The UPC-lsi Discriminative Phrase Selection System: NIST MT Evaluation 2008

Cristina España-Bonet, Jesús Giménez and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
cespana@am.ub.es, {jgimenez,lluism}@lsi.upc.edu

Abstract

This document describes the system developed by the Empirical MT Group at the Technical University of Catalonia, LSI Department, for the Arabic-to-English task at the 2008 NIST MT Evaluation Campaign. Our system explores the application of discriminative learning to the problem of phrase selection in Statistical Machine Translation. Instead of relying on Maximum Likelihood estimates for the construction of translation models, we use local classifiers which are able to take further advantage of contextual information. Local predictions are softly integrated into a global log-linear phrase-based statistical MT system as an additional feature. Automatic evaluation results according to a heterogeneous set of metrics operating at different linguistic levels are presented. These show a low level of agreement between metrics. Improvements over the baseline are either inexistent or not significant, except for the case of semantic metrics based on discourse representations and several syntactic metrics based on constituent and dependency parsing.

1 Introduction

This document describes the system developed by the Empirical MT Group at the Technical University of Catalonia (UPC), LSI Department for the Arabic-to-English task at the 2008 NIST MT Evaluation Campaign. Our system is a phrase-based SMT system extended with discriminative phrase translation models (Giménez and Màrquez, 2007a; Giménez and Màrquez, 2008).

The paper is organized as follows. In Section 2, we describe the baseline system. Then, in Section 3, we describe our approach to Discriminative Phrase Translation (DPT), and how DPT predictions are softly integrated into the baseline system. Finally, in Section 4, internal evaluation results are presented.

2 Baseline System

Our baseline system follows the standard phrase-based SMT architecture, in which models are combined in a log-linear fashion (Och and Ney, 2002):

$$\hat{e} = \text{argmax}_e \left\{ \sum_{m=1}^{M} \lambda_m h_m(e, f) \right\} \qquad (1)$$

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 besides the language model probability $P(e)$ and the maximum likelihood estimation both generative and discriminative ($P_{\text{MLE}}(f|e)$ and $P_{\text{MLE}}(e|f)$), the lexical translation probability in both directions ($P_{lex}(f|e)$ and $P_{lex}(e|f)$), the distortion model $P_{di}(e, f)$, and a word penalty and phrase penalty models.

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 (Och and Ney, 2003) to obtain the alignments, and the tools available with the Moses package (Koehn et al., 2006; Koehn et al., 2007) 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 optimize the weights of every probability table by optimizing translation performance on a development set. For this optimization we use a minimum error rate training (MERT) (Och, 2003) where BLEU is the reference score.

2.1 Data

In order to run our experiments, we compile a training set from six corpora supplied by the Linguistic Data Consortium (LDC):

- Arabic English Parallel News Part 1 (61,000 lines)
- Arabic News Translation Text Part 1 (18,000 lines)
- Arabic Treebank English Translation (23,800 lines)
- eTIRR Arabic English News Text (4,000 lines)
- Multiple-Translation Arabic (Parts 1 & 2) (15,533 lines)
- TIDES MT2004 Arabic evaluation data (1,329 lines)

From the whole corpus, lines[1] with a length shorter than 100 words and not more than nine times longer in one language than in the other one are used in the compilation. That is the optimal length for training with the Moses decoder and the length ratio limit for obtaining the alignments with GIZA++. With this, 123,662 lines, a 99% of the total, have been obtained. Both the translation model and the language model are estimated from this compilation.

For the development and test sets we selected 500 lines from the same corpora with the exception of the *Multiple-Translation Arabic* and the *TIDES MT2004 Arabic evaluation data*. The number of lines from each corpus is proportional to the one in the training set.

[1]Each line corresponds to the minimum aligned unit. The aligments are given at a fragment level, which is in most cases larger than one sentence.

2.2 Pre-processing

The input data have been pre-processed and converted to a unique codification. Afterwards, we tokenize both of the input languages and annotate them with the lemma, part-of-speech (PoS) and chunk label for each word. The concrete tools depend on the language.

The standard Buckwalter transliteration[2] has been a prerequisite necessary to annotate the Arabic part of the corpus using the ASVMTools (Mona Diab and Jurafsky, 2004). This software uses the Yamcha SVM tools (Kudo and Matsumoto, 2003) to tokenize, PoS tag and Base Phrase Chunk the input text. The process of tokenization segments the words in proclitics, stems+affixes, and enclitics, although the determiner *Al-* has not been segmentated. Punctuation is considered as an independent token as well. For PoS tagging, the package uses the 24 PoS tags from the collapsed tag set of the Arabic TreeBank distribution. For chunking, phrases have been labeled according to IOB tagging scheme (Inside-Outside-Beginning). Notice, that we do not obtain Arabic lemmas with ASVMTools.

English sentences have been lowercased and tokenized. Lemma and PoS have been obtained with SVMTool (Giménez and Màrquez, 2004), and Yamcha (Kudo and Matsumoto, 2003) has been used afterwards for BP chunking. PoS labels correspond to full tagset from the Wall Street Journal with 36 labels, and, as before, chunk labels follow the IOB tagging scheme.

3 Dedicated Discriminative Phrase Selection

In standard phrase-based SMT systems, like that described by Koehn et al. (2003), translation models are built on the basis of relative frequency counts, i.e., Maximum Likelihood Estimates (MLE). Thus, all the occurrences of the same source phrase are assigned, no matter what the context is, the same set of translation probabilities. For that reason, recently, there is a growing interest on the construction of dedicated discriminative models to the problem of lexical selection which are able to take into account a wider feature context (Vickrey et al., 2005;

[2]`http://www.ldc.upenn.edu/myl/morph/buckwalter.html`

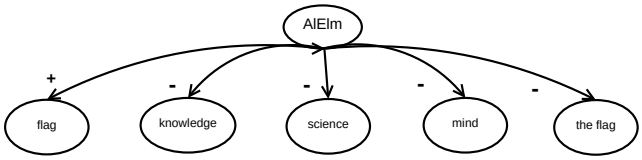

Figure 1: Phrase Translation. An Example

Giménez and Màrquez, 2007a; Carpuat and Wu, 2007; Bangalore et al., 2007; Stroppa et al., 2007; Giménez and Màrquez, 2008). Lexical selection is addressed as a multi-class classification task. For each possible source word (or phrase) according to a given bilingual lexical inventory (e.g., the translation model), a distinct classifier is trained to predict lexical correspondences based on local context. Thus, during decoding, for every distinct instance of every source phrase a distinct context-aware translation probability distribution is potentially available.

As an illustration, in Figure 1, we show an example of Arabic-to-English phrase translation, in which the source phrase *AlElm*, in this case translated as *flag*, has several possible candidate translations such as *knowledge*, *science* or *mind*. This is an example where the lack of diacritics in written texts causes a unique transliteration to have different meanings. The context of the phrase is then important to disambiguate the meaning, and that will be used when training the classifiers.

3.1 Learning

There exist a wide variety of learning algorithms which can be applied to the multiclass classification scenario defined. In this work we have focused on local linear binary Support Vector Machines[3] (Vapnik, 1995; Cristianini and Shawe-Taylor, 2000)[4]. We have applied a simple *one-vs-all* binarization scheme, i.e., a binary classifier is learned for every possible translation candidate e_j in order to distinguish between examples of this class and all the rest. Training examples are extracted from the same training data as in the case of conventional MLE-based

[3]SVMs have been learned using the SVM^{light} package by Thorsten Joachims, which is freely available at `http://svmlight.joachims.org` (Joachims, 1999).

[4]Because adjusting the C parameter for each binary classifier was impractical, it has been set to the default value, $\frac{\sum(\vec{x_i}\vec{x_i})^{-1}}{N}$, where $\vec{x_i}$ is a sample vector and N corresponds to the number of samples.

Sentence :

w tAbE mr$d AlIxwAn " In AlElm AlmTlwb fy dyn nA hw kl Elm nAfE tbqY l AlnAs vmrt h , swA' kAn ElmAF $rEyAF Ow ElmAF tjrybyAF .

Phrase features :

word n-grams	AlElm
PoS n-grams	NN
coarse PoS n-grams	N
chunk n-grams	B-NP

Sentence features :

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

Table 1: Phrase Translation Features. An Example

models, i.e., a phrase-aligned parallel corpus (see Section 2). Each occurrence of each source phrase f_i is used to generate a positive example for the actual class (or classes) corresponding to the aligned target phrase (or phrases), and a negative example for the classes corresponding to the other possible translations of f_i.

3.2 Feature Engineering

We have built a feature set which considers different kinds of information, always from the source sentence. Each example has been encoded on the basis of the *local context* of the phrase to be disambiguated and the *global context* represented by the whole source sentence.

As for the local context, we use n-grams ($n \in \{1, 2, 3\}$) of: word forms, parts-of-speech, and base phrase chunking IOB labels, in a window of 5 tokens to the left and to the right of the phrase to disambiguate. We also exploit part-of-speech and chunk information inside the source phrase, because, in contrast to word forms, these may vary and thus report very useful information. Text has been automatically annotated as explained in Section 2.2. However, for the case of parts-of-speech, because tag sets take into account fine morphological distinctions, we have additionally defined several coarser classes grouping morphological variations of nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, determiners and punctuation marks.

As for the global context, we collect topical information by considering content words (i.e., nouns, verbs, adjectives and adverbs) in the source sentence as a bag of words. We distinguish between words at the left and right of the source phrase being disambiguated.

As an illustration, Table 1 shows the feature representation for the example depicted in Figure 1. At the top, the reader may find the source sentence containing the phrase *AlElm* which has been annotated at the level of shallow syntax. The corresponding source phrase and source sentence features are shown below.

3.3 Soft Integration of DPT Predictions

We consider every instance of f_i as a separate classification problem. In each case, we collect the classifier outcome for all possible phrase translations e_j of f_i. However, SVMs outcomes are not probabilities but unbounded real numbers. We transform them into probabilities by applying the *softmax function* described by Bishop (1995).

At translation time, we do not constrain the decoder to use the translation e_j with highest probability. Instead, we make all predictions available and let the decoder choose. We have avoided implementing a new decoder by pre-computing all DPT predictions for all possible translations of all source phrases appearing in the test set. The input text is conveniently transformed into a sequence of identifiers[5], which allows us to uniquely refer to every distinct instance of every distinct word and phrase in the test set. Translation tables are accordingly modified so that each distinct occurrence of every single source phrase has a distinct list of phrase translation candidates with their corresponding DPT predictions.

f_i	e_j	$P_{DPT}(e\|f)$	$P_{MLE}(e\|f)$
...			
AlElm$_1$	flag	0.1986	0.3241
AlElm$_1$	the	0.0419	0.0207
AlElm$_1$	mind	0.0401	0.0620
AlElm$_1$	the flag	0.0397	0.0414
AlElm$_1$	flag during	0.0394	0.0138
AlElm$_1$	knowledge	0.0392	0.1103
AlElm$_1$	flag caused	0.0387	0.0138
AlElm$_1$	science	0.0377	0.1793
AlElm$_1$	education	0.0377	0.0138
AlElm$_1$	in mind	0.0371	0.0138
...			

Table 2: Translation table. An Example

As an illustration, Table 2 shows a fragment of the translation table corresponding to the phrase *"AlElm"* in the running example. Notice how this concrete instance has been properly identified by indexing the phrase ("AlElm" $\rightarrow$ "AlElm$_1$"). We show DPT predictions and MLE-based (columns 3 and 4, respectively) for several phrase candidate trans-

[5] In our case a sequence of w_i tokens, where w is a word and i corresponds to the number of occurrences of word w seen in the test set before the current one. For instance, the source sentence in the example depicted in Figure 1 is transformed into "$wywm\ AlAHd_8\ ,_{371}\ \$hdt_3\ Edp_8\ mdn_1\ AfgAnyp\ tZAhrAt\ AHtjAj\ ElY_{456}\ Alrswm_{39}\ Alms\}yp\ l_{873}\ Alnby\ (_{186}\ S\ )_{186}\ ,_{372}\ Hyv_{28}\ tm_{22}\ AHrAq\ AlElm_1\ AldnmArky\ ._{1128}$".

lations sorted in decreasing DPT probability order. The first observation is that both methods agree on the top-scoring candidate translation, "flag". However, the distribution of the probability mass is significantly different. While, in the case of the MLE-based model, there are three candidate translations clearly outscoring the rest, concentrating more than 60% of the probability mass, in the case of the DPT model predictions give a clear advantage to the top-scoring candidate although with less probability, and the rest of candidate translations obtain a very similar score.

4 Internal System Evaluation

In the following, we will evaluate two systems that include DPT predictions and compare them with the baseline which only considers MLE estimations. These three system are identified as:

SMT baseline system.

DPT baseline system in which discriminative probabilities based on MLE are replaced with DPT predictions.

DPT$^+$ baseline system extended with DPT predictions as an additional feature.

4.1 Settings

We have focused on the Arabic-to-English task. The training set consists of 123,622 parallel sentences. After performing phrase extraction over the training data, and discarding source phrases ocurring only once, translation candidates for 585,307 source phrases were obtained.

All of these phrases are used to construct the translation tables by frequency counts, but we consider only those appearing more than 100 times in the corpus to be representative enough to train the classifiers. That represents about 1% of the total amount of phrases, but since they are the most frequent ones they will cover most of the test set if it belongs to the same domain. Besides, because phrase alignments have been obtained automatically and, therefore, include many errors, source phrases may have, in their turn, a large number of associated possible phrase translations, most of them occurring very few times. We have discarded many of them by considering only as possible phrase translations those which are selected more than 0.5% of the times as the actual translation. The resulting training set consists of 5,321 Arabic source phrases.

Let us note that, because not all phrase pairs which have a MLE-based prediction have also a DPT prediction, but only those with a sufficient number of training examples, in order to provide equal opportunities to both models, we have incorporated translation probabilities for these phrases pairs into the DPT model. Up to now, this incorporation has been done in a naïve way by duplicating the MLE probability.

4.2 A Heterogeneous Set of Metrics for Automatic MT Evaluation

Most existing metrics limit their scope to the lexical dimension. However, recently, there have been several attempts to take into account deeper linguistic levels. For instance, ROUGE (Lin and Och, 2004) and METEOR (Banerjee and Lavie, 2005) may consider stemming. Additionally, METEOR may perform a lookup for synonymy in WordNet (Fellbaum, 1998). We may find as well several syntax-based metrics (Liu and Gildea, 2005; Amigó et al., 2006; Owczarzak et al., 2007; Mehay and Brew, 2007), and even metrics operating at the level of shallow semantics (Giménez and Màrquez, 2007b) and semantics (Giménez, 2007). These metrics have been showed to produce more reliable system evaluations than metrics based on lexical similarity alone.

For the purpose of performing heterogeneous automatic MT evaluations, we use the IQ$_{\text{MT}}$ package (Giménez and Amigó, 2006), which provides a rich set of more than 500 metrics at different linguistic levels[6]. We have selected a representative set of metrics, based on different similarity criteria:

- Lexical n-gram similarity (on word forms).
- Shallow-syntactic similarity (on part-of-speech tags and base phrase chunks).
- Syntactic similarity (on dependency and constituent trees).
- Shallow-semantic similarity (on named entities and semantic roles)

[6]The IQ$_{\text{MT}}$ software is available at `http://www.lsi.upc.edu/~nlp/IQMT`.

- Semantic similarity (on discourse representations).

A deeply detailed description of the metric set may be found in the IQ$_{MT}$ technical manual (Giménez, 2007).

4.3 Results

Level	Metric	SMT	DPT	DPT$^+$
Lexical	1-PER	**0.5248**	0.5224	0.5221
	1-WER	**0.3166**	0.3075	0.3081
	1-TER	**0.3679**	0.3606	0.3613
	BLEU	0.2388	0.2387	**0.2396**
	NIST	**6.4044**	6.3263	6.3225
	GTM (e=1)	0.5708	**0.5730**	0.5705
	GTM (e=2)	**0.2166**	0.2154	0.2161
	GTM (e=3)	**0.1756**	0.1743	0.1750
	RG-L	0.5290	**0.5305**	0.5276
	RG-S$\star$	**0.3442**	**0.3443**	0.3410
	RG-SU$\star$	**0.3634**	**0.3635**	0.3604
	RG-W-1.2	0.3085	**0.3111**	0.3091
	MTR-exact	0.4948	**0.4991**	0.4974
	MTR-stem	0.5142	**0.5164**	0.5153
	MTR-wnstm	0.5183	**0.5207**	0.5193
	MTR-wnsyn	0.5396	**0.5430**	0.5413
Shallow Syntactic	SP-Op-$\star$	0.4150	**0.4218**	0.4185
	SP-Oc-$\star$	0.4193	**0.4237**	0.4214
	SP-NIST$_l$	**6.5745**	6.4771	6.4790
	SP-NIST$_p$	**5.6618**	5.6225	5.6161
	SP-NIST$_{iob}$	**4.7187**	4.6627	4.6795
	SP-NIST$_c$	**4.1460**	4.0858	4.1047
Syntactic	DP-O_l-$\star$	0.2019	**0.2057**	0.2049
	DP-O_c-$\star$	**0.3344**	0.3314	0.3318
	DP-O_r-$\star$	**0.2347**	0.2319	0.2319
	DP-HWC$_w$	**0.0575**	0.0556	**0.0574**
	DP-HWC$_c$	0.2118	0.2168	**0.2181**
	DP-HWC$_r$	0.1422	0.1474	**0.1484**
	CP-O_p-$\star$	0.4133	**0.4183**	0.4158
	CP-O_c-$\star$	0.3823	**0.3868**	0.3847
	CP-STM	**0.2150**	**0.2144**	0.2128
Shallow Semantic	NE-M_e-$\star$	0.2963	**0.2979**	0.2933
	NE-O_e-$\star$	**0.3518**	**0.3515**	0.3472
	NE-O_e-$\star\star$	0.4161	**0.4217**	0.4185
	SR-M_r-$\star$	**0.0868**	0.0841	0.0848
	SR-O_r-$\star$	**0.2073**	0.2059	0.2048
	SR-O_r	**0.4143**	0.4076	0.4104
Semantic	DR-O_r-$\star$	0.2101	**0.2192**	0.2157
	DR-O_{rp}-$\star$	0.3139	**0.3272**	0.3204
	DR-STM	0.1563	0.1508	**0.1591**

Table 3: Automatic evaluation of MT results

Table 3 provides automatic evaluation results. Metrics are grouped according to the linguistic level at which they operate.

At the lexical level, while metrics based on rewarding longer n-gram matchings tend to prefer the 'SMT' baseline, variants of ROUGE and METEOR tend to prefer the 'DPT' system. Interestingly, the 'DPT$^+$' attains the highest score only according to BLEU, although not significantly.

At the shallow-syntactic level, metrics based on lexical overlapping over parts-of-speech and base chunk phrases prefer the 'DPT' and 'DPT$^+$' alternatives, with a slight advantage in favour of the 'DPT' system. However, NIST variants over sequences of lemmas, parts-of-speech, chunk labels and chunk types consistently prefer the 'SMT' baseline.

At the properly syntactic level, metrics exhibit very different behaviors. For instance, with respect to metrics based on lexical overlapping over dependency trees, while the *'DP-O_l-$\star$'* metric (i.e., overlapping between lexical items hanging at the same level of the tree) gives a clear advantage to DPT systems, the *'DP-O_c-$\star$'* (i.e., lexical overlapping between grammatical categories) and *'DP-O_r-$\star$'* (i.e., lexical overlapping between grammatical relations) metrics prefer the 'SMT' baseline. In contrast, metrics based on head-word chain matching (HWC) over dependency trees and metrics based on lexical overlapping over constituent trees clearly prefer the DPT alternatives. Finally, the syntactic tree matching (STM) metric confers a similar score to the three systems.

At the shallow-semantic level, whereas metrics based on lexical overlapping and matching between named entities (NE) seem to prefer the 'DPT' system, metrics based on semantic roles (SR) prefer the 'SMT' baseline.

Finally, at the semantic level, metrics based on lexical overlapping between discourse representations (DR) confer a significant advantage to the DPT alternatives, specially in the case of the 'DPT' system. The semantic tree matching (STM) metric gives a slight advantage to the 'DPT$^+$' system.

5 Further Steps

This work has been our first approach to the Arabic-to-English translation task, so, although results do not considerably improve those of a standard SMT system, we believe there is room for improvement.

Besides a better processing of Arabic and adding

Arabic lemmas to our feature set, we plan to refine our system. That involves a better integration of DPT predictions into the SMT system by completing DPT probabilities for those phrases without prediction applying a discounting. Other aspects such as the optimization of the weights for every translation model will be also explored.

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02). We are recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government. CEB has been funded by the Spanish Ministry of Education and Science under the research grant BES-2004-4435.

References

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. 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)*, pages 152–159.

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

Marine Carpuat and Dekai Wu. 2007. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–72.

N. Cristianini and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press.

C. Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Jesús Giménez and Lluís Màrquez. 2008. Discriminative Phrase Selection for Statistical Machine Translation. In *Learning Machine Translation*, NIPS Workshop. MIT Press.

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

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th LREC*, pages 43–46.

Jesús Giménez and Lluís Màrquez. 2007a. Context-aware Discriminative Phrase Selection for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 159–166.

Jesús Giménez and Lluís Màrquez. 2007b. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 256–264.

Jesús Giménez. 2007. Iqmt v 2.1. technical manual (lsi-07-29-r). Technical report, TALP Research Center. LSI Department. http://www.lsi.upc.edu/ nlp/IQMT/IQMT.v2.1.pdf.

T. Joachims. 1999. Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. The MIT Press.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, and Christine Moran. 2006. Open Source Toolkit for Statistical Machine Translation. Technical report, Johns Hopkins University Summer Workshop. http://www.statmt.org/jhuws/.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. Technical report, (ACL 2007) demonstration session.

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

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram

Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Dennis Mehay and Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Kadri Hacioglu Mona Diab and Daniel Jurafsky. 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.

Franz Josef Och and Hermann Ney. 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)*, pages 295–302.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*.

Nicolas Stroppa, Antal van den Bosch, and Andy Way. 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)*, pages 231–240.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag. ISBN 0-387-98780-0.

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. 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)*.