Discriminative Phrase-Based Models for Arabic Machine Translation

CRISTINA ESPAÑA-BONET, JESÚS GIMÉNEZ and LLUÍS MÀRQUEZ
TALP Research Center, LSI Department, Universitat Politècnica de Catalunya

A design for an Arabic-to-English translation system is presented. The core of the system implements a standard Phrase-Based Statistical Machine Translation architecture, but it is extended by incorporating a local discriminative phrase selection model to address the semantic ambiguity of Arabic. Local classifiers are trained using linguistic information and context to translate a phrase, and this significantly increases the accuracy in phrase selection with respect to the most frequent translation traditionally considered. These classifiers are integrated into the translation system so that the global task gets benefits from the discriminative learning. As a result, we obtain significant improvements in the full translation task at the lexical, syntactic and semantic levels as measured by an heterogeneous set of automatic evaluation metrics.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Machine translation*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Experimentation

Additional Key Words and Phrases: Arabic, Discriminative learning, English, Statistical machine translation

1. INTRODUCTION

Nowadays, one of the most common paradigms for Machine Translation (MT) is the statistical approach, above all when there is a large amount of parallel texts available as it is the case of the Arabic–English language pair. From the first works on Statistical Machine Translation (SMT) by Brown et al. [1990; 1993], the field has experienced several enhancements. It was soon noticed that translation is not a word to word process, that the information of surrounding words would help and that one word could be translated into more than one element. This motivated the usage of *phrases*[1] as translation units and consequently the birth of Phrase-Based SMT [Och and Ney 2004; Koehn et al. 2003]. Further enhancements involve the incorporation of syntactic structure. Syntax-Based SMT [Yamada and Knight

[1] Within this context and the context of this paper, a phrase is a sequence of words that appear together in the source sentence, but it is not necessarily defined according to the syntactic structure of the sentence.

Authors' address: TALP Research Center, LSI Department, Universitat Politècnica de Catalunya, Jordi Girona Salgado 1–3, E-08034, Barcelona, Catalunya.
Authors' e-mails: cristinae@lsi.upc.edu, jgimenez@lsi.upc.edu and lluism@lsi.upc.edu

2001; Chiang 2005] is currently an active field of research too, but we stick to Phrase-Based SMT in this work.

In SMT, the best translation for a given source sentence f is the most probable one, the target sentence e, and the probability is expressed as the sum of different components. The log-linear model [Och and Ney 2002], a generalisation of the original noisy-channel approach, estimates the probability as the logarithmic sum of several terms. Two of them, the language model $P(e)$ and the translation model $P(f|e)$, are the core of the approach. The former is a way to assign probabilities to word sequences in the target language which take care of the fluency of the output. The latter is the term taking into account the correspondence between the two languages, that is, the probability that a sentence f is translated into the sentence e, and it is usually splitted as the sum of the probabilities $P_i(f_i|e_i)$ for every phrase i that make up the sentence.

Within the standard framework, the probabilities of the translation model are calculated via frequency counts in a training corpus at the phrase level. Therefore, the probability score associated to the translation of a phrase f_i into e_i does not include any information on the context of the phrase or on the grammar of the sentence; it is just a lexical translation of the isolated phrase. The language model somehow takes care of the context in the target language but at a short distance (usually from three to five words), and, besides, independently from the translation model.

It seems clear that using linguistic information and the surrounding context of each phrase should help the translation so, a first question to ask is how this can be included in the statistical approach. In that respect one may think of translation as a *phrase selection*, and treat it as a classification problem instead of assigning a translation probability given by relative frequency counts. Machine learning techniques can then be used to score the translations using various features that encode the information of the phrase context. One could understand the different translations of a phrase as different senses of that phrase, and try to identify which is the intended sense for each word in a sentence. This interpretation shows an analogy between treating phrase selection as a classification problem and word sense disambiguation (WSD) techniques, where classifiers are used to select the correct sense of a word. Several works exploit this idea for MT on different language pairs –see for instance [Carpuat and Wu 2005; Vickrey et al. 2005; Carpuat and Wu 2007; Bangalore et al. 2007; Giménez and Màrquez 2008; Stroppa et al. 2007; Specia et al. 2008] and references therein. Most important differences among these works are briefly outlined in Section 3.

Although using discriminative learning methods for SMT can be useful for any language pair, those source languages with especially ambiguous semantics where words tend to have a larger number of lexical translations could get more benefits from the procedure. The non-diacritisation of Arabic written documents is one of the major causes for the increment of the ambiguity with respect to other languages. Since short vowels, for instance, are written as diacritics, its absence makes that sometimes the only way to know the meaning of a written word is by its context. Arabic is therefore an appropriate language to test the power of the discriminative phrase selection.

In this paper, based on our previous experience on the case of Spanish-to-English translation [Giménez and Màrquez 2008], we have trained a dedicated lexical selection model for Arabic-to-English translation. Our model deals with the translation of every source phrase as a multi-class classification problem, where every possible translation of the given phrase is a class. These local phrase translation classifiers rely on Support Vector Machines (SVM) as learning paradigm. Local predictions are then softly integrated into a SMT architecture so they can interact with other models without modifying the basic architecture.

The outline of the paper is as follows. First of all, in Section 2, we point at some peculiarities of Arabic that will be relevant for our system. Section 3 explains the discriminative phrase selection method and Section 4 the data we use in the analysis and the pre-process we apply. Next, in Section 5, we study the local task of phrase selection and afterwards in Section 6 we explore its extension to the full task of translation. Finally, we draw our conclusions.

2. ARABIC LANGUAGE IN THE CONTEXT OF SMT

The Arabic script is an alphabet with allographic variants, diacritics and ligatures. Each character has four allographs depending on its position within the word: initial, medial, final or as stand alone. The alphabet is composed by 25 consonants, 3 semi-consonants, 3 short vowels, 3 long vowels and 2 diphthongs. The short vowels, *fatha*, *kasra* and *damma*, are not letters themselves but diacritics written above or below consonants. Other diacritics are also used as a non-vowel mark (*sukun*), as a double consonant mark (*shadda*), or as a letter itself (*hamza*). For example, عِلْم, عَلَم and عَلَّمَ are three different vocalisations for the consonants علم.

However, diacritics are not usually seen in written texts. They appear in the Koran, in some other religious texts, classical poetry, textbooks or in complex texts to avoid ambiguity. In most cases, when pronunciation is not especially important, texts are non-vocalised and non-diacritised. This is mostly the case of the corpora used for MT and that increases the ambiguity of written texts, being the context sometimes the only way of choosing among the different meanings. The three possible vocalisations of علم seen before must be distinguished so that they can be translated as "science" or "knowledge" (عِلْم), "flag" (عَلَم) or "teach" (عَلَّمَ). These three words are perfectly distinguishable when speaking but not when reading. This kind of ambiguity is to be added to homonyms in Arabic. Besides, verbal declinations can further increase the number of meanings.

In general, the codification of Arabic script is different from Latin script. Since we deal here with a language pair that mixes both scripts, it is useful to unify the codification. There exist several transliterations to convert Arabic characters to the Latin alphabet. In NLP, the original texts encoded in ISO-8859-6 or CP-1256 for example are usually converted to the Buckwalter transliteration[2]. This is a one to one correspondence between Unicode and UTF-8 codification. Once all of our data are in UTF-8 they can be treated homogeneously by machines. Besides, the romanisation eases the understanding for those not familiarised with the Arabic

[2] The Buckwalter transliteration can be found at `http://www.qamus.org/transliteration.htm`

phonetics. This way, the previous example can be read as *Eilom* (عِلْم), *Ealam* (عَلَم) or *Eallama* (عَلَّمَ).

Arabic is a morphologically rich language, and another characteristic to take into account in our system is the fact that words are formed by combination of several elements sometimes joined together by ligatures. A full word agglutinates to the root affixes and clitics. Affixes mark tense, gender and number. Clitics are divided into proclitics (before the root) and enclitics (at the end of the word). Proclitics are prepositions, conjunctions and determiners; enclitics are pronouns and possessives.

Let us see an example. The Arabic token وعلمهم (or *wElmhm* using Buckwalter's transliteration) is translated into three English words: "and their knowledge". It can be morphologically segmented as:

enclitic	stem	proclitic
hm	**Elm**	w
(their)	(knowledge)	(and)

where it is taken into account that Arabic is read from right to left. It is clear from this example that the segmentation of *wElmhm* in *w Elm hm* will ease the translation by improving the alignments and reducing the original sparsity, since the number of occurrences in the corpus of every segment by itself will be higher than the occurrences of the full Arabic word.

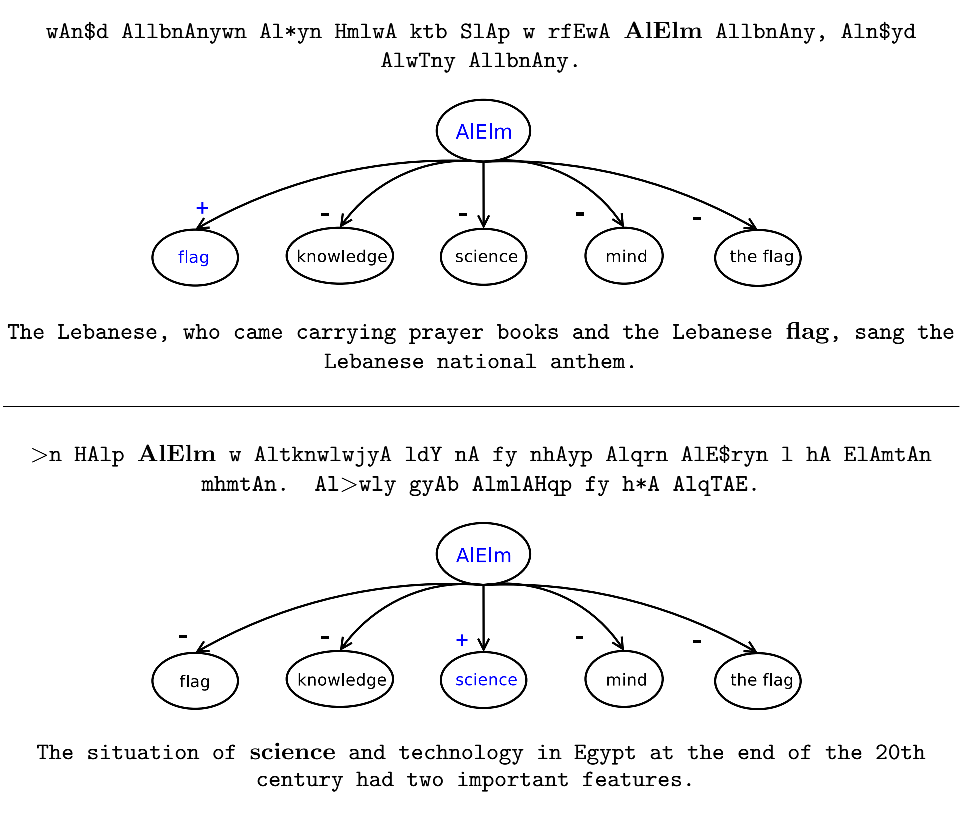
3. DISCRIMINATIVE PHRASE TRANSLATION MODEL

There are several recent methods in the literature to integrate discriminative learning techniques into the translation process. In 2005, Carpuat and Wu [2005] used WSD predictions, as a pre-process, to constrain the possible translations available at decoding time. The same year, Vickrey et al. [2005] applied discriminative models for word selection but they were used in a blank-filling task instead of full translation. This work was first extended to the full translation task and afterwards to translate phrases instead of words (see [Carpuat and Wu 2007] and references therein).

The related works dedicated to full translation differ on the learning algorithm and have been applied to different language pairs and developed with different evaluation metrics, being a direct comparison difficult. Carpuat and Wu [2007] used a WSD system which combined naïve Bayes, maximum entropy, boosting and kernel PCA-based models. Simultaneously, Bangalore et al. [2007] relied on a maximum entropy model, Stroppa et al. [2007] applied memory based learning and a bit later Specia et al. [2008] used Inductive Logic Programming. Here, we use the model of Giménez and Màrquez [2008] based on SVMs to solve the multi-class classification problem where every possible translation is a class.

In that model, the phrases are extracted from the alignments estimated from the parallel corpus. Therefore, the candidate phrases to be used for the discriminative phrase selection are not syntactic phrases but word n-grams, and are the same as the collection used in the translation model of the SMT system. The translation table obtained from the alignments is then our classification problem being the translation of each source phrase a multi-class classification problem.

For the discriminative learning, each occurrence of a phrase is taken as a positive example for its current translation and negative for the rest of possible translations.



wAn$d AllbnAnywn Al*yn HmlwA ktb SlAp w rfEwA **AlElm** AllbnAny, Aln$yd AlwTny AllbnAny.

The Lebanese, who came carrying prayer books and the Lebanese **flag**, sang the Lebanese national anthem.

>n HAlp **AlElm** w AltknwlwjyA ldY nA fy nhAyp Alqrn AlE$ryn l hA ElAmtAn mhmtAn. Al>wly gyAb AlmlAHqp fy h*A AlqTAE.

The situation of **science** and technology in Egypt at the end of the 20th century had two important features.

Fig. 1. Example of the translation of the phrase العلم (*AlElm* in Buckwalter transliteration) in two different contexts. A linear SVM is trained for each possible translation using a different translation as a positive example ("flag" or "science") and the rest as negative ones.

This way the multi-class problem is binarised and converted into a one-vs-all decision. Let us see an example. The word *Elm* is found in the corpus together with the article: *AlElm*. This token is seen in 114 examples with 10 possible translations, being the most frequents:

AlElm:

Translations	flag	science	knowledge	mind	the flag
# examples	47	26	15	9	6

When training the pair (AlElm, science) we find 26 positive examples and 88 negative ones in the corpus, while there are 6 positive examples for (AlElm, the flag) for instance (see Figure 1).

As another example we show the translation of the Arabic phrase وقع (*wqE*), which will be further considered in Section 6 to illustrate the full translation task. The word *wqE* appears in the corpus 289 times with 30 different translations such as:

wqE:

Translations	signed	took place	was signed	occurred	happened	fell
# examples	70	36	30	23	16	5

Although the translation "signed" is much more frequent than "fell", the context of both translations should be different enough so that it can distinguish the cases where the verb fallen is more appropriate. This information can be encoded as features of the phrase.

Linear SVMs are trained with tens of features all of them coming from the source sentence. That is, our vector of features is $\phi(f_i)$. Since we are interested in including linguistic information in the learning process, the Arabic part of the parallel corpus must be annotated so that the feature set for each example can contain information of the source phrase. For this purpose we consider the part-of-speech (PoS), a coarser version of the PoS, and the BIO label resulting of a base phrase chunking for the phrase itself[3]. We also include information of the local context, five words to the left and five to the right, by taking 3-grams of the linguistic information. A bag of words of the whole sentence is used to take into account the global context of the phrase. All of this collection of features make up $\phi(f_i)$. As an example, Table I shows the set of features to be used in the learning process for a case where *AlElm* is translated as "knowledge".

We train a classifier for every possible translation e_j of a phrase f_i given the previous set of features and the examples in the training parallel corpus. For instance, for the words *AlElm* and *wqE* discussed previously, 10 and 30 SVMs must be trained, respectively. When translating one phrase, the classifiers for every translation are considered, resulting into a collection of SVM scores that can be converted into probabilities using a softmax function [Bishop 1995]. We do not only obtain the most adequate translation but which is the probability of all of them. In other words, we do not select one translation but make all predictions available to the decoder as an alternative translation model.

However, not every phrase will have a DPT (Discriminative Phrase Translation) prediction. We require a minimum number of examples in order to train the classifiers, let us say 100 in our experiments. For those phrases with fewer examples we extend the probability $P_{\text{DPT}}(e|f)$ with the standard MLE (Maximum Likelihood Estimation) prediction. Even for a phrase with more than these 100 occurrences in the training corpus, there might be some of the translations with a representation in the corpus too small to be learned satisfactorily. As we will see in Section 6, we do not train a classifier for translation options that represent less than a 0.5% of the total number of examples of the given phrase; these cases are also completed with the MLE score. Whenever we combine both predictions, DPT and MLT, we normalise the probabilities to the percentage of examples estimated with each method so that the scores sum 1.

The final probability is included in the translation system as a component of a log-linear model. A standard SMT system estimates the probability of a translation

[3] Meaning *B* Beginning of a phrase, *I* Inside a phrase and *O* Outside a phrase.

Annotated sentence (word$_{PoS|coarsePoS|chunk}$) :

$\texttt{w}_{CC\|C\|O}$ $\texttt{tAbE}_{VBD\|V\|B-VP}$ $\texttt{mr\$d}_{NN\|N\|B-NP}$ $\texttt{AlIxwAn}_{NN\|N\|B-NP}$ $``_{PUNC\|P\|O}$ $\texttt{In}_{IN\|I\|B-SBAR}$ $\mathbf{AlElm}_{NN\|N\|B-NP}$ $\texttt{AlmTlwb}_{JJ\|J\|I-NP}$ $\texttt{fy}_{IN\|I\|B-PP}$ $\texttt{dyn}_{NN\|N\|B-NP}$ $\texttt{nA}_{PRP\$\|P\|I-NP}$ $\texttt{hw}_{PRP\|P\|B-NP}$ $\texttt{kl}_{NN\|N\|B-NP}$ $\texttt{Elm}_{NN\|N\|I-NP}$ $\texttt{nAfE}_{NN\|N\|B-NP}$ ...

Phrase features:

PoS	NN
coarse PoS	N
chunk	B-NP

Sentence features:

word *n*-grams	(AlmTlwb)$_1$, (fy)$_2$, (dyn)$_3$, (nA)$_4$, (hw)$_5$, (In)$_{-1}$, (”)$_{-2}$, (AlIxwAn)$_{-3}$, $(\text{mr\$d})_{-4}$, (tAbE)$_{-5}$, (AlmTlwb fy)$_1$, (fy dyn)$_2$, (dyn nA)$_3$, (nA hw)$_4$, (In AlmTlwb)$_{-1}$, (” In)$_{-2}$, (AlIxwAn ”)$_{-3}$, $(\text{mr\$d AlIxwAn})_{-4}$, $(\text{tAbE mr\$d})_{-5}$, (AlmTlwb fy dyn)$_1$, (fy dyn nA)$_2$, (dyn nA hw)$_3$, (In AlmTlwb fy)$_{-1}$, (” In AlmTlwb)$_{-2}$, (AlIxwAn ” In)$_{-3}$, $(\text{mr\$d AlIxwAn ”})_{-4}$, $(\text{tAbE mr\$d AlIxwAn})_{-5}$
PoS *n*-grams	(JJ)$_1$, (IN)$_2$, (NN)$_3$, $(\text{PRP\$})_4$, (PRP)$_5$, (IN)$_{-1}$, (PUNC)$_{-2}$, (NN)$_{-3}$, (NN)$_{-4}$, (VBD)$_{-5}$ (JJ IN)$_1$, (IN NN)$_2$, $(\text{NN PRP\$})_3$, $(\text{PRP\$ PRP})_4$, (IN JJ)$_{-1}$, (PUNC IN)$_{-2}$, (NN PUNC)$_{-3}$, (NN NN)$_{-4}$, (VBD NN)$_{-5}$ (JJ IN NN)$_1$, $(\text{IN NN PRP\$})_2$, $(\text{NN PRP\$ PRP})_3$, (IN JJ IN)$_{-1}$, (PUNC IN JJ)$_{-2}$, (NN PUNC IN)$_{-3}$, (NN NN PUNC)$_{-4}$, (VBD NN NN)$_{-5}$,
coarse PoS *n*-grams	(J)$_1$, (I)$_2$, (N)$_3$, (P)$_4$, (P)$_5$, (I)$_{-1}$, (P)$_{-2}$, (N)$_{-3}$, (N)$_{-4}$, (V)$_{-5}$ (J I)$_1$, (I N)$_2$, (N P)$_3$, (P P)$_4$, (I J)$_{-1}$, (P I)$_{-2}$, (N P)$_{-3}$, (N N)$_{-4}$, (V N)$_{-5}$ (J I N)$_1$, (I N P)$_2$, (N P P)$_3$, (I J I)$_{-1}$, (P I J)$_{-2}$, (N P I)$_{-3}$, (N N P)$_{-4}$, (V N N)$_{-5}$
chunk *n*-grams	(I-NP)$_1$, (B-PP)$_2$, (B-NP)$_3$, (I-NP)$_4$, (B-NP)$_5$, (B-SBAR)$_{-1}$, (O)$_{-2}$, (B-NP)$_{-3}$, (B-NP)$_{-4}$, (B-VP)$_{-5}$ (I-NP B-PP)$_1$, (B-PP B-NP)$_2$, (B-NP I-NP)$_3$, (I-NP B-NP)$_4$, (B-SBAR I-NP)$_{-1}$, (O, B-SBAR)$_{-2}$, (B-NP O)$_{-3}$, (B-NP B-NP)$_{-4}$, (B-VP B-NP)$_{-5}$ (I-NP B-PP B-NP)$_1$, (B-PP B-NP I-NP)$_2$, (B-NP I-NP B-NP)$_3$, (B-SBAR I-NP B-PP)$_{-1}$, (O B-SBAR I-NP)$_{-2}$, (B-NP O B-SBAR)$_{-3}$, (B-NP B-NP O)$_{-4}$, (B-VP B-NP B-NP)$_{-5}$
bag-of-words	left: AlIxwAn, $\text{mr\$d}$, tAbE right: $\text{\$rEyAF}$, AlmTlwb, AlnAs, Elm, ElmAF, dyn, kAn, kl, nAfE, swA’, tbqY, tjrybyAF, vmrt

Table I. Set of features used for the given example to train a classifier for the phrase *AlElm*.

as the sum of several terms:

$$
\begin{aligned}
\log P_{\text{SMT}}(e|f) \;=\; & \lambda_{lm} \log P(e) + \lambda_{lg} \log lex(f|e) + \lambda_{ld} \log lex(e|f) \\
& + \lambda_g \log P_{\text{MLE}}(f|e) + \lambda_d \log P_{\text{MLE}}(e|f) \\
& + \lambda_{di} \log P_{di}(e,f) + \lambda_{ph} \log ph(e) + \lambda_w \log w(e)\,, \qquad (1)
\end{aligned}
$$

where $P(e)$ is the language model probability, $lex(f|e)$ and $lex(e|f)$ are the generative and discriminative lexical translation probabilities respectively, $P_{\text{MLE}}(f|e)$ the MLE generative translation model, $P_{\text{MLE}}(e|f)$ the discriminative one, $P_{di}(e,f)$ the distortion model and $ph(e)$ and $w(e)$ correspond to the phrase and word penalty models.

The log-linear model admits the addition of new scores, so that we consider our final translation probability to be:

$$\log P(e|f) = \log P_{\text{SMT}}(e|f) + \lambda_{\text{DPT}} \log P_{\text{DPT}}(e|f)\,, \qquad (2)$$

where $P_{\text{SMT}}(e|f)$ is the full sum of log-probabilities. As an alternative, we also use the original form of $P_{\text{SMT}}(e|f)$ with the substitution of $P_{\text{MLE}}(e|f)$ by $P_{\text{DPT}}(e|f)$. In both cases, $P_{\text{DPT}}(e|f)$ interacts with the rest of components to select the final translation.

4. CORPUS AND PRE-PROCESSING

We apply the discriminative phrase translation model to the Arabic-to-English translation task. In the following, we describe the data we use for that purpose and the pre-processing needed.

4.1 Corpus

The training set is a compilation of six corpora supplied by the Linguistic Data Consortium (LDC) for the *2008 NIST Machine Translation Open Evaluation*[4]. The sources for these corpora are the Agence France Press News Service, An Nahar, Assabah, Xinhua News Service, Language Weaver News, and Ummah Press Service.

From the whole corpus, we select those segments with a length shorter than 100 words. A segment is the minimum aligned unit in the parallel corpus and corresponds to one or more sentences. The length ratio limit for obtaining the alignments with `GIZA++` [Och and Ney 2003] forces to discard segments that are more than nine times longer in one language than in the other. This filtering selects 123,662 lines, a 99% of the total, resulting a medium size corpus under the point of view of collecting alignments.

For development and test of the full translation task, we selected 500 lines from the same corpora proportionally to the training set. These small sets serve us for a fast development process. Besides, we use a larger test set with 1,357 sentences as subministrated for the *2008 NIST MT Evaluation Campaign* to report our final analysis and results.

4.2 Pre-processing and Annotation

The use of linguistic information in disambiguating the phrases makes it necessary to annotate the corpus beforehand. A minimal standard pre-processing in the corpus has been applied too, and it differs across languages.

Since we only include linguistic information of the source sentences, there is no need to annotate the English part of the corpora. The only pre-processing has been to lowercase and tokenise the sentences.

[4] `http://www.nist.gov/speech/tests/mt/2008/`

The pre-process for Arabic is a bit wider. First of all, it is useful to change the codification of the texts. We romanise the original corpus with Buckwalter transliteration. As minor details, we alter the standard transliteration by using the *XML-friendly* version which changes the characters <, > and & to I, O and W respectively. That allows to generate the XML files necessary for the discriminative learning without problems. Note, as well, that actual presentation glyphs vary with context as well as entering into various ligatures. The ligature of the letters *lam* and *alif* (ل + ا) with the corresponding diacritics, لا, لآْ, لَا or لأ, have not been detected in the automatic transliteration but converted afterwards.

The standard Buckwalter transliteration has been a prerequisite necessary to annotate the Arabic part of the corpus using the `AMIRA package` [Diab et al. 2004]. This software uses the `Yamcha SVM tools` [Kudo and Matsumoto 2003] to apply the three steps we are interested in: tokenisation, PoS tagging and base phrase chunking of the input text. `AMIRA` includes models trained on news domain with the Arabic Penn TreeBank ATB 1 v3.0, ATB 2 v2.0 and ATB 3 v2.0. Finally, since the public version of `AMIRA` does not separate the determiner ال (*Al-*), we have separated it after the annotation process. We have looked for all the words beginning with *Al-* in the Arabic WordNet[5] [Fellbaum et al. 2006], and when a word does not appear we segment out the determiner and adapt the chunk label as appropriate. Words and PoS remain the same. We have identified 309 words from the Arabic WordNet beginning with *Al-*. The Arabic WordNet contains word lemmas but we do not count with an Arabic lemmatiser. Therefore, we are comparing the stem and not the lemma with those words and there can be a loss in the precision of the segmentation.

Notice that the separation of determiners increases the length of the sentence. Before any processing, the mean length of a sentence in our corpus has 27.4 tokens. The number grows up to 31.8 when the clitics are segmented out, and up to 38.2 when also are the determiners. This has consequences when cleaning the corpus because the length of the English sentence remains the same, a mean of 34.5 tokens per sentence. The limit of `GIZA++` for the ratio between the lengths of the sentences for calculating the alignments eliminates more sentences the more we segment the original text. In the case where both clitics and determiners have been separated from the stem, we have kept sentences shorter than 120 words instead of the 100 words limit of the other cases. With this we obtain three corpora in the news domain differentiated by the level of segmentation with the main characteristics described in the left part of Table II.

5. DISCRIMINATIVE PHRASE SELECTION, THE LOCAL TASK

Before approaching the full task of translation we show some details of the subtask of phrase selection. The strength of this method is its capability of using the context of each phrase and the linguistic information available in order to select the best translation. And this is especially useful to solve ambiguities, as we have seen a very common semantic phenomenon in Arabic.

We have trained linear SVMs to solve this problem. On the one hand, the features for training the classifier are extracted from both the source phrase and

[5] Arabic WordNet: `http://www.globalwordnet.org/AWN/`

	lines	tokens	toks/line	BLEU	
				dev	test
punct.	124,154	3,402,824	27.4	25.76	23.46
punct.+clitics	123,662	3,939,726	31.8	26.25	23.81
punct.+clitics+Al	123,498	4,718,933	38.2	25.28	23.21

Table II. Definition of three corpora according to the degree of segmentation of the Arabic words as shown in the left column. The two right columns show the BLEU scores for the translation of the NIST's news compilation.

source sentence in Arabic but not from the target in English. From the phrase, we consider word, PoS, coarse PoS and chunk label n-grams. The same features are extracted from the full sentence with the addition of the bag-of-words.

On the other hand, the candidate phrases are those extracted from the word alignments obtained with `GIZA++`. The input corpus is the same training set used for training the translation system (Section 4). 588,220 phrase pairs are extracted from this corpus, but most of them are not frequent enough to train a classifier based on their number of examples. If we restrict our analysis to phrases appearing more than 100 times in the training set and with more than one possible translation, a collection of 5,321 phrases is selected. Even if we are considering less than a 1% of the total, we are keeping the most frequent ones and, so, they cover most of the corpus with more than half of the occurrences. For each of these phrases, we learn a binary SVM for every translation, unless for those which do not have a representative number of positive examples. A low number of examples of a given phrase translation can be an evidence of a bad alignment for instance. We minimise this effect by discarding translations that occur less than a 0.5% of the phrase occurrences.

For training the SVMs we use the `SVM`^{light} `package` [Joachims 1999] and for efficiency reasons we work with linear kernels. The regularisation parameter of SVMs (C), the trade-off between the training error and the margin, is adjusted in the learning process for each phrase.

Table III shows the comparison of the accuracy for the phrase selection task obtained by SVMs and labelled as the Discriminative Phrase Translation (DPT), and that given by the Most Frequent Translation (MFT). Most of the phrases appear less than 500 times in the corpus, and for them an improvement in accuracy of 7.8 percentual points is obtained. The highest accuracies are for the most frequent phrases, but this already happens with the MFT. Besides, since the number of such phrases is small, the mean accuracy is not much influenced by them. The improvement in accuracy for all the phrases obtained with the DPT is of 7.5 percentual points, which is comparable to some previous results on Spanish-to-English translation [Giménez and Màrquez 2008].

We can take a look at some particular examples. Let us go back to our running example, *AlElm*. When training the classifiers with a set of features similar to that shown in Table I, we obtain, after a 10-fold cross-validation, an accuracy of 71.3%. The most frequent translation does it well 49.6% of times. That is, attains a 40% relative improvement on the selection of the phrase translation.

As another example we comment the learning for the translation of the Arabic phrase *wqE*. As before, the accuracy of the most frequent translation (30.6%) is

Training set occurrences	#	Acc.MFT (%)	Acc.DPT (%)
100-500	4,310	58.7	66.5
501-1,000	565	62.3	68.8
1,001-5,000	393	66.7	73.0
5,001-10,000	27	72.2	79.5
10,001-50,000	19	66.6	74.8
> 50,000	7	76.2	80.7
Total:	5,321	59.8	67.3

Table III. Mean accuracy obtained in the phrase translation task by the most frequent translation (MFT) and with SVMs (DPT) for the set of extracted phrases. Results are also given for subsets of phrases grouped according to its frequency.

beaten by the accuracy given by the SVMs (42.6%). This is the general trend, the accuracy in the translation of phrases is improved with respect to that corresponding to the most frequent translation, but the amount of improvement depends on the phrase, the number of translations and the number of examples. In the next section, we define a measure of phrase translation accuracy for these phrases within translations given by the full machine translation system.

6. FULL TRANSLATION TASK

In the following, we investigate whether the improvement obtained for the local task of phrase selection has a positive repercussion on the global translation task.

6.1 Baseline System

Our baseline system follows the standard phrase-based SMT architecture, in which models are combined in a log-linear fashion. This architecture has the main advantage of allowing for considering additional *feature functions* further than the language and translation probability models typically used. Here, we use the standard features for an SMT system, i.e., those in Equation 1.

We build a 5-gram language model by interpolated Kneser-Ney discounting using the `SRILM Toolkit` [Stolcke 2002]. As for the translation models, we use the `GIZA++ Toolkit` to obtain the alignments, and the tools available with the `Moses` [Koehn et al. 2006; Koehn et al. 2007] package for phrase extraction and estimations of maximum likelihood probabilities.

In order to speed up the translation process, we have limited the number of candidate translations to 20 and set the distortion limit to 6 positions. Using these settings, the final search in the space of translations is accomplished by the `Moses` decoder.

Finally, we optimise the weights of every probability table by optimising translation performance on a development set. For this optimisation we use a minimum error rate training (MERT) [Och 2003] where BLEU [Papineni et al. 2002] is the reference score.

6.2 Segmentation in Arabic

As a first step we determine which is the adequate degree of segmentation in the Arabic words. This is independent of the analysis of phrase selection, but will provide us with a higher baseline quality. For this, we use the three data sets introduced in Section 4.2 with three different levels of tokenisation. With the coarser tokenisation the sparsity of the vocabulary increases and the mean length of an Arabic sentence is 0.80 times the English one. The first level of clitic segmentation diminishes the sparsity and equals the ratio between lengths to 0.92. With the second level, Arabic sentences are already longer than the English ones with ratio 1.11. In all these cases we use a language model computed from each training set without the addition of out of domain data.

We see in the right part of Table II that the best results are obtained when the sentence length in both languages is comparable, where punctuation marks and all the clitics except *Al-* are segmented. The additional separation of the determiner worsens the BLEU score by several possible reasons. First, because the method used to segment out *Al-* can be segmenting true full words. Second, because Arabic has some determiners which have no analogy in English such as those before adjectives that are added when the noun is determined as well. Finally, the difference in the sentence length and the errors on the *Al-* segmentation can make worse the quality of the alignments.

Our results agree with those of El Isbihani et al. [2006]. They tested different segmentation methods and obtained the best results for the segmentation obtained with `AMIRA`, that is without separating *Al-*, for a corpus built from the corpora of the Arabic-to-English NIST task. The worst results in their case correspond to the method that most segmentates the corpus with a ratio between the mean Arabic sentence length and the English one of 1.20.

With these results in mind, we use in the following the Arabic part of the corpus with the clitic segmentation of `AMIRA`.

6.3 Discriminative Phrase Translation

Finally, we integrate DPT predictions into the SMT system. To do this, we precalculate the DPT predictions for all possible translations of all source phrases appearing in the test (or development) set. Calculating these probabilities beforehand allows us to use a standard decoder without any modification to estimate them online, but a small trick is needed to distinguish every distinct instance of every distinct phrase. So, the input text is transformed by introducing identifiers which correspond to the number of occurrences of the word seen in the test set before the current one. For instance, the second time the transliterated word *AlElm* appears in the set is annotated as $AlElm_1$:

$$\ldots\ \ \texttt{Hyv}_{28}\ \ \texttt{tm}_{22}\ \ \texttt{AHrAq}\ \ \texttt{AlElm}_{1}\ \texttt{AldnmArky}\ \ ._{1128}$$

For those words without DPT prediction there is not subindex.

In a similar way and for the same reason, translation tables must be modified. Now, each occurrence of every source phrase has a distinct list of phrase translation candidates with their DPT predictions. DPT predictions are only estimated for the phrases appearing in the test set. Still, indexing increments tremendously the size of the translation table, and we only keep the first 50 translations for every phrase. This is not a problem since, as we said when explaining the baseline system, we limit the decoder to use the first 20 translations. Translations are sorted by weighting all the scores. Being the scores different, every system (baseline and DPT) already

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(f\|e)$	$lex(f\|e)$	$P_{MLE}(e\|f)$	$lex(e\|f)$
AlElm_1	flag	0.1986	0.6438	0.5417	0.3241	0.2826
AlElm_1	the	0.0419	0.0001	0.0001	0.0207	0.0217
AlElm_1	mind	0.0401	0.0608	0.0425	0.0620	0.0543
AlElm_1	the flag	0.0397	0.4000	0.5417	0.0414	0.0786
AlElm_1	flag during	0.0394	0.6667	0.5417	0.0138	0.0001
AlElm_1	knowledge	0.0392	0.0846	0.0798	0.1103	0.0924
AlElm_1	flag caused	0.0387	1.0000	0.5417	0.0138	0.0001
AlElm_1	science	0.0377	0.1529	0.1477	0.1793	0.1413
AlElm_1	education	0.0377	0.0018	0.0029	0.0138	0.0163
AlElm_1	in mind	0.0371	0.0571	0.0425	0.0138	0.0004
AlElm_1	...					

Table IV. Example of a fragment of the translation table indexed in order to take into account DPT predictions.

differs in the translation candidates list available to the decoder.

In case we do not have a DPT prediction for a phrase because it did not have the minimum number of examples required (100 in our experiments), we complete the translation table by using the MLE prediction for that phrase. For those phrases with only some of the translation probabilities obtained with the DPT method (the others having less than a 0.5% of positive examples in our experiments), we normalise the probabilities of each method, MLE and DPT, to their number of examples with respect to the total.

Table IV shows all the translations available for the phrase *AlElm* the second time it appears in the test set. In this case, the preferred translation would be the same both according to $P_{DPT}(e|f)$ and to $P_{MLE}(e|f)$, but one can already see in the table that the distribution of the probability mass is different for both predictions and that can alter the best choice.

Notice that we make available to the decoder several scores. Therefore, the decoder does not always use the DPT prediction as the best translation. DPT is competing with the MLE prediction and the remaining features shown in Equation 2. The weight of every score is determined during the MERT tuning process. In our results, the DPT prediction always has a larger weight than the MLE one, being $\lambda_{DPT} \sim 3\lambda_{MLE}$. We checked another configuration as well, where the discriminative probabilities $P_{DPT}(e|f)$ replace $P_{MLE}(e|f)$ instead of being added as an additional feature. We denote by *DPT* this last system where the DPT prediction replaces the MLE one, and by DPT^+ the system where the DPT prediction is added.

In order to study the impact of DPT predictions we perform a deep analysis by using an heterogeneous set of metrics for evaluation. In previous sections, we used a lexical metric to evaluate the quality of the translation, BLEU, which besides of being one of the standard approaches allows for a fast development process. Here, for the final analysis, we use the $\texttt{IQ}_{\text{MT}}$ package [Giménez and Amigó 2006], which provides a rich set of metrics at different linguistic levels. We have selected a representative set of metrics based on different similarity criteria:

(1) **Lexical n-gram similarity** on word forms (including the following individual metrics: $\mathrm{A_{pt}}$, PER, TER, WER, BLEU, General Text Matching –GTM–, ROUGE –RG–, and METEOR –MTR–).
(2) **Shallow-syntactic similarity** on part-of-speech tags and base phrase chunks (Shallow Parsing –SP– family).
(3) **Syntactic similarity** on dependency and constituent trees (Dependency Parsing –DP– and Constituency Parsing –CP– families).
(4) **Shallow-semantic similarity** on semantic roles (Semantic Roles –SR– family).

Most of the metrics measure the amount of matching or overlap of linguistic elements, where a linguistic element can be any of the constituents just mentioned: words, parts-of-speech, dependency relations, syntactic phrases, named enties, semantic roles, etc. We have also defined a phrase translation accuracy ($\mathrm{A_{pt}}$) that measures the percentage of phrases with the demanded conditions to estimate a DPT prediction with a correct translation. By a correct translation it is understood a translation that is found in the reference too or at least in one of them in case of having multiple references. Metrics beyond the lexical ones are named as follows. The first two letters denote the family, then it is shown an *O* for *o*verlapping or an *M* for *m*atching and the corresponding linguistic element. The $\star$ symbol is added for averages over all the types of a given linguistic element. With this definition 'SR-O_r-$\star$' represents the average lexical overlap among semantic roles of the same type. A more detailed description of the metric set may be found in the $\mathtt{IQ}_{\mathrm{MT}}$ technical manual [Giménez 2007].

We translate the test set supplied for the *2008 NIST MT Evaluation Campaign* and evaluate the translations against four references so that the derived conclusions are reliable. The results of our automatic evaluation can be read in Table V, where we show in boldface the score for the preferred system. The set of metrics is calculated for the two systems with DPT prediction (*DPT* and *DPT*$^+$) together with the baseline where there is no DPT prediction (indicated by *SMT* in the table). In general, improvements are obtained with the DPT systems at the three linguistic levels: lexical, syntactic and semantic.

The phrase translation accuracy $\mathrm{A_{pt}}$ is 1.5 percentual points higher for the *DPT* system than for the *SMT* system; the difference is of 1.1 percentual points if we consider the *DPT*$^+$ instead. This accuracy refers only to phrases with a DPT prediction, but there is a loss in the gain obtained in the isolated task due to the interaction of the DPT model with the rest, i.e. translation and language models.

At the lexical level, all the metrics but TER and WER prefer the *DPT* system over the baseline. The *DPT*$^+$ is of the same order or slightly better than the *SMT* system as well, but the substitution of the MLE predictions by the DPT ones seems to be more effective, probably because of the minor number of parameters to optimise. For this system, the BLEU score increases from 31.0 to 32.4 and the NIST one from 8.7 to 8.9.

In order to check whether these results are statistically significant, we generate 1,000 sets by pair bootstrap resampling of the original test sets [Koehn 2004]. With the previous values, the *DPT* system shows to be statistically better than both *DPT*$^+$ and *SMT* systems, and *DPT*$^+$ statistically better than *SMT*.

Level	Metric	SMT	DPT	DPT^{+}
Lexical	A_{pt}	0.8256	**0.8399**	0.8370
	1-PER	0.5814	**0.5892**	0.5852
	1-TER	**0.4493**	0.4482	0.4454
	1-WER	**0.4161**	0.4102	0.4078
	BLEU-4	0.3103	**0.3243**	0.3175
	NIST-5	8.7113	**8.9053**	8.7920
	GTM-1	0.6974	**0.7159**	0.7107
	GTM-2	0.2234	**0.2267**	0.2247
	GTM-3	0.1721	**0.1745**	0.1728
	RG-L	0.4986	**0.4993**	0.4968
	RG-S⋆	0.3185	**0.3229**	0.3188
	RG-SU⋆	0.3395	**0.3437**	0.3395
	RG-W-1.2	0.2662	**0.2675**	0.2659
	MTR-exact	0.4909	**0.5001**	0.4958
	MTR-stem	0.5098	**0.5174**	0.5135
	MTR-wnstm	0.5147	**0.5222**	0.5186
	MTR-wnsyn	0.5352	**0.5426**	0.5391
Shallow Syntactic	SP-Oc-⋆	0.4376	**0.4448**	0.4407
	SP-Op-⋆	0.4195	**0.4271**	0.4235
	SP-NISTc-5	5.5783	5.6684	**5.6703**
	SP-NISbioT-5	5.9931	**6.1318**	6.1172
	SP-NISTl-5	8.8869	**9.0547**	8.9523
	SP-NISTp-5	6.9679	**7.1610**	7.1117
Syntactic	CP-Oc-⋆	0.3943	**0.3995**	0.3962
	CP-Op-⋆	0.4220	**0.4296**	0.4265
	CP-STM-9	**0.2396**	**0.2394**	0.2380
	DP-Oc-⋆	0.3852	**0.3949**	0.3892
	DP-Ol-⋆	0.3051	**0.3164**	0.3115
	DP-Or-⋆	0.2523	**0.2557**	0.2534
	DP-HWC-c-4	**0.2986**	0.2975	0.2970
	DP-HWC-r-4	0.2023	0.2023	**0.2029**
	DP-HWC-w-4	**0.0835**	0.0826	0.0831
Shallow Semantic	SR-Mr-⋆	0.0224	0.0227	**0.0262**
	SR-Mrv-⋆	0.0123	**0.0129**	**0.0129**
	SR-Or	0.3686	**0.3792**	0.3609
	SR-Or-⋆	0.1160	0.1209	**0.1234**
	SR-Orv	0.0685	**0.0815**	0.0765
	SR-Orv-⋆	0.0284	0.0325	**0.0349**

Table V. Automatic evaluation of the translated test set supplied for the *2008 NIST MT Evaluation* using lexical, syntactic and semantic metrics.

Source	لكن الجّزء الأكبر من هذه الأسلحة **وقع** في يد حماس في قطاع غزة. lkn Aljz' AlAkbr mn h*h AlAslHp **wqE** fy yd HmAs fy qTAE gzp.
Baseline	But the largest part of these weapons **signed** in the hands of Hamas in Gaza Strip.
DPT	However, the largest part of these weapons **fell** in the hands of Hamas in Gaza Strip.
Refs.	But most of these weapons **have fallen** into the hands of Hamas in the Gaza Strip. But most of these weapons **fell** into the hands of Hamas in the Gaza Strip. However, the largest part of these weapons **landed** in the hands of Hamas in the Gaza Strip. But most of these weapons **fell** into the hands of Hamas in Gaza Sector.

Table VI. Example A. The translation obtained with the DPT system selects the correct word in the given context although being the least frequent translation.

Source	و كان صرح صباحا ل هيئة الأذاعة البريطانية (بي بي سي) أن **مصدرا في روسيا** أبلغه بأمر الاغتيال. w kAn SrH SbAHA l hy}p AlA*AEp AlbryTAnyp (by by sy) An **mSdrA fy rwsyA** Ablgh bAmr AlAgtyAl.
Baseline	And had announced in the morning to the British Broadcasting Corporation (BBC) that **source in Russia** informed them about assassination.
DPT	And had announced in the morning to the British Broadcasting Corporation (BBC) that **source** informed them about assassination **in Russia**.
Refs.	In the morning he told the British Broadcasting Corporation (BBC) that a **Russian source** had told him about the assassination. In the morning, he told the British Broadcasting Corporation (BBC) that a **source in Russia** had informed him about the assassination order. In the morning he told the British Broadcasting Cooperation that a **source in Russia** had informed him of the assassination order. In the morning he told the British Broadcasting Corporation (BBC) that a **source in Russia** informed him about the assassination order.

Table VII. Example B. Sentence where the inclusion of the DPT prediction alters the final order of the phrases. In this case, it degrades the quality of the translation.

On the other hand, the syntax of the translations is improved as well. Metrics based on shallow parsing (SP) and constituent parsing (CP) behave as the lexical metrics and favour the *DPT* system. The only scores indifferent to the discriminative learning are those reflecting similarities among dependency trees (DP), since two out of six metrics prefer the baseline.

Finally, the quality of the semantics as measured by the similarities between the semantic roles (SR) of the translation and the target increases for the discriminative methods. The metrics which do not take into account the lexical realisation of the linguistic element favour the *DPT* system, those considering the lexical realisation prefer the DPT^+ one.

6.3.1 *Analysis at the sentence level.* So far, we quantified the improvement of the translations at the system level, i.e. for all the test set, but one can also study the nature of the improvement by checking how concrete translations are modified. Of course, there is not a one-to-one correspondence between a particular

Source	و قال بوش الثلاثاء ان الاستراتيجية التي ترتكز على ارسال...
	w qAl bw$ AlvlAvA' An AlAstrAtyjyp Alty trtkz ElY ArsAl...
Baseline	Bush said ∅ Tuesday that ∅ strategy based on sending...
DPT	Bush said **on** Tuesday that **the** strategy based on sending...
Refs.	**On** Tuesday, Bush said that **the** strategy focusing on sending...
	Bush said **on** Tuesday that **the** strategy based on sending...
	Bush said **on** Tuesday that **the** strategy that focuses on sending...
	Bush said **on** Tuesday that **the** strategy based on sending...

Table VIII. Example C. The DPT system favours in general more fluent translations by increasing the number of functional words as seen in the example.

translation preferred by the discriminative method and such modification because all the components play a role in the final election of the full translation, but anyway one can extract some general ideas.

We randomly selected 50 sentences from the test set that contain at least one of the phrases disambiguated by the discriminative method with a frequency ν ($100 < \nu < 500$). As seen at the beginning of this section with the example sentence for the phrase *AlElm*, several phrases with DPT prediction coexist in a same sentence. We calculate all the set of metrics shown in Table IV at a sentence level for this small subset and analyse the results.

Although the mean effect is the improvement reflected in Table V, individual sentences get both benefits and damages from the discriminative phrase translation. Tables VI, VII and VIII show the translations of three of the sentences: Example A, B and C respectively. In those sentences, some general characteristics are outlined.

Example A accomplishes the main objective of the method. In this case, a phrase that according to its frequency in the corpus has a probability one order of magnitude lower than the most frequent translation gets promoted due to the DPT prediction (the isolated task of this phrase selection has been analysed in Section 5). This way *wqE* is translated as “fell” instead of the MFT “signed” being in agreement with 2 of the 4 references. Lexical metrics are the ones that reflect better this type of improvements.

Since, as we have said, all the probabilities interact among them in order to determine the translation of the whole sentence, the addition of the DPT prediction can alter the structure of the output. For instance, Example B in Table VII shows a case where the effect is a reordering of the phrases. In the given example, the reorder damages the final translation and the meaning of the original sentence is modified.

Finally, Example C allows us to comment the gain in fluency in the translations. Articles and prepositions are more frequent in the translations obtained with the DPT method. In fact, the mean length of these translations is one word larger than the ones with the baseline. In the sentence of Table VIII that corrects the output from “Monday” to “on Monday” and from “strategy” to “the strategy”. In this case, this has a positive repercussion specially with the BLEU metric since the length of the matching n-grams is larger, but it damages the translation of headlines which are common in news corpora such as the one we use.

7. CONCLUSIONS

We have shown the positive impact of including a discriminative phrase translation model in a SMT architecture designed for the Arabic-to-English translation task.

First of all, we have studied the task of phrase selection independently of the full translation. By training a classifier to choose the adequate phrase translation for every instance of a phrase, we have obtained a gain of 7.5 percentual points in accuracy with respect to the answer that would give the most frequent translation. These classifiers are informed of the context of the source phrase and its part-of-speech and chunk label. Information on the target phrase would further improve the results, but the integration in a SMT system would not be straightforward and one would need a new architecture.

Taking into account that the probabilities used in SMT are estimated from relative frequency counts, we study how the gain in accuracy achieved by the DPT predictions affects the translation quality according to automatic evaluation metrics. Improvements are obtained at the three linguistic levels analysed: lexical, syntactic and semantic. The DPT system that substitutes the probability score from the maximum likelihood estimate $P_{MLE}(e|f)$ by the discriminative prediction $P_{DPT}(e|f)$ is preferred by a 74% of the calculated metrics, that is, 28 out of 38. Just in 5 cases the baseline is not improved; for the remaining ones, the best system is that combining both $P_{MLE}(e|f)$ and $P_{DPT}(e|f)$.

These encouraging results have also been found for the Spanish-English language pair [Giménez and Màrquez 2008], but as expected from the semantic ambiguities of Arabic, the gain is larger for this language. The Arabic phrases are translated locally with more accuracy (2.5 percentual points more) than the Spanish ones, and this is captured by all lexical metrics in the full translation. Contrary to the Spanish case, the improvement in lexical selection in Arabic has a positive repercussion not only on semantics but on syntax as well. Also, corpus heterogeneity –several sources in our case as compared to one source in our previous experiments on the case of Spanish-English (EuroParl)– may have contributed positively. Recall that our methods have been designed to model lexical selection based on source context, thus, in principle, they should be far more robust to domain shifts.

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02) and the DOI/REFLEX-NBCHC050031 program.

REFERENCES

Bangalore, S., Haffner, P., and Kanthak, S. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*. 152–159.

Bishop, C. M. 1995. 6.4: Modeling conditional distributions. In *Neural Networks for Pattern Recognition*. Oxford University Press, 215.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. 1990. A statistical approach to machine translation. *Computational Linguistics 16,* 2, 79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics 19,* 2, 263–311.

Carpuat, M. and Wu, D. 2005. Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation. In *Proceedings of IJCNLP*.

Carpuat, M. and Wu, D. 2007. How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, Ann Arbor, Michigan, 263–270.

Diab, M., Hacioglu, K., and Jurafsky, D. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*. Boston, MA.

El Isbihani, A., Khadivi, S., Bender, O., and Ney, H. 2006. Morpho-syntactic arabic preprocessing for arabic to english statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, 15–22.

Fellbaum, C., Alkhalifa, M., Black, W. J., Elkateb, S., Pease, A., Rodríguez, H., and Vossen, P. 2006. Introducing the arabic wordnet project. In *Proceedings of the 3rd Global Wordnet Conference, Jeju Island, Korea.*

Giménez, J. 2007. IQMT v 2.1. Technical Manual (LSI-07-29-R). Tech. rep., TALP Research Center. LSI Department.
http://www.lsi.upc.edu/∼nlp/IQMT/IQMT.v2.1.pdf.

Giménez, J. and Amigó, E. 2006. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*. 685–690.

Giménez, J. and Màrquez, L. 2008. *Discriminative Phrase Selection for SMT*. NIPS Workshop Series. MIT Press, 205–236.

Joachims, T. 1999. Making large-scale support vector machine learning practical. 169–184.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*. Barcelona, Spain.

Koehn, P., Hoang, H., Mayne, A. B., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*. 177–180.

Koehn, P., Och, F. J., and Marcu, D. 2003. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*. Edomonton, Canada.

Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., and Moran, C. 2006. Open Source Toolkit for Statistical Machine Translation. Tech. rep., Johns Hopkins University Summer Workshop. http://www.statmt.org/jhuws/.

Kudo, T. and Matsumoto, Y. 2003. Fast methods for kernelbased text analysis. In *Proceedings of ACL-2003. Sapporo, Japan.*

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*. Sapporo, Japan.

Och, F. J. and Ney, H. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. 295–302.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics 29,* 1, 19–51.

Och, F. J. and Ney, H. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics 30,* 4, 417–449.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics.* 311–318.

Specia, L., Sankaran, B., and das Graças Volpe Nunes, M. 2008. n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In *Computational Linguistics and Intelligent Text Processing.* Lecture Notes in Computer Science, vol. 4919/2008. Springer Berlin / Heidelberg, 399–410.

Stolcke, A. 2002. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing.*

Stroppa, N., van den Bosch, A., and Way, A. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI).* 231–240.

Vickrey, D., Biewald, L., Teyssier, M., and Koller, D. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP).*

Yamada, K. and Knight, K. 2001. A syntax-based statistical translation model. In *ACL.* 523–530.