Discriminative learning within Arabic Statistical Machine Translation

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
{cristinae,jgimenez,lluism}@lsi.upc.edu

Abstract

Written Arabic is a especially ambiguous due to the lack of diacritisation of texts, and this makes the translation harder for automatic systems that do not take into account the context of phrases. Here, we use a standard Phrase-Based Statistical Machine Translation architecture to build an Arabic-to-English translation system, but we extend it by incorporating a local discriminative phrase selection model which addresses this semantic ambiguity. Local classifiers are trained using both 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 improvements in the full translation of Arabic documents at the lexical, syntactic and semantic levels as measured by an heterogeneous set of automatic metrics.

1 Introduction

Nowadays, one of the most common paradigms for Machine Translation (MT) is the statistical approach, above all when one has at his disposal a large amount of parallel texts as it is the case of Arabic and English. From the first works on Statistical Machine Translation (SMT) by Brown *et al.* [2], 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* as translation units and consequently the birth of Phrase-Based SMT [9, 6]. 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.

In SMT, the best translation for a given source sentence is the most probable one, and the probability is expressed as the sum of different components. The log-linear model [8], 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 collection of probability scores of word sequences in the target language that take care of the fluency of the output. The latter is the one taking into account the correspondence between the two languages.

Usually, 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).

It seems clear that using linguistic information and the surrounding context of each phrase should help the translation. One can 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 classify the translations using various features that encode the information of the phrase context. Here, 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. That shows an analogy between this idea 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 Refs. [3, 11, 4, 1, 5] and references therein).

Although using discriminative learning methods for SMT can be useful for any language pair, those source languages with especially ambiguous semantics 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 then a perfect language to test the power of the discriminative phrase selection.

In this paper, we use Support Vector Machines (SVMs) to select the adequate translation for every instance of a phrase. These local results are included into a SMT architecture so that the discriminative learning is incorporated in the global Arabic-to-English translation system without modifying the basic structure.

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[1]. That 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 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, genus 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 syntactic phrase "and their knowledge" is written in Arabic as a single word وعلمهم (or *wElmhm* using Buckwalter's transliteration). The word 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: it will improve the alignments and reduce 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

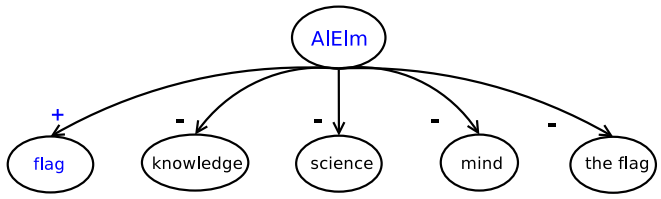
There are several recent methods in the literature to integrate discriminative learning techniques into the translation process. In 2005, Carpuat and Wu [3] used WSD predictions to constrain the possible translations available in decoding time. The same year, Vickrey *et al.* [11] applied discriminative models for word selection but 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 ([4] and references therein).

Carpuat and Wu, the authors of Ref. [4], used a WSD system which combined naïve Bayes, maximum entropy, boosting and kernel PCA-based models. Bangalore *et al.* [1] relied on a maximum entropy model. Here, we use the model of Giménez and Màrquez [5] 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

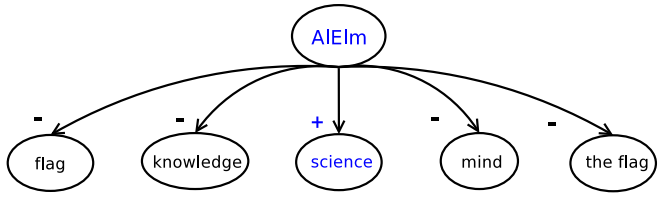
[1]The Buckwalter transliteration can be found at `http://www.qamus.org/transliteration.htm`

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.

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

translation model of the SMT system. The translation table obtained from the alignments is then our 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. This way the multi-class problem is binarised and converted in a one-vs-all decision as it is graphically seen in Figure 1 for two examples of the phrase العلم (*AlElm*).

The linear SVMs are fed with tens of features all of them coming from the source sentence. 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 IOB label resulting of a base phrase chunking for the phrase itself. 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. We show a more detailed example in Section 5.

After training the classifier for every possible translation phrase one obtains a SVM score for each instance of a phrase, and that score is converted into a probability using a softmax

function. The result of this model as to its application to machine translation is then a probability table $P_{\text{DPT}}(e|f)$. 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 less examples we extend the $P_{\text{DPT}}(e|f)$ table 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.

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 as the sum of several terms:

$$\begin{aligned}\log P_{\text{SMT}}(e|f) &= \lambda_{lm}\log P(e) + \lambda_{lg}\log lex(f|e) + \lambda_{ld}\log lex(e|f) \\ &\quad +\lambda_{g}\log P_{\text{MLE}}(f|e) + \lambda_{d}\log P_{\text{MLE}}(e|f) \\ &\quad +\lambda_{di}\log P_{di}(e,f) + \lambda_{ph}\log ph(e) + \lambda_{w}\log w(e)\,, \end{aligned} \tag{1}$$

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:

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

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 $P_{\text{DPT}}(e|f)$.

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. 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, segments 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 the `Moses` decoder[2] and the length ratio limit for obtaining the alignments with `GIZA++`[3]. These segments or lines are the minimum aligned unit in the parallel corpus and correspond to one or more sentences. The filtering selects 123,662 lines, a 99% of the total, which is equivalent

[2]Moses decoder: `http://www.statmt.org/moses/`
[3]Giza++ package: `http://www.fjoch.com/GIZA++.html`

to about $4 \cdot 10^6$ tokens, resulting a medium size corpus under the point of view of collecting alignments.

For the development in the full translation task, we selected 500 lines from the same corpora proportionally to the training set. Results are given for the test set from NIST08 evaluation.

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.

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

The pre-process for Arabic is a bit wider. First of all, it is useful to change the codification of the texts and we romanise the orginal 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. The character for *madda*, | , is a reserved character in the `Moses` decoder that separates the different factors for a word. Therefore, it has been substituted by *L* after the annotation process. 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, ﻷ, ﻹ or ﻵ, 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`[4]. This software uses the `Yamcha SVM tools`[5] 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 the Arabic Penn TreeBank ATB 1 v3.0, ATB 2 v2.0 and ATB 3 v2.0, therefore on a news domain.

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. 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 source sentence in Arabic but not from the target in English. From the phrase we consider word, part-of-speech, coarse part-of-speech and chunk labels n-grams. The same features are extracted from the full sentence with the addition of the bag-of-words which keeps the words at the right and at the left of the phrase.

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 phrases 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

[4]The Arabic processing tools AMIRA can be found at `http://www.cs.columbia.edu/~mdiab/`

[5]Yamcha SVM tools: `http://chasen.org/~taku/software/yamcha/`

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 1: 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.

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. For each of these phrases, we learn a 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 times.

For training the SVMs we use the `SVM`^{light} `package`[6]. The free parameter of SVMs, the trade-off between the training error and the margin, is adjusted in the learning process for each phrase.

Table 1 shows the comparison of the accuracy for the phrase selection task obtained by SVMs and labeled 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 a 7.8% is obtained. A larger gain is get for more frequent phrases, but these are a minority, and the mean in the whole training set reflects an improvement of a 7.5%.

We can take a look at some particular examples. Our running 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

With 114 examples, all translations appear more than a 0.5% of the times and are learned. For one example where *AlElm* is translated as "knowledge" the set of features to be used in the learning process would be that of Table 2. Since this phrase is an only word the phrase features are just unigrams. As for the sentence, we consider up to trigrams of features for tokens ranging from the position of the phrase minus five to the position plus five.

When training the classifiers with the help of the previous features, we obtain, after a 10-fold cross-validation, an accuracy of 71.3%. The most frequent translation does it well the 49.6% of times. That is, one gets a 40% of relative improvement on the selection of the phrase translation.

As another example we comment the learning for the translation of the Arabic phrase وقع (*wqE*), an example that 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:

[6]SVM^{light} package: http://svmlight.joachims.org/

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

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

Phrase features:

PoS	NN
coarse PoS	N
chunk	B-NP

Sentence features:

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

Table 2: Set of features used for the given example to train a classifier for the phrase *AlElm*.

wqE:

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

As before, the accuracy of the most frequent translation (30.6%) is beated 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.

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 task of translation.

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`[7]. As for the translation models, we use the `GIZA++ Toolkit` to obtain the alignments, and the tools available with the `Moses` 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 optimizing translation performance on a development set. For this optimisation we use a minimum error rate training (MERT) [7] where BLEU [10] is the reference score.

6.2 Discriminative Phrase Translation

Finally, we integrate DPT predictions into the SMT system. To do this, we pre-calculate 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 $AlElm_1$:

$$\ldots \quad \texttt{Hyv}_{28} \quad \texttt{tm}_{22} \quad \texttt{AHrAq} \quad \texttt{AlElm}_1 \quad \texttt{AldnmArky} \quad .{}_{1128}$$

For those words without subindex there is not DPT prediction.

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, even when filtered for only the phrases in the test set, the resulting tables become larger than 1GB and do not fit into memory at decoding time. Therefore, we only keep the first 50 translations[8] for every phrase. Translations are sorted by weighting all the scores. Being the scores differents, every system (baseline and DPT) already 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

[7]SRILM Toolkit: `http://www.speech.sri.com/projects/srilm/`

[8]Using more than 20 translations per phrase during decoding was found to provide no improvement when applied to our baseline with respect to the case where only 20 translations are available.

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

using the MLE prediction. 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 to the number of examples of each method with respect to the total.

Table 3 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, BLEU, to evaluate the quality of the translation. Here, we use the IQ$_{\text{MT}}$ package[9], which provides a rich set of more than 500 metrics at different linguistic levels. We have selected a representative set of metrics, based on different similarity criteria:

- Lexical n-gram similarity on word forms (PER, TER, WER, BLEU, General Text Matching -GTM-, ROUGE -RG-, and METEOR -MTR-).
- Shallow-syntactic similarity on part-of-speech tags and base phrase chunks (Shallow Parsing -SP- family).
- Syntactic similarity on dependency and constituent trees (Dependency Parsing -DP- and Constituency Parsing -CP- families).
- Shallow-semantic similarity on semantic roles (Semantic Roles -SR- family).

A deeply detailed description of the metric set may be found in the IQ$_{\text{MT}}$ technical manual.

[9] IQ$_{\text{MT}}$ software: `http://www.lsi.upc.edu/~nlp/IQMT`.

We translate the test set supplied for the *2008 NIST MT Evaluation* and evaluate the translations against four references. The results of our automatic evaluation can be read in Table 4, where we show in boldface numbers 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.

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. We generate 1000 sets by bootstrap resampling of the original test set to check whether these results are statistically significant. With the previous values, the *DPT* system shows to be statistically better than both DPT^+ and *SMT* systems, and DPT^+ statistically better than *SMT*.

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).

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.

6.2.1 Analysis at the sentence level

So far we quantified the improvement of the translations at the system level, 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 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 $100 < \nu < 500$. 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 3 at a sentence level for this small subset and analyse the results.

Although the mean effect is the improvement reflected in Table 3, individual sentences get both benefits and damages from the discriminative phrase translation. Tables 5, 6 and 7 show the translations of three of the sentences: Example A, B and C respectively; there, 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 get more benefits from this improvement.

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 6 shows a case where the effect is a reorder of

Level	Metric	SMT	DPT	DPT$^+$
Lexical	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-cNIST-5	5.5783	5.6684	**5.6703**
	SP-iobNIST-5	5.9931	**6.1318**	6.1172
	SP-lNIST-5	8.8869	**9.0547**	8.9523
	SP-pNIST-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 4: Automatic evaluation of the translated test set supplied for the *2008 NIST MT Evaluation* using lexical, syntactic and semantic metrics.

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 7 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

Source	لكن الجزء الاكبر من حذح الاسلهة **وقع** في يد هماس 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 5: 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 صره صباها ل حية الاذاعة البريطانية (بي بي سي) ان **مصدرا في روسيا** ابلغح بامر 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 6: 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.

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 a 7.5% 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 preditions 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

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

discriminative prediction $P_{DPT}(e|f)$ is preferred by a 73% of the calculated metrics, that is, 27 out of 37. 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 [5], but as expected from the semantic ambiguities of Arabic, the gain is larger for this language. The Arabic phrases are translated locally with 2.5% more accuracy than the Spanish ones, and that 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.

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

[1] S. Bangalore, P. Haffner, and S. Kanthak. 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, 2007.

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

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

[4] M. Carpuat and D. Wu. 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)*, 2007.

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

[6] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1 2003.

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

[8] F. J. Och and H. Ney. 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, 2002.

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

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

[11] D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. 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)*, 2005.