequence and presence of the two reordering rules

NN@ JJ@1 → ... and

NN@0 JJ@1 RB@2 → ...

the latter pattern will be applied.

(2) The rule containing the maximum lexical information is applied; that is, when there is more than one alternative pattern from different groups, the lexicalized rules have preference over the partially lexicalized, and partially lexicalized rules have preference over general ones.

Figure 5.9 shows the reordered source-side tree corresponding to the Example 1 with the applied pattern :

ADVP@0 VP@1 → VP@1 ADVP@0 | ADVP@0 << 都 >> VP@1 << 包括 了 >>

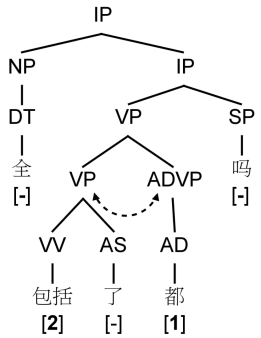


Figure 5.9: *Reordered source-side parse tree (Example 1).*

The resulting reordered Chinese phrase more closely matches the order of the target language and is considered as a result of the sub-tree transfer (see phrase 5.7):

Plain Zh:	全	都	包括 了	吗	
Reordered Zh:	全	包括 了	都	吗	(5.7)
Gloss:	whole include	all	SfP		
Eng. ref.:	' includes everything '				

Another parse tree with swapped branches is shown in Figure 5.10. This is an illustration of a long-distance source sentence reordering performed by the SBR algorithm corresponding to the Example 5.5 with the applied lexicalized pattern

NP@0 VP@1 → VP@1 NP@0 | NP@0 << 他 的 电话号码 和 住址 >> VP@1 << 给 你 >>

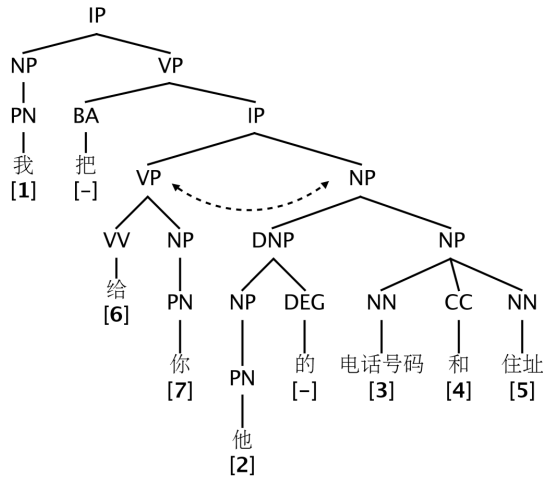


Figure 5.10: *Reordered source-side parse tree (Example 2).*

Two blocks of Chinese words "给 你/*give you*" and "他 的 电话号码 和 住址/*his telephone and address*" are swapped, which leads to a monotone mutual word order (see phrase 5.8):

Plain Zh:	我	把	他 的 电话号码 和 住址	给 你	
Reordered Zh:	我	把	给 你	他 的 电话号码 和 住址	(5.8)
Gloss:	I	*BA*	give you his telephone and address		
Eng. ref.:	' I give you his telephone number and address '				

Once the reordering of the training corpus is ready, it is realigned, and a new, more monotonic alignment is passed to the SMT system. In theory, the word links from the original alignment can be used; however, from our experience, running GIZA++ again results in a better word alignment since it is easier to learn on the modified training example.

5.2 Coupling SBR and decoding

To improve the reordering power of the translation system, we implemented an additional reordering as described in [Cre08a]. Multiple word segmentations are encoded in a word lattice, which is then passed to the input of the decoder, containing reordering alternatives consistent with the previously extracted rules.

A word lattice is defined as a direct acyclic graph

$$G = (V, E) \tag{5.9}$$

with one root node $n_0 \in V$ and one goal node $n_N \in V$. V and E are, respectively, the set of nodes and edges of the graph G. Edges are labeled with source-side words.

The decoder takes the n-best reordering of a source sentence coded in the form of a word lattice. As stated before, this approach is in line with recent research tendencies in SMT, as described, for example, in [Hil08, Xu05].

In other words, the TM described in 5.1 is transformed as represented by Equation 5.10:

$$S \rightarrow S' \rightarrow n \times S' \rightarrow T \tag{5.10}$$

where $n \times S'$ is a word lattice, compactly representing the n-best reorderings of the source-side sentences S'.

Originally, word lattice algorithms did not involve syntax in the reordering process; therefore, their reordering power is limited at representing long-distance reordering. Our approach is designed in the spirit of hybrid MT, integrating syntax transfer approach and statistical word lattice methods to achieve better MT performance on the basis of the latest TMs used in the field.

During training, a set of word permutation patterns is automatically learned following given word-to-word alignment. Since the original and monotonized (reordered) alignments may vary, different sets of reordering patterns are generated. Note that no information

about the syntax of the sentence is used: the reordering permutations are motivated by the crossed links found in the word alignment, and, consequently, the generalization power of this framework is limited to local permutations.

In the step prior to decoding, the system generates a word reordering graph for every source sentence, expressed in the form of a word lattice. The decoder processes the word lattice input instead of the single-best hypothesis, extending the monotonic search graph with alternative paths.

Figure 5.11 illustrates that local reordering permutations "都/*all*" ↔ "包括 了/*include*" from Example 1 can be captured with an input graph exploiting morpho-syntactic information (POS tags). The path in a lattice 5.11 "1 → 2 → 6 → 7 → 14 → 15" generates a reordered sequence that is equivalent to the permutation proposed by the SBR method, and this particular sequence can be correctly reordered, thereby extending the monotonic search graph with a word lattice for the monotonic train.

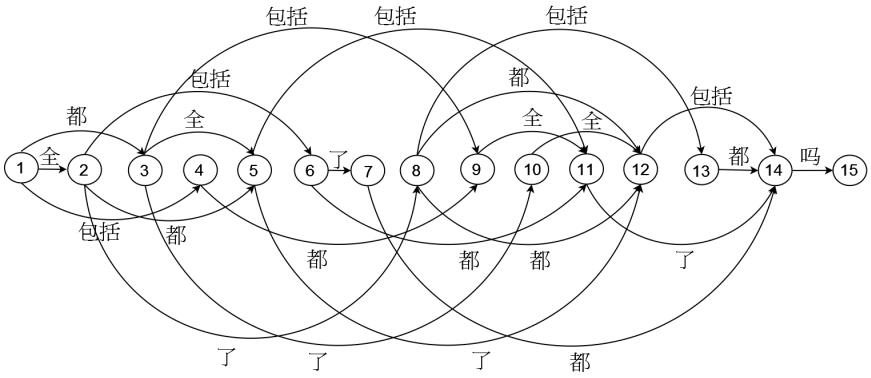


Figure 5.11: *Word lattice for Example 1.*

By contrast, Figure 5.12 shows an example of a word lattice that does not capture the $NP \leftrightarrow VP$ reordering. There are no paths presented in the graph that could model the correct word order.

However, if SBR is applied, the generated word lattice produces a different and better set of reordering hypotheses, as shown in Figure 5.13. The "给 你/*give you*" clump is placed just after the second word in the sentence, and the set of reordering hypotheses includes the correct sequence, leading to monotonic decoding (path "1 → 2 → 3 → 5 → 8 → 13 → 18 → 22 → 26 → 30").

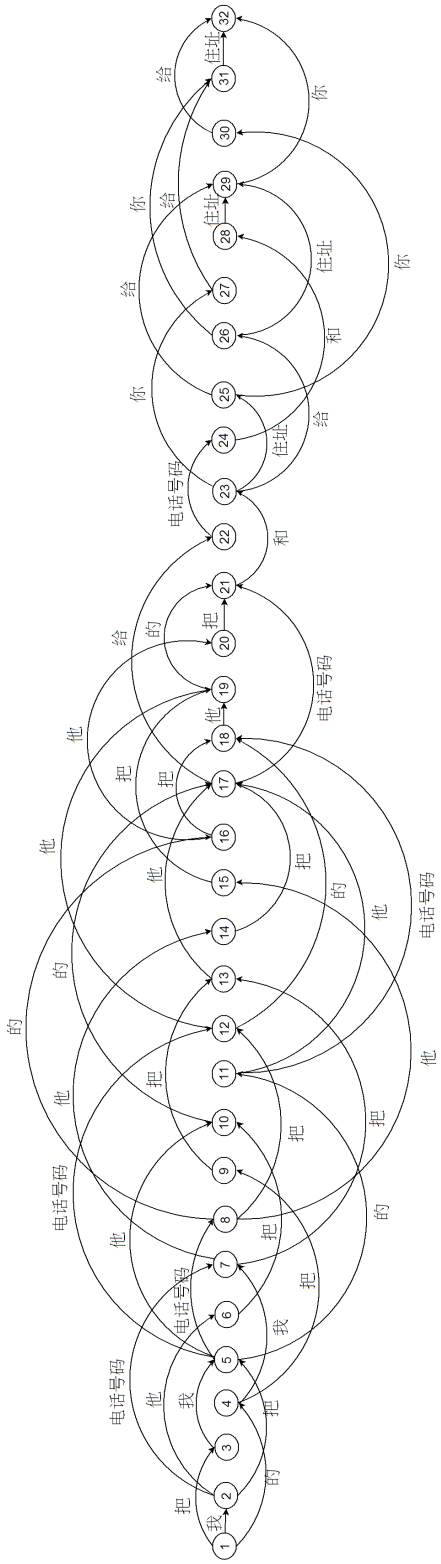


Figure 5.12: *Word lattice without SBR reordering applied (Example 2).*

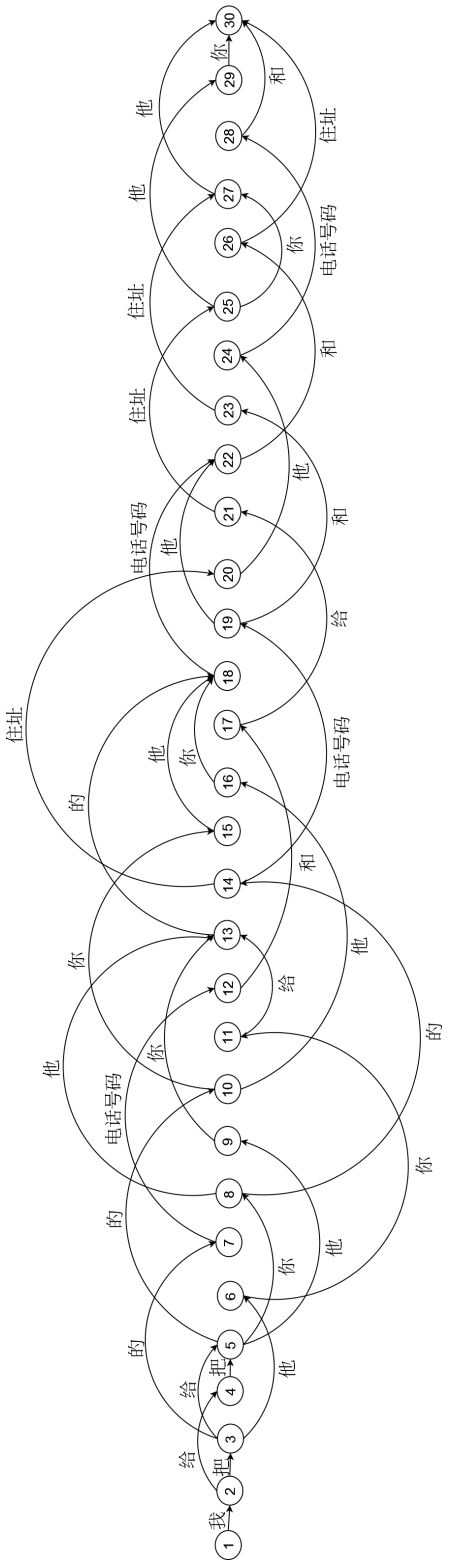


Figure 5.13: *Word lattice with SBR reordering applied (Example 2).*

5.3 Experiments and results

This section details the experiments carried out to evaluate the performance of the SBR approach. To understand the value in terms of accuracy and efficiency of the proposed reordering framework, two directions and four translation tasks have been employed, namely, small and large Chinese-to-English tasks (details can be found in §A.5 and §A.4, respectively), and small and large Arabic-to-English tasks (§A.7 and §A.6).

The main reason that the Chinese-to-English and Arabic-to-English translation tasks were chosen as a main experimental field is because European languages are not so crucial for the global (long-distance) reordering problem as the translation between Asian/Semitic languages and English. Both Chinese and Arabic languages differ from English in many cases, and the corresponding translation tasks can be characterized by a high level of demand on word reordering. On the other hand, the reordering needs for Arabic-to-English and Chinese-to-English translation are very different and can shed light on the universality of the proposed reordering scheme.

We report on two sets of experiments with phrase-based and N-gram-based models driving the translation process.

5.3.1 Data

The major motivation to conduct experiments on corpora of different size is to obtain empirical results on unseen (or partially seen) examples using a word reordering method derived from the generalization model. Application of SMT to language pairs lacking parallel data is an interesting challenge, widely discussed in recent literature [Pop06b, CB6a]. In the framework of the study, we intend to understand the impact of the proposed generalization algorithm on the translation accuracy using the limited resources and discuss implications for efficient reordering on the example of the BTEC corpus from the tourist domain, which is traditionally proposed in IWSLT evaluation campaigns.

We also perform experiments on large corpora from the news domain (NIST translation tasks) and thereby show that the SBR method scales for a larger training set and that the improvement is maintained.

In particular, Chinese-to-English experiments were performed on the BTEC'07 corpus (§A.5) and NIST'06 (§A.4) material. In Arabic-English experiments we used the

BTEC'08 corpus (§A.7) along with a 1-million-sentence extraction from the NIST'08 corpus (§A.6), which was provided to the NIST evaluation campaign in 2008.

Apart from the quantity of training data, the corpora under consideration differ in average sentence length (*ASL*), which is the key corpus characteristic in global reordering studies. We expect that the need for longer distance reorderings would be clearer in longer sentences, as in the NIST corpus, but we also include the BTEC corpus to see whether there is an effect on shorter sentences as well.

Comparative sentence length statistics for BTEC and NIST training material are provided in Table 5.1.

	Chinese-to-English				Arabic-to-English			
	BTEC		NIST		BTEC		NIST	
	Arabic	English	Arabic	English	Chinese	English	Chinese	English
ASL	8.46	9.05	31.62	31.77	6.66	7.22	25.80	26.71

Table 5.1: *Average sentence length for BTEC and NIST corpora.*

5.3.2 Common details

The following paragraphs describe parameters that are common for experiments with phrase- and N-gram-based systems.

Preprocessing. Language independent preprocessing consists of standard filtering procedures. Some sentence pairs are removed from the training data to allow for better performance of the alignment tool and monolingual parsers.

Sentence pairs are removed according to the following two criteria:

- *Fertility filtering*: removes sentence pairs with a word ratio larger than a predefined threshold value (3 in all the experiments).
- *Length filtering*: removes sentence pairs with at least one sentence of more than 100 words in length. This helps maintain bounded alignment computational times.

English data preprocessing. English preprocessing included lower-casing; splitting off contractions such as "*'s*", "*'ve*", "*'ll*", "*'m*"; and separation of punctuation marks attached to words. English POS tags (48 POS tags) were obtained with the Penn treebank Stanford parser [Kle03], which was used in this step as a POS tagger.

Chinese data preprocessing. The Chinese portion of the corpora was re-segmented with the ICT-CLAS tool [Zha03]. The Chinese POS tags (32 tags) were obtained with the Stanford parser.

Arabic data preprocessing. Arabic preprocessing was performed using the combination of morphological analysis and disambiguation (MADA toolkit) [Hab05] and a general tokenizer TOKAN [Hab06] for disambiguation and tokenization. For disambiguation, only diacritic unigram statistics were employed. For tokenization, we used the D3 scheme. The scheme splits the following set of clitics: *w+, f+, b+, k+, l+, Al+*, and pronominal clitics. The Arabic POS tags (26 POS tags) were obtained with the Stanford parser on all taggable tokens.

Word alignment. Word-to-word alignment for both directions was generated with the standard GIZA++ package and 50 statistical word classes. The intersection of direct and inverse word alignments is used in SBR training.

Tuples and phrase translation probabilities, along with lexical weights, are calculated on the extended alignment matrix using the **grow-diag-final** method. Rather than using preface forms of English words, we used English stems following the point from [dG06].

Parsing. The SBR system requires access to source and target language parse trees, along with the intersection of source-to-target and target-to-source word alignments. In the framework of the study we used the Stanford parser as an NLP parsing engine in Arabic-to-English and Chinese-to-English experiments[36].

The Stanford parser was trained on the respective treebank sets: the English treebank is provided with 14 syntactic tags, the Arabic treebank has 23 syntactic categories, and the Chinese treebank operates with 44 constituent categories.

Language model. The target-side LM was estimated using the SRI language modeling toolkit to calculate all N-gram LMs (including the N-gram-based TM).

The values and features that vary from those that are provided in this section, as well particular tools and parameters for each set of experiments, are summarized in the

[36] Generally speaking, the system permits using any other natural language parser allowing for different formal grammars for the source and the target languages.

corresponding section.

5.3.3 Experiments with phrase-based SMT

In the first set of experiments, we assessed the impact of the proposed reordering method on phrase-based SMT performance. We trained a translation system with the 2008 version of Moses, following the guidelines provided on the Moses site in all the points except for reordering model usage.

Experimental setup

Baseline system characteristics that were used in the set of phrase-based experiments are summarized in Table 5.2.

	Chinese-to-English	Arabic-to-English
Word alignment	GIZA++	
Symmetrization	GDF (intersection for SBR)	
TM	Phrase-based, max.length=20 (Moses default)	
LM	4-gram, modified Kneser-Ney discounting with interpolation	
Decoding	Monotonic, beam=50 (10 during decoding)	
Reordering	No (*plain* configuration) Lexicalized reordering [Til04](*baseline* configuration)	
Optimization criteria	BLEU	

Table 5.2: *Phrase-based system parameters. SBR experiments.*

The scores considered are: BLEU scores obtained for the development set as the final point of the MERT procedure (*Dev*); and BLEU and METEOR scores obtained on the test dataset (*Test*). Automatic evaluation conditions were case-insensitive and punctuation marks were taken into account.

Core experiments

As stated earlier, we experimented on Chinese-to-English and Arabic-to-English translations in the news and tourism domains, which, apart from the topic, differ in amount of training material and average sentence length. Both translation tasks convey local and global reorderings.

Five SMT systems were contrasted considering the set of experiments carried out on the phrase-based system:

1. *Baseline* refers to the Moses baseline system; the training data are not reordered, and a lexicalized reordering (MSD) model [Til04] is applied with default parameters;

2. The *Plain+WL* system is the standard Moses configuration with lexicalized reordering capabilities turned off; the input of the decoder is represented as a POS-based word lattice, trained and built in accordance with the reordering framework presented in [Cre08a];

3. *Plain+SBR* refers to the monotonic system configuration with syntactically reordered source parts, only isomorphic sub-tree transfer ("main" rules) are considered; the development and test datasets are monotonically decoded;

4. In the *Plain+SBR+NI* configuration SBR implements both isomorphic ("main" rules) and non-isomorphic ("secondary" rules) sub-tree transfer; development and test datasets are monotonically decoded;

5. The *Plain+SBR+NI+WL* system employs a complete circle of the SBR reordering including syntax-based source-side monotonization (isomorphic and non-isomorphic sub-tree transfer considered); a word graph input sentence representation provides the decoder with various reordering paths.

Table 5.3 presents BTEC results for the Chinese-to-English and Arabic-to-English tasks. Table 5.4 shows the results obtained on the NIST corpus. Best scores are placed in cells filled with grey.

The *Plain+WL* configuration shows the effect of Crego's algorithm applied on an unreordered data and can be seen as an alternative baseline system.

Accurate selection of specific values for the cut-off thresholds for each group's set of rules is of critical importance in the SBR system. Apart from its use as an instrument to eliminate noisy patterns, which can appear as a result of alignment or parsing errors, in some cases a pruning mechanism drives the process of reordering rule application and establishes a balance between more specific and less specific patterns. Aggressive pruning of general and partially lexicalized rules increases the system's accuracy. On the other hand, such pruning decreases its generalization capability.

Results of comparative experiments conducted to empirically determine the optimal combination of cut-off thresholds for BTEC and NIST tasks are provided in §5.3.6. In all BTEC experiments, fully lexicalized rules are not pruned ($k_{lex} = 0$); rather, the threshold

k_{gener} is set to 3 and $k_{part} = 1$. In Arabic-to-English NIST experiments the cut-off parameters are set to the values: $k_{lex} = 1$, $k_{part} = 5$, $k_{gener} = 5$; and in Chinese-to-English NIST experiments they are set to $k_{lex} = 1$, $k_{lex} = 7$, $k_{gener} = 9$.

Discussion of BTEC results. The results for the Chinese-to-English BTEC task show a promising potential for the SBR algorithm. Non-isomorphic tree mapping impacts negatively on translation performance to a negligible degree (-0.05 BLEU and -0.17 METEOR points on the test set). However, the introduction of word lattice results significantly improves translation accuracy and allows the gain of ≈ 1.3 BLEU points on unseen data. We explain such a marked improvement in translation quality by the effect introduced by additional local reordering dependencies captured by the word lattice trained on the pre-reordered text, which were not considered when the input graph was trained on the unreordered text.

	dev	test BLEU	test METEOR
BTEC ZhEn experiments			
Baseline	48.52	47.21	68.33
Plain+WL	48.31	47.07	68.14
Plain+SBR	48.75	47.52	68.59
Plain+SBR+NI	48.79	47.47	68.56
Plain+SBR+NI+WL	48.90	48.78	68.85
BTEC ArEn experiments			
Baseline	48.46	47.10	68.10
Plain+WL	48.17	46.71	67.62
Plain+SBR	48.91	47.76	67.75
Plain+SBR+NI	48.71	47.49	67.33
Plain+SBR+NI+WL	48.65	47.43	67.27

Table 5.3: *Summary of the BTEC experimental results carried out on the phrase-based SMT system.*

	dev	test BLEU	test METEOR
NIST ZhEn experiments			
Baseline	55.52	49.29	65.10
Plain+WL	55.54	49.18	65.12
Plain+SBR	55.49	49.87	65.71
Plain+SBR+NI	55.60	50.03	65.82
Plain+SBR+NI+WL	55.65	50.12	65.91
NIST ArEn experiments			
Baseline	48.84	48.91	59.64
Plain+WL	48.21	48.50	59.18
Plain+SBR	49.09	49.22	59.75
Plain+SBR+NI	49.13	49.38	60.19
Plain+SBR+NI+WL	49.86	49.43	60.17

Table 5.4: *Summary of the NIST experimental results carried out on the phrase-based SMT system.*

Arabic-to-English results for the BTEC task are not so definite; the best system configuration is *Plain+SBR*, and degradation of translation performance is observed simultaneously with introduction of *NI* and *WL* features. However, the statistical significance test[37] of BLEU scores reveals that translation scores shown by all Arabic-English BTEC systems including SBR are very similar and do not differ statistically (in this case, the statistical significance interval is about ±0.5 BLEU points). By contrast, the performance shown by the best SBR system (*Plain+SBR*) statistically differs from both baseline configurations in terms of BLEU score measured on the test dataset.

A phenomenon of lack of correlation between development and test results similar to the one described in §3.4.5 can be observed by analyzing Table 5.3 (see *Plain+SBR* and *Plain+SBR+NI* systems for the Chinese-to-English task). Although the simplex optimization algorithm converges at a higher point (*Plain+SBR+NI*), test results are negligibly worse.

[37] All statistical significance calculations supporting a better comparison among different systems are done for a 95% confidence interval and 1000 resamples, following the guidelines from [Koe04].

Discussion of NIST results. NIST results are clearer and thus easier to analyze. As in the case of the BTEC task, application of the SBR technique demonstrates an improvement in translation quality according to automatic scores.

Plain+SBR+NI+WL is found to be the best system configuration analyzing both sets of NIST experiments. Although results shown by SBR systems are very similar, the most significant observation is that introduction of SBR reordering (with non-isomorphic tree transfer capabilities) leads to a performance improvement over the *baseline* configuration by about 0.7 BLEU points for Chinese-to-English (1.4 %) and by about 0.5 (1.1%) for Arabic-to-English. These observations are statistically significant for the NIST tasks (the thresholds are ±0.6 BLEU points and ±0.5 BLEU for Chinese-to-English and Arabic-to-English NIST tasks, respectively). The METEOR score also increases with an increase in reordering system complexity, supporting the BLEU results.

Another important conclusion that can be drawn from the comparison of non-deterministic systems trained on the unreordered and reordered data (*Plain+WL* and *Plain+SBR+NI+WL*) is that the introduction of syntactically motivated reordering capabilities in a non-deterministic way leads to a significant improvement in the score of about 1 BLEU point and 0.8-1 METEOR points in both directions, which is a numerical expression of an aggregate effect of isomorphic and non-isomorphic *SBR* introduction.

Syntax-based rewrite rules. As mentioned above, SBR operates with three groups of reordering rules, which are the product of complete or partial delexicalization of the originally extracted patterns. The groups are processed and pruned independently. Basic rules statistics for both translation tasks can be found in Table 5.5 and 5.6 (all the numbers are provided after pruning and employing cut-off values from the previous paragraph; for further information concerning pruning strategy refer to §5.3.6).

The majority of the reordering rules consist of two or three elements. In the case of the Arabic-to-English BTEC task no patterns comprise more than three nodes. Considering Chinese-to-English BTEC translation, a few rules contain more than 3 elements (3.2 % of lexicalized patterns, 5.5 % of partially lexicalized and 19.4 % of general rules).

Rules for the NIST task involve more elements in the reordering process (up to 8). In addition, there are some long lexicalized rules (7-8), generating a high number of partially lexicalized patterns.

This rule distribution can be explained by the fact that the Stanford parser tends to

Group	# of rules	Voc	2-element	3-element	4-element	[5-8]-element
BTEC experiments						
Specific rules	703	413	406	7	-	-
Part. lex. rules	1,306	432	382	50	-	-
General rules	259	5	5	-	-	-
NIST experiments						
Specific rules	9,896	6,715	5,891	472	72	259
Part. lex. rules	172,374	52,945	41,878	7,816	1,010	2,241
General rules	1,053	515	180	90	72	30

Table 5.5: *Basic reordering rules statistics (Arabic-to-English).*

Group	# of rules	Voc	2-element	3-element	4-element	[5-8]-element
BTEC experiments						
Specific rules	2,874	1,663	1,609	45	9	-
Part. lex. rules	3,688	2,912	2,750	158	4	-
General rules	519	103	83	18	2	-
NIST experiments						
Specific rules	22,395	14,802	10,173	2,495	1,827	307
Part. lex. rules	210,604	77,590	53,652	10,821	7,720	5,397
General rules	928	433	361	39	21	12

Table 5.6: *Basic reordering rules statistics (Chinese-to-English).*

generate a deep tree structure with few nodes on each level. Longer rules are principally produced by the non-isomorphic mapping algorithm, tending to reconstruct the target-side word order structure by placing source-side sub-trees in the proper order. On the other hand, very few multi-node sub-trees are generated by the Stanford parser, which are more likely to appear when longer sentences are parsed.

Repeatability of rules is about 10 % higher for the BTEC tasks; in particular, 59.3 % and 57.8 % of all extracted rules are unique to the case of BTEC Arabic-English and Chinese-English translation tasks, respectively, versus 67.8 % and 66.09 % for the corresponding

tasks using the NIST corpus.

Tables 5.7 and 5.8 show some examples of reordering rules from each group.

	Specific rules
1A	*NN@0 NP@1 → NP@1 NN@0 | NN@0 << Asm >> NP@1 << +y >> | 2.7E-2*
2A	*DTNN@0 DTJJ@1 → DTJJ@1 DTNN@0 | DTNN@0 << AlAmm >>DTJJ@1 << AlmtHdp >> | 5.2E-2*
3A	*NN@0 NP@1 → NP@1 NN@0 | NN@0 << Asm >> NP@1 << +k >> | 2.6E-2*
4A	*NN@0 JJ@1 → JJ@1 NN@0 | NN0@ << ymyn >> JJ1@ << +k >> | 2.4E-2*
	Partially lexicalized rules
5A	*NN@0 NP@1 → NP@1 NN@0 | NN@0 << NON >> NP@1 << +k >> | 9.0E-2*
6A	*NN@0 JJ@1 → JJ@1 NN@0 | NN@0 << NON >> JJ@1 << +k >> | 5.9E-2*
7A	*DTNN@0 DTJJ@1 → DTJJ@1 DTNN@0 | DTNN@0 << NON >>DTJJ@1 << AlmtHdp >> | 1.7E-3*
8A	*NN@0 NNP@1 → NNP@1 NN@0 | NN@0 << NON >>NNP@1 << $rm >> | 1.7E-3*
	General rules
9A	*NN@0 JJ@1 → JJ@1 NN@0 | 3.9E-1*
10A	*JJ@0 NN@1 PP@2 → NN@1 JJ@0 PP@2 | 9.6E-2*
11A	*RT@0 VBP@1 NP@2 → PRT@0 NP@2 VBP@1 | 6.5E-2*
12A	*NN@0 NNP@1 → NNP@1 NN@0 | 3.4E-2*

Table 5.7: *Examples of Arabic-to-English reordering rules.*

Reordering rules can be driven by POS tags only (see lines 2A, 6A, or 9A, for example), syntactic constituents only (2B, 3B, 8B), or by both categories simultaneously (3A, 4A, 8A). Patterns can be monotone, without violating source-side order (3B, 4B, 10B) or swapping all or some of the rule elements (5A, 2B, 8B). They can include a single token as a reordering element (1A, 1B) or involve a sequence of tokens in the reordering process (2B).

It is worth noting that 7A is a delexicalized version of the rule provided in 2A with generalized determiner-noun box, and 12A is a fully delexicalized version of 8A.

Translation examples and discussion. Examples 5.11-5.13 demonstrate how two reordering techniques interact within a sentence taken from the NIST corpus with a need for both global and local word permutations for an Arabic-to-English translation task. Examples of SBR application are **highlighted in bold**, while translated clumps that need local reorderings are underlined.

Ar. plain: ***AElnt*** *Ajhzp AlAElAm l* *bEvp AlAmm AlmtHdp fy syrAlywn An ...*
Gloss: **announced** press release by mission nations united in sierra leone that ... (5.11)
Eng. ref.: ' a press release by the united nations mission to sierra leone **announced** that ... '
Baseline: ' the media and the united nations of mission in sierra leone that ... '

	Specific rules
1B	*NP@0 LC@1 → LC@1 NP@0* | *NP@0 << 饭>> LC@1 << 后>>* | *2.8E-4*
2B	*ADVP@0 VP@1 → VP@1 ADVP@0* | *ADVP@0 << 然后>> VP@1 << 请 填写 这个 通关 申报 表格>>* | *7.8E-5*
3B	*QP@0 CP@1 NP@2 → QP@0 CP@1 NP@2* | *QP@0 << 个>> CP@1 << 安静>> NP@2 << 房间>>* | *1.8E-4*
4B	*IP@0 VP@1 → IP@0 VP@1* | *IP@0 << 高兴>> VP@1 << 见到你>>* | *3.4E-4*
	Partially lexicalized rules
5B	*NP@0 VP@1 → VP@1 NP@0* | *NP@0 << NON >> VP@1 << 哪>>* | *7.9E-2*
6B	*ADJP@0 NP@1 → ADJP@0 NP@1* | *ADJP@0 << NON >> NP@1 << 邮件>>* | *3.1E-4*
7B	*NP@0 LC@1 → LC@1 NP@0* | *NP@0 << NON >> LC@1 << 前>>* | *5.8E-4*
8B	*NP@0 PP@1 VP@2 → VP@2 PP@1 NP@0* | *NP@0 << NON >> PP@1 << NON >> VP@2 << 打电话>>* | *6.5E-5*
	General rules
9B	*NP@0 LC@1 → LC@1 NP@0* | *8.2E-3*
10B	*DNP@0 ADJP@1 NP@2 → DNP@0 ADJP@1 NP@2* | *1.1E-3*
11B	*IP@0 LC@1 → LC@1 IP@0* | *1.6E-3*
12B	*QP@0 DNP@1 ADJP@2 NP@3 → QP@0 ADJP@2 DNP@1 NP@3* | *8.7E-5*

Table 5.8: *Examples of Chinese-to-English reordering rules.*

As can be seen from Example 5.11, the Moses baseline system omits translation of "*AElnt/announced*" placed in the original position at the beginning of the sentence.

Another issue is a complex reordering within the clump "*bEvp AlAmm AlmtHdp/mission nations united*", requiring two local permutations to get the correct word order ("*AlAmm AlmtHdp/nations united*" → "*AlmtHdp AlAmm/united nations*" and "*bEvp AlmtHdp AlAmm/ mission united nations*" → "*AlmtHdp AlAmm bEvp/united nation mission*"). In the course of the performance of permutations, two extra words are embedded in the clump, distorting its meaning.

By contrast, if the source sentence is syntactically reordered in the preprocessing step and the word "*AElnt/announced*" is moved to the ninth position, the generated translation contains the content verb "*announced*", and the output is more fluent and preserves the adequate meaning (Example 5.12).

Ar. reord.:	*Ajhzp*	*AlAElAm*	*l*	*bEvp AlmtHdp AlAmm fy syrAlywn*	***AElnt*** *An ...*
Gloss:	press	release	by	mission united nations in sierra leone	**announced** that ...

(5.12)

Eng. ref.: ' a press release by the united nations mission to sierra leone **announced** that ... '
Plain+SBR+NI: ' the media and the mission of united nations in sierra leone **announced** that ... '

Example 5.13 illustrates the handling of local reordering permutations done with a word lattice algorithm. In this case, the reordering leads to better clump translation motivated by the rule predicted word order. Although the deterministic approach translates the underlined clump as "*mission of united nations*", integration of the pre-estimated word lattice allows the decoder to find another translation "*united nations mission*" that is much better in terms of BLEU score (because the latter translation is closer to the reference translation). For clarity's sake, it must be mentioned that this translation can be considered a paraphrasing of the fluent translation shown in Example 5.12.

Ar. reord.:	*Ajhzp*	*AlAElAm*	*l*	*bEvp AlmtHdp AlAmm fy syrAlywn*	***AElnt***	*An ...*
Gloss:	press	release	by	mission nations united in sierra leone	**announced**	that ...

(5.13)

Eng. ref.: ' a press release by the united nations mission to sierra leone **announced** that ... '
Plain+SBR+NI+WL: ' the media and the united nations mission in sierra leone **announced** that ...'

5.3.4 Experiments with N-gram-based SMT

The proposed approach was also evaluated on the N-gram-based SMT system of [Mar06b], which is an alternative to the phrase-based translation approach, and which has proved to be competitive with the state-of-the-art systems in recent evaluation campaigns [Kha08, Lam07b]. A detailed description of the N-gram-based approach can be found in chapter 2.

The experiments were conducted on the BTEC'07 (§A.5) and the NIST'06 (§A.4) Chinese-English corpora. The BTEC'08 Arabic-English corpus (§A.7) was also used in experiments applying a translation unit blending strategy.

Experimental setup

The baseline system characteristics that were used in the set of N-gram-based experiments can be found in Table 5.9.

In N-gram-based experiments, we used the **MaxJumps** constraint distortion model, as briefly described in chapter 4. Given that the bilingual n-gram is estimated over the reordered set of tuples (unfolded tuples), two parameters are used to restrict the search during decoding time:

- A distortion limit (m): any source word (or tuple) is allowed to be reordered only if it does not exceed a distortion limit, measured in number of source words.

	Chinese-to-English BTEC	Chinese-to-English NIST
Word alignment	GIZA++	
Symmetrization	Union (intersection for SBR)	
Tuples	Unfolded	
NULL-source tuples	IBM1 model	
Embedded words	No	
TM	4-gram, modified Kneser-Ney discounting with interpolation	
LM	4-gram, modified Kneser-Ney discounting	
Other features	WP, LEX1, LEX2, POS LM	
POS LM	4-gram, Good-Turing discounting	
Decoding	Non-monotonic, distance-based distortion	
Pruning	Histogram, tnb=10	Histogram, tnb=20
Reordering	A word distance-based distortion model [Cj06c], m=5, j=5	
Optimization criteria	100BLEU+4NIST	

Table 5.9: *N-gram-based system parameters. SBR experiments.*

- A reordering limit (j): any translation path is allowed to perform only j reordering jumps.

The use of these constraints implies a necessary trade-off between quality and efficiency, depending on the difficulty of the task. For more details, refer to [Cre05d, Cre06a].

For all system configurations, apart from monotone experiments, parameters of the distance-based reordering model were set to $m = 5$ and $j = 5$ for a fair trade-off between efficiency and accuracy.

Core experiments

The set of core experiments with N-gram-based SMT contrast systems was performed considering the BTEC'07 and NIST'06 Chinese-English corpora.

As in the previous section, we reported the final scores obtained as a result of model weights tuning for a development dataset done with the simplex algorithm (*dev*), along with BLEU and METEOR scores for the test dataset (*test*). Evaluation conditions were case-insensitive, and punctuation marks are taken into account.

We contrasted three N-gram-based system configurations comparing the SBR results with the **MaxJumps** distortion model:

- *Baseline*: the training data are not reordered and allow for constrained distortion ($m = 5, j = 5$) during decoding;
- *Baseline+SBR*: SBR is applied in the preprocessing step involving "main" rules only; the development and test sets are monotonically decoded and constrained distortion is considered;
- *Baseline+SBR+NI*: SBR is applied involving isomorphic ("main" rules) and non-isomorphic ("secondary rules") sub-tree transfer; the constrained distortion ($m = 5, j = 5$) is allowed.

The automatic evaluation scores for BTEC corpus are reported in Table 5.10 and those for the NIST task in Table 5.11.

	dev	test BLEU	test METEOR	# tuples	voc tuples
BTEC ZhEn experiments					
Baseline	48.17	46.02	66.98	150,378	36,643
Baseline+SBR	48.55	46.67	67.92	157,345	36,936
Baseline+SBR+NI	48.83	46.52	67.82	151,430	36,501

Table 5.10: *Summary of the BTEC experimental results carried out on the TALP-UPC N-gram-based SMT system.*

	dev	test BLEU	test METEOR	# tuples	voc tuples
NIST ZhEn experiments					
Baseline	58.28	47.04	62.38	4,761,412	1,692,397
Baseline+SBR	58.80	47.31	62.54	5,374,819	1,801,515
Baseline+SBR+NI	58.96	47.57	62.77	5,379,902	1,805,789

Table 5.11: *Summary of the NIST experimental results carried out on the TALP-UPC N-gram-based SMT system.*

Note that development scores obtained from the NIST'06 corpus are not comparable to the corresponding results from the phrase-based experiments due to a difference in optimization metric (BLEU vs. 100BLEU+4NIST).

Columns 5 and 6 in both tables show the number and the vocabulary of tuples extracted from the training corpus (unfolded algorithm).

Discussion of results. For both sets of Chinese-English experiments, application of the SBR technique demonstrates improvement in translation quality according to the automatic scores.

Results for Chinese-to-English BTEC task show consistent improvement in terms of translation quality when SBR features are incorporated (0.65 BLEU points (of 1.4 %), which is statistically significant for the BTEC task). As in phrase-based experiments, introduction of non-isomorphic sub-tree transfer has no positive effect on the MT scores (≈ 0.15 points of BLEU degradation; the METEOR score is slightly worse for the $Baseline + SBR + NI$ configuration).

N-gram-based NIST systems demonstrate a very similar behavior when compared with phrase-based translation. The *Baseline+SBR+NI* system is found to be the best configuration, outperforming the *Baseline* configuration by about 0.53 BLEU points (1.15 %) and reaching a statistical significance threshold (±0.5 BLEU points). The METEOR score also increases with a rise in reordering system complexity, which supports the BLEU results.

Although the maximum theoretical number of tuples for a given word alignment can be extracted with the unfolding algorithm, re-running of the alignment procedure on the pre-reordered training data helps to locate additional correct links between far tokens. Number and vocabulary of extracted tuples are very important indicators of an N-gram-based system performance. The system, which has at its disposal a higher number of elementary bilingual units, is able to construct a more complete and fluent translation of the unseen data.

Translation examples. Examples 5.14 and 5.15 illustrate source monotonization with the use of the SBR algorithm. Standard N-gram-based SMT cannot generate a fluent English translation of the Chinese sentence, primarily due to the system's inability to move the sequence "从 史密斯 先生 那儿/*from smith mr there*" to the end of the sentence.

The clump subject to SBR reordering is **highlighted in bold**.

Zh. plain:	我	真	高兴		格林	先生	我	**从**	**史密斯**	**先生**	**那儿**	听到	很多	有关	你的	情况	
Gloss:	I	so	pleased		green	mr	I	**from**	**smith**	**mr**	**there**	hear	a lot	about	your	situation	(5.14)

Eng. ref.: ' the pleasure is all mine mr green i 've heard a lot about you from mr smith '

Baseline: ' I am very glad mr green **mr smith where** I heard a lot about your circumstances '

Example 5.15 shows how the N-gram-based translation system can benefit from long-range permutation of the Chinese clump "从 史密斯 先生 那儿" such that it better matches the structure of the English counterpart.

Zh. reord.: 我 真 高兴 格林 先生 我 听到 很多 有关 你的 情况 **从 史密斯 先生 那儿**
Gloss: I so pleased green mr I hear a lot about your situation **from smith mr there** (5.15)
Eng. ref.: ' the pleasure is all mine mr green i 've heard a lot about you from mr smith '
Baseline: ' mr green I am very glad I heard a lot of your circumstances there **from mr smith** '

Local reorderings, namely, 格林 先生/*green mr* → 先生 格林/*mr green* and 史密斯 先生/*smith mr* → 先生 史密斯/*mr smith*, are tackled by regular tuples exploiting alignment crossings.

The example shows a nice symbiosis of the SBR method and tuple-based translation. It also shows that SBR does not bias against tuple internal reordering

5.3.5 Deriving benefit from a purely generalized SBR

This section describes an alternative way to use SBR, namely, for a *blending* strategy of translation units that combines original[38] and reordered bilingual tuples extracted from the parallel corpora with unreordered and reordered source parts. The augmented set of tuples is next passed to the N-gram-based SMT system, increasing its capability to find a better translation of unseen data.

Method

In terms of this study, we operate exclusively with generalized (i.e., unlexicalized) reordering rules that can cause reordering errors induced by a specific number of grammatical exceptions, which can be easily found in any language. As described in the previous sections, one way to address this problem is through full or partial lexicalization of the reordering patterns. An alternative approach, described here, accounts for a finite-state transducer architecture of the N-gram-based SMT and is called the *tuple blending* model.

Once the corpus with reordered source part is aligned, two sets of tuples are extracted based on the reordered and original alignment matrices. In the final stage of the TM construction, the bilingual units from these sets are combined following the criterion of maximizing the number of tuples at the sentence level. This technique entails more tuples involvement in TM construction, which provides better bilingual generalization (shorter translation units have higher probability of appearance in the translating corpus than longer ones).

Figure 5.14 illustrates the process of tuple derivation from the aligned bilingual sentence following the regular unit extraction method. Given a word alignment along with the

[38] Here, we refer to the tuples extracted with the "regular" algorithm.

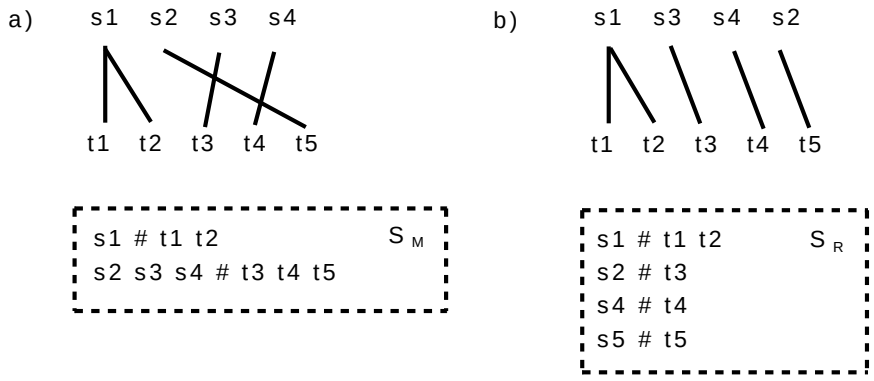


Figure 5.14: *Tuple extraction from an unreordered and a correctly reordered bilingual sentences.*

bilingual sentence with original and reordered source parts, two alternative sets of tuples can be extracted (S_M and S_R). The decision of which set will be considered during TM generation is made based on the simple criterion of number of extracted units. In the example provided in 5.14, S_R will be taken since it consists of four shorter tuples, in contrast with the plain configuration, which allows for only two units.

Parse errors, redundant generalization and lack of capability to consider lexical exceptions are handled, as shown in Figure 5.15. S_M set of tuples is generated from the unreordered sentence (a) and provides the decoder with four unique tuples. Example (b) shows a situation in which erroneous source-side reordering leads to a permutation $s2\,s3 \leftrightarrow s4\,s5$ and generation of the two-element set of tuples (S_R). In this case, S_M will be taken into account during TM construction following the criterion of maximizing the number of tuples at the sentence level.

Experiments.

The experiments were performed on the Arabic-English BTEC'08 (§A.7) and Chinese-English BTEC'07 corpora (§A.5). We reported final BLEU scores obtained on the development set as the final point of the optimization procedure, with automatic translation results measured on the test set and the total number of extracted tuples.

We considered four translation systems:

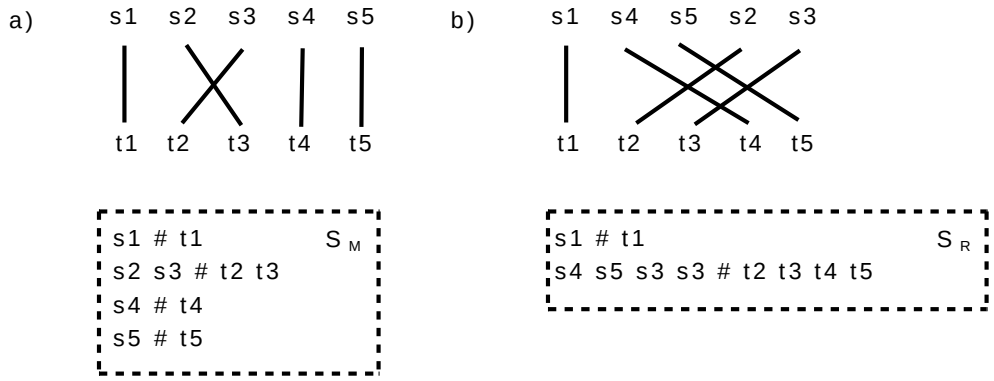


Figure 5.15: *Tuple extraction from an unreordered and an erroneously reordered bilingual sentences.*

- A *monotonic* model corresponding to the N-gram-based systems as described in Table 5.9 but with the tuples extracted using a "regular" scheme. The decoding is monotonic. In this case the parallel corpora with monotonic source part are involved in TM construction, and no distortion model is provided.

- A *reordered* model refers to the N-gram-based system trained and tuned on the data with reordered source part and allowing for monotonic decoding. Only generalized rules are taken into account ($k_{gener} = 3$) during SBR.

- A *blending* model as previously described in this section.

- The alternative *MaxJumps* reordering, which includes the "unfold" algorithm of tuples extraction and constrained distance-based distortion model used in the decoding step (as described in [Cj06c]) with distortion and reordering limits set to 5 ($m = j = 5$).

Experimental results for both tasks are provided in Table 5.12.

The Arabic-to-English N-gram-based *blending* SMT system outperformed the distance-based constrained search technique by about 0.4 BLEU points and 0.2 METEOR points. However, because the *MaxJumps* final optimization point is ≈ 0.4 BLEU above the *blending* system, we assume that it can be due in part to inconsistency of development and test results that are typical for small translation tasks. Results obtained with the *blending* algorithm outperformed both *monotonic* and *reordered* systems by 4.14 BLEU points (4.1 %) and

	dev	test BLEU	test METEOR	# tuples
BTEC ArEn experiments				
Monotonic	40.55	43.78	57.17	135,855
Reordered	41.05	45.15	58.28	143,934
Blending	43.20	47.92	59.33	170,572
MaxJumps	43.61	47.46	59.14	163.755
BTEC ZhEn experiments				
Monotonic	45.77	43.81	64.52	103,714
Reordered	47.90	45.54	66.25	124,955
Blending	48.02	45.93	66.84	150,378
MaxJumps	48.17	46.02	66.98	142,802

Table 5.12: Summary of the tuple blending experimental results.

2.77 BLEU points (5.8 %), respectively. These results show that sentence-level combining of tuples extracted from the corpus with original and reordered source parts is useful in terms of translation accuracy.

Chinese-to-English results show that the bilingual units *blending* method is competitive with the alternative reordering algorithm. In this case, *blending* configuration outperforms the *monotonic* system by about 2.1 BLEU points when translating an unseen dataset and about 0.4 BLEU points in comparison with the system trained on a syntactically pre-reordered corpus. For the Chinese-to-English translation task, the *blending* and *MaxJumps* reordering algorithms showed comparable performance, significantly outperforming both the *monotonic* and the syntactically *reordered* SMT systems.

5.3.6 Effect of rule pruning

The pruning strategy applied to each of the sets of reordering rules had a strong influence in the final reordering decision taken in the deterministic step. In this section we present the results of an empirical study dedicated to finding the optimal combination of pruning strategies for SBR. In particular, we evaluate the quality gain that can be achieved via accurate selection of cut-off thresholds for each group of reordering rules on the Arabic-to-English BTEC and NIST tasks.

Results are shown in Table 5.13. M_{lex}, M_{part} and M_{geber} refer to the size of lexicalized, partially lexicalized, and general reordering models, respectively. BLEU and METEOR scores are estimated on the test corpus. Evaluation conditions were case-insensitive with punctuation marks taken into account.

k_{lex}	M_{lex}	k_{part}	M_{part}	k_{gener}	M_{gener}	BLEU	METEOR	sec./sent.
BTEC ArEn experiments								
0	703	0	19,034	0	515	46.88	68.07	4.7E-2
0	703	1	7,519	1	259	46.86	68.03	1.1E-2
0	703	2	4,913	1	259	47.12	68.20	6.4E-3
0	703	3	1,306	1	259	47.52	68.39	3.1E-3
0	703	4	1,107	1	259	47.51	68.36	3.0E-3
0	703	3	1,306	2	125	47.49	68.31	3.1E-3
1	104	3	1,306	1	259	47.41	68.22	2.2E-3
2	52	3	1,306	1	259	47.13	68.18	1.9E-3
NIST ArEn experiments								
0	68,901	0	830,026	0	11,027	-	-	2.7
0	68,901	3	244,902	0	11,027	-	-	2.1
0	68,901	5	172,374	0	11,027	48.92	59.27	1.7
0	68,901	7	159,416	0	11,027	48.58	59.02	1.4
1	9,896	5	172,374	0	11,027	48.99	59.35	1.2
2	3,660	5	172,374	0	11,027	48.81	59.29	1.1
1	9,896	5	172,374	1	5,262	49.17	59.69	1.2
1	9,896	5	172,374	3	1,692	49.20	59.71	1.1
1	9,896	5	172,374	5	1,053	49.22	59.75	1.1
1	9,896	5	172,374	7	973	49.15	59.68	1.0
1	9,896	5	172,374	9	901	49.12	59.68	1.0

Table 5.13: *Effect of pruning strategy on phrase-based system translation quality, on reordering model size, and on processing time.*

The experiments were done on a dual-processor Pentium IV Intel Xeon Quad Core X5355 2.66 GHz machine with 24 G of RAM. All computational time results are approximated. MT scores for minimally pruned NIST configurations are not provided due to time-consumption of the reordering procedure.

In the case of the BTEC task, values of $k_{lex} = 0$, $k_{part} = 3$, and $k_{gen} = 1$ produced a translation BLEU improvement of ≈ 0.3 points with respect to unpruned configuration. Active pruning of partially lexicalized rules is mandatory for large translation tasks due to

an exponential growth of model size when the corpus and average sentence length increase. For the BTEC task, this impacts positively on the automatic translation scores: changing the k_{part} value from 3 to 2 leads to a translation BLEU reduction of ≈ 0.4 BLEU points and ≈ 0.2 METEOR points. On the other hand, the system is not highly sensitive to the changes of k_{gener} value, which can be explained by the fact that many reorderings are done with lexicalized or partially lexicalized patterns.

The optimal set of thresholds for the NIST task is $k_{lex} = 1$, $k_{part} = 5$, and $k_{gen} = 5$. Despite the cut-off values for partially lexicalized and general rules being equal, the motivation to set them to a relatively high value of 5 is different.

Generalized patterns can be wrong for some examples. That is why we consider an aggressive pruning strategy of general patterns reasonable and maintain the higher value of thresholds for general patterns. This approach limits the number of the patterns that do not capture a certain number of grammatical exceptions, which can be easily found in any language.

The increase in k_{part} value leads to a very important decrease in the size of the partially lexicalized reordering model. The latter varies within the range of 89-95 % of the total size of the SBR model, depending on the other threshold values. On the other hand, setting the threshold to a value higher than 5 causes a degradation in translation quality.

As in the case of the BTEC task, the reordering system is less sensitive to general rule threshold variations, possibly due to the wide coverage of reordering instances by other groups of rules.

5.4 Discussion and conclusions

In this chapter, we have described a new approach to word reordering in SMT, which successfully integrates syntax-based reordering in phrase-based and N-gram-based SMT. This approach correlates with the human intuitive notion of translation. At its best, a successful translation should read as if it were originally written in the new language. When translating a sentence, a human defines the target-language word order according to an extensive set of grammatical and semantic rules, along with abundant exceptions determined by certain lexical features, thereby neutralizing the "natural" word order [EZ81].

The SBR model described in this chapter similarly models reordering transformations in the pre-translation step:

- First, the system automatically learned a set of syntactic reordering patterns that exploit systematic differences between word order of source and target languages. Intending to avoid the pitfall of overgeneralizing, which can result in the loss of a relatively solid body of lexically motivated dependencies, we introduced three sets of reordering rules characterized by different generalization levels.

- In the next step, the rules were applied to the source part of the same training corpus changing the source sentence structure such that it more closely matches the word order of the target language.

- Finally, we showed that the translation quality can be improved by coupling the SBR algorithm with a state-of-the-art word alignment-based reordering framework, which was applied during decoding.

We provided a qualitative analysis of the extracted reordering patterns and studied the effect that reordering rules learned under different circumstances have on translation.

We also examined the idea that the proposed syntactically motivated reordering coupled with the bilingual units blending method applicable to the N-gram-based SMT impacts positively on the translation accuracy and shows competitive performance in comparison with an alternative reordering model.

The method was tested on tasks with a high need for word reorderings, namely, translating from Chinese and Arabic into English. We achieved significant improvement in translation quality when that SBR technique was applied. We also showed that employing a non-deterministic algorithm to extend a search graph with reordering hypotheses significantly outperforms state-of-the-art distortion models.

We started the evaluation on smaller translation corpora with a high need for local reordering; we then showed that the method scales to a large training set requiring long-distance permutations caused by longer sentences. We showed that the improvement measured by automatic evaluations is maintained for larger corpora.

Chapter 6

Conclusions and future work

In this chapter, we summarize the main results and most significant achievements of this Ph.D. dissertation and point out several research aspects that could be addressed in future work.

6.1 Conclusions

Translating from one language to another is one of the most complex higher order activities of the human brain. Despite the success of the statistical approach in the task of MT, the problem of efficient and effective word and phrase reordering is still far from being solved. This thesis extends the state-of-the-art in SMT, continuing the tradition of introducing linguistic and syntactic knowledge into SMT [Elm08, dG06]. We also address the problem of accurate and efficient target language modeling for SMT, which we consider as a secondary direction of the research work presented in this thesis. The experimental field, which in the majority of alternative works is limited to phrase-based translation models, is extended with a joint-probability approach estimating an n-gram of bilingual tuples.

In particular, the scientific contributions of the thesis include the following:

- We showed that existing and optimized algorithms of LM domain adaptation can be expanded to the wider problem of translation task adaptation, considering the translation of the unseen dataset as a monolingual text-classification problem. We reported on experiments that build an adaptive N-gram-based SMT system that translates slightly edited output of the speech recogniser (verbatim transcription).

We have shown that our proposed technique of general and specific target-side LMs interpolation provides better generalisation for English-to-Spanish and Spanish-to-English translation directions than a standard n-gram LM trained on a clean corpus.

- We identified an efficient LM pruning strategy that reveals a fair trade-off between translation quality and efficiency of the translation process. We described the impact of accurate threshold cut-off selection on the LM size, its noisiness, and final SMT performance.

- We presented an innovative technique of SMT enhancement with a continuous-space LM based on a neural network. This approach exploits the ability of non-linear statistical data models to learn distributed representations to reduce the impact of the curse of dimensionality. The main advantage of this approach to language modeling is that a continuous-space LM estimates the posterior probabilities as an interpolation of any possible context of length $n-1$ instead of backing-off to shorter contexts. The proposed alternative LM was introduced to the N-gram-based SMT system as a feature function and was used in the re-ranking step to re-score the k-best list of translation hypotheses. We reported results for an Arabic-to-English translation task, demonstrating a significant improvement in translation quality through a better target-side language model representation in contrast with the state-of-the-art LM approach.

- We proposed a novel word reordering approach that alleviates the word order challenge involving morpho-syntactical and statistical information in the context of a pre-translation reordering framework.

  In the framework of this part of the thesis, we provided a basic functional building block that can be exploited by a higher-level SMT system to efficiently address structural differences between source and target languages. Here, our interest was the value of syntax in word reordering for SMT, and the major statement was that morpho-syntactic information is useful in handling global reordering for the languages with high word order disparity.

  Features distinguishing the SBR approach from the alternative methods are:

  1. A two-step reordering approach addresses the reordering problem through integration of the deterministic (harmonizing source and target word order) and

non-deterministic approaches (using POS information to construct and extend an input graph model). We showed that the former is responsible for long-range reordering dependencies, while the latter deals better with local word permutations.

2. The use of a target-side tree as a filter in the rule extraction step limits reordering rules to those that conform to the target-side tree. We implemented the strategy to address discrepancies between source and target parse trees through a source-side tree traversal and target-side multi-level subtree reconstruction, aiming to put the source words in desirable (and, we hope, correct) word order.
3. Different from most current reordering systems that assume essentially isomorphic trees [Als00, Yam01], we used an idea similar to the extended domain of locality [Jos97] and implemented an extended tree-to-string transducer spanning multi-level subtrees on the source side of the corpus with competitive results.
4. We proposed a special technique applicable to N-gram-based SMT, consisting in sentence-level combination of bilingual tuples extracted from the parallel corpora with monotone and syntactically reordered source parts.

We have shown that the proposed reordering algorithm achieves better MT performance on the basis of the phrase-based and N-gram-based translation models than the standard distance-based reordering model. The experiments were performed for Chinese-to-English and Arabic-to-English translation tasks. The algorithm also demonstrated that the SBR model shows competitive performance for both small and large training sets with requirements for long- and short-range reorderings.

We believe that the work presented in this thesis provides a flexible concept, which can be generalized to a more complete framework of other cross-language NLP tasks.

6.2 Future work

The SBR algorithm and related framework present many opportunities for future work. In this section, we explore some possible extensions to the work in the thesis and describe some of the paths we wish to investigate in the future.

- The most obvious step of the SBR system evolution is to weight the local reordering suggestions with various probabilities based on the relative frequency of rule appearance. In the next stage we would like to enhance the proposed technique by introducing a weighted non-deterministic approach to address global reordering dependencies using the SBR algorithm in a non-deterministic way. It will prevent the irrevocable decisions made within the deterministic approach through complete integration of the reordering decisions in the SMT system.

- Although the SBR algorithm provides an accurate way to extract and apply reordering rules, we believe that a semi-supervised approach to reordering, including intermediate human-made revision of rewrite patterns, can be beneficial for the SMT. Word order harmonization performed on the basis of a set of automatically learned rules (see [Cj06b, Xia04]) has gained many adherents over the past few years. However, as far as we know, no research has been presented on the semi-supervised approach to word reordering, in which the automatically extracted reordering patterns are revised by a human, thereby penalizing or facilitating a series of transformations proposed by the machine.

- Moving beyond a reordering transduction scheme, we propose an alternative implementation of the SBR idea with a generative model utilizing synchronous transduction grammar to generate pairs of reordered and monotone strings. Generally, a recent trend in SMT has been toward the use of synchronous grammar-based formalisms in translation [Chi05, Ven06, Vil09]. We propose to extend the standard formal grammar mechanism to describe the admissible orderings. This approach seems to be theoretically more appealing, since it can lead to a more satisfactory model of reordering with superior integration into the SMT system.

- A crucial parameter of the models that employ source- and target-side syntactical information in the translation or reordering process is the quality and accurate tuning of the syntactical models for both languages. We plan to pursue the improvement of reordering by conducting a set of empirical experiments to tune the parsing parameters. We also speculate that adaptation of special parsing schemes can result in significant improvement in reordering quality and provide the system with more robust patterns.

- Finally, a possible line of future research would be to analyzing alternative ways of reordering rules generalization. We would like to investigate the possibility of employing dependency trees and chunk bracketing in the reordering process.

Appendices

Appendix A

Corpora description

A.1 EuroParl Spanish-English corpus. Version 2.

Table A.1 shows the basic statistics of the second version of the Spanish-English *EuroParl*[39] (European Parliament Plenary Session transcription) corpus [Koe05a].

Set	Language	Sentences	Words	Voc	ASL	References
			FTE			
Train	Spanish	1.28M	36.57M	153K	28.55	-
Train	English	1.28M	34.91M	106K	27.25	-
Dev	Spanish	430	15.33K	3.2K	35.66	2
Dev	English	735	18.76K	3.2K	25.35	2
Test	Spanish	840	22.77K	4.0K	27.11	2
Test	English	1,094	26.92K	3.9K	24.60	2
			Verbatim			
Dev	Spanish	792	25.6K	3.2K	32.35	2
Dev	English	1,194	30.2K	3.9K	25.30	2
Dev 5K	Spanish	500	15.48K	2.3K	30.97	2
Dev 5K	English	500	11.78K	2.2K	23.57	2
Test	Spanish	897	30.10K	4.7K	33.63	2
Test	English	1,155	30.48K	4.0K	26.39	2

Table A.1: *EuroParl corpus. Version 2. Basic statistics.*

Note: ASL refers to average sentence length

[39] In some tasks, this corpus is called *EPPS*.

This data have been prepared by the RWTH and proposed as training data to the participants of the second evaluation campaign in the context of the European Project TC-STAR.

A 5K development subset (*Dev 5K*) was used in the LM adaptation experiments (see §3.2) and consists of the first 500K sentences of the entire development set.

A.2 EuroParl verbatim corpus.

A monolingual corpus of EuroParl Spanish and English verbatim transcriptions was provided by ELDA and transcribed by UPC. Table A.2 shows the basic statistics of this corpus. Only training data are provided.

Set	Language	Sentences	Words	Voc	ASL	References
Train	Spanish	70K	512K	20K	29.16	-
Train	English	73K	781K	17K	26.03	-

Table A.2: *EuroParl verbatim corpus. Basic statistics.*

A.3 Italian-English BTEC corpus.

Italian-English Basic Travel Expression Corpus (BTEC) [Tak02] includes data from a tourist domain and was proposed to the participants within the open data track of IWSTL 2006 evaluation campaign. BTEC corpus is characterized by extremely limited amount of training data and models a real situation in which foreign tourist appears in an English-speaking country and needs simple explanations and other practical information useful for travelers.

Set	Language	Sentences	Words	Voc	ASL	References
Train	Italian	24.5K	166.3K	10.2K	6.54	-
Train	English	24.5K	155.4K	7.3K	6.15	-
Dev	Italian	489	5.2K	1.2K	10.18	7
Test	Italian	500	6K	7.3K	6.94	7

Table A.3: *BTEC ItEn corpus. Basic statistics.*

A.4 Chinese-English NIST'06 corpus.

Table A.4 presents a corpus that we call "NIST'06". The training data consists of news and broadcast data proposed to the participants within the NIST[40] 2006 evaluation campaign. The training data used is available from the Linguistic Data Consortium[41].

Set	Language	Sentences	Words	Voc	ASL	References
Train	Chinese	1.02M	26.23M	157K	25.80	-
Train	English	1.02M	27.25M	216K	26.71	-
Dev	Chinese	500	14.05K	3.7K	28.10	4
Test	Chinese	3,940	85.2K	10.4K	6.94	4

Table A.4: *NIST'06 ZhEn corpus. Basic statistics.*

To tune system parameters we used a 500-line extraction from the NIST'04 test dataset as suggested by the evaluation organizers.

A.5 Chinese-English BTEC'07 corpus.

The 2007 version of the BTEC corpus, consisting of short tourism-related sentences and which was used in reordering experiments, is presented in Table A.5.

Set	Language	Sentences	Words	Voc	ASL	References
Train	Chinese	44.9K	299.0K	11.4K	6.66	-
Train	English	44.9K	324.4K	9.0K	7.22	-
Dev	Chinese	489	5.2K	1.1K	10.66	7
Test	Chinese	500	5.5K	1.3K	11.10	7

Table A.5: *BTEC'07 ZhEn corpus. Basic statistics.*

A.6 Arabic-English NIST'08 corpus (extraction).

Table A.6 presents basic statistics of the 1M-line extraction from the corpus that was provided to the NIST 2008 evaluation campaign and belongs to the news domain. We did not used the complete training set, which includes more than 3M running sentences, because

[40] the National Institute of Standards and Technology
[41] http://www.ldc.upenn.edu/

we found this amount of data redundant for reordering experiments. Besides, experiments with such a huge corpus are extremely costly and time-consuming.

Set	Language	Sentences	Words	Voc	ASL	References
Train	Arabic	1.0M	31.62M	189.59K	31.62	-
Train	English	1.0M	31.77M	156.31K	31.77	-
Dev	Arabic	1,043	29.72K	5.9K	28.49	4
Test	Arabic	2,040	61.62K	9.9K	30.21	4

Table A.6: *NIST'08 ArEn corpus. Basic statistics.*

A.7 Arabic-English BTEC'08 corpus.

Basic statistics for the 2008 version of the Arabic-English BTEC corpus from the tourist domain can be found in Table A.7.

Set	Language	Sentences	Words	Voc	ASL	References
Train	Arabic	24.9K	225K	11.4K	9.05	-
Train	English	24.9K	210K	7.6K	8.46	-
Dev	Arabic	489	5.9K	1.2K	12.10	6
Test	Arabic	500	6.5K	1.4K	13.16	6

Table A.7: *BTEC'08 ArEn corpus. Basic statistics.*

Appendix B

Project framework

B.1 TC-STAR project

The TC-STAR (Technology and Corpora for Speech to Speech Translation) project[42], financed by European Commission within the Sixth Program[43], is envisaged as a long-term effort to advance research in all core technologies for speech-to-speech translation (SST), which is a combination of automatic speech recognition, spoken language translation, and speech synthesis.

The objective of the TC-STAR project is to significantly reduce the gap between human and machine performance for SST. The focus is on the development of new, possibly revolutionary, algorithms and methods, integrating the relevant human knowledge available at translation time into a data-driven framework. Examples of such new approaches are the integration of linguistic knowledge in the statistical approach of spoken-language translation, the statistical modeling of pronunciation of unconstrained conversational speech in automatic speech recognition, and new acoustic, and prosodic models for generating expressive speech in synthesis.

TC-STAR began in 2005 and was completed in 2007. During this time new approaches to SST were explored and evaluated. In addition, the infrastructure needed for accelerating the rate of progress in the field was created. To foster significant advances in all SST technologies, periodic competitive evaluations were conducted in 2005, 2006 (see §C.1.1), and 2007 (see §C.1.2). A measure of success of the project was the involvement of external

[42] http://www.tc-star.org/
[43] http://ec.europa.eu/research/fp6/index_en.cfm

participants in the evaluation campaigns. Results were presented and discussed in a series of TC-STAR evaluation workshops.

The project participants are listed below:

Istituto Trentino di Cultura IRC - IRST (Tech. coord.)	Italy
RWTH Aachen - ISL	Germany
CNRS - LIMSI	France
Universitat Politècnica de Catalunya	Spain
Universit at Karlsruhe (TH) - IPD	Germany
IBM	Germany
Nokia	Finland
Siemens	Germany
SRIT	France
Sony	Germany
ELDA	France
RU-SPEX	Netherlands

B.2 AVIVAVOZ project

AVIVAVOZ is a three-year project funded by the Spanish Government. It started in January 2007 and is devoted to advanced research in all key technologies related to speech-translation systems (e.g., speech recognition, machine translation, and speech synthesis).

The goal of the project is to achieve actual improvements in all speech translation system components in order to provide a speech mediating system for human communications among the official languages of the Spanish State (Spanish, Catalan, Basque, and Galician), and between Spanish and English.

The project considers both improving and integrating each of three involved technologies. In speech recognition, a robust system for a wide application domain (e.g., broadcast news and parliamentary sessions) and large vocabulary will be developed. In MT, improvements will be achieved for statistical translation techniques by including different sources of linguistic knowledge (event detection, syntactic and semantic analysis). In speech synthesis, new acoustic, and prosodic models for expressive speech generation will be developed. The final issue to be considered in this project is related to the interaction and integration of the three involved technologies.

Within the framework of the project, the Albayzín evaluation campaign was organized

in 2008. This event is described in §C.5.

The project participants are listed below:

Speech Processing Group (Universitat Politècnica de Catalunya)	Barcelona
Signal Processing Group (Universidade de Vigo)	Vigo
Aholab - Signal Processing Laboratory (Euskal Herriko Unibertsitatea)	Bilbao

Appendix C

International evaluation campaigns

The purpose of evaluation campaigns is to measure the quality of MT algorithms and systems to determine to what extent the system answers the goals of accurate automatic translation and meets the needs of the final users of the MT tools. Another goal of evaluation campaigns is to promote research on automatic translation and to encourage collaboration among research teams. Furthermore, translation systems can be compared on an equal and objective basis, and knowledge is shared among researchers from several sites.

International evaluation campaigns play an important role in promotion of the progress for MT technologies. These evaluations support MT research and help advance the state-of-the-art in MT technology. A precise set of evaluation criteria, which includes principally evaluation data and evaluation metrics, enables several teams to compare their solutions to a given NLP problem. Evaluation campaigns are typically organized by various institutions, consortiums, conferences, or workshops and serve as the perfect instrument to assess the translation improvements of SMT systems.

Evaluation conditions vary depending on the objectives of the particular campaign. Participants' submissions are usually evaluated automatically using various translation evaluation measures; however, human evaluation metrics are also sometimes used to compare systems[44]. Training and testing materials are taken from various sources and are explicitly specified. Precise datasets enable several research teams to compare their MT solutions depending on the goals of the particular MT test and can be oriented toward speech translation, news texts translation, etc. The participating groups cannot be allowed to use any

[44]Human evaluation strategy requires a certain degree of human intervention that is a quite high-cost process. Consequently, this evaluation is usually conducted only on a limited amount of submissions

other data for developing their systems (restricted track conditions) or be unlimited in terms of data for training (open track conditions).

During this Ph.D. research, the MT group at the UPC-TALP research center participated in **eleven** MT evaluations. In this Appendix we report evaluation results, specifications of the evaluation events, and a brief description of the SMT system(s) submitted by the UPC-TALP translation group.

C.1 TC-STAR evaluations

The European project TC-STAR[45] presented in §B.1 organized its first internal evaluation in 2005 (for members of the project, including TALP-UPC) and two open evaluations in 2006 and 2007. Three translation tasks were proposed to the participants: Chinese-to-English broadcast news translation, and Spanish-to-English and English-to-Spanish European Parliament plenary speech translations.

To study the effect of recognition errors and spontaneous speech phenomena, particularly for the task of European Parliament transcription (*EuroParl* or *EPPS*), three types of input to the translation system were studied and compared within the evaluations:

- **ASR**: the output of automatic speech recognizers, without using punctuation marks; the data are automatically segmented at syntactic or semantic breaks.
- **Verbatim**: the verbatim (i.e., correct) transcription of the spoken sentences including phenomena of spoken language such as false starts, ungrammatical sentences, etc.
- **FTE**: the so-called final text editions of official transcriptions of the European Parliament, which do not include the artefacts of spoken language. These text transcriptions differ slightly from the verbatim ones. Some sentences are rewritten.

Roughly speaking, parallel training data consisted of the European Parliament corpus [Koe05a]. In addition to the European Parliament translation tasks (*EPPS* track), a complementary Spanish-to-English task was included in this evaluation for portability assessment. This data consisted of transcriptions from Spanish Parliament, for which no parallel training was provided (*Cortes* track).

[45] http://www.tc-star.org/

C.1.1 TC-STAR 2006 evaluation

The second TC-STAR evaluation took place in February 2006; a detailed description of the shared task can be accessed through http://www.elda.org/en/proj/tcstar-wp4/tcs-run2.htm.

TALP-UPC participated in the Spanish-to-English and English-to-Spanish translation directions considering three different tasks: *EuroParl* English-to-Spanish, *EuroParl* Spanish-to-English, and *Cortes* Spanish-to-English. For each of these tasks, three different translation conditions were considered: *FTE*, *verbatim* transcriptions, and *ASR*.

We presented the original version of the N-gram-based SMT system [Mar06b] enhances with some novel feature functions and reordering strategies that consider POS information. A detailed description of the TALP-UPC N-gram-based system submitted to the evaluation can be found in the following publication:

J.B. Mariño, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, M.R. Costa-jussà and M. Khalilov, *UPC's Bilingual N-gram Translation System.* Proceedings of the TC-Star Speech to Speech Translation Workshop, pp. 43-48, Barcelona (Spain), June 2006.

Table C.1 shows official results of the second evaluation of the TC-STAR project ranked according to the case-sensitive BLEU scores. For brevity's sake (1), we provide results exclusively for the FTE condition; and (2) we report the best participants' submission results when multiple results were available. Hereafter, UPC scores are placed in table cells filled with grey.

Considering all translation tasks, the UPC system achieved very competitive results when compared to other participants.

C.1.2 TC-STAR 2007 evaluation

The third and the last TC-STAR evaluation was organized one year after the previous campaign, in February 2007. UPC-TALP participated in the same shared task of Spanish-to-English and English-to-Spanish translations and all the conditions proposed. A description of the run can be obtained from http://www.elda.org/en/proj/tcstar-wp4/tcs-run3.htm.

The N-gram-based SMT system presented in the evaluation was built from unfold translation units, and made use of POS-tag rules to account for reorderings. A set of six additional models was used: a target LM, a word bonus, a target tagged LM, a source (reordered)

Spanish-to-English				English-to-Spanish	
EuroParl		Cortes		EuroParl	
System	BLEU	System	BLEU	System	BLEU
IBM	54.06	IBM	42.08	IRST	49.81
RWTH	53.10	RWTH	40.92	UED	49.50
UW	53.80	UPC	40.37 (3)	RWTH	49.44
IRST	52.40	IRST	39.66	UPC-TALP	48.85 (4)
UPC-TALP	52.30 (5)	UED	39.04	UW	48.75
UED	51.87	UW	39.04	IBM	47.71
UKA	47.05	UKA	35.17	UKA	44.04
SYSTRAN	45.72	SYSTRAN	35.02	DFKI	36.32
DFKI	43.04	DFKI	31.10	SYSTRAN	36.29

Table C.1: *Case-sensitive BLEU scores for TC-STAR'06 evaluation (FTE condition).*

LM, and two lexicon models computed on the basis of word-to-word translation probabilities. At the preprocessing step, we employed the same Spanish morphology reduction strategy as shown in [Cj07b]. Unfortunately, detailed system descriptions are not publicly disseminated and are available exclusively to TC-STAR workshop participants.

Table C.2 shows the summary of BLEU scores for Spanish-to-English and English-to-Spanish translation directions under the FTE condition.

Spanish-to-English				English-to-Spanish	
EuroParl		Cortes		EuroParl	
System	BLEU	System	BLEU	System	BLEU
UKA	52.96	IBM	45.42	UKA	54.11
IBM	52.08	UKA	45.34	UPC-TALP	53.34 (2)
RWTH	51.20	RWTH	44.18	RWTH	52.49
IRST	51.08	UPC-TALP	43.79 (4)	IRST	52.26
UPC-TALP	50.88 (5)	IRST	43.39	IBM	50.91
NICT	49.79	JHU	39.98	-	-
JHU	47.88	TSL	33.89	-	-
TSL	43.67	UDS	30.88	-	-

Table C.2: *Case-sensitive BLEU scores for TC-STAR'07 evaluation (FTE condition).*

C.2 IWSLT evaluations

Since 2004, the C-STAR[46] consortium has organized the International Workshop on Spoken Language Translation (IWLST) on a yearly basis. This workshop includes an evaluation campaign oriented toward speech translation with the goal of providing a framework for the validation of existing evaluation methodologies when applied to the evaluation of spoken language translation technologies. Thus, the consortium hopes to open new avenues for improving existing methods.

Training material is a multi-lingual BTEC corpus [Tak02], traditionally limited to an extremely small amount of data - no more than $60K$ lines. BTEC corpus models a real situation when foreign tourists appear in an English-speaking country and need simple explanations and other practical information useful to travelers.

C.2.1 IWSLT'06 evaluation

The third IWSLT evaluation took place in fall 2006. A detailed description of the evaluation can be found in [Pau06]. The main goal of the IWSLT 2006 testing was to provide a framework to validate existing evaluation methodologies as applied to the evaluation of spoken language translation technologies, and thereby open new roads for improving current methods. That year the organisers proposed to concentrate efforts on the translation of data that were different in topic, style, and nature from that in the training material. Therefore, development and test datasets proposed contained out-of-domain data.

UPC-TALP participated in Chinese-to-English, Arabic-to-English, Italian-to-English, and Japanese-to- English tasks with a standard phrase-based system (**TALPphr**), enriched with the deterministic SMR technique [Cj09]. We also applied an additional distance-based reordering model (the so-called **MaxJumps**, as described in §4.1.1), for Chinese-to-English, Arabic-to-English, and Japanese-to-English translations.

In addition, we participated in the combination of **TALPphr** and the N-gram-based SMT system (**TALPtup**). The system combination (**TALPcomb**) used several n-gram LMs, a word bonus, and the IBM Model 1 for the whole sentence. The combination appeared to obtain clear improvements in BLEU score.

Out participation is detailed in the following publications:

[46] http://www.c-star.org/

- J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, M.R. Costa-jussà, J.B. Mariño, R. Banchs and J.A.R. Fonollosa *TALP Ngram-based SMT System for IWSLT 2006.* Proceedings of the 3rd International Workshop on Spoken Language Translation, pp. 116-122, Kyoto (Japan), November 2006.

- M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J.A.R. Fonollosa, J.B. Mariño and R. Banchs *TALP Phrase-based System and TALP System Combination for the IWSLT 2006* Proceedings of the 3rd International Workshop on Spoken Language Translation, pp. 116-122, Kyoto (Japan), November 2006.

Table C.3 shows details of the UPC-TALP participation in the IWSLT'06 evaluation, along with comparative results of other participants.

Zh2En		Ar2En		It2En		Ja2En	
System	BLEU	System	BLEU	System	BLEU	System	BLEU
RWTH	21.1	IBM	22.74	NICT	29.89	RWTH	21.41
JHU	18.63	TALPtup	21.36 (2)	TALPcom	28.37 (2)	NTT	19.84
MIT/AF	18.61	NICT	21.17	TALPtup	28.18 (3)	NICT	18.99
NTT	18.34	TALPcom	21.01 (4)	MIT/AF	27.98	MIT/AF	18.91
NICT	17.75	NTT	20.71	ITC	27.97	UKA/CMU	18.91
UKA/CMU	17.10	UKA/CMU	19.95	UW	27.87	ITC	16.04
TALPcom	16.50 (8)	TALPphr	19.08 (7)	NTT	27.69	SLE	15.99
TALPtup	16.24 (9)	ITC	17.23	TALPphr	26.84 (8)	HKUST	15.23
TALPphr	15.99 (10)	HKUST	14.77	DCU	25.98	KU	14.18
XMU	15.79	DCU	14.50	UKA/CMY	23.88	TALPcom	13.90 (10)
ITC	15.60	CLIPS	4.90	HKUST	23.74	TALPtup	13.70 (11)
HKUST	15.45	-	-	CLIPS	13.68	NAIST	13.11
ATT	12.26	-	-	-	-	TALPphr	12.80 (13)
NLPR	10.37	-	-	-	-	CLIPS	7.55

Table C.3: *Case-sensitive BLEU scores for IWSLT'06 evaluation.*

It is questionable to label our results satisfactory for the Japanese-to-English and Chinese-to-English tasks, mainly because we did not introduce the development set as training before doing the final translation. On the other hand, our results for the Italian-to-English and Arabic-to-English translation tasks were more competitive.

C.2.2 IWSLT'07 evaluation

The fourth IWSLT evaluation was organized in October 2007. Evaluation details are outlined in [For07]. Human evaluation for some of the submissions was included in the program. The input data consisted of the output of ASR systems for read speech and clean text. The exception was the Chinese English task, which used only clean text.

UPC-TALP participated in Chinese-to-English and Arabic-to-English translation tasks with an N-gram-based system using POS-tag reordering rules and implementing tuple extraction in accordance with an unfold algorithm. Apart from n-gram TM, the system introduced six additional feature functions: a target LM, a word bonus, two lexicon models, a target tagged LM, and a source tagged (reordered) LM.

We used language-specific preprocessing schemes for Arabic and Chinese using the method described in chapter 5. Our submission also included a continuous-space neural network LM provided by LIMSI-CNRS[47], which is believed to be particularly important for tasks with limited resources, as it is the case for IWSLT evaluations. Although all publicly available data were allowed, we used only the provided data to train the system.

The UPC-TALP translation system is described in the following publications:

P. Lambert, M.R. Costa-jussà, J.M. Crego, M. Khalilov, J.B. Mariño, R. Banchs, J.A.R. Fonollosa and H. Schwenk *The TALP Ngram-based SMT System for IWSLT 2007.* Proceedings of the 4th International Workshop on Spoken Language Translation, pp. 169-174, Trento (Italy), October 2007.

Table C.4 shows comparative results of IWSLT'07 evaluation ranked by the human evaluation score. The human evaluation (*%Best*) consisted of the average number of times that a system was judged to be better than any other system [CB07]. For each task, 300 sentences out of the 724 sentences in the evaluation set were randomly selected and presented to at least 3 evaluators. Since the ranking metric required that each submission be compared to the other system outputs, each sentence might be presented multiple times but in the company of different sets of systems.

Considering the Arabic-English pair, the UPC-TALP SMT system attained outstanding results, ranked in both cases (by human and automatic measures) as one of the best systems.

[47] http://www.limsi.fr/

System	Ar2En				Zh2En		
	Clean		ASR			Clean	
	%Better	BLEU	%Better	BLEU	System	%Better	BLEU
DCU	45.1	47.09	28.1	39.42	CASIA	37.6	36.48
UPC-TALP	42.9 (2)	48.04 (3)	31.8 (1)	44.45 (1)	I2R	37.0	40.77
UEKAE	36.4	49.23	19.8	36.79	ITC	34.8	37.50
UMD	36.0	48.58	25.0	39.08	RWTH	32.4	37.08
UW	35.4	41.61	26.9	40.92	FBK	30.6	34.72
MIT	35.1	45.53	31.4	44.29	CMU	30.6	34.44
CMU	33.9	44.63	25.5	37.56	UPC-TALP	28.3 (7)	29.91 (11)
LIG	33.9	41.35	24.2	38.04	XMU	28.1	28.88
NTT	25.3	34.03	25.5	36.26	HKUST	25.5	34.26
GREYC	21.7	32.90	-	-	MIT	25.0	36.31
HKUST	13.1	19.51	11.2	14.20	NTT	24.6	27.89
-	-	-	-	-	ATR	24.2	31.33
-	-	-	-	-	UMD	23.6	32.11
-	-	-	-	-	DCU	18.6	27.37
-	-	-	-	-	NUDT	16.1	19.34

Table C.4: *Case-sensitive BLEU scores and human evaluation results for IWSLT'07 evaluation.*

Especially relevant is the performance achieved in the ASR task, where state-of-the-art results were obtained. Notice that our system did not take multiple ASR output hypotheses into account except for the single-best one. This gave additional relevance to the results achieved in the ASR task when compared to other systems.

The UPC-TALP SMT system showed a reduction in performance when considering the Chinese-to-English task. One explanation for this situation is that our system might be less robust for noisy alignments (in particular, under scarce data availability) than for standard phrase-based systems. The important reordering needs, the complexity of the Chinese vocabulary, and the small amount of data available rendered the alignment process significantly more difficult in this translation task.

C.2.3 IWSLT'08 evaluation

IWSLT 2008 evaluation took place in Hawaii in October 2008. Details about the evaluation campaign can be found in [Pau08]. In addition to a traditional task of the translation of spontaneous speech recorded in a real situation, the feasibility of pivot-language-based translation approaches was studied in 2008.

Human (subjective) evaluation was carried out with respect to the *fluency* and *adequacy*

of the translation. Moreover, a paired-comparison evaluation based on the obtained ranking results was carried out to compare two MT systems directly. In other words, given two MT system translations of the evaluation data set, the first system was compared with the second system output on a sentence-by-sentence basis according to the *ranking* grades where both systems were ranked together.

This year we focused on the Arabic-to-English, Chinese-to-Spanish, and pivot Chinese-(English)-Spanish translation tasks and presented phrase-based SMT systems. The novelties that were introduced in the 2008 systems included SMR method, linear combination of translation and reordering models, and new technique dealing with insertion of punctuation marks for a phrase-based SMT system.

UPC-TALP participated in collaboration with the Institute for Infocomm Research[48] (I2R) in Singapore. Specifically, we collaborated with I2R in Chinese-(English)-Spanish translation. I2R provided us with a Chinese-to-English SMT system, and the UPC team was responsible for an English-to-Spanish translation.

Comparative results of the IWSLT 2008 evaluation can be found in Tables C.5 and C.6. We report BLEU scores and human ranking for all the participant systems. The results are ranked by BLEU score obtained on the *Clean* run.

System	Ar2En		
	Clean	ASR	
	BLEU	BLEU	Rank.
MITTL	34.25	30.50	44.15
RWTH	33.54	27.45	38.22
LIUM	31.81	25.62	37.41
UPC-TALP	31.31 (4)	25.63 (3)	39.01 (3)
DCU	29.23	24.03	36.34
LIG	28.38	25.45	37.56
TTK	27.78	25.19	35.74
PT	24.93	19.35	19.77
QMUL	19.04	16.13	22.89
GREYC	15.32	13.66	14.98

Table C.5: *Case-sensitive BLEU scores and human evaluation ranking for IWSLT'08 evaluation (Arabic-to-English results).*

We found our results satisfactory, especially for the pivot and Arabic-to-English translation tasks.

[48] http://www.i2r.a-star.edu.sg

System	Zh2Es			System	Zh2(En)2Es		
	Clean	ASR			Clean	ASR	
	BLEU	BLEU	Rank.		BLEU	BLEU	Rank.
TCH	34.57	30.52	47.73	TCH	40.42	35.43	49.32
FBK	29.60	24.24	33.42	FBK	39.41	32.51	39.90
DCU	27.10	23.89	28.99	UPC-TALP	38.09 (3)	32.51 (3-4)	39.01 (3)
TTK	26.62	24.40	28.99	NICT	37.11	32.81	30.88
NICT	26.41	23.31	29.79	DCU	32.42	28.47	31.72
PT	25.72	20.10	19.77	TTK	31.88	28.15	34.16
UPC-TALP	25.65 (7)	22.14 (6)	26.42 (6)	GREYC	15.80	15.05	15.46
GREYC	19.70	18.91	15.46	QMUL	2.87	11.59	17.72

Table C.6: *Case-sensitive BLEU scores and human evaluation ranking for IWSLT'08 evaluation (Chinese-to-Spanish and Chinese-(English)-Spanish results).*

C.3 WMT evaluations

Beginning in 2005, Workshops on Statistical Machine Translation (WMT), organized at the Annual Meetings of the Association for Computational Linguistics[49] have hosted a shared translation task. The main goals of the WMT evaluation campaigns are to promote MT performance for European languages on large-scale (about 30 million words in the training corpus) and over a range of relatively wide political domains. WMT shared tasks are traditionally characterized by a wide variety of MT systems, including statistical, rule-based, and hybrid systems.

In contrast to many other evaluations, translation tasks proposed in the WMT evaluation are normally mutual, including directions from and into English. Shared task organizers provide a parallel corpus as training data, a baseline system, and additional linguistic resources that entail a low barrier of entry to the evaluation campaign. Participants may augment the baseline system or use their own system. Training includes the *EuroParl* corpus; however, under unconstrained conditions participants may use any additional resources. Automatic evaluation is normally done on both case-sensitive and case-insensitive bases, case-insensitive considered as primary.

[49] http://www.aclweb.org

C.3.1 WMT'07 evaluation

In June 2007 WMT invited research groups and industrial institutions from all over the world to participate in the international MT evaluation campaign. Evaluation details are reported in [CB07].

UPC-TALP participated in three translation directions, namely, Spanish-to-English, French-to-English, and German-to-English, considering both direct and inverse translation tasks.

The shared task participants were provided with a common set of training and test data for all language pairs. The considered data were part of the European Parliament dataset, and included "News Commentary" data as well. In addition to the *EuroParl* test set, editorials from the Project Syndicate website[50] were collected and employed as an secondary test set (*news* domain). Human evaluation was used regarding *adequacy* and *fluency* of system output.

UPC-TALP participated with a standard N-gram-based SMT system enhanced with an augmented version of the SMR method following weighted non-deterministic approach to word reordering. In addition, the presented system introduced a target LM based on statistical classes, a feature for out-of-domain units and an improved optimization procedure [Lam07a].

UPC-TALP participation is outlined in the following publications:

M.R. Costa-jussà, J.M. Crego, P. Lambert, M. Khalilov, J.A.R. Fonollosa, J.B. Mariño and R. Banchs *Ngram-Based Statistical Machine Translation Enhanced with Multiple Weighted Reordering Hypotheses.* Proceedings of the Second Workshop of Statistical Machine Translation (WMT) ACL, pp. 167-170, Prague (Czech Republic), July 2007.

Table C.7 summarizes the results of the WMT'07 evaluation campaign for Spanish-to-English and vice versa translation tasks[51] ranked by the BLEU score.

Considering the Spanish-to-English results, the UPC SMT system obtained very competitive results, especially for the out-of-domain task, where the human and automatic measures rewarded the system with the best results.

[50] http://www.project-syndicate.com/

[51] For brevity's sake we do not provide the results for other translation pairs. Details can be found in [CB07].

Spanish-to-English				English-to-Spanish			
System	BLEU	Adequacy	Fluency	System	BLEU	Adequacy	Fluency
EuroParl							
UEDIN	32.4	0.593	0.610	UEDIN	31.6	0.586	0.638
CMU	32.3	0.552	0.568	UPC-TALP	31.2 (2)	0.584 (2)	0.578 (4)
UPC-TALP	32.2 (3)	0.587 (2)	0.604 (2)	CMU-UKA	31.1	0.563	0.581
CMU-UKA	32.0	0.557	0.564	UPV	30.4	0.573	0.587
UPV	31.5	0.562	0.573	NRC	29.9	0.546	0.548
NRC	31.3	0.477	0.489	SYSTRAN	21.2	0.495	0.482
SYSTRAN	29.0	0.525	0.566	-	-	-	-
SAAR	24.5	0.328	0.542	-	-	-	-
News							
UPC-TALP	34.6 (1)	0.566 (1)	0.543 (1)	UCB	33.1	0.449	0.414
UEDIN	32.7	0.546	0.534	UPC-TALP	32.8 (2)	0.510 (1-2)	0.488 (3)
UPV	28.3	0.435	0.459	CMU-UKA	32.7	0.510	0.492
CMU-UKA	29.9	0.522	0.495	UEDIN	32.2	0.429	0.419
NRC	29.9	0.479	0.464	NRC	31.1	0.408	0.392
SYSTRAN	25.9	0.525	0.503	UPV	28.5	0.405	0.418
SAAR	24.4	0.446	0.460	SYSTRAN	28.1	0.501	0.507

Table C.7: *Case-insensitive BLEU scores and human evaluation ranking for WMT'07 evaluation (Spanish-to-English and English-to-Spanish results).*

In the case of the English-to-Spanish results, although they were also highly competitive, the UPC system slightly lost performance in comparison with other systems. The preprocessing step reducing the Spanish vocabulary seemed to help more in the Spanish-to-English direction than in the English-to-Spanish direction.

C.3.2 WMT'08 evaluation

The shared translation task of the 2007 ACL Workshop on SMT took place in June 2008 [CB08]. One difference between the 2008 evaluation and those of the previous years' workshops was a refined manual evaluation strategy. Evaluators were asked to assess the systems' output in three ways: (1) ranking translated sentences relative to each other, (2) ranking the translations of syntactic constituents drawn from the source sentence, and (3) assigning absolute yes or no judgements to the translation of syntactic constituents.

UPC-TALP participated in the evaluation campaign with the N-gram-based SMT system differing from the 2007 version in the introduction of a target LM, based on linguistic

classes (POS), morphology reduction for an inflectional language (Spanish), and a new version of an extended monotone reordering model based on automatically learned reordering rules. We constructed systems for the Spanish-to-English and English-to-Spanish translations for both the traditional EuroParl and a challenging news stories tasks. In each case, we used only the supplied data for each language pair for models training and optimization.

The 2009 system is outlined in the following publication:

M. Khalilov, C.A. Henríquez Q., M. R. Costa-jussà, J.M. Crego, A. Hernández H., P. Lambert, J.A.R. Fonollosa, J.B. Mariño and R. Banchs, *The TALP-UPC Ngram-based statistical machine translation system for ACL-WMT 2008,* Proceedings of the Association for Computational Linguistics, Third Workshop on Statistical Machine Translation (ACL'08-SMT), pp. 127-130 , Columbus (USA), June 2008.

Evaluation results for Spanish-to-English and English-to-Spanish tasks can be found in Tables C.8 and C.9, respectively.

Spanish-to-English				
System	BLEU	SR	CR	YN
Europarl				
CMU	0.33	0.714	0.847	0.882
CUED	0.33	0.676	0.846	0.857
LIMSI	0.33	0.780	0.854	0.902
UEDIN	0.33	0.660	0.865	0.879
DCU	0.32	0.677	0.868	0.854
SAAR	0.32	0.671	0.893	0.869
UPC-TALP	0.32 (2-4)	0.687 (3)	0.870 (2)	0.857 (5-6)
UCL	0.25	0.425	0.646	0.730
RBMT	0.19	0.427	0.455	0.648
News				
CUED	0.21	0.674	0.818	0.638
LIMSI	0.20	0.583	0.739	0.675
RBMT	0.20	0.577	0.699	0.594
SAAR	0.19	0.669	0.760	0.697
UCB	0.19	0.543	0.706	0.635
UPC-TALP	0.19 (4-6)	0.602 (3)	0.763(2)	0.707 (1)
CMU	0.18	0.567	0.715	0.635
UEDIN	0.18	0.561	0.758	0.622

Table C.8: *Case-sensitive BLEU scores and human evaluation results for WMT'08 evaluation (Spanish-to-English results).*

Notice that only primary participants' submission results are presented. Hereafter, the submissions are ranked by BLEU score. Human evaluation metrics are: *SR* - sentence ranking judgements, *CR* - constituent ranking judgements, and *YN* - Yes/No judgements for constituent translations judgements. The numbers indicate the percent of time that each system was judged to be greater than or equal to any other system. Among multiple rule-based translation systems (RMBT) the system demonstrating the best BLEU scores is presented.

English-to-Spanish				
System	BLEU	SR	CR	YN
Europarl				
CMU	0.32	0.667	0.825	0.804
LIMSI	0.31	0.737	0.855	0.872
UW	0.32	0.735	0.790	0.785
SAAR	0.31	0.717	0.849	0.806
UEDIN	0.30	0.714	0.818	0.888
UPC-TALP	0.30 (5-6)	0.593 (6)	0.775 (6)	0.903 (1)
UCL	0.25	0.500	0.592	0.714
RBMT	0.21	0.554	0.561	0.582
News				
RMBT	0.21	0.724	0.570	0.599
SAAR	0.20	0.548	0.696	0.639
UCB	0.20	0.586	0.653	0.568
CMU	0.19	0.494	0.721	0.459
LIMSI	0.19	0.537	0.694	0.532
UEDIN	0.18	0.481	0.625	0.493
UPC-TALP	0.18 (6-7)	0.601 (2)	0.595 (6)	0.366(7)

Table C.9: *Case-insensitive BLEU scores and human evaluation results for WMT'08 evaluation (English-to-Spanish results).*

Our Spanish-to-English submissions are ranked better than English-to-Spanish ones. The UPC-TALP Spanish-to-English system was one of the best, according to the human evaluations.

C.3.3 WMT'09 evaluation

The WMT 2009 evaluation campaign took place in April 2009. The evaluation campaign is detailed in [CB09]. Unlike the previous year's event, only one test dataset was proposed to the participants, namely, a challenging news stories tasks (out-of-domain), while the traditional Europarl set was eliminated. Unconstrained submissions were admitted but

marked with a special label indicating that additional training and tuning data were used. Human evaluation included: (1) ranking of translated sentences relative to each other and (2) editing the output of systems without displaying the source or a reference translation, and then later judging whether edited translations were correct.

UPC-TALP was a permanent participant of the WMT shared translations tasks, traditionally concentrated on the Spanish-to-English and vice versa language pairs. Unlike the previous years, in 2009 we concentrated on the investigation of the translation model interpolation for a standard phrase-based translation system based on the Moses toolkit.

Description of our participation can be found in the following publication:

J.A.R. Fonollosa, M. Khalilov, M. R. Costa-jussà, J.B. Mariño, C.A. Henríquez Q., A. Hernández H. and R. Banchs, *The TALP-UPC phrase-based translation system for EACL-WMT 2009,* Proceedings of the 4th Workshop on Statistical Machine Translation (WMT'09) , pp. 85-89 , Athens (Greece), March 2009.

Comparative results for Spanish-to-English and English-to-Spanish tasks are provided in Table C.10. *Constr.?* indicates constrained condition.

Spanish-to-English				English-to-Spanish			
System	BLEU	Constr.?	Rank.	System	BLEU	Constr.?	Rank.
GOOGLE	0.29	NO	.70	GOOGLE	0.28	NO	.65
UEDIN	0.26	YES	.56	NUS	0.25	YES	.59
UPC-TALP	0.26 (2-3)	YES	.59 (2)	UEDIN	0.25	YES	.66
NICT	0.22	YES	.37	UPC-TALP	0.25 (2-4)	YES	.58 (5)
RBMT	0.20	NO	.55	RBMT	0.22	NO	.64
SAAR	0.20	NO	.51	RWTH	0.22	YES	.51
-	-	-	-	SAAR	0.20	NO	.48

Table C.10: *Case-insensitive BLEU scores and human evaluation results (ranking translations relative to each other) for WMT'09 evaluation (Spanish-to-English and English-to-Spanish results).*

Our submission was a best-performing system under constrained conditions for both tasks and was ranked fairly high, according to the human evaluation. During the post-evaluation period, we have performed additional word-reordering experiments, comparing the results obtained with a SMR and SBR algorithms. Furthermore, the outputs of the systems were combined selecting the translation with the MBR technique [Kum04] that allowed significant outperformance of the baseline configuration. Results obtained using

two reordering methods and a combination of monotone, SBR and SMR systems' output, are presented in Table C.11.

System	BLEU Es2En	BLEU En2Es
Primary	26.04	25.16
SMR	24.95	24.09
SBR	24.24	23.52
System combination	26.44	25.39

Table C.11: WMT'09 post-evaluation experiments.

Unfortunately, the promising reordering techniques and the combination of their outputs were not applied within the evaluation deadline, which could have improved our primary results by about 0.4 BLEU points for Spanish-to-English translation and about 0.2 BLEU points for the opposite direction.

C.4 NIST evaluations

With extensive experience in automatic speech recognition benchmark tests, the National Institute of Standards and Technology (NIST), an entity of the government of the United States, has organized yearly MT tests since the early 2000s. Focused on producing a breakthrough in translation quality, these tests are usually unlimited in terms of data for training. The target language is English, and sources include Arabic and Chinese. Further information can be accessed through http://www.nist.gov/speech/tests/mt/index.htm.

C.4.1 NIST'06 evaluation

The UPC-TALP SMT team participated in the NIST MT evaluation for the first time in 2006. The 2006 evaluation considered Arabic and Chinese the source languages under test, and English the target language. The text data consisted of newswire text documents, web-based newsgroup documents, human transcription of broadcast news, and human transcription of broadcast conversations. Performance was measured using the BLEU metric. Human assessments were also taken into account on the evaluation, but only for the six best-performing systems (in terms of BLEU).

The evaluation conditions were called "*Large Data Track*" (limited the training data to data in the LDC[52] public catalogue existing before February 1st, 2006) and "*Unlimited Data*

[52] http://www.ldc.upenn.edu/

Track" (extended the training data to any publicly available data existing before February 1, 2006).

UPC participated only on the larger condition of both tasks (Chinese-to-English and Arabic-to-English) with an original N-gram-based performing unfold units, using heuristic constraints to allow for reordering (**MaxJumps** model). Four additional models were employed: a target LM, a word bonus and two lexicon models (direct and inverse).

Evaluation results are summarized in Table C.12. The results are sorted by the BLEU scores and reported separately for the GALE subset and the NIST subset.

Arabic-to-English			Chinese-to-English		
System	NIST subset	GALE subset	System	NIST subset	GALE subset
GOOGLE	42.81	18.26	ISI	33.93	14.13
IBM	39.54	16.74	GOOGLE	33.16	14.70
ISI	39.08	17.14	LW	32.78	12.99
RWTH	39.06	16.39	RWTH	30.22	11.87
APPTEK	38.74	19.18	ICT	29.13	11.85
LW	37.41	15.94	UEDIN	28.30	11.99
BBN	36.90	14.61	BBN	27.81	11.65
NTT	36.80	15.33	NRC	27.62	11.94
ITCIRST	34.66	14.75	ITCIRST	27.49	11.94
UKA/CMU	33.69	13.92	UMD/JHU	27.04	11.40
UMD/JHU	33.33	13.70	NTT	25.95	11.16
UEDIN	33.03	13.05	NICT	24.49	11.06
SAKHR	32.96	16.48	CMU	23.48	11.35
NICT	29.30	11.92	MSR	23.14	9.72
QMUL	28.96	13.45	QMUL	22.76	9.43
LCC	27.78	11.29	HKUST	20.80	9.84
UPC-TALP	27.41 (17)	11.49 (16)	UPC-TALP	20.71 (17)	9.31 (17)
COL	24.65	9.60	UPENN	19.58	9.23
UCB	19.78	7.32	ISCAS	18.16	8.60
AUC	15.31	6.35	LCC	18.14	8.13
DCU	9.47	3.20	XMU	15.80	7.47
KCSL	5.22	1.76	LINGUA	13.41	6.63
-	-	-	KCSL	5.12	1.99
-	-	-	KSU	4.01	2.18

Table C.12: *Case-sensitive BLEU scores for NIST'06 evaluation.*

UPC-TALP results for both tasks were far from the best system's results, due in large part to a very poor level of data preprocessing of the huge number of corpora available.

C.4.2 NIST'08 evaluation

The 2008 NIST Open Machine Translation evaluation continued the ongoing series of evaluations of human language translation technology. Again, constrained and unconstrained

conditions were proposed to the participants.

The UPC-TALP SMT team participated in Arabic-to-English and Urdu-to-English translation tasks. Official evaluation results for the Arabic-to-English task are reported in Table C.13. UPC-TALP participated with an N-gram-based system enhanced with a minimum-translation-error discriminative alignment training [Lam07a]. Unconstrained results are provided for informative purposes.

System	BLEU-4	IBM BLEU
Constrained		
GOOGLE	45.57	45.26
IBM-UMD	45.25	43.00
IBM	45.07	42.76
BBN	43.40	42.90
LIUM	42.98	41.05
ISI-LW	42.48	42.27
CUED	42.38	40.18
SRI	42.49	40.31
UEDIN	40.28	38.33
UMD	39.06	37.84
UPC-TALP	37.43 (11)	35.76 (12)
COL	37.40	35.94
NTT	36.71	35.40
CMU	34.81	34.79
QMUL	33.08	31.81
SAKHR	31.33	31.33
UPC-LSI	30.32	28.76
BASISTECH	25.29	24.23
AUC	14.15	13.59
Unconstrained		
GOOGLE	47.72	47.39
IBM	47.17	45.27
APPTEK	44.83	44.74
CMU	43.13	41.14

Table C.13: *Case-sensitive BLEU-4 and IBM BLEU scores for NIST'08 evaluation (Arabic-to-English results).*

Our system was outperformed by the groups following orthodox phrase-based and hierarchical phrase-based translation, yet our results were state-of-the-art.

C.5 Albayzín evaluation

The purpose of the Albayzín evaluation campaign that took place in 2008 was to promote the research in MT between Spanish and Basque languages, and to foster collaboration among Spanish research groups, participating in the AVIVAVOZ project (see §B.2). The goal of this evaluation was to promote the exchange of ideas, to stimulate creativity, to favor collaboration among research teams that focus their research on speech technologies, and to compare different techniques employed. Apart from an automatic text translation task, the evaluation also included language recognition and speech synthesis tasks. Evaluation results are presented in [Sai08].

Few bilingual and monolingual resources exist related to the Basque language, so a translation task with Basque as a target language is very problematic. Two different conditions were proposed: one using limited resources and another with no resources limitation. The final test consisted of the translation of some texts from the same domain as the training material.

TALP-UPC presented a phrase-based system based on Moses, which, apart from a standard set of feature models, introduced two target LMs: one based on lemmas and the second based on linguistic classes (POS). The details can be found in the following publication:

C.A. Henríquez Q., M. Khalilov, J.B. Mariño and N. Ezeiza N., *The AVIVAVOZ phrase-based statistical machine translation system for ALBAYZÍN 2008,* Proceedings of of las V Jornadas en Tecnologi'a del Habla - the V Biennial Workshop on Speech Technology , pp. 123-125 , Bilbao (Spain), November 2008.

Results of the evaluation are presented in Table C.14.

System	BLEU	NIST
ALBAYZÍN (UPC-TALP)	8.12 (1)	3.90 (2)
IXA (EHU)	8.10	3.98
PRHLT (UPV)	7.11	3.65

Table C.14: *Case-sensitive BLEU and NIST scores for Albayzín evaluation (Spanish-to-Basque results).*

The UPC-TALP system was ranked the first among three participants by BLEU score

and second by NIST score. Results showed how different translation technologies handle a challenging translation into Basque under conditions of limited resources. As has been stated previously, UPC-TALP presented a phrase-based SMT system; the IXA-EHU system was a combination of two outputs (Matrex [Str06] and Seg [Agi06]); and the PRHLT-UPV system followed the approach based on stochastic inversion transduction grammar with five non-terminal symbols.

C.6 Acronyms

APPTEK	Applications Technology Inc.	USA
ATT	AT&T Inc.	USA
AUC	The American University in Cairo	Egypt
BASISTECH	Basis Technology	USA
BBN	BBN Technologies	USA
CLIPS	Institut d'informatique el Mathématiques Aplliqueés de Grenoble	France
CMU	Carnegie Mellon	USA
COL	Columbia University	USA
CUED	University of Cambridge	UK
DCU	Dublin City University	Ireland
DFKI	German Research Center for Artificial Intelligence	Germany
EHU	Universidad del País Vasco / Euskal Herriko Unibertsitatea	Spain
FBK	Fondazione Bruno Kesler	Italy
GOOGLE	Google	USA
GREYC	University of Caen	France
HKUST	Hong Kong University of Science and Technology	Hong Kong
I2R	Institute for Infocomm Research	Singapore
IBM	IBM	USA
ICT	Institute of Computing Technology Chinese Academy of Sciences	China
ISCAS	Institute of Software	China
ISI	USC-ISI	USA
ITC	ITC-irst	Italy
JHU	John Hopkins University	USA
KCSL	KCSL Inc.	Canada
KSU	Kansas State University	USA
KU	Kyoto University	Japan
LCC	Language Computer	USA

LIG	University J. Fourier	France
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Science de l'Ingenièieur	France
LINGUA	Lingua Technologies Inc.	Canda
LIUM	University du Maine (Le Mans)	France
LW	Language Weaver Inc.	USA
MIT/AF	Massachesetts Institute of Technology and Air Force	USA
MITTL	MIT Lincoln Laboratory	USA
MSR	Microsoft Research Asia	China
NAIST	Nara Institute of Science and Technology	Japan
NICT	National Institute of Information and Communications Technology	Japan
NLPR	National Laboratory of Pattern Recognition	China
NRC	National Research Council	Canada
NTT	NTT Communication Science Laboratories	Japan
NUDT	National University of Defense Technology	China
PT	Pohang University of Science and Technology	Corea
QMUL	Queen Mary University of London	UK
RWTH	Rheinish-Westphalian Technical University	Germany
SAAR	Saarland University	Geemany
SAKHR	Sakhr Software Co.	Egypt
SLE	SHARP Laboratories of Europe	UK
SYSTRAN	SYSTRAN Language Translation Software	France
TCH	Toshiba China R&D Center	China
TSL	Translendium SL	Spain
TTK	TÜBİTAK-UEKAE	Turkey
UCB	University of California Berkeley	USA
UEDIN	University of Edinburgh	Scotland
UEKAE	National Research Institute of Electronics and Cryptology	Turkey
UKA	Universitaet Karlsruhe	Germany
UMD	University of Maryland	USA
UPC-LSI	Universitat Politècnica de Catalunya, LSI	Spain
UPC-TALP	Universitat Politècnica de Catalunya, TALP	Spain
UPENN	University of Pennsylvania	USA
UPV	Universidad Politécnica de Valencia	Spain
UW	University of Washington	USA
XMU	Xiamen University	China

Appendix D

Publications by the author

The next is a list of major publications by the author:

1. M. Khalilov, J.A.R. Fonollosa and M. Dras **Coupling hierarchical word reordering and decoding in phrase-based statistical machine translation**. Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL-HLT'09, pages - to appear, Boulder, Colorado (USA), June 2009.
2. M. Khalilov, J.A.R. Fonollosa and M. Dras **A new subtree-transfer approach to syntax-based reordering for statistical machine translation**. Proceedings of EAMT'09, pp. 197-204, Barcelona (Spain), May 2009.
3. M. Khalilov and J.A.R. Fonollosa **N-gram-based Statistical Machine Translation versus Syntax Augmented Machine Translation: comparison and system combination**. Proceedings of EACL'09, pp. 424-432, Athens (Greece), April 2009.
4. M. Khalilov, J.A.R. Fonollosa, F. Zamora-Martínez, M.J. Castro-Bleda and S. España-Boquera **Arabic-English translation improvement by target-side neural network language modeling**. Proceedings of HLT&NLP within the Arabic World International Workshop at LREC'08, Marrakech (Morocco), May 2008.
5. M. Khalilov, J.A.R. Fonollosa, F. Zamora-Martínez, M.J. Castro-Bleda and S. España-Boquera **Neural Network Language Models for Translation with Limited Data**. Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence, pp. 445-451, Dayton, Ohio (USA), November 2008.
6. M. Khalilov, J.A.R. Fonollosa and M. Dras **Deriving benefit from a generalized syntax-based reordering**[53]. Proceedings of las V Jornadas en Tecnologi'a del Habla - the V Biennial Workshop on Speech Technology, pp. 269-272, Bilbao (Spain), November, 2008.
7. M. Khalilov and J.A.R. Fonollosa **Comparación y combinación de los sistemas de traducción**

[53] Best Student Paper Award.

automática basados en n-gramas y en sintaxis[54]. Proceedings of SEPLN'08, pp. 259-266, Madrid (Spain), September 2008.

8. M. Khalilov **Target language modeling improvement techniques for statistical machine translation**. Proceedings of the Doctoral Consortium at the 8th EUROLAN Summer School, pp. 39-45, Iasi (Romania), July-August 2007.
9. M. Khalilov and J.A.R. Fonollosa **Language modeling for verbatim translation task**. Proceedings. of the IV Jornadas en Tecnología del Habla - the IV Biennial Workshop on Speech Technology, pp. 83-87, Zaragoza (Spain), November, 2006.

Other publications:

1. M. Khalilov, A.H. Hernández, M.R. Costa-jussà, J.M. Crego, C.A. Henríquez, P. Lambert, J.A.R. Fonollosa, J.B. Mariño J.B. and R. Banchs **The TALP-UPC Ngram-based statistical machine translation system for ACL-WMT 2008**. Proceedings of the Third Workshop on Statistical Machine Translation at ACL'08, pp. 127-130, Columbus (USA), June 2008.
2. M. Khalilov, M.R. Costa-jussà, C.A. Henríquez, J.A.R. Fonollosa, A.H. Hernández, J.B. Mariño, R. Banchs, C. Boxing, M. Zhang, A. Aw and H. Li **The TALP&I2R SMT Systems for IWSLT 2008**. Proceedings of IWSLT'08, pp. 116-123, Hawaii (USA), October 2008.
3. M.R. Costa-jussà, P. Lambert, J.M. Crego, M. Khalilov, J.A.R. Fonollosa, J.B. Mariño and R. Banchs **Ngram-based system enhanced with multiple weighted reordering hypotheses**. Proceedings of ACL'07, Second Workshop on Statistical Machine Translation, pp. 167-170, Prague (Czech Republic), June 2007.
4. P. Lambert, M.R. Costa-jussà, J.M. Crego, M. Khalilov, J.B. Mariño, R. Banchs, J.A.R. Fonollosa and H. Schwenk **The TALP ngram-based SMT system for IWSLT 2007**. Proceedings of IWSLT'07, pp. 169-174, Trento (Italy), October 2007.
5. M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, R. Banchs, J.B. Mariño and J.A.R. Fonollosa **TALP Phrase-based statistical translation system for European language pairs**. Proceedings of the HLT-NAACL Workshop on Statistical Machine Translatiion, New York (USA), June 2006.
6. J.M. Crego, A. de Gispert, P. Lambert, M.R. Costa-jussà, M. Khalilov, R. Banchs, J.B. Mariño and J.A.R. Fonollosa **N-gram-based SMT System Enhanced with Reordering Patterns**. Proceedings of the HLT-NAACL Workshop on Statistical Machine Translatiion, New York (USA), June 2006.
7. J.B. Mariño, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, M.R. Costa-jussà and M. Khalilov, **UPC's Bilingual N-gram Translation System**. Proceedings of the TC-Star Speech to Speech Translation Workshop, Barcelona (Spain), June 2006.
8. M.R. Costa-jussà, J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J.A.R. Fonollosa, J.B. Mariño and R. Banchs **TALP Phrase-based System and TALP System Combination for the IWSLT 2006**. Proceedings of IWSLT'06, Kyoto (Japan), November 2006.

[54]Publication in Spanish.

9. J.M. Crego, A. de Gispert, P. Lambert, M. Khalilov, M.R. Costa-jussà, J.B. Mariño, R. Banchs and J.A.R. Fonollosa **TALP Ngram-based SMT System for IWSLT 2006**. Proceedings of IWSLT'06, Kyoto (Japan), November 2006.

Bibliography

[Agi06] E. Agirre, A.Díaz de Ilarraza, G. Labaka, and K. Sarasola. Uso de infromación morfológica en el alineamiento espanñol-euskara. In *Proceedings of the XXII COngreso de la SEPLN*, Zaragoza, Spain, 2006.

[Als00] H. Alshawi, S. Bangalore, and S. Douglas. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1):45–60, 2000.

[AO99] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D.Melamed, F. Och, D. Purdy, N.A. Smith, and D. Yarowsky. Statistical machine translation: Final report. Technical report, 1999. Johns Hopkins University Summer Workshop.

[AO06] Y. Al-Onaizan and K. Papineni. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 529–536, Sydney, NSW, Australia, July 2006.

[Aya06] N. F. Ayan and B. J. Dorr. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 9–16, Morristown, NJ, USA, 2006.

[Ban00] S. Bangalore and G. Riccardi. Stochastic finite-state models for spoken language machine translation. In *Proceedings of Workshop on Embedded Machine Translation Systems*, pages 52–59, April 2000.

[Ban05] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[Ben03] Y. Bengio, R. E. Ducharme, and P. Vincent. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[Ben04] O. Bender, R. Zens, E. Matusov, and H. Ney. Alignment templates: the RWTH SMT system. In *Proceedings of IWLST'04*, pages 79–84, Kyoto, Japan, 2004.

[Ber96a] A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 1(22):39–72, March 1996.

[Ber96b] A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillet, A. S. Kehler, and R. L. Mercer. Language translation apparatus and method of using context-based translation models. United States Patent 5,510,981, 1996.

[Ber05] N. Bertoldi and M. Federico. A new decoder for spoken language translation based on confusion networks. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workhsop (ASRU'05)*, December 2005.

[Ber06] N. Bertoldi. Minimum error training (updates). Technical report, 2006. Slides of the JHU Summer Workshop.

[Ber07] N. Bertoldi, R. Zens, and M. Federico. Speech translation by confusion network decoding. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*, April 2007.

[BH60] Y. Bar-Hillel. The present state of automatic translation of languages. *Advances in Computers*, 1:91–163, 1960.

[Bil03] J. Bilmes and K. Kirchhoff. Factored language models and generalized parallel backoff. In *proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL) 2003*, pages 4–6, Edmonton, Canada, May-June 2003.

[Bis95] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[Boj07] O. Bojar, S. Cinková, and J. Ptácek. Towards english-to-czech mt via tectogrammatical layer. In *Proceedings of the Sixth International Workshop on Treebnaks and Linguisti Theories*, Bergen, Norway, 2007.

[Bon94] B. Bonnie. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 4(20):597–663, 1994.

[Bra00] T. Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*, 2000.

[Bro90a] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[Bro90b] P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J.D. Lafferty, R. Mercer, and P.S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

[Bro93] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[Buc94] T. Buckwalter. Issues in arabic orthography and morphology analysis. In *Proceedings of COLING 2004*, pages 31–34, Geneva, Switzerland, August 1994.

[Car98] M. Carl. A constructivist approach to machine translation. In *Proceedings of NeMLaP3/CoNLL98*, pages 247–256, Sydney, NSW, Australia, 1998.

[Car04] X. Carreras, I. Chao, L. Padrò, and M. Padrò. Freeling: An open-source suite of language analyzers. In *LREC'04*, Lisbon, Portugal, May 2004.

[Cas02] F. Casacuberta, E. Vidal, and J. M. Vilar. Architectures for speech-to-speech translation using finite-state models. In *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, pages 39–44, 2002.

[Cas03] M. J. Castro and F. Prat. New directions in connectionist language modeling. *Computational Methods in Neural Modeling*, 2686, 2003.

[Cas04] F. Casacuberta and E. Vidal. Machine translation with inferred stochastic finite-state transducers. *Computational Linguistics*, 30(2):205–225, 2004.

[CB6a] C. Callison-Burch, Ph. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL 2006*, pages 17–24, June 2006a.

[CB04] C. Callison-Burch, D. Talbot, and M. Osborne. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 175–182, Barcelona, Spain, July 2004.

[CB07] C. Callison-Burch, C. Fordyce, Ph. Koehn, Ch. Monz, and J. Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, 2007.

[CB08] C. Callison-Burch, C. Fordyce, P. Koehn, Ch. Monz, and J. Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008.

[CB09] C. Callison-Burch, Ph. Koehn, C. Monz, and J. Schroeder. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece, april 2009.

[Cha00] E. Charniak. A maximum entropy-inspired parser. In *Proceedings of the North American Association for Computational Linguistics Conference (NAACL) 2000*, pages 132–139, 2000.

[Cha02] P. Charoenpornsawat, V. Sornlertlamvanich, and T. Charoenporn. Improving translation quality of rule-based machine translation. In *Proceedings of COLING-02 on Machine translation in Asia*, pages 1–6, Morristown, NJ, USA, 2002.

[Cha03] E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation. In *Proceedings of the MT Summit IX. Intl. Assoc. for Machine Translation*, 2003.

[Che98] C. Chelba and F. Jelinek. Exploiting syntactic structure for language modeling. In *Proceedings of the 36th annual meeting on Association for Computational Linguistics*, pages 225–231, Morristown, NJ, USA, 1998.

[Che99] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 4(13):359–394, 1999.

[Che05] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. The ITC-irst SMT system for IWSLT-2005. In *Proceedings of IWSLT 2005*, page 98–104, 2005.

[Che06] B. Chen, M. Cettolo, and M. Federico. Reordering rules for phrase-based statistical machine translation. In *Proceedings of IWSLT'06*, pages 105–112, 2006.

[Che08] Boxing Chen, Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. I^2R Multi-Pass Machine Translation System for IWSLT 2008. In *Proc. of the International Workshop on Spoken Language Translation*, pages 46–51, Hawaii, USA, 2008.

[Chi05] D. Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL) 2005*, pages 263–270, 2005.

[Chi07] D. Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 2(33):201–228, 2007.

[Civ06] J. Civera, A. L. Lagarda, E. Cubel, F. Casacuberta, E. Vidal, J. M. Vilar, and S. Barrachina. A computer-assisted translation tool based on finite-state technology. In *Proceedings of EAMT'06*, 2006.

[Cj06a] M. R. Costa-jussà, J. M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J.B. Mari no, R. Banchs, and J.A.R. Fonollosa. TALP phrase-based statistical translation system for european language pairs. In *Human Language Technology Conference (HLT-NAACL)'06: Proceedings of the Workshop on Statistical Machine Translation*, pages 142–145, New York, USA, June 2006.

[Cj06b] M. R. Costa-jussà and J. A. R. Fonollosa. Statistical machine reordering. In *Proceedings of the HLT/EMNLP 2006)*, 2006.

[Cj06c] M. R. Costa-jussà, J. M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J. A. Fonollosa, J. B. Mariño, and R. E. Banchs. TALP phrase-based system and TALP system combination for IWSLT 2006. In *Proceedings of the IWSLT 2006*, pages 123–129, 2006.

[Cj07a] M. R. Costa-jussà and J. A. R. Fonollosa. Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT))*, pages 171–176, Prague, Czech Republic, July 2007.

[Cj07b] M. R. Costa-jussà, P. Lambert, J. M. Crego, M. Khalilov, J. A. R. Fonollosa, J.B. Mari no, and R. Banchs. Ngram-based system enhanced with multiple weighted reordering hypotheses. In *Proceedings of the Association for Computational Linguistics, Second Workshop on Statistical Machine Translation*, pages 167–170, Prague,Czech Republic, June 2007.

[Cj08a] M. R. Costa-jussà and J. A. R. Fonollosa. Computing multiple weighted reordering hypotheses for a statistical machine translation phrase-based system. In *Proceedings of AMTA'08*, Honolulu, USA, October 2008.

[Cj08b] M. R. Costa-jussà, J. A. R. Fonollosa, and E. Monte. Using reordering in statistical machine translation based on alignment block classification. In *Proceedings of the LREC'08 Conference*, Marrakech, Morocco, May 2008.

[Cj09] M. R. Costa-jussà. *New reordering and modeling approaches for statistical machine translation*. PhD thesis, Universitat Politècnica de Catalunya, July 2009.

[Col99] M. Collins. *Head-driven syntactical models for natural language processing*. PhD thesis, University of Pennsylvania, 1999.

[Col05] M. Collins, P. Koehn, and I. Kučerová. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on ACL 2005*, pages 531–540, 2005.

[Cow08] A. Cowan. *A Tree-to-Tree Model for Statistical Machine Translation*. PhD thesis, Massachusetts Institute of Technology, June 2008.

[Cre05a] J. M. Crego, J. B. Mariño, and A. de Gispert. An ngram-based statistical machine translation decoder. In *Proceedings of INTERSPEECH05*, 2005.

[Cre05b] J. M. Crego, J. B. Mariño, and A. de Gispert. An ngram-based statistical machine translation decoder. In *Proceedings of INTERSPEECH05*, 2005.

[Cre05c] J. M. Crego, J. B. Mariño, and A. de Gispert. Reordered search and tuple unfolding for ngram-based smt. In *In Proc. of MT Summit X*, pages 283–289, September 2005.

[Cre05d] J. M. Crego, J. B. Mariño, and A. de Gispert. Reordered search and tuple unfolding for ngram-based smt. In *In Proc. of MT Summit X*, pages 283–289, September 2005.

[Cre06a] J. Crego, A. de Gispert, P. Lambert, M. Khalilov, M. Costa-jussà, J. Marino, R. Banchs, and J.A.R. Fonollosa. The TALP Ngram-based SMT System for IWSLT 2006. In *Proceedings of IWSLT 2006*, pages 116–122, 2006.

[Cre06b] J. M. Crego and J. B. Mariño. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215, 2006.

[Cre07a] J. M. Crego and J. B. Mariño. Improving statistical MT by coupling reordering and decoding. *Machine Translation*, 20(3):199–215, 2007.

[Cre07b] J.M. Crego and Mari no J.B. Syntax-enhanced n-gram-based sm. In *Proceedings of the 11th Machine Translation Summit (MT summit XI)*, pages 111–118, Copenhagen, Denmark, September 2007.

[Cre08a] J. M. Crego. *Architecture and modeling for N-gram-based Statistical Machine Translation*. PhD thesis, Universitat Politècnica de Catalunya, February 2008.

[Cre08b] J. M. Crego and N. Habash. Using shallow syntax information to improve word alignment and reordering for smt. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT 2008)*, Columbus, Ohio, USA, 2008.

[DÓ8] D. Déchelotte, G. Adda, A. Allauzen, H. Bonneau-Maynard, O. Galibert, J. Gauvain, Ph. Langlais, and F. Yvon. Limsi's statistical translation systems for WMT'08. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 107–110, Columbus, Ohio, June 2008.

[Dag94] I. Dagan and K. Church. *Termight*: Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP'94)*, pages 34–40, October 1994.

[Dav01] Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya. Interlingua-based english–hindi machine translation and language divergence. *Machine Translation*, 16(4):251–304, 2001.

[Dem77] A. E. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B(39):1–38, 1977.

[Den05] E. Denoual. The influence of data homogeneity on nlp system performance. In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP'05)*, pages 228–233, 2005.

[DeN07] S. DeNeefe, K. Knight, W. Wang, and D. Marcu. What can syntax-based MT learn from phrase-based MT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 755–763, 2007.

[dG02] A. de Gispert and J. B. Mariño. Using X-grams for speech-to-speech translation. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'02)*, September 2002.

[dG06] A. de Gispert. *Introducing Linguistic Knowledge into Statistical Machine Translation*. PhD thesis, Universitat Politècnica de Catalunya, October 2006.

[Dod02] G. Doddington. Automatic evaluation of machine translation quality using n-grams co-occurrence statistics. In *In HLT 2002 (Second Conference on Human Language Technology)*, pages 128–132, 2002.

[Dye08] C. Dyer, S. Muresan, and P. Resnik. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, USA, June 2008.

[Eck05] M. Eck and C. Hori. Overview of the IWSLT 2006 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT'05*, 2005.

[Eis03] J. Eisner. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 2 (companion volume), pages 205–208, Sapporo, Japan, 2003.

[Elm08] J. Elming. *Syntactic reordering in Statistical Machine Translation*. PhD thesis, Copenhagen Business School, 2008.

[EZ81] I. Even-Zohar. Translation theory today: A call for transfer theory. *Poetics Today*, 2(4):1–7, 1981.

[For07] C. Fordyce. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the 4th International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy, October 2007.

[Fos07] G. Foster and R. Kuhn. Mixture-model adaptation for SMT. In *In Annual Meeting of the Association for Computational Linguistics: Proc. of the Second Workshop on Statistical Machine Translation (WMT)*, pages 128–135, Prague, Czech Republic, June 2007.

[Fra06] A. Fraser and D. Marcu. Measuring word alignment quality for statistical machine translation. Technical report, ISI/University of Southern California, California, USA, 2006.

[Gal04] M. Galley, M. Hopkins, K. Knight, and D. Marcu. What's in a translation rule? In *Proceedings of the HLT/NAACL-04*, 2004.

[Gal06] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thaye. Scalable inference and training of context-rich syntactic translation models. In *Proc. of the 21st International Conference on COmputational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 961–968, July 2006.

[Gao02] J. Gao and M. Zhang. Improving language model size reduction using better pruning criteria. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 176–182, Philadelphia, PE, USA, July 2002.

[Ger02] Daniel Gervais. The full-text multilingual corpus: Breaking the translation memory bottleneck, April 2002.

[Gim06] J. Giménez and E. Amigò. Iqmt: A framework for automatic machine translation evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, 2006.

[Goo00a] J. Goodman and J. Gao. Language model size reduction by pruning and clustering. In *In proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 110–113, Beijing, China, October 2000.

[Goo00b] J. T. Goodman. Language model size reduction by pruning and clustering. In *In ICSLP'00*, pages 110–113, 2000.

[Gra04] J. Graehl and K. Knight. Training tree transducers. In *Proceedings of NAACL-HLT'04*, pages 105–112, 2004.

[Hab05] N. Habash and O. Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 573–580, Ann Arbor, Michigan, USA, June 2005.

[Hab06] N. Habash and F. Sadat. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 49–52, 2006.

[Has07] H. Hassan, K. Sima'an, and A. Way. Supertagged phrase-based statistical machine translation. In *Proceedings 45th Annual Meeting of the Assoc. for Comp. Linguistics*, pages 288–295, Prague, Czech Republic, June 2007.

[Has08] H. Hassan, K. Sima'an, and A. Way. Syntactically lexicalized phrase-based statistical translation. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1260–1273, September 2008.

[Hil05] A. S. Hildebrand, E. Matthias, S. Vogel, and A. Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval. In *In Proceedings of EAMT 2005*, pages 133–142, 2005.

[Hil08] A. S. Hildebrand, K. Rottmann, M. Noamany, Q. Gao, S. Hewavitharana, N. Bach, and S. Vogel. Recent improvements in the CMU large scale chinese-english SMT system. In *Proceedings of ACL-08: HLT (Companion Volume)*, pages 77–80, 2008.

[Hir93] U. Hiroshi and Z. Meiying. Interlingua for multilingual machine translation. In *Proceedings of MT Summit IV*, pages 157–169, July 1993.

[Hua06] L. Huang, K. Knight, and A. Joshi. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-06)*, pages 223–226, New York, USA, 2006.

[Hut86] W.J. Hutchins. *Machine translation: past, present, future*. Chichester: Ellis Horwood, 1986.

[Igl09] G. Iglesias, A. de Gispert, E.R. Banga, and W. Byrne. Hierarchical phrase-based translation with weighted finite state transducers. In *Proceedings of NAACL-HLT 2009*, page to appear, 2009.

[Ima05] K. Imamura, H. Okuma, and E. Sumita. Practical approach to syntax-based statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 267–274, 2005.

[Jam00] F. James. Modified kneser-ney smoothing of n-gram models. Technical report, Research Institute for Advanced Computer Science (RIACS), October 2000.

[J.E03] J.Eisner. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the Associaction for COmputational Linguistics (ACL) 2003 (companion volume)*, pages 205–208, 2003.

[Jel97] F. Jelinek. *Statistical Methods for Speech Recognition*. Language, Speech, and Communication. The MIT Press, 1997.

[Jos97] A. Joshi and Y. Schabes. Tree-adjoining grammars. *G. Rozenberg and A. Salomaa, editors, Handbook of formal languages*, 3, 1997.

[Jur00] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, 2000.

[Kan05] S. Kanthak, D. Vilar, E. Matusov, R. Zens, and H. Ney. Novel reordering approaches in phrase-based statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 167–174, 2005.

[Kay92] M. Kay, J. Gawron, and P Norvig. *Verbmobil: a translation system for face-to-face dialog*. CSLI, 1992.

[Kha07] M. Khalilov. Target language modeling improvement techniques for statistical machine translation. In *Proceedings of the Doctoral Consortium at the 8th EUROLAN Summer School*, pages 39–45, Iasi, Romania, 2007.

[Kha08] M. Khalilov, A. H. Hernández, M. R. Costa-jussà, J. M. Crego, C. A. Henríquez, P. Lambert, J. A. R. Fonollosa, J. B. Mariño, and R. Banchs. The talp-upc ngram-based statistical machine translation system for acl-wmt 2008. In *Proceedings of the ACL 2008 Third Workshop on Statistical Machine Translation (WMT'08)*, pages 127–131, 2008.

[Kha09] M. Khalilov, J.A.R. Fonollosa, and M. Dras. A new subtree-transfer approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT)*, page to appear, Barcelona, Spain, May 2009.

[Kir05] K. Kirchjoff and M. Yand. Improved language modeling for statistical machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 125–128, Ann Arbor, MI, USA, June 2005.

[Kle03] D. Klein and C. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the ACL 2003*, pages 423–430, 2003.

[Kni98] K. Knight and Y. Al-Onaizan. Translation with finite-state devices. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation (AMTA) and the Information Soup*, pages 421–437, London, UK, October 1998.

[Kni99] K. Knight. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4), December 1999.

[Koe03] Ph. Koehn, F. Och, and D. Marcu. Statistical phrase-based machine translation. In *Proceedings of the HLT-NAACL 2003*, pages 48–54, 2003.

[Koe04] Ph. Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2004*, pages 388–395, 2004.

[Koe05a] Ph. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*, pages 79–86, September 2005.

[Koe05b] Ph. Koehn, A. Amittai, A. Birch, C. Callison-Burch, M. Osborne, D. Talbot, and M. White. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. of the International Workshop on Spoken Languages Translation*, Pittsburgh, October 2005.

[Koe07a] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open-source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL) 2007*, pages 177–180, 2007.

[Koe07b] Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open-source toolkit for statistical machine translation. In *Proceedings of the Association for Computational Linguistics (ACL) 2007*, pages 177–180, 2007.

[Koe07c] Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Open-source toolkit for statistical machine translation: factored translation models and confusion network decoding. Technical report, Johns Hopkins University, 2007.

[Kum04] S. Kumar and W. Byrne. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, 2004.

[Lam06] P. Lambert and R. E. Banchs. Tuning machine translation parameters with SPSA. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 190–196, Kyoto, Japan, 2006.

[Lam07a] P. Lambert, R. E. Banches, and J.M. Crego. Discriminative alignment training without annotated data for machine translation. In *Proceedings of the Human Language Technology Conference (HLT-NAACL'07)*, pages 199–215, Rochester, USA, 2007.

[Lam07b] P. Lambert, M.R. Costa-jussà, J. M. Crego, M. Khalilov, J. Mariño, R.E. Banchs, J.A.R. Fonollosa, and H. Shwenk. The talp ngram-based smt system for iwslt 2007. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT07)*, pages 169–174, 2007.

[Lam08] P. Lambert. *Exploiting Lexical Information and Discriminative Alignment Training in Statistical Machine Translation*. PhD thesis, Universitat Politècnica de Catalunya, July 2008.

[Lan05] P. Langlais, S. Gandrabur, T. Leplus, and G. Lapalme. The long-term forecast for weather bulletin translation. *Machine Translation*, 19(1):83–112, 2005.

[Lar06] H. Larochelle and Y. Bengio. Distributed representation prediction for generalization to new words. Technical Report 1284, Université de Montréal, 2006. Département d'Informatique et Recherche Opérationnelle.

[Lau93] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models using maximum likelihood estimation of exponential distributions. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1993*, Minneapolis, MN, USA, April 1993.

[Lav07] A. Lavie and A. Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second ACL Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007.

[Lav08] A. Lavie and A. Agarwal. METEOR, M-BLEU and M-TER: automatic metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation at the 46th Meeting of the Association for Computational Linguistics (ACL-2008)*, pages 115–118, Columbus, OH, June 2008.

[Li07] C. Li, L. Minghui, D. Zhang, M. Li, M. Zhou, and Y. Guan. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45rd Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 720–727, Prague, Czech Republic, 2007.

[Lin04a] C. Y. Lin and F. Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of Coling 2004*, pages 501–507, Geneva, Switzerland, 2004.

[Lin04b] D. Lin. A path-based transfer model for machine translation. In *Proceedings of the COLING'04*, 2004.

[Mar99] S. Martin, C. Hamacher, J. Liermann, F.Wessel, and H. Ney. Assessment of smoothing methods and complex stochastic language modeling. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 5, pages 1939–1942, Budapest, Hungary, September 1999.

[Mar02] D. Marcu and W. Wong. A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of EMNLP02*, pages 133–139, 2002.

[Mar06a] D. Marcu, W. Wong, A. Echihabi, and K. Knight. SPMT: statistical machine translation with syntactified target language phrases. In *EMNLP'06*, pages 44–52, Sydney, NSW, Australia, July 2006.

[Mar06b] J. B. Mariño, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. R. Fonollosa, and M. R. Costa-jussà. N-gram based machine translation. *Computational Linguistics*, 32(4):527–549, December 2006.

[Mat06] E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popovic', S. Hasan, and H. Ney. The RWTH machine translation system. In *Proceedings of TC-STAR Workshop on Speech-to-Speech Translation*, pages 31–36, Barcelona, Spain, June 2006.

[McC04] I. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, and H. Bourlard. On the use of information retrieval measures for speech recognition evaluation. IDIAP-RR 73, IDIAP, Martigny, Switzerland, 2004.

[Mel04] I.D. Melamed. Statistical machine translation by parsing. In *Proceedings of the Association for Computational Linguistics (ACL) 2004*, pages 111–114, 2004.

[Men05] A. Menezes and C. Quirk. Microsoft research treelet translation system: IWSLT evaluation. In *Proceedings of IWSLT 2006*, pages 105–108, 2005.

[Mil91] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. Five papers on wordnet. *Special Issue of International Journal of Lexicography*, 3(4):235–312, 1991.

[Mn06] J.B. Mariño, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, Costa-jussà, M.R., and M. Khalilov. Upc's bilingual n-gram translation system. In *Proceedings of the TC-Star Speech to Speech Translation Workshop*, Barcelona, Spain, June 2006.

[Moo06] R. C. Moore, W. Yih, and A. Bode. Improved discriminative bilingual word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 513–520, Sydney, NSW, Australia, 2006.

[Nag06] M. Nagata, K. Saito, K. Yamamoto, and K. Ohashi. A clustered global phrase reordering model for statistical machine translaton. In *Proceedings of ACL'06*, pages 713–720, 2006.

[Nel65] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer organization*, 7:308–313, 1965.

[Nes06] R. Nesson, S. Shieber, and A. Rush. Induction of probabilistic synchronous tree-insertion grammars for machine translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 8–12, Boston, Massachusetts, USA, August 2006.

[Nie04] S. Nießen and H. Ney. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204, 2004.

[Nie09] J. Niehues and M. Kolss. A pos-based model for long-range reorderings in smt. In *Proceedings of the Fourth Workshop on Statistical Machine Translation at EACL 2009*, pages 206–214, March 2009.

[O.07] Bojar O. and Cmejrek. Mathematical model of tree transformations, 2007. Project Euromatrix - Deliverable 3.2.

[Och99] F. Och. An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76, June 1999.

[Och00a] F. Och and H. Ney. A comparison of alignment models for statistical machine translation. In *Proc. of the 18th conference on Computational Linguistics*, pages 1086–1090, Morristown, NJ, USA, 2000.

[Och00b] F. Och and H. Ney. Improved statistical alignment models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, October 2000.

[Och02a] F. Och. *Statistical Machine Translation: From Single Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, 2002.

[Och02b] F. Och and H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)02*, pages 295–302, 2002.

[Och03a] F. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, July 2003.

[Och03b] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[Och03c] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 1(29):19–51, March 2003.

[Och03d] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[Och04] F. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational Linguistics*, 2004.

[Pap98] K. Papineni, S. Roukos, and R. Ward. Maximum likelihood and discriminative training of direct translation models. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 189–192, May 1998.

[Pap02] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) 2002*, pages 311–318, 2002.

[Par99] Richard Parkinson. *Cracking Codes: The Rosetta Stone and Decipherment*. Berkeley & Los Angeles: University of California Press, 1999.

[Pau06] M. Paul. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of IWSLT06*, pages 1–15, 2006.

[Pau08] M. Paul. Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of IWSLT08*, pages 1–17, Hawaii, USA, 2008.

[Pie66] J. R. Pierce and J. B. Carroll. Language and machines: Computers in translation and linguistics. Technical report, National Academy of Sciences/National Research Council, Washington, DC, USA, 1966.

[Pla05] E. Planas. Similis: Second-generation translation memory software. In *Proceedings of the International Conference Translating and the Computer 27*, London, November 2005.

[Pop06a] M. Popovic and H. Ney. Pos-based word reorderings for statistical machine translation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 1278–1283, May 2006.

[Pop06b] M. Popovic' and H. Ney. Statistical machine translation with a small amount of bilingual training data. In *Proceedings of the 5th LREC'06 SALTMIL Workshop on Minority Languages*, pages 25–29, May 2006.

[Qui05] C. Quirk, A. Menezes, and C. Cherry. Dependency treelet translation: syntactically informed phrasal smt. In *Proceedings of EACL 2005*, 2005.

[Rap02] William J. Rapaport. Word sense disambiguation using target language corpus in a machine translation system. *Minds and Machines*, 12/1:3–59, 2002.

[Roa07] B. Roark, M. Saraclar, and M. Collins. Discriminative n-gram language modeling. *Comput. Speech Lang.*, 21(2):373–392, 2007.

[Ros94] R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. PhD thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, April 1994.

[Rot07] K. Rottmann and S. Vogel. Word reordering in statistical machine translation with a pos-based distortion model. In *Procceding of TMI 2007: 11th International Conference on Theoretical and Methodological Issues in MT*, Skvde, Sweden, 2007.

[Sai08] I. Sainz. Análysis de los resultados de la evaluación ALBAYZIN-TTS 2008. In *Proceedings of las V Jornadas en Tecnologi'a del Habla - the V Biennial Workshop on Speech Technology*, Bilbao, Spain, November 2008.

[Sar07] R. Sarikaya and Y. Deng. Joint morphological-lexical language modeling for machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)'07*, pages 145–148, Rochester, USA, April 2007.

[Sch06] H. Schwenk, M. R. Costa-jussà, and J. A. R. Fonollosa. Continuous space language models for the IWSLT 2006 task. In *Proceedings of IWSLT 2006*, pages 166–173, 2006.

[Sch07a] H. Schwenk. Continuous space language models. *Computer Speech and Language*, 21(3):492–518, 2007.

[Sch07b] H. Schwenk, M. R. Costa-jussà, and J. A. R. Fonollosa. Smooth bilingual translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 430–438, Prague, Czech Republic, 2007.

[Sch08] H. Schwenk and Y. Estève. Data selection and smoothing in an open-source system for the 2008 nist machine translation evaluation. In *Proceedings of the Interspeech'08*, pages 2727–2730, Brisbane, Australia, September 2008.

[Sha48] C. E. Shannon. *A Mathematical Theory of Communication*. CSLI Publications, 1948.

[Sha51] C. E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30:50–64, January 1951.

[Shi90] S. Shieber and Y. Schabes. Synchronous tree-adjoining grammars. In *Proceedings of COLING'90*, pages 253–258, 1990.

[Shi07] S. Shieber. Probabilistic synchronous tree-adjoining grammars for machine translation: The argument from bilingual dictionaries. In *Proceedings of SSST Workshop on Syntax and Structure in Statistical Translation, NAACL-HLT*, pages 88–95, Rochester, New York, USA, April 2007.

[Sii07] V. Siivola, T. Hirsimäki, and S. Virpioja. On growing and pruning kneser-ney smoothed n-gram models. *IEEE Transactions on Speech, Audio and Language Processing*, 15(5):1617–1624, 2007.

[SM07] N. Singh-Miller and M. Collins. Trigger-based language modeling using a loss-sensitive perceptron algorithm. In *Proceedings of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2007*, pages IV–25 – IV–28, Honolulu, Hawaii, USA, April 2007.

[Sno05] M. Snover, B. Dorr, R. Schwartz, J. Makhoul, L. Micciula, and R. Weischedel. A study of translation error rate with targeted human annotation. Technical report, University of Maryland, July 2005. LAMPTR-126, CS-TR-4755, UMIACS-TR-2005-58.

[Son99] F. Song and W. B. Croft. A general language model for information retrieval (poster abstract). In *Proceedings of the eighth international conference on Information and knowledge management*, pages 279–280, 1999.

[Ste00] M. Steedman. *The Syntactic Process*. The MIT press, 2000. Cambridge, Massachusetts.

[Sto02] A. Stolcke. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904, 2002.

[Str98] O. Streiter. A semantic description language for multilingual nlp. In *In paper presented at the Tuscan Word Centre - Institut für Deutsche Sprache Workshop on Multilingual Lexical Semantics*, june 1998.

[Str06] N. Stroppa and A. Way. MaTrEx: DCU machine translation system for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36, Kyoto, Japan, 2006.

[Tak02] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of LREC 2002*, pages 147–152, 2002.

[Til97] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based search for statistical translation. In *Proceedings of European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece, September 1997.

[Til03] C. Tillmann and H. Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 1(29):93–133, 2003.

[Til04] C. Tillman. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*, 2004.

[Til05] C. Tillmann and T. Zhang. A localized prediction model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on ACL 2005*, pages 557–564, 2005.

[Tom70] P. Toma, I.A. Kozlik, and D.G. Perwin. Systran machine translation system. final technical report. Technical report, Griffiss AFB: RADC, September 1970.

[Vau68] B. Vauquois. A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *Proceedings of International Federation for Information Processing (IFIP) Congress*, pages 254–260, Edinburg, August 1968.

[Ven06] A. Venugopal and A. Zollmann. Syntax augmented machine translation via chart parsing with integrated language modeling. Technical report, 2006.

[Ven09] A. Venugopal and A. Zollmann. Grammar based statistical mt on hadoop: An end-to-end toolkit for large scale pscfg based mt. *The Prague Bulletin of Mathematical Linguistics*, (91):67–78, 2009.

[Vid97a] E. Vidal. Finite-state speech-to-speech translation. In *Proceedings of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 111–114, Munich, April 1997.

[Vid97b] E. Vidal. Finite-state speech-to-speech translation. In *Proceedings of 1997 IEEE Int. Conf. on Acoustic, Speech and Signal Processing*, pages 111–114, 1997.

[Vil09] D. Vilar and H. Ney. On lm heuristics for the cube growing algorithm. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 242–249, Barcelona, Spain, May 2009.

[Vog96] S. Vogel, H. Ney, and C. Tillmann. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics*, pages 836–841, August 1996.

[Wan07] C. Wang, M. Collins, , and P. Koehn. Chinese syntactic reordering for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 737–745, Prague, Czech Republic, June 2007.

[Wat06] T. Watanabe, H. Tsukada, and H. Isozaki. Left-to-right target generation for hierarchical phrase-based translation. In *Proceedings of ACL'06*, pages 777–784, 2006.

[Wea55] W. Weaver. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23, Cambridge, MA, 1949/1955. MIT Press. Reprinted from a memorandum written by Weaver in 1949.

[Wu96] D. Wu. A polynomial-time algorithm for statistical machine translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, 1996.

[Wu97] D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 3(23):377–403, 1997.

[Wu99] J. Wu and S. Khudanpur. Combining nonlocal, syntactic and n-gram dependencies in language modeling. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, volume 5, pages 2179–2182, Budapest, Hungary, September 1999.

[Wu06] H. Wu, H. Wang, and Zh. Liu. Boosting statistical word alignment using labeled and unlabeled data. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 913–920, Sydney, Australia, July 2006.

[Xia04] F. Xia and M. McCord. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the COLING 2004*, 2004.

[Xio06] D. Xiong, Q. Liu, and S. Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 521–528, 2006.

[Xu03] P. Xu, A. Emami, and F. Jelinek. Training connectionist models for the structured language model. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 160–167, Morristown, NJ, USA, 2003.

[Xu04] P. Xu and F. Jelinek. Random forest in language modeling. In *Proceedings of EMNLP 2004*, pages 325–332, 2004.

[Xu05] J. Xu, E. Matusov, R. Zens, and H. Ney. Integrated chinese word segmentation in statistical machine translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT) 2005*, pages 141–147, Pittsburgh, PA, USA, October 2005.

[Xu07] J. Xu, Y. Deng, Y. Gao, and H. Ney. Domain dependent statistical machine translation. In *Proceedings of the MT Summit XI*, Copenhagen, Denmark, September 2007.

[Yam01] K. Yamada and K. Knight. A syntax-based statistical translation model. In *Proceedings of the Association for Computational Linguistics (ACL) 2001*, pages 523–530, 2001.

[Zen02] R. Zens, F. Och, and H. Ney. Phrase-based statistical machine translation. In *Proceedings of the 25th Annual German Conference on AI: Advances in Artificial Intelligence*, pages 18–32, 2002.

[Zen03] R. Zens and H. Ney. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Sapporo, Japan, 2003.

[Zha01] Y. Zhang, R. Brown, and R. Frederking. Adapting an example-based translation system to chinese. In *Proceedings of Human Language Technology Conference 2001*, pages 7–10, San Diego, California, USA, March 2001.

[Zha03] H. Zhang, H. , Yu, D. Xiong, and Q. Liu. HHMM-based chinese lexical analyzer ICTCLAS. In *Proceedings of the 2nd SIGHAN Workshop On Chinese Language Processing affiliated with 41st ACL*, pages 184–187, July 2003.

[Zha07a] D. Zhang, M. Li, C. Li, and M. Zhou. Phrase reordering model integrating syntactic knowledge for SMT. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 533–540, Prague, Czech Republic, June 2007.

[Zha07b] Y. Zhang, R. Zens, and H. Ney. Improved chunk-level reordering for statistical machine translation. In *Proceedings International Workshop on Spoken Language Translation (IWSLT)*, Trento, Italy, October 2007.

[Zip49] G. K. Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, 1949.

[Zol06] A. Zollmann and A. Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings of the North American Association for Computational Linguistics Conference (NAACL) 2006*, 2006.

[Zol08] A. Zollmann, A. Venugopal, F. Och, and J. Ponte. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of Coling 2008*, pages 1145–1152, Manchester, August 2008.

[Zwa07] S. Zwarts and M. Dras. Syntax-based word reordering in phrase-based statistical machine translation: Why does it work? In *Proceedings of the MT Summit XI*, Copenhagen, Denmark, 2007.