

Jesús Giménez  
Lluís Màrquez

*This chapter 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 suggest using local classifiers which are able to take further advantage of contextual information. Local predictions are softly integrated into a factored phrase-based statistical MT system leading to a significantly improved lexical choice, according to a heterogeneous set of metrics operating at different linguistic levels. However, automatic evaluation has also revealed that improvements in lexical selection do not necessarily imply an improved sentence grammaticality. This fact evinces that the integration of dedicated discriminative phrase translation models into the statistical framework requires further study. Besides, the lack of agreement between metrics based on different similarity assumptions indicates that more attention should be paid to the role of automatic evaluation in the context of MT system development.*

---

## 1.1 Introduction

Traditional Statistical Machine Translation (SMT) architectures, like the one implemented in this work, address the translation task as a search problem (Brown et al., 1990). Given an input string in the source language, the goal is to find the output string in the target language which maximizes the product of a series of probability models over the search space defined by all possible partitions of the source string and all possible reorderings of the translated units. This search process implicitly decomposes the translation problem into two separate but interrelated subproblems:

**Word Selection**, also referred to as *lexical choice*, is the problem of deciding, given a word (or phrase)  $f$  in the source sentence, which word (or phrase)  $e$  in

the target language is the most appropriate translation. This problem is mainly addressed by translation models, which serve as probabilistic bilingual dictionaries, typically accounting for  $P(f|e)$ ,  $P(e|f)$  or  $P(e, f)$ . Translation models provide, for each word (or phrase) in the source vocabulary, a list of translation candidates with associated translation probabilities. During the search there is another component which addresses word selection, the language model. This component helps the decoder to move towards translations which are more appropriate, in terms of grammaticality, in the context of what is known so far about the target sentence being generated.

**Word ordering** refers to the problem of deciding which position must the translation candidate  $e$  occupy in the target sentence. This problem is mainly addressed by the reordering model which allows for certain word movement inside the sentence. Again, the language model helps the decoder, in this case to move towards translations which preserve a better word ordering according to the rules of the target language.

In standard phrase-based SMT systems, like that described by Koehn et al. (2003), the estimation of these models is fairly simple. For instance, 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 in the application of discriminative learning, both for word ordering (Chang and Toutanova, 2007; Cowan et al., 2006) and, specially, for word selection (Bangalore et al., 2007; Carpuat and Wu, 2007a; Giménez and Màrquez, 2007b; Stroppa et al., 2007; Vickrey et al., 2005).

Interest in discriminative word selection has also been motivated by recent results in Word Sense Disambiguation (WSD). The reason is that SMT systems perform an implicit kind of WSD, except that instead of working with word senses, SMT systems operate directly on their potential translations. Indeed, recent semantic evaluation campaigns have treated word selection as a separate task, under the name of *multilingual lexical sample* (Chklovski et al., 2004; Jin et al., 2007). Therefore, the same discriminative approaches which have been successfully applied to WSD, should be also applicable to SMT. In that spirit, instead of relying on MLE for the construction of the translation models, approaches to discriminative word selection suggest building dedicated discriminative translation models which are able to take into account a wider feature context. Lexical selection is, therefore, addressed as a 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.

In this chapter, we extend the work presented in (Giménez and Màrquez, 2007b). First, in Section 1.2, we describe previous and current approaches to dedicated word selection. Then, in Section 1.3 our approach to Discriminative Phrase Translation

(DPT) is fully described. We present experimental results on the application of DPT models to the Spanish-to-English translation of European Parliament Proceedings. In Section 1.4, prior to considering the full translation task, we measure the local accuracy of DPT classifiers at the isolated *phrase translation* task in which the goal is not to translate the whole sentence but only individual phrases without having to integrate their translations in the context of the target sentence. We present a comparative study on the performance of four different classification settings based on two different learning paradigms, namely Support Vector Machines and Maximum Entropy models.

In Section 1.5, we tackle the full translation task. We have built a state-of-the-art factored phrase-based SMT system based on linguistic data views at the level of shallow parsing (Giménez and Màrquez, 2005, 2006). We compare the performance of DPT and MLE-based translation models built on the same parallel corpus and phrase alignments. DPT predictions are integrated into the SMT system in a *soft* manner, by making them available to the decoder as an additional log-linear feature so they can fully interact with other models (e.g., language, distortion, word penalty and additional translation models) during the search. We separately study the effects of using DPT predictions for all phrases as compared to focusing on a small set of very frequent phrases.

This chapter has also served us to study the problem of MT evaluation. We have applied a novel methodology for *heterogeneous* automatic MT evaluation which allows for separately analyzing quality aspects at different linguistic levels, e.g., lexical, syntactic, and semantic (Giménez and Màrquez, 2007a). This methodology also offers a robust mechanism to combine different similarity metrics into a single measure of quality based on *human likeness* (Giménez and Màrquez, 2008). We have complemented automatic evaluation results through error analysis and by conducting a number of manual evaluations. Main conclusions are summarized in Section 1.6.

---

## 1.2 Approaches to Dedicated Word Selection

Brown et al. (1991a,b) were first to suggest using dedicated WSD models in SMT. In a pilot experiment, they integrated a WSD system based on mutual information into their French-to-English word-based SMT system. Results were limited to the case of binary disambiguation, i.e., deciding between only two possible translation candidates, and to a reduced set of very common words. A significantly improved translation quality was reported according to a process of manual evaluation. However, apparently, they abandoned this line of research.

Some years passed until these ideas were recovered by Carpuat and Wu (2005a), who suggested integrating WSD predictions into a phrase-based SMT system. In a first approach, they did so in a *hard* manner, either for decoding, by constraining the set of acceptable word translation candidates, or for post-processing the SMT system output, by directly replacing the translation of each selected word with the

WSD system prediction. However, they did not manage to improve MT quality. They encountered several problems inherent to the SMT architecture. In particular, they described what they called the *language model effect* in SMT: “*The lexical choices are made in a way that heavily prefers phrasal cohesion in the output target sentence, as scored by the language model*”. This problem is a direct consequence of the hard interaction between their WSD and SMT systems. WSD predictions cannot adapt to the surrounding target context. In a later work, Carpuat and Wu (2005b) analyzed the converse question, i.e., they measured the WSD performance of SMT systems. They showed that dedicated WSD models significantly outperform the WSD ability of current state-of-the-art SMT models. Consequently, SMT should benefit from WSD predictions.

Simultaneously, Vickrey et al. (2005) studied the application of *context-aware* discriminative word selection models based on WSD to SMT. Similarly to Brown et al. (1991a), they worked with translation candidates instead of word senses, although their models were based on maximum entropy and dealt with a larger set of source words and higher levels of ambiguity. However, they did not approach the full translation task but limited to the *blank-filling* task, a simplified version of the translation task, in which the target context surrounding the word translation is available. They did not encounter the language model effect because: (i) the target context was fixed a priori, and (ii) they approached the task in a soft way, i.e., allowing WSD-based models to interact with other models during decoding.

Following similar approaches to that of Vickrey et al. (2005), Cabezas and Resnik (2005) and Carpuat et al. (2006) used WSD-based models in the context of the full translation task to aid a phrase-based SMT system. They reported a small improvement in terms of BLEU score, possibly because they did not work with phrases but limited to single words. Besides, they did not allow WSD-based predictions to interact with other translation probabilities. More recently, a number of authors, including us, have extended these works by moving from words to phrases and allowing discriminative models to fully cooperate with other phrase translation models. Moderately improved MT quality results have been obtained (Bangalore et al., 2007; Carpuat and Wu, 2007a,b; Giménez and Márquez, 2007b; Stroppa et al., 2007; Venkatapathy and Bangalore, 2007). All these works were being elaborated at the same time, and were presented in very near dates with very similar conclusions. We further discuss the differences between them in Section 1.6.

In a different approach, Chan et al. (2007) used a WSD system to provide additional features for the hierarchical phrase-based SMT system based on bilingual parsing developed by Chiang (2005, 2007). These features were intended to give a bigger weight to the application of rules that are consistent with WSD predictions. A moderate but significant improvement in terms of BLEU was reported.

As another alternative direction, Specia et al. (2007) has suggested using Inductive Logic Programming techniques to the problem of word selection. They have presented very promising results on a small set of words from different grammatical categories. They have not yet approached the full translation task.

Overall, apart from evincing that this is a very active research topic, some of the works listed in this section show clear evidence that dedicated word selection models might be useful for the purpose of MT.

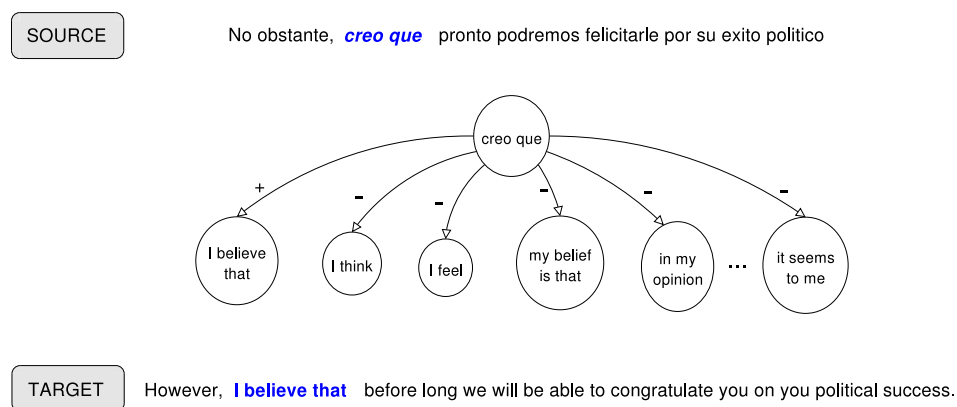
## 1.3 Discriminative Phrase Translation

Instead of relying on MLE estimation to score the phrase pairs  $(f_i, e_j)$  in the translation table, DPT models deal with the translation of every source phrase  $f_i$  as a multiclass classification problem, in which every possible translation of  $f_i$  is a class. As an illustration, in Figure 1.1, we show a real example of Spanish-to-English phrase translation, in which the source phrase “*creo que*”, in this case translated as “*I believe that*”, has several possible candidate translations.

### 1.3.1 Problem Setting

Training examples are extracted from the same training data as in the case of conventional MLE-based models, i.e., a phrase-aligned parallel corpus (see Section 1.5.1). We use each occurrence of each source phrase  $f_i$  to generate a positive training example for the class corresponding to the actual translation  $e_j$  of  $f_i$  in the given sentence, according to the automatic phrase alignment. Let us note that phrase translation is indeed a multilabel problem. Since word alignments allow words both in the source and the target sentence to remain unaligned, the phrase extraction algorithm employed allows each source phrase to be aligned with more than one target phrase, and viceversa, with the particularity that all possible phrase translations are embedded or overlap. However, since the final goal of DPT

**Figure 1.1** An example of phrase translation



classifiers is not to perform local classification but to provide a larger system with translation probabilities, in our current approach no special treatment of multilabel cases has been performed.

### 1.3.2 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 two families, namely Support Vector Machines (SVM)<sup>1</sup> (Vapnik, 1995; Cristianini and Shawe-Taylor, 2000), and Maximum Entropy (ME)<sup>2</sup> (Jaynes, 1957; Berger et al., 1996). We have tried four different learning settings, respectively based on linear binary SVMs (SVMlinear), degree-2 polynomial binary SVMs (SVMpoly2), linear multiclass SVMs (SVMmc), and multiclass ME models (MaxEnt).

#### *Binary vs. Multiclass Classification*

While approaches 3 and 4 implement by definition a multiclass classification scheme, approaches 1 and 2 are based on binary classifiers, and, therefore, the multiclass problem must be binarized. We have applied *one-vs-all* binarization, 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. 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  $f_i$ . At classification time, given a source phrase  $f_i$ , SVMs associated to each possible candidate translation  $e_j$  of  $f_i$  will be applied, and the most confident candidate translation will be selected as the phrase translation.

#### *Support Vector Machines vs. Maximum Entropy*

The SVM and ME algorithms are based on different principles. While the SVM algorithm is a linear separator which relies on margin maximization, i.e. on finding the hyperplane which is more distant to the closest positive and negative examples, ME is a probabilistic method aiming at finding the least biased probability distribution that encodes certain given information by maximizing its entropy. An additional interest of comparing the behavior of SVM and ME classifiers is motivated by the nature of the global MT system architecture. While the outcomes of ME classifiers are probabilities which can be easily integrated into the SMT framework, SVM

---

1. SVMs have been learned using the SVM<sup>light</sup> and SVM<sup>struct</sup> packages by Thorsten Joachims, which are freely available at <http://svmlight.joachims.org> (Joachims, 1999).  
 2. ME models have been learned using the MaxEnt package by Zhang Le, which is freely available at [http://homepages.inf.ed.ac.uk/s0450736/maxent\\_toolkit.html](http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html).

predictions are unbounded real numbers. This issue will be further discussed in Section 1.5.2.

### *Linear vs Polynomial Kernels*

Although SVMs allow for a great variety of kernel functions (e.g., polynomial, gaussian, sigmoid, etc.), in this work, based on results published in recent WSD literature (Lee and Ng, 2002; Màrquez et al., 2006), we have focused on linear and polynomial kernels of degree-2 (see Section 1.4). The main advantage of using linear kernels, over other kernel types, is that this allows for working in the primal formulation of the SVM algorithm and, thus, to take advantage of the extreme sparsity of example feature vectors. This is a key factor, in terms of efficiency, since it permits to considerably speed up both the training and classification processes (Giménez and Màrquez, 2004a). The usage of linear kernels requires, however, the definition of a rich feature set.

#### 1.3.3 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, lemmas, and base phrase chunking IOB labels<sup>3</sup>, in a window of 5 tokens to the left and to the right of the phrase to disambiguate. We also exploit part-of-speech, lemmas 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 using the following tools: SVMTool for PoS tagging (Giménez and Màrquez, 2004b), Freeling for lemmatization (Carreras et al., 2004), and Phreco for base phrase chunking (Carreras et al., 2005). These tools have been trained on the WSJ Penn Treebank (Marcus et al., 1993), for the case of English, and on the 3LB Treebank (Navarro et al., 2003) for Spanish, and, therefore, rely on their tag sets. 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 lemmas. We distinguish between lemmas at the left and right of the source phrase being disambiguated.

---

3. IOB labels are used to denote the position (Inside, Outside, or Beginning of a chunk) and, if applicable, the type of chunk.

**Table 1.1** An example of phrase translation features

<b>Source Sentence</b>	creo <sub>[creer,VMI:B-VP]</sub> que <sub>[que:CS:B-CONJP]</sub> pronto <sub>[pronto,AQ,O]</sub> podremos <sub>[podremos,VMS:B-VP]</sub> felicitarle <sub>[felicitarle,VMN,I-VP]</sub> por <sub>[por,SP,B-PP]</sub> su <sub>[su,DP,B-NP]</sub> éxito <sub>[éxito,NC,I-NP]</sub> político <sub>[politico,AQ,I-NP]</sub> ·[.,FP,O]
<b>Source phrase features</b>	
Lemma <i>n</i> -grams	(creer) <sub>1</sub> , (que) <sub>2</sub> , (creer,que) <sub>1</sub>
PoS <i>n</i> -grams	(VMI) <sub>1</sub> , (CS) <sub>2</sub> , (VMI,CS) <sub>1</sub>
Coarse PoS <i>n</i> -grams	(V) <sub>1</sub> , (C) <sub>2</sub> , (V,C) <sub>1</sub>
Chunk <i>n</i> -grams	(B-VP) <sub>1</sub> , (B-CONJP) <sub>2</sub> , (B-VP,B-CONJP) <sub>1</sub>
<b>Source sentence features</b>	
Word <i>n</i> -grams	(pronto) <sub>1</sub> , (podremos) <sub>2</sub> , (felicitarle) <sub>3</sub> , (por) <sub>4</sub> , (su) <sub>5</sub> , (–,pronto) <sub>–1</sub> , (pronto,podremos) <sub>1</sub> , (podremos,felicitarle) <sub>2</sub> , (felicitarle,por) <sub>3</sub> , (por,su) <sub>4</sub> , (–,–,pronto) <sub>–2</sub> , (–,pronto,podremos) <sub>–1</sub> , (pronto,podremos,felicitarle) <sub>1</sub> , (podremos,felicitarle,por) <sub>2</sub> , (felicitarle,por,su) <sub>3</sub>
Lemma <i>n</i> -grams	(pronto) <sub>1</sub> , (poder) <sub>2</sub> , (felicitar) <sub>3</sub> , (por) <sub>4</sub> , (su) <sub>5</sub> , (–,pronto) <sub>–1</sub> , (pronto,poder) <sub>1</sub> , (poder,felicitar) <sub>2</sub> , (felicitar,por) <sub>3</sub> , (por,su) <sub>4</sub> , (–,–,pronto) <sub>–2</sub> , (–,pronto,poder) <sub>–1</sub> , (pronto,poder,felicitar) <sub>1</sub> , (poder,felicitar,por) <sub>2</sub> , (felicitar,por,su) <sub>3</sub>
PoS <i>n</i> -grams	(AQ) <sub>1</sub> , (VMS) <sub>2</sub> , (VMN) <sub>3</sub> , (SP) <sub>4</sub> , (DP) <sub>5</sub> , (–,AQ) <sub>–1</sub> , (AQ,VMS) <sub>1</sub> , (VMS,VMN) <sub>2</sub> , (VMN,SP) <sub>3</sub> , (SP,DP) <sub>4</sub> , (–,–,AQ) <sub>–2</sub> , (–,AQ,VMS) <sub>–1</sub> , (AQ,VMS,VMN) <sub>1</sub> , (VMS,VMN,SP) <sub>2</sub> , (VMN,SP,DP) <sub>3</sub>
Coarse PoS <i>n</i> -grams	(A) <sub>1</sub> , (V) <sub>2</sub> , (V) <sub>3</sub> , (S) <sub>4</sub> , (D) <sub>5</sub> (–,A) <sub>–1</sub> , (A,V) <sub>1</sub> , (V,V) <sub>2</sub> , (V,S) <sub>3</sub> , (S,D) <sub>4</sub> (–,A,V) <sub>–1</sub> , (–,–,A) <sub>–2</sub> , (A,V,V) <sub>1</sub> , (V,V,S) <sub>2</sub> , (V,S,D) <sub>3</sub>
Chunk <i>n</i> -grams	(O) <sub>1</sub> , (B-VP) <sub>2</sub> , (I-VP) <sub>3</sub> , (B-PP) <sub>4</sub> , (B-NP) <sub>5</sub> , (–,O) <sub>–1</sub> , (O,B-VP) <sub>1</sub> , (B-VP,I-VP) <sub>2</sub> , (I-VP,B-PP) <sub>3</sub> , (B-PP,B-NP) <sub>4</sub> , (–,–,O) <sub>–2</sub> , (–,O,B-VP) <sub>–1</sub> , (O,B-VP,I-VP) <sub>1</sub> , (B-VP,I-VP,B-PP) <sub>2</sub> , (I-VP,B-PP,B-NP) <sub>3</sub>
Bag-of-lemmas	left = ∅ right = { pronto, poder, felicitar, éxito, político }

As an illustration, Table 1.1 shows the feature representation for the example depicted in Figure 1.1. At the top, the reader may find the sentence annotated at the level of shallow syntax (following a ‘word<sub>[lemma:PoS:IOB]</sub>’ format). The corresponding source phrase and source sentence features are shown below. We have not extracted any feature from the target phrase, nor the target sentence, neither the correspondence (i.e., word alignments) between source and target phrases. The reason is that, in this work, the final purpose of DPT models is to aid an existing SMT system to make better lexical choices during decoding, and using these type of features would have forced us to build a more complex decoder.

---

## 1.4 Local Phrase Translation

Analogously to the *word translation* task definition by Vickrey et al. (2005), rather than predicting the sense of a word according to a given sense inventory, in *phrase translation* the goal is to predict the correct translation of a *phrase*, for a given target language, in the context of a sentence. This task is simpler than the full translation task in that phrase translations of different source phrases do not have

**Table 1.2** Numerical description of the set of ‘all’ phrases

#occurrences	#phrases	Phrase		Phrase	
		length	#phrases	entropy	#phrases
(100, 500]	23,578	1	7,004	[0, 1)	6,154
(500, 1000]	3,340	2	12,976	[1, 2)	11,648
(1000, 5000]	2,997	3	7,314	[2, 3)	8,615
(5000, 10000]	417	4	2,556	[3, 4)	3,557
(10000, 100000]	295	5	799	[4, 5)	657
> 100000	22			[5, 6)	18

to interact in the context of the target sentence. However, it provides an insight to the gain prospectives.

#### 1.4.1 Data Sets and Settings

We have used the data from the Openlab 2006 Initiative<sup>4</sup> promoted by the TC-STAR Consortium<sup>5</sup>. This test suite is entirely based on European Parliament Proceedings<sup>6</sup> covering April 1996 to May 2005. We have focused on the Spanish-to-English task. The training set consists of 1,281,427 parallel sentences. After performing phrase extraction over the training data (see details in Section 1.5.1), also discarding source phrases occurring only once (around 90%), translation candidates for 1,729,191 source phrases were obtained. In principle, we could have built classifiers for all these source phrases. However, in many cases learning could be either unfruitful or not necessary at all. For instance, 27% of these phrases are not ambiguous (i.e., have only one associated possible translation), and most phrases count on few training examples. Based on these facts, we decided to build classifiers only for those source phrases with more than one possible translation and 100 or more occurrences. Besides, due to the fact that phrase alignments have been obtained automatically and, therefore, include many errors, source phrases may have a large number of associated phrase translations. Most are wrong and occur 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<sup>7</sup>. The resulting training set consists of 30,649 Spanish source phrases. Table 1.2 presents a brief numerical description of the phrase set. For instance, it can be observed that most phrases are trained on less than 5,000 examples. Most of them are length-2 phrases and most have an entropy lower than 3.

4. <http://tc-star.itc.it/openlab2006/>

5. <http://www.tc-star.org/>

6. <http://www.euoparl.eu.int/>

7. This value was empirically selected so as to maximize the local accuracy of classifiers on a small set of phrases of varying number of examples.

**Table 1.3** Evaluation scheme for the local phrase translation task

#examples	evaluation scheme	
	development and test	test only
<b>2 – 9</b>	leave-one-out	
<b>10..99</b>	10-fold cross validation	
<b>100..499</b>	5-fold cross validation	
<b>500..999</b>	3-fold cross validation	
<b>1000..4999</b>	train(80%)–dev(10%)–test(10%)	train(90%)–test(10%)
<b>5000..9999</b>	train(70%)–dev(15%)–test(15%)	train(80%)–test(20%)
<b>&gt; 10000</b>	train(60%)–dev(20%)–test(20%)	train(75%)–test(25%)

As to feature selection, we discarded features occurring only once in the training data, and constrained the maximum number of dimensions of the feature space to 100,000, by discarding the less frequent features.

### 1.4.2 Evaluation

Local DPT classifiers are evaluated in terms of accuracy against automatic phrase alignments, which are used as gold standard. Let us note that, in the case of multilabel examples, we count the prediction by the classifier as a hit if it matches any of the classes in the solution. Moreover, in order to maintain the evaluation feasible, a heterogeneous evaluation scheme has been applied (see Table 1.3). Basically, when there are few examples available we apply cross-validation, and the more examples available the fewer folds are used. Besides, because cross-validation is costly, when there are more than 1,000 examples available we simply split them into training, development and test sets, keeping most of the examples for training and a similar proportion of examples for development and test. Also, as the number of examples increases, the smaller proportion is used for training and the bigger proportion is held out for development and test. In all cases, we have preserved, when possible, the proportion of samples of each phrase translation so folders do not get biased.

### 1.4.3 Adjustment of Parameters

Supervised learning algorithms are potentially prone to overfit training data. There are, however, several alternatives in order to fight this problem. In the case of the SVM algorithm, the contribution of training errors to the objective function of margin maximization is balanced through the  $C$  regularization parameter of the soft margin approach (Cortes and Vapnik, 1995). In the case of the ME algorithm, the most popular method is based on the use of a gaussian prior on the parameters of the model, whose variance,  $\sigma^2$ , may be balanced (Chen and Rosenfeld, 1999). Learning parameters are typically adjusted so as to maximize the accuracy of local classifiers over held-out data. In our case, a greedy iterative strategy has been

**Table 1.4** Numerical description of the representative set of 1,000 phrases selected

#occurrences	#phrases	Phrase		Phrase	
		length	#phrases	entropy	#phrases
(100, 500]	790	1	213	[1, 2)	467
(500, 1000]	100	2	447	[2, 3)	362
(1000, 5000]	92	3	240	[3, 4)	139
(5000, 10000]	11	4	78	[4, 5)	31
(10000, 50000]	7	5	22	[5, 6)	1

followed. In the first iteration several values are tried. In each following iteration,  $n$  values around the top scoring value of the previous iteration are explored at a resolution of  $\frac{1}{n}$  the resolution of the previous iteration, and so on, until a maximum number of iterations  $I$  is reached<sup>8</sup>.

#### 1.4.4 Comparative Performance

We present a comparative study of the four learning schemes described in Section 1.3.2. The  $C$  and  $\sigma^2$  parameters have been adjusted. However, because parameter optimization is costly, taking into account the large number of classifiers involved, we have focused on a randomly selected set of 1,000 representative source phrases with a number of examples in the [100, 50,000] interval. Phrases with a translation entropy lower than 1 have been also discarded. A brief numerical description of this set is available in Table 1.4.

Table 1.5 shows comparative results, in terms of accuracy. The local accuracy for each source phrase is evaluated according to the number of examples available, as described in Table 1.3. DPT Classifiers are also compared to the *most frequent translation* baseline (MFT), which is equivalent to selecting the translation candidate with highest probability according to MLE. The ‘macro’ column shows macro-averaged results over all phrases, i.e., the accuracy for each phrase counts equally towards the average. The ‘micro’ column shows micro-averaged accuracy, where each test example counts equally<sup>9</sup>. The ‘optimal’ columns correspond to the accuracy computed on optimal parameter values, whereas the ‘default’ columns correspond to the accuracy computed on default  $C$  and  $\sigma^2$  parameter values. In the case of SVMs, we have used the SVM<sup>light</sup> default value for the  $C$  parameter<sup>10</sup>. In the case of ME, we have set  $\sigma^2$  to 1 for all classifiers. The reason is that this

8. In our case,  $n = 2$  and  $I = 3$ . In the case of the  $C$  parameter of SVMs first iteration values are set to  $10^i$  (for  $i \in [-4, +4]$ ), while for the  $\sigma^2$  of ME prior gaussians, values are  $\{0, 1, 2, 3, 4, 5\}$ .

9. The contribution of each phrase to micro-averaged accuracy has been conveniently weighted so as to avoid the extra weight conferred to phrases evaluated via cross-validation.

10. The  $C$  parameter for each binary classifier is set to  $\frac{\sum (\bar{x}_i \bar{x}_i)^{-1}}{N}$ , where  $\bar{x}_i$  is a sample vector and  $N$  corresponds to the number of samples. In the case of multiclass SVMs, the default value is 0.01.

**Table 1.5** Phrase translation accuracy over a selected set of 1,000 phrases based on different learning types vs. the MFT baseline

model	optimal		default	
	macro (%)	micro (%)	macro (%)	micro (%)
<b>MFT</b>	64.72	67.63	64.72	67.63
<b>SVMlinear</b>	70.59	74.12	69.31	73.51
<b>SVMpoly2</b>	<b>71.10</b>	<b>74.70</b>	<b>69.79</b>	<b>73.86</b>
<b>SVMmc</b>	69.93	73.39	57.56	63.15
<b>MaxEnt</b>	71.08	74.34	67.38	70.69

was the most common return value, with a frequency over 50% of the cases, of the parameter tuning process on the selected set of 1,000 phrases.

When the  $C$  and  $\sigma^2$  are properly optimized, all learning schemes, except linear multiclass SVMs, exhibit a similar performance, with a slight advantage in favour of polynomial SVMs. The increase with respect to the MFT baseline is comparable to that described by Vickrey et al. (2005). These results are, taking into account the differences between both tasks, also coherent with results attained in WSD (Agirre et al., 2007). However, when default values are used, ME models suffer a significant decrease, and multiclass SVMs fall even below the MFT baseline. Therefore, in these cases, a different parameter adjustment process for every phrase is required.

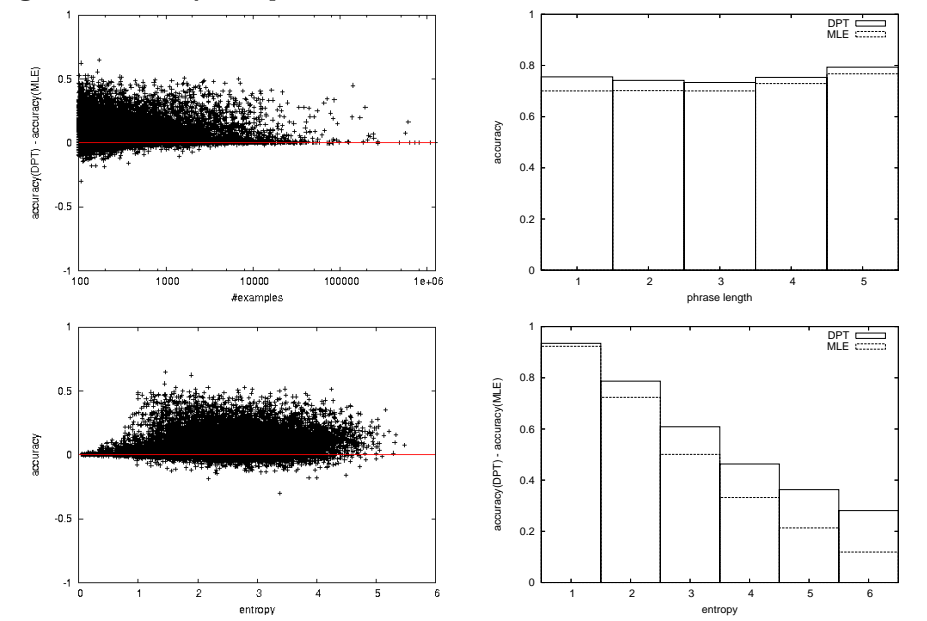
#### 1.4.5 Overall Performance

The aim of this subsection is to analyze which factors have a bigger impact on the performance of DPT classifiers applied to the set of *all* phrases. In this scenario, no matter how greedy the process is, the adjustment of the  $C$  and  $\sigma^2$  becomes impractical. For that reason we have used fixed default values. In the case of SVMs, for the sake of efficiency, we have limited to the use of linear kernels.

Phrase translation results are shown in Table 1.6. Again, phrases are evaluated according to the number of examples available, as described in Table 1.3. We distinguish between the case of using *all* the 30,649 phrases counting on 100 or more examples (columns 1 and 2), and the case of considering only a small subset of 317 very *frequent* phrases occurring more than 10,000 times (columns 3 and 4). The first observation is that both DPT learning schemes outperform the MFT baseline when default learning parameters are used, being linear SVMs clearly the best. The second observation is that the difference, in terms of micro-averaged accuracy gain with respect to the MFT baseline, between using all phrases and focusing on a set of very frequent ones is very small. The reason is that the set of frequent phrases dominates indeed the evaluation with 51.65% of the total number of test cases. In contrast, macro-averaged results confer a significantly wider advantage to DPT models applied to the set of frequent phrases, specially in the case of linear SVMs. This result is significant taking account the high results of the MFT baseline

**Table 1.6** Overall phrase translation accuracy

model	all		frequent	
	macro (%)	micro (%)	macro (%)	micro (%)
MFT	70.51	80.49	79.77	86.12
SVMlinear	74.52	<b>85.48</b>	86.32	<b>91.33</b>
MaxEnt	72.73	82.53	82.31	87.94

**Figure 1.2** Analysis of phrase translation results

on this set. A third, marginal, observation is that frequent phrases are easier to disambiguate, presumably because of their lower entropy (see MFT performance).

In Figure 1.2 we analyze several factors which have a direct influence on the behavior of DPT classifiers. All plots correspond to the case of linear SVMs. For instance, the top-left plot shows the relationship between the local accuracy gain and the number of training examples, for all source phrases. As expected, DPT classifiers trained on fewer examples exhibit the most unstable behavior, yielding a maximum accuracy gain of 0.65 and a maximum decrease of 0.30. However, in general, with a sufficient number of examples (over 10,000), DPT classifiers outperform the MFT baseline. It can also be observed that for most of the phrases trained on more than around 200,000 examples the accuracy gain is very low. The reason, however, is in the fact that these are phrases with very low translation entropy, mostly stop words, such as punctuation marks (“.”, “,”), determiners (“el”, “la”, “los”, “las”, “un”, “una”), or conjunctions and prepositions (“y”, “de”, “en”, “a”). There is a very interesting positive case, that of phrase “que”, which acts

mostly as a conjunction or relative pronoun, and that most often gets translated into “*that*” or “*which*”. This phrase, which appears more than 600,000 times in the data with a translation entropy of 1.68, attains an accuracy gain of 0.16.

The top-right plot shows the relationship between micro-averaged accuracy and source phrase length. There is improvement across all phrase lengths, but, in general, the shorter the phrase the larger the improvement. This plot also indicates that phrases up to length-3 are on average harder to disambiguate than longer phrases. Thus, there seems to be a trade-off between phrase length, level of ambiguity (i.e., translation entropy), and number of examples. Shorter phrases are harder because they exhibit higher ambiguity. DPT is a better model for these phrases because it is able to properly take advantage of the large number of training examples. Longer phrases are easier to model because they present a lower ambiguity. Middle length phrases are hardest because they present a high ambiguity and not many examples.

We further investigate this issue in the two bottom plots. The bottom-left plot shows the relationship between the local accuracy gain and translation entropy, for all source phrases. It can be observed that for phrases with entropy lower than 1 the gain is very small, while for higher entropy levels the behavior varies. In order to clarify this scenario, we analyze the relationship between micro-averaged accuracy and phrase translation entropy at different intervals (bottom-right plot). As expected, the lower the entropy the higher the accuracy. Interestingly, it can also be observed that as the entropy increases the accuracy gain in favour of DPT models increases as well.

---

## 1.5 Exploiting Local DPT Models for the Global Task

In this section, we analyze the impact of DPT models when the goal is to translate the whole sentence. First, we describe our phrase-based SMT baseline system and how DPT models are integrated into the system. Then, some aspects of evaluation are discussed, with special focus on the adjustment of the parameters governing the search process. Finally, MT results are evaluated and analyzed, and several concrete cases are commented.

### 1.5.1 Baseline System

Our system follows a standard phrase-based SMT architecture. This involves three main components:

- **Translation Model.** For translation modeling, we follow the approach by Koehn et al. (2003), in which phrase pairs are automatically induced from word alignments. In our case, however, we have built richer word alignments by working with *linguistic data views* up to the level of shallow syntax, as described in (Giménez and Màrquez, 2005, 2006). This can be seen as a particular case of the recently emerged factored

MT models (Koehn and Hoang, 2007). Phrase alignments are extracted from a word-aligned parallel corpus linguistically enriched with part-of-speech information, lemmas, and base phrase chunk labels. Text has been automatically annotated using the tools described in Section 1.3.3. We have used the *GIZA++ SMT Toolkit*<sup>11</sup> to generate word, PoS, lemma, and chunk label alignments (Och and Ney, 2003). We have followed the *global phrase extraction* strategy described in (Giménez and Márquez, 2005), i.e., a single translation table is built on the union of alignments corresponding to different linguistic data views. Phrase extraction is performed following the *phrase-extract* algorithm described by Och (2002). This algorithm takes as input a word aligned parallel corpus and returns, for each sentence, a set of phrase pairs that are *consistent* with word alignments. A phrase pair is said to be consistent with the word alignment if all the words within the source phrase are only aligned to words within the target phrase, and viceversa. We have worked with the union of source-to-target and target-to-source alignments, with no heuristic refinement. Only phrases up to length five are considered. Also, phrase pairs appearing only once are discarded, and phrase pairs in which the source/target phrase is more than three times longer than the target/source phrase are ignored. Phrase pairs are scored on the basis of relative frequency (i.e., Maximum Likelihood Estimates, MLE).

■ **Language Model.** We use the *SRI Language Modeling Toolkit* (Stolcke, 2002). Language models are based on word trigrams. Linear interpolation and Kneser-Ney discounting have been applied for smoothing.

■ **Search Algorithm.** We use the *Pharaoh* stack-based beam search decoder (Koehn, 2004), which naturally fits with the previous tools. Keeping with usual practice, in order to speed up the translation process, we have fixed several of the decoder parameters. In particular, we have limited the number of candidate translations to 30, the maximum beam size (i.e., stack size) to 100, and used a beam threshold of  $10^{-5}$  for pruning the search space. We have also set a distortion limit of 4 positions.

This architecture was extended by Och and Ney (2002) for considering additional *feature functions* further than the language and translation probability models. Formally:

$$\hat{e} = \operatorname{argmax}_e \{\log P(e|f)\} \approx \operatorname{argmax}_e \left\{ \sum_{m=1}^M \lambda_m h_m(e, f) \right\}$$

The integration of DPT predictions into this scheme is straightforward. Models are combined in a log-linear fashion:

---

11. <http://www.fjoch.com/GIZA++.html>

$$\begin{aligned} \log P(e|f) \approx & \lambda_{lm} \log P(e) + \lambda_g \log P_{\text{MLE}}(f|e) + \lambda_d \log P_{\text{MLE}}(e|f) \\ & + \lambda_{\text{DPT}} \log P_{\text{DPT}}(e|f) + \lambda_d \log P_d(e, f) + \lambda_w \log w(e) \end{aligned}$$

$P(e)$  stands for the language model probability.  $P_{\text{MLE}}(f|e)$  corresponds to the MLE-based generative translation model, whereas  $P_{\text{MLE}}(e|f)$  corresponds to the analogous discriminative model.  $P_{\text{DPT}}(e|f)$  corresponds to the DPT model which uses DPT predictions in a wider feature context. Finally,  $P_d(e, f)$  and  $w(e)$ , correspond to the distortion and word penalty models<sup>12</sup>. The  $\lambda$  parameters controlling the relative importance of each model during the search must be adjusted. We further discuss this issue in subsection 1.5.5.

### 1.5.2 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  $f_i$ . In the case of ME classifiers, outcomes are directly probabilities. However, in the case of SVMs, outcomes are unbounded real numbers. We transform them into probabilities by applying the *softmax function* described by Bishop (1995):

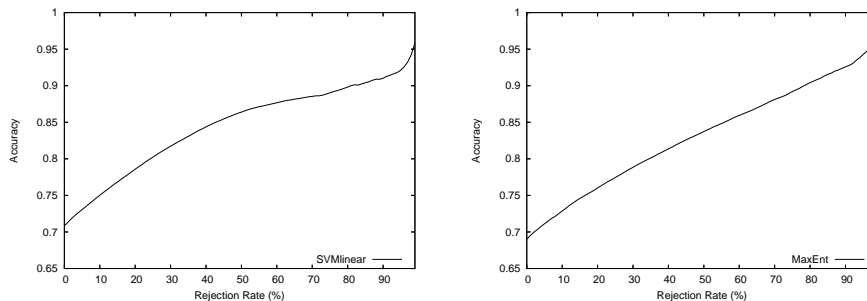
$$P(e_j|f_i) = \frac{e^{\gamma \text{score}_{ij}}}{\sum_{k=1}^K e^{\gamma \text{score}_{ik}}}$$

where  $K$  denotes the number of possible target phrase translations for a given source phrase  $f_i$ , and  $\text{score}_{ij}$  denotes the outcome for target phrase  $e_j$  according to the SVM classifier trained for  $f_i$ . In order to verify the suitability of this procedure, we computed rejection curves for the estimated output probabilities with respect to classification accuracy. For that purpose, we have used the representative set of 1,000 phrases from subsection 1.4.4. This set offers almost 300,000 predictions. In order to calculate rejection curves, the probability estimates for these predictions are sorted in decreasing order. At a certain level of rejection (n%), the curve plots the classifier accuracy when the lowest scoring n% subset is rejected. We have collected values for 100 rejection levels at a resolution of 1%. We tested different values for the  $\gamma$  parameter of the softmax function. The selected final value is  $\gamma = 1$ . In Figure 1.3 (left) we plot the rejection curve for linear SVMs. For the sake of comparison, the rejection curve for ME classifiers is also provided (right plot). It can be observed that both rejection curves are increasing and smooth, indicating a good correlation between probability estimates and classification accuracy<sup>13</sup>.

---

12. We have used default *Pharaoh's* word penalty and distortion models.

13. Other transformation techniques can be found in recent literature. For instance, Platt (2000) suggested using a sigmoid function.



**Figure 1.3** Rejection curves. Linear SVMs + softmax (left) vs. ME (right)

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 pre-computed 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<sup>14</sup>, 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. Let us note that, as described in Section 1.4.1, for each source phrase, not all associated target translations 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 into the DPT model by applying linear discounting.

As an illustration, Table 1.7 shows a fragment of the translation table corresponding to the phrase “*creo que*” in the running example. Notice how this concrete instance has been properly identified by indexing the words inside the phrase (“*creo que*”  $\rightarrow$  “*creo<sub>14</sub> que<sub>441</sub>*”). We show MLE-based and DPT predictions (columns 3 and 4, respectively) for several phrase candidate translations sorted in decreasing MLE probability order. The first observation is that both methods agree on the top-scoring candidate translation, “I believe that”. 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 70% 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.

14. 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 occurrence number. For instance, the source sentence in the example depicted in Figure 1.1 is transformed into “*creo<sub>14</sub> que<sub>441</sub> pronto<sub>0</sub> podremos<sub>0</sub> felicitarle<sub>0</sub> por<sub>109</sub> su<sub>0</sub> éxito<sub>3</sub> político<sub>4</sub> .366*”.

**Table 1.7** An example of translation table

$f_i$	$e_j$	$P_{MLE}(e f)$	$P_{DPT}(e f)$
...			
creo <sub>14</sub> que <sub>441</sub>	i believe that	0.3624	0.2405
creo <sub>14</sub> que <sub>441</sub>	i think that	0.1975	0.0506
creo <sub>14</sub> que <sub>441</sub>	i think	0.1540	0.0475
creo <sub>14</sub> que <sub>441</sub>	i feel that	0.0336	0.0511
creo <sub>14</sub> que <sub>441</sub>	i think it	0.0287	0.0584
creo <sub>14</sub> que <sub>441</sub>	i believe that it	0.0191	0.0487
creo <sub>14</sub> que <sub>441</sub>	i think that it	0.0114	0.0498
creo <sub>14</sub> que <sub>441</sub>	believe that	0.0108	0.0438
creo <sub>14</sub> que <sub>441</sub>	i believe that this	0.0077	0.0482
creo <sub>14</sub> que <sub>441</sub>	i believe it	0.0060	0.0439
...			

Using this technique for integrating DPT predictions into the system we have avoided having to implement a new decoder. However, because translation tables may become very large, it involves a severe extra cost in terms of memory and disk consumption. Besides, it imposes a serious limitation on the kind of features the DPT system may use. In particular, features from the target sentence under construction and from the correspondence between source and target (i.e., alignments) can not be used.

### 1.5.3 Settings

We use the data sets described in Section 1.4.1. Besides, for evaluation purposes we count on a separate set of 1,008 sentences. Three human references per sentence are available. We have randomly split this set in two halves, which are respectively used for development and test.

### 1.5.4 Evaluation

Evaluating the effects of using DPT predictions in the full translation task presents two serious difficulties. In first place, the actual room for improvement caused by a better translation modeling is smaller than estimated in Section 1.4. This is mainly due to the SMT architecture itself which relies on a search over a probability space in which several models cooperate. For instance, in many cases errors caused by a poor translation modeling may be corrected by the language model. In a recent study over the same data set (Spanish-to-English translation of the Openlab 2006 corpus), Vilar et al. (2006) found that only around 28% of the errors committed by their SMT system were related to word selection. In half of these cases errors are caused by a wrong word sense disambiguation, and in the other half the word sense is correct but the lexical choice is wrong. In second place, most conventional

automatic evaluation metrics have not been designed for this purpose and may, therefore, not be able to reflect possible improvements attained due to a better word selection. For instance,  $n$ -gram based metrics such as BLEU (Papineni et al., 2001) tend to favour longer string matchings, and are, thus, biased towards word ordering. In order to cope with evaluation difficulties we have applied several complementary actions, which are described below.

### *Heterogeneous 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, 2004b) 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, 2007a) and semantics (Giménez, 2007). For the purpose of performing heterogeneous automatic MT evaluations, we use the IQ<sub>MT</sub> package (Giménez and Amigó, 2006), which provides a rich set of more than 500 metrics at different linguistic levels<sup>15</sup>. For our experiments, we have selected a representative set of around 50 metrics, based on different similarity criteria:

#### ■ Lexical Similarity

- **BLEU- $n$  | BLEU <sub>$i$</sub> - $n$** : Accumulated and individual BLEU scores for several  $n$ -gram levels ( $n = 1..4$ ) (Papineni et al., 2001).
- **NIST- $n$  | NIST <sub>$i$</sub> - $n$** : Accumulated and individual NIST scores for several  $n$ -gram levels ( $n = 1..5$ ) (Doddington, 2002).
- **GTM- $e$** : General Text Matching F-measure, for several values of the  $e$  parameter controlling the reward for longer matchings ( $e = 1..3$ ) (Melamed et al., 2003).
- **METEOR**: F-measure based on unigram alignment (Banerjee and Lavie, 2005):
  - \* **METEOR<sub>exact</sub>**: only ‘exact’ module.
  - \* **METEOR<sub>porter</sub>**: ‘exact’ and ‘porter\_stem’.
  - \* **METEOR<sub>wnstm</sub>**: ‘exact’, ‘porter\_stem’ and ‘wn\_stem’.
  - \* **METEOR<sub>wnsyn</sub>**: ‘exact’, ‘porter\_stem’, ‘wn\_stem’ and ‘wn\_synonymy’.
- **ROUGE**: Recall oriented measure (Lin and Och, 2004b):
  - \* **ROUGE <sub>$n$</sub>** : for several  $n$ -grams ( $n = 1..4$ )
  - \* **ROUGE<sub>L</sub>**: longest common subsequence.

---

15. The IQ<sub>MT</sub> software is available at <http://www.lsi.upc.edu/~nlp/IQMT>.

- \* **ROUGE<sub>w,1.2</sub>**: weighted longest common subsequence ( $w = 1.2$ ).
- \* **ROUGE<sub>S\*</sub>**: skip bigrams with no max-gap-length.
- \* **ROUGE<sub>SU\*</sub>**: skip bigrams with no max-gap-length, including unigrams.
- **WER**: Word error rate (Nießen et al., 2000). We use 1-WER.
- **PER**: Position-independent word error rate (Tillmann et al., 1997). We use 1-PER.
- **TER**: Translation edit rate (Snover et al., 2006). We use 1-TER.
- **Shallow Syntactic Similarity (SP)**
  - **SP- $O_p$ -\*** Average lexical overlapping over parts-of-speech.
  - **SP- $O_c$ -\*** Average lexical overlapping over base phrase chunk types.

At a more abstract level, we use the NIST metric to compute accumulated/individual scores over sequences of:

- **SP-NIST<sub>l</sub>** Lemmas.
  - **SP-NIST<sub>p</sub>** Parts-of-speech.
  - **SP-NIST<sub>c</sub>** Base phrase chunks.
  - **SP-NIST<sub>ioB</sub>** Chunk IOB labels.
  - **Syntactic Similarity**
    - **On Dependency Parsing (DP)**
      - \* **DP-HWC** These metrics correspond to variants of the head-word chain matching (HWCM) metric presented by Liu and Gildea (2005) slightly modified so as to consider different head-word chain types:
        - **DP-HWC<sub>w</sub>** words.
        - **DP-HWC<sub>c</sub>** grammatical categories.
        - **DP-HWC<sub>r</sub>** grammatical relations.
- In all cases only chains up to length 4 are considered.
- \* **DP- $O_l|O_c|O_r$**  These metrics correspond exactly to the LEVEL, GRAM and TREE metrics introduced by Amigó et al. (2006):
    - **DP- $O_l$ -\*** Average overlapping among words according to the level of the dependency tree they hang at.
    - **DP- $O_c$ -\*** Average overlapping among words *directly hanging* from terminal nodes (i.e. grammatical categories) of the same type.
    - **DP- $O_r$ -\*** Average overlapping among words ruled by non-terminal nodes (i.e. grammatical relationships) of the same type.
  - **On Constituency Parsing (CP)**
    - \* **CP-STM** These metric correspond to a variant of the syntactic tree matching (STM) metric presented by Liu and Gildea (2005), which considers subtrees up to length 9.

- \* **CP- $O_p$ -★** Similarly to ‘*SP- $O_p$ -★*’, this metric computes average lexical overlapping over parts-of-speech, which are now consistent with the full parsing.
- \* **CP- $O_c$ -★** Analogously, this metric computes average lexical overlapping over base phrase chunk types.
- **Shallow-Semantic Similarity**
  - **On Named Entities (NE)**
    - \* **NE- $O_e$ -★** Average lexical overlapping among NEs of the same type. This metric includes the NE type ‘O’ (i.e., Not-a-NE). We introduce another variant, ‘**NE- $O_e$ -★★**’, which considers only actual NEs.
    - \* **NE- $M_e$ -★** Average lexical matching among NEs of the same type.
  - **On Semantic Roles (SR)**
    - \* **SR- $O_r$ -★** Average lexical overlapping among SRs of the same type.
    - \* **SR- $M_r$ -★** Average lexical matching among SRs of the same type.
    - \* **SR- $O_r$**  This metric reflects ‘role overlapping’, i.e., overlapping among semantic roles independently from their lexical realization.
- **Semantic Similarity**
  - **On Discourse Representations (DR)**
    - \* **DR-STM** This metric is similar to the ‘*CP-STM*’ variant referred above, in this case applied to discourse representation structures, i.e., discourse referents and discourse conditions (Kamp, 1981), instead of constituent trees.
    - \* **DR- $O_r$ -★** Average lexical overlapping among discourse representation structures of the same type.
    - \* **DR- $O_{rp}$ -★** Average morphosyntactic overlapping (i.e., among grammatical categories –parts-of-speech– associated to lexical items) among discourse representation structures of the same type.

A detailed description of these metrics can be found in the IQ<sub>MT</sub> technical manual (Giménez, 2007). Let us only explicitly note that most of these metrics rely on automatic linguistic processors, which are not equally available for all languages, and which exhibit different levels of performance (i.e., effectiveness and efficiency). This implies several important limitations on their applicability.

### *MT Evaluation based on Human Likeness*

Heterogeneous MT evaluations might be very informative. However, a new question arises. Since metrics are based on different similarity criteria, and, therefore, biased towards different aspects of quality, scores conferred by different metrics may be controversial. Thus, as system developers we require an additional tool, a meta-evaluation criterion, which allows us to select the most appropriate metric or set of

metrics for the task at hand. Most often, metrics are evaluated on the basis of *human acceptability*, i.e., according to their ability to capture the degree of acceptability to humans of automatic translations, usually measured in terms of correlation with human assessments. However, because human assessments are expensive to acquire, a prominent alternative meta-evaluation criterion, referred to as *human likeness*, has been recently suggested. Metrics are evaluated according to their ability to capture the features that distinguish human translations from automatic ones. This can be measured in terms of discriminative power (Corston-Oliver et al., 2001; Lin and Och, 2004a; Kulesza and Shieber, 2004; Amigó et al., 2005; Gamon et al., 2005). The underlying assumption is that, given that human translations are gold standard, a *good* metric should never rank automatic translations higher than human translations. Then, when a system receives a high score according to such a metric, we can ensure that the system is able to emulate the behaviour of human translators.

We follow the approach suggested by Amigó et al. (2005) in *QARLA*, a probabilistic framework originally designed for the case of Automatic Summarization, but adapted by Giménez and Amigó (2006) to the MT scenario. Following their methodology, we use *QARLA* in two complementary steps. First, we determine the set of metrics with highest discriminative power by maximizing over the KING measure. Second, we use QUEEN to measure overall MT quality according to the optimal metric set<sup>16</sup>. Given a set of test cases  $A$ , a set of similarity metrics  $X$ , and sets of human references  $R$ :

■ **QUEEN** $_{X,R}(A)$  operates under the *unanimity* principle, i.e., the assumption that a ‘good’ translation must be similar to all human references according to all metrics. QUEEN is defined as the probability, over  $R \times R \times R$ , that, for every metric in  $X$ , the automatic translation  $a$  is more similar to a human reference  $r$  than two other references,  $r'$  and  $r''$ , to each other. Formally:

$$\text{QUEEN}_{X,R}(a) = \text{Prob}(\forall x \in X : x(a, r) \geq x(r', r''))$$

where  $x(a, r)$  stands for the similarity between  $a \in A$  and  $r \in R$  according to the metric  $x \in X$ . Thus, QUEEN is able to capture the features that are common to *all* human references, and accordingly reward those automatic translations which share them, and penalize those which do not. Besides, QUEEN exhibits several properties which make it really practical for the purpose of our task. First, since QUEEN focus on unanimously supported quality distinctions, it is is a measure of high precision. Second, QUEEN provides a robust means of combining several metrics into a single measure of quality; it is robust against metric redundancy, i.e., metrics devoted to very similar quality aspects, and with respect to metric scale properties.

---

16. The KING and QUEEN measures are available inside IQ<sub>MT</sub>.

▪  $\mathbf{KING}_{A,R}(X)$  represents the probability that, for a given set of human references  $R$ , and a set of metrics  $X$ , the QUEEN quality of a human reference is not lower than the QUEEN quality of *any* automatic translation in  $A$ . Formally:

$$\mathbf{KING}_{A,R}(X) = \text{Prob}(\forall a \in A : \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\}}(a))$$

Thus, KING accounts for the proportion of cases in which a set of metrics has been able to fully distinguish between automatic and manual translations.

MT evaluation based on human likeness has been successfully applied to the optimization of SMT system parameters (Lambert et al., 2006). Besides, it has been shown to be a robust means of integrating different linguistic quality dimensions together (Giménez and Màrquez, 2008).

### *A Measure of Phrase Translation Accuracy*

For the purpose of evaluating the changes related only to a specific set of phrases (e.g., ‘all’ vs. ‘frequent’ sets), we introduce a new measure,  $\mathbf{A}_{pt}$ , which computes *phrase translation accuracy* for a given list of source phrases. For every test case,  $\mathbf{A}_{pt}$  counts the proportion of phrases from the list appearing in the source sentence which have a valid<sup>17</sup> translation both in the target sentence and in at least one reference translation. Cases in which no valid translation is available in any reference translation are not taken into account. Moreover, in order to avoid using the same target phrase more than once for the same translation case, when a phrase translation is used, source and target phrases are discarded. In fact, because in general source-to-target alignments are either unknown or automatically acquired,  $\mathbf{A}_{pt}$  calculates an approximate solution. Current  $\mathbf{A}_{pt}$  implementation inspects phrases from left to right in decreasing length order.

#### 1.5.5 Adjustment of Parameters

As we have seen in Section 1.4, DPT models provide translation candidates only for specific subsets of phrases. Therefore, in order to translate the whole test set, alternative translation probabilities for all the source phrases in the vocabulary which do not have a DPT prediction must be provided. We have used MLE-based predictions to complete DPT tables. However, interaction between DPT and MLE models is problematic. Problems arise when, for a given source phrase,  $f_i$ , DPT predictions must compete with MLE predictions for larger source phrases  $f_j$  overlapping with or containing  $f_i$  (See Section 1.5.6). We have alleviated these problems by splitting DPT tables in 3 subtables: (1) phrases with DPT prediction, (2) phrases with DPT prediction only for subphrases of it, and (3) phrases with no DPT prediction for any subphrase. Formally:

---

17. Valid translations are provided by the translation table.

$$P_{\text{DPT}}(e|f) = \begin{cases} \lambda_d P_{\text{DPT}}(e|f) & \text{if } \exists P_{\text{DPT}}(e|f) \\ \lambda_o P_{\text{MLE}}(e|f) & \text{if } (\neg \exists P_{\text{DPT}}(e|f)) \wedge (\exists P_{\text{DPT}}(e'|f') \wedge (f' \cap f \neq \emptyset)) \\ \lambda_{\neg} P_{\text{MLE}}(e|f) & \text{otherwise} \end{cases}$$

In order to perform fair comparisons, all  $\lambda$  parameters governing the search must be adjusted. We have simultaneously adjusted these parameters following a greedy iterative strategy similar to that applied for the optimization of the  $C$  and  $\sigma^2$  parameters of local DPT classifiers (see subsection 1.4.3)<sup>18</sup>. The parameter configuration yielding the highest score, according to a given automatic evaluation measure  $x$ , over the translation of the development set will be used to translate the test set. Let us remark that, since metrics are based on different similarity assumptions, optimal parameter configurations may vary very significantly depending on the metric used to guide the optimization process. Most commonly, the BLEU metric, widely accepted as a de facto standard, is selected. However, in this work, we additionally study the system behavior when  $\lambda$  parameters are optimized on the basis of human likeness, i.e, by maximizing translation quality according to the QUEEN measure over the metric combination  $X^+$  of highest discriminative power according to KING. This type of tuning has proved to lead to more robust system configurations than tuning processes based on BLEU alone (Lambert et al., 2006).

For the sake of efficiency, we have limited to the set of lexical metrics provided by IQ<sub>MT</sub>. Metrics at deeper linguistic levels have not been used because their computation is currently too slow to allow for massive evaluation processes as it is the case of parameter adjustment. Moreover, due to the fact that exploring all possible metric combinations was not viable, for the KING optimization we have followed a simple algorithm which performs an *approximate suboptimal* search. The algorithm proceeds as follows. First, individual metrics are ranked according to their KING quality<sup>19</sup>. Then, following that order, metrics are individually added to the set of optimal metrics only if the global KING increases.

The resulting optimal set is:  $X^+ = \{ \text{METEOR}_{w\text{nsyn}}, \text{ROUGE}_{w-1.2} \}$ , which includes variants of METEOR and ROUGE, metrics which, interestingly, share a common ability to capture lexical and morphological variations (use of stemming, and dictionary lookup).

### 1.5.6 Results

We compare the performance of DPT and MLE-based models in the full translation task. Since the adjustment of internal parameters ( $C$  and  $\sigma^2$ ) is impractical, based

---

18. In order to keep the optimization process feasible, in terms of time, the search space is pruned as described in Section 1.5.1.

19. KING has been computed over a representative set of baseline systems based on different non-optimized parameter configurations.

on the results from the Section 1.4, we have limited to test the behavior of binary SVMs. Also, for the sake of efficiency, we have limited to linear kernels.

We have used a system which relies on MLE for the estimation of translation models (*'MLE'*) as a baseline. We separately study the case of (i) using DPT for the set of *'all'* phrases and that of (ii) using DPT predictions for the reduced set of *'frequent'* phrases. This latter set exhibits a higher local accuracy. However, most phrases in this set are single words<sup>20</sup>. Thus, it constitutes an excellent material to analyze the interaction between DPT and MLE-based probabilities in the context of the global task. Besides, this set covers 67% of the words in the test, whereas the *'all'* set covers up to 95% of the words. In both cases, DPT predictions for uncovered words are provided by the MLE model.

Table 1.8 shows automatic evaluation results according to different metrics, including BLEU and QUEEN. For the sake of informativeness, METEOR<sub>*w<sub>n</sub>syn*</sub> and ROUGE<sub>*w<sub>-1.2</sub>*</sub> scores used in QUEEN computations are provided as well. Phrase translation accuracy is evaluated by means of the  $A_{pt}$  measure, both over the set of *'all'* and *'frequent'* phrases. We have separately studied the cases of parameter optimizations based on BLEU (rows 1 to 3) and QUEEN (rows 4 to 6). The first observation is that in the two cases DPT models yield an improved lexical choice according to the respective evaluation metric guiding the adjustment of parameters. However, for the rest of metrics there is not necessarily improvement. Interestingly, in the case of BLEU-based optimizations, DPT predictions as an additional feature report a significant BLEU improvement over the MLE baseline only when all phrases are used (see rows 2 and 3). In contrast, in the case of QUEEN-based optimizations, improvements take place in both cases, although with less significance. It is also interesting to note that the significant increase in phrase translation accuracy ( $A_{pt}$ ) only reports a very modest improvement in the rest of metrics (see rows 5 and 6). This could be actually revealing a problem of interaction between DPT predictions and other models.

### BLEU vs QUEEN

Table 1.8 illustrates the enormous influence of the metric selected to guide the optimization process. A system adjusted so as to maximize the score of a specific metric does not necessarily maximize the scores conferred by other metrics. In that respect, BLEU and QUEEN exhibit completely opposite behaviors. Improvements in BLEU do not necessarily imply improvements in QUEEN, and viceversa. We have further analyzed this controversial relationship by comparing optimal parameter configurations, and observed that  $\lambda$ 's are in a very similar range, except for the weight of the word penalty model ( $\lambda_w$ ), close to 0 in the case of BLEU, whereas in the case of QUEEN, it takes negative values around -1, thus, favouring longer

---

20. The *'frequent'* set consists of 240 length-1 phrases, 64 length-2 phrases, 12 length-3 phrases and 1 length-4 phrase.

**Table 1.8** Automatic evaluation of MT results. DPT predictions as an additional feature

System Config.	QUEEN (lexical)	METEOR (wnsyn)	ROUGE (w_1.2)	$A_{pt}$ (all)	$A_{pt}$ (frq)	BLEU
<b>BLEU-based optimization</b>						
MLE	0.4826	0.7894	0.4385	0.7099	0.7915	0.6331
DPT <sub>all</sub>	0.4717	0.7841	0.4383	0.7055	0.7823	<b>0.6429</b>
DPT <sub>frq</sub>	0.4809	0.7863	0.4386	0.7102	0.7941	0.6338
<b>QUEEN-based optimization</b>						
MLE	0.4872	0.7924	0.4384	0.7158	0.8097	0.6149
DPT <sub>all</sub>	0.4907	<b>0.7949</b>	0.4391	0.7229	0.8115	0.6048
DPT <sub>frq</sub>	<b>0.4913</b>	0.7934	<b>0.4404</b>	<b>0.7245</b>	<b>0.8251</b>	0.6038

translations. This seems to indicate that the heuristically motivated brevity penalty factor of BLEU could be responsible for the ‘BLEU vs QUEEN’ puzzle observed. We have verified this hypothesis by inspecting BLEU values before applying the penalty factor. These are on average 0.02 BLEU points higher (0.605  $\rightarrow$  0.625), which explains part of the puzzle. The other part must be found in the fact that, while BLEU is based on  $n$ -gram precision, QUEEN is a meta-metric which combines different quality aspects, in this case borrowed from ROUGE and METEOR.

### *Beyond Lexical Similarity*

In order to analyze other quality aspects beyond the lexical dimension, in Table 1.9 we provide automatic evaluation results according to several metric representatives from different linguistic levels. Metrics are grouped according to the level at which they operate (i.e, lexical, shallow-syntactic, syntactic, shallow-semantic and semantic). We have also computed two different QUEEN values, namely QUEEN( $X^+$ ) and QUEEN( $X_{LF}^+$ ). The first value corresponds to the application of QUEEN to the optimal metric combination based on lexical features only, whereas the second value corresponds to QUEEN applied to the optimal metric combination considering linguistic features at different levels. In this latter case, the optimal metric combination, obtained following the procedure described in subsection 1.5.5, is:  $X_{LF}^+ = \{ \text{METEOR}_{wnsyn}, \text{SP-NIST}_p, \text{SR-}M_r\text{-}\star \}$ , which includes metrics at the lexical, morphosyntactic, and shallow-semantic levels, respectively based on unigram alignment precision, part-of-speech  $n$ -gram matching, and average lexical matching over semantic roles.

First of all, metrics are evaluated according to their ability to distinguish between manual and automatic translations, as computed by KING. Broadly speaking, KING is a measure of discriminative power. For instance, if a metric obtains a KING of 0.5, it means that in 50% of the test cases, it is able to explain by itself the differences in quality between manual and automatic translations. Thus, KING

**Table 1.9** Automatic evaluation of MT results. Linguistic features and meta-evaluation

Metric	KING	BLEU-based optim.			QUEEN-based optim.		
		MLE	DPT <sub>all</sub>	DPT <sub>frq</sub>	MLE	DPT <sub>all</sub>	DPT <sub>frq</sub>
1-WER	0.1521	0.6798	<b>0.6908</b>	0.6842	0.6652	0.6504	0.6542
1-PER	0.1422	0.7679	<b>0.7764</b>	0.7701	0.7571	0.7397	0.7481
1-TER	0.1508	0.7031	<b>0.7135</b>	0.7062	0.6882	0.6737	0.6777
BLEU	0.1164	0.6331	<b>0.6429</b>	0.6338	0.6149	0.6048	0.6038
NIST	0.1488	11.2205	<b>11.3403</b>	11.2398	10.9525	10.7508	10.8012
GTM.e1	0.1151	0.8971	0.8954	0.8977	0.8988	0.8948	<b>0.9013</b>
GTM.e2	0.1257	0.4364	<b>0.4391</b>	0.4348	0.4321	0.4296	0.4259
ROUGE <sub>L</sub>	0.1270	0.6958	<b>0.6984</b>	0.6962	0.6914	0.6887	0.6907
ROUGE <sub>W</sub>	<b>0.1594</b>	0.4385	0.4383	0.4386	0.4384	<b>0.4391</b>	<b>0.4404</b>
MTR <sub>exact</sub>	0.1601	0.7324	0.7278	0.7306	0.7332	<b>0.7376</b>	<b>0.7381</b>
MTR <sub>wsyn</sub>	<b>0.1786</b>	0.7894	0.7841	0.7863	0.7924	<b>0.7949</b>	<b>0.7934</b>
QUEEN( $X^+$ )	0.1806	0.4826	0.4717	0.4809	0.4872	<b>0.4907</b>	<b>0.4913</b>
SP- $O_p$ -*	0.1217	<b>0.6915</b>	0.6904	<b>0.6914</b>	0.6901	0.6834	0.6868
SP- $O_c$ -*	0.1263	0.6878	<b>0.6895</b>	0.6883	0.6884	0.6855	0.6857
SP-NIST <sub>l</sub>	0.1455	11.3156	<b>11.4405</b>	11.3335	11.0455	10.8469	10.9004
SP-NIST <sub>p</sub>	<b>0.1865</b>	9.9739	<b>10.0359</b>	9.9399	9.7361	9.5766	9.5174
SP-NIST <sub>ioB</sub>	0.1772	7.6315	<b>7.6583</b>	7.6174	7.4876	7.3850	7.3525
SP-NIST <sub>c</sub>	0.1680	6.9357	<b>7.0100</b>	6.9507	6.8205	6.6931	6.7000
DP-HWC <sub>w</sub>	0.1071	0.2755	<b>0.2823</b>	0.2712	0.2778	0.2742	0.2743
DP-HWC <sub>c</sub>	0.1475	<b>0.5048</b>	0.4980	0.4994	<b>0.5051</b>	0.4848	0.5014
DP-HWC <sub>r</sub>	0.1481	<b>0.4491</b>	0.4443	0.4433	<b>0.4492</b>	0.4292	0.4435
DP- $O_l$ -*	0.1382	0.5100	0.5089	0.5073	<b>0.5149</b>	0.5034	0.5063
DP- $O_c$ -*	0.1561	<b>0.6078</b>	0.6032	0.6034	0.6043	0.6053	0.6039
DP- $O_r$ -*	0.1693	<b>0.4672</b>	0.4642	0.4627	0.4653	0.4597	0.4610
CP- $O_p$ -*	0.1217	<b>0.6883</b>	<b>0.6893</b>	<b>0.6886</b>	0.6872	0.6807	0.6839
CP- $O_c$ -*	0.1349	0.6530	<b>0.6559</b>	0.6541	0.6520	0.6444	0.6485
CP-STM	0.1224	<b>0.4723</b>	0.4686	0.4674	<b>0.4712</b>	0.4606	0.4573
NE- $O_e$ -**	0.1131	0.7067	<b>0.7100</b>	0.7073	0.7046	0.6954	0.7020
NE- $O_e$ -*	0.0079	0.2049	0.2047	<b>0.2069</b>	0.2017	0.2004	0.2042
NE- $M_e$ -*	0.0079	0.2001	0.1991	<b>0.2021</b>	0.1964	0.1984	0.1994
SR- $O_r$ -*	0.1138	0.4347	0.4361	<b>0.4400</b>	0.4324	0.4158	0.4234
SR- $M_r$ -*	<b>0.1323</b>	<b>0.2892</b>	0.2861	<b>0.2892</b>	<b>0.2901</b>	<b>0.2895</b>	0.2846
SR- $O_r$	0.1230	0.6359	0.6416	<b>0.6482</b>	0.6378	0.6219	0.6329
DR- $O_r$ -*	0.1369	0.4766	0.4788	0.4758	<b>0.4813</b>	<b>0.4819</b>	0.4781
DR- $O_{rp}$ -*	0.1501	0.5840	0.5847	0.5814	0.5877	<b>0.5923</b>	0.5825
DR-STM	0.0688	0.3526	0.3510	0.3539	<b>0.3633</b>	0.3570	0.3543
QUEEN( $X^+_{LF}$ )	0.2011	0.3018	0.2989	<b>0.3058</b>	0.3008	0.2937	0.3005

serves as an estimate of the impact of specific quality aspects on the overall system performance. In that respect, it can be observed that highest KING values are obtained by metrics based on lexical, shallow-syntactic and syntactic similarities.

As to system evaluation, quality aspects are diverse, and as such, it is not always the case that all aspects improve together. For instance, at the lexical and shallow-syntactic levels, most metrics prefer the ‘DPT<sub>all</sub>’ system optimized over BLEU. Only some ROUGE and METEOR variants prefer the DPT systems optimized over QUEEN. After all, the  $X^+$  set, used in the QUEEN computation, consists of these metrics, so this result was expected. In any case, the fact that all metrics based on lexical similarities consistently prefer DPT over MLE confirms that DPT predictions yield an improved lexical choice.

At the syntactic level, however, most metrics prefer the ‘MLE’ systems. Only the shallowest metrics, e.g., DP-HWC<sub>w</sub> (i.e., lexical head-word matching over dependency trees), CP- $O_p-\star$  and CP- $O_e-\star$  (i.e., lexical overlapping over parts-of-speech and phrase constituents) seem to prefer DPT systems, always optimized over BLEU. This is a very interesting result since it reveals that an improved lexical similarity does not necessarily lead to an improved syntactic structure.

At the shallow-semantic level, while NE metrics are not very informative<sup>21</sup>, SR metrics seem to prefer the ‘DPT<sub>freq</sub>’ system optimized over BLEU, whereas at the properly semantic level, metrics based on discourse representations prefer the ‘DPT<sub>all</sub>’ and ‘MLE’ systems optimized over QUEEN. Therefore, no clear conclusions can be made on which model or optimization strategy leads to a better semantic structure.

Finally, combining metrics from all linguistic levels on the basis of human likeness, i.e., QUEEN( $X_{LF}^+$ ), the best system is ‘DPT<sub>freq</sub>’ optimized over BLEU. This would indicate that focusing on a set of frequent phrases is more productive in terms of overall quality.

Several conclusions must be drawn from these results. First, the fact that an improved lexical and semantic similarity does not necessarily lead to an improved sentence grammaticality might be revealing problems of interaction between DPT predictions and the other models in the SMT system. We have verified this hypothesis through a number of manual evaluations. These have revealed that gains are mainly related to the adequacy dimension, whereas for fluency there is no significant improvement. See, for instance, manual evaluations reported in (Giménez and Màrquez, 2007b). These correspond to the case of pairwise system comparisons. Two different judges evaluated a subset of translation test cases in terms of adequacy and fluency. Results reported an adequacy improvement in 39% of the cases, while in 17% there was a decrement. In the case of fluency, results reported improvement in 30% of the cases, while in 37% there was a decrement.

---

21. Observe the low KING values attained, close to zero, except for the case of the ‘NE- $O_e-\star$ ’ metric, which also considers overlapping among tokens which are not named entities.

**Table 1.10** Case of Analysis #1. DPT models help

<b>Source</b>	yo quisiera que el incumplimiento institucional del consejo <b>fuera sancionado</b>
<b>Ref 1</b>	i would like the council 's institutional infringement <b>to be penalised</b>
<b>Ref 2</b>	i would like the council 's institutional non-fulfilment of its obligations <b>to be sanctioned</b>
<b>Ref 3</b>	i would like to see the institutional non-compliance of the council <b>punished</b>
<b>BLEU-based optimizations</b>	
<b>MLE</b>	i would like to see the failure to comply with institutional <b>outside</b> of the council <b>sanctioned</b>
<b>DPT<sub>all</sub></b>	i would like to see the institutional breach of the council <b>was sanctioned</b>
<b>DPT<sub>freq</sub></b>	i would like to see the institutional breach of the council <b>outside sanctioned</b>
<b>QUEEN-based optimizations</b>	
<b>MLE</b>	i would like to see the failure to comply with the institutional councils <b>outside sanctioned</b>
<b>DPT<sub>all</sub></b>	i would like to see the failure to comply with the institutions of the council <b>were to be sanctioned</b>
<b>DPT<sub>freq</sub></b>	i would like to see the failure to comply with the institutional councils <b>outside sanctioned</b>

Second, the lack of consensus between metrics based on different similarity criteria reinforces the need for evaluation methodologies which allow system developers to take into account a heterogeneous set of quality aspects.

### Error Analysis

Tables 1.10, 1.11 and 1.12 show three sentence fragments illustrating the different behavior of the system configurations evaluated. We start, in Table 1.10, by showing a positive case in which the DPT predictions help the system to find a better translation for *'fuera sancionado'*. Observe how baseline SMT systems, whose translation models are based on MLE, all wrongfully translate *'fuera'* as *'outside'* instead of as an auxiliary verb form (e.g., *'was'* or *'were'*) or past form of the accompanying verb *'sancionado'* (e.g., *'sanctioned'* or *'penalised'*). In contrast, *'DPT<sub>all</sub>'* systems are able to provide more appropriate translations for this phrase, regardless of the metric guiding the parameter optimization process. Observe also, how *'DPT<sub>freq</sub>'* systems, which, unfortunately, do not count on DPT predictions for this not frequent enough phrase, commit all the same mistake than MLE-based systems.

Tables 1.11 and 1.12 present two cases in which the metric guiding the optimizations has a stronger influence. In Table 1.11, all MLE baseline systems wrongfully translate *'cuyo nombre'* into *'whose behalf'*. Only the *'DPT<sub>all</sub>'* system optimized over BLEU is able to find a correct translation (*'whose name'*). In Table 1.12, while MLE-based systems provide all fairly correct translations of *'van a parar a'* into *'go to'*, DPT predictions may cause the system to wrongfully translate *'van a parar a'* into *'are going to stop to'*. Only the *'DPT<sub>freq</sub>'* system optimized over BLEU is able to find a correct translation. The underlying cause behind these two cases is

**Table 1.11** Case of Analysis #2. DPT models may help

<b>Source</b>	aquel diputado <b>cuyo nombre</b> no conozco
<b>Ref 1</b>	the member <b>whose name</b> i do not know
<b>Ref 2</b>	the honourable member , <b>whose name</b> i can not recall
<b>Ref 3</b>	that member <b>whose name</b> i ignore
BLEU-based optimizations	
<b>MLE</b>	that member <b>whose behalf</b> i do not know
<b>DPT<sub>all</sub></b>	that member <b>whose name</b> i do not know
<b>DPT<sub>frq</sub></b>	that member <b>whose behalf</b> i do not know
QUEEN-based optimizations	
<b>MLE</b>	that member <b>on whose behalf</b> i am not familiar with
<b>DPT<sub>all</sub></b>	that member <b>on whose behalf</b> i am not familiar with
<b>DPT<sub>frq</sub></b>	that mep <b>whose behalf</b> i am not familiar with

**Table 1.12** Case of Analysis #3. DPT models may not help

<b>Source</b>	poco más del 40 % de los fondos <b>van a parar a</b> esos países .
<b>Ref 1</b>	only slightly more than 40 % of the money <b>ends up in</b> those countries .
<b>Ref 2</b>	little more than 40 % of these funds <b>end up in</b> these countries .
<b>Ref 3</b>	little more than 40 % of the funds <b>are going to</b> those countries .
BLEU-based optimizations	
<b>MLE</b>	little more than 40 % of the funds <b>go to them</b> .
<b>DPT<sub>all</sub></b>	little more than 40 % of the funds <b>will stop to these countries</b> .
<b>DPT<sub>frq</sub></b>	little more than 40 % of the funds <b>go to these countries</b> .
QUEEN-based optimizations	
<b>MLE</b>	just a little more than 40 % of the money <b>goes to those countries</b> .
<b>DPT<sub>all</sub></b>	little more than 40 % of the funds <b>are going to stop to these countries</b> .
<b>DPT<sub>frq</sub></b>	little more than 40 % of the funds <b>are going to stop to these countries</b> .

that there is no DPT prediction for ‘*cuyo nombre*’ and ‘*van a parar a*’, two phrases of very high cohesion, but only for subphrases of it (e.g., ‘*cuyo*’, ‘*nombre*’, ‘*van*’, ‘*a*’, ‘*parar*’, ‘*van a*’, ‘*a parar*’). DPT predictions for these subphrases must compete with MLE-based predictions for larger phrases, which may cause problems of interaction.

---

## 1.6 Conclusions

In this work, we have shown that discriminative phrase translation may be successfully applied to SMT. Despite the fact that measuring improvements in word selection is a very delicate issue, experimental results, according to several well-known metrics based on lexical similarity, show that dedicated DPT models yield a significantly improved lexical choice over traditional MLE-based ones. However, by evaluating linguistic aspects of quality beyond the lexical level (e.g., syntactic, and semantic), we have found that an improved lexical choice and semantic structure does not necessarily lead to improved grammaticality. This result has been veri-

fied through a number of manual evaluations, which have revealed that gains are mainly related to the adequacy dimension, whereas for fluency there is no significant improvement.

As we have seen in Section 1.2, other authors have recently conducted similar experiments. Although tightly related, there exist several important differences between the works by Carpuat and Wu (2007a), Bangalore et al. (2007), Stroppa et al. (2007), and ours. Some are related to the context of the translation task, i.e., language-pair and task domain. For instance, while we work in the Spanish-to-English translation of European Parliament proceedings, Carpuat and Wu (2007a) and Bangalore et al. (2007) work on the Chinese-to-English translation of basic travel expressions and newswire articles, and Stroppa et al. (2007) work on the Chinese-to-English and Italian-to-English translation of basic travel expressions. Additionally, Bangalore et al. (2007) present results on Arabic-to-English translation of proceedings of the United Nations and on French-to-English translation of proceedings of the Canadian Parliament.

Other differences are related to the disambiguation system itself. While we rely on SVM predictions, Carpuat and Wu (2007a) use an ensemble of four combined models (naïve Bayes, maximum entropy, boosting, and Kernel PCA-based models), Stroppa et al. (2007) rely on memory-based learning, and Bangalore et al. (2007) use maximum entropy. Besides, Bangalore et al. (2007) employ a slightly different SMT architecture based on stochastic finite-state transducers which addresses the translation task as two separate processes: (i) global lexical selection, i.e., dedicated word selection, and (ii) sentence reconstruction. Moreover, their translation models are indeed bilingual language models. They also deal with reordering in a different manner. Prior to translation, the source sentence is reordered so as to approximate the right order of the target language. This allows them to perform a monotonic decoding.

There are also significant differences in the evaluation process. Bangalore et al. (2007) rely on BLEU as the only measure of evaluation, Stroppa et al. (2007) additionally rely on NIST, and Carpuat and Wu (2007a) show results according to eight different standard evaluation metrics based on lexical similarity including BLEU and NIST. In contrast, in this work, we have used a set of evaluation metrics operating at deeper linguistic levels. We have also relied on the QUEEN measure, which allows for non-parametric combinations of different metrics into a single measure of quality.

Besides, this study has also served us to start a discussion on the role of automatic metrics in the development cycle of MT systems and the importance of meta-evaluation. We have shown that basing evaluations and parameter optimizations on different metrics may lead to very different system behaviors. For system comparison, this may be solved by conducting manual evaluations. However, this is impractical for the adjustment of parameters, where hundreds of different configurations are tried. Thus, we argue that more attention should be paid to the meta-evaluation process. In our case, metrics have been evaluated on the basis of human likeness. Other solutions exist. The main point, in our opinion, is that sys-

tem development is *metricwise*. In other words, for the sake of robustness, it is crucial that the metric (or set of metrics) guiding the development process is able to capture the possible quality variations induced by system modifications. This is specially important given the fact that, most often, system improvements focus on partial aspects of quality, such as word selection or word ordering, which can not always be expected to improve together.

Finally, the fact that improvements in adequacy do not lead to an improved fluency evinces that the integration of local DPT probabilities into the statistical framework requires further study. We believe that if DPT models considered features from the target side under generation and from the correspondence between source and target, phrase translation accuracy would improve and cooperation with the decoder would be even softer. Although, still, predictions based on local training may not always be well suited for being integrated in the target translation. Thus, we also argue that if phrase translation classifiers were trained in the context of the global task their integration would be more robust and translation quality could further improve. The possibility of moving towards a new global DPT architecture in the fashion, for instance, of those suggested by Tillmann and Zhang (2006) or Liang et al. (2006) should be considered.

---

## 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. Authors are thankful to the TC-STAR Consortium for providing such very valuable data sets. We are also grateful to David Farwell and Sandra Fontanals who helped us as judges in the numerous processes of manual evaluation. We would also like to thank the anonymous reviewers for their valuable comments and suggestions.

---

## References

- Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic, June 2007.
- Enrique Amigó, Julio Gonzalo, Anselmo Pe nas, and Felisa Verdejo. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, 2005.
- Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 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)*, 2006.
- Satanjeev Banerjee and Alon Lavie. 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*, 2005.
- Srinivas Bangalore, Patrick Haffner, and Stephan 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.
- A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1): 39–72, 1996.
- Christopher M. Bishop. *Neural Networks for Pattern Recognition*, chapter 6.4: Modeling Conditional Distributions. Oxford University Press, 1995.
- Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 76–85, 1990.
- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. Word-sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 264–270, Morristown, NJ, USA, 1991a. Association for Computational Linguistics.

- Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. A statistical approach to sense disambiguation in machine translation. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 146–151, Morristown, NJ, USA, 1991b. Association for Computational Linguistics.
- Clara Cabezas and Philip Resnik. Using WSD Techniques for Lexical Selection in Statistical Machine Translation (CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42). Technical report, University of Maryland, College Park. [http://lampsrv01.umiacs.umd.edu/pubs/TechReports/LAMP\\_124/-LAMP\\_124.pdf](http://lampsrv01.umiacs.umd.edu/pubs/TechReports/LAMP_124/-LAMP_124.pdf), 2005.
- Marine Carpuat and Dekai Wu. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2005a.
- Marine Carpuat and Dekai Wu. Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation. In *Proceedings of IJCNLP*, 2005b.
- Marine Carpuat and Dekai Wu. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–72, 2007a.
- Marine Carpuat and Dekai 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)*, 2007b.
- Marine Carpuat, Yihai Shen, Yu Xiaofeng, and Dekai Wu. Toward Integrating Semantic Processing in Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 37–44, 2006.
- Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC*, pages 239–242, 2004.
- Xavier Carreras, Lluís Màrquez, and Jorge Castro. Filtering-ranking perceptron learning for partial parsing. *Machine Learning*, 59:1–31, 2005.
- Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word Sense Disambiguation Improves Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 33–40, 2007.
- Pi-Chuan Chang and Kristina Toutanova. A Discriminative Syntactic Word Order Model for Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9–16, 2007.
- Stanley F. Chen and Ronald Rosenfeld. A Gaussian Prior for Smoothing Maximum Entropy Models. Technical report, Technical Report CMUCS -99-108, Carnegie

- Mellon University, 1999.
- David Chiang. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, 2005.
- David Chiang. Hierarchical Phrase-based Translation. *Computational Linguistics*, 33(2):201–228, 2007.
- Timothy Chklovski, Rada Mihalcea, Ted Pedersen, and Amruta Purandare. The Senseval-3 Multilingual EnglishHindi lexical sample task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 5–8, Barcelona, Spain, July 2004. Association for Computational Linguistics.
- Simon Corston-Oliver, Michael Gamon, and Chris Brockett. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147, 2001.
- Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
- Brooke Cowan, Ivona Kucerova, and Michael Collins. A Discriminative Model for Tree-to-Tree Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
- George Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145, 2002.
- Christiane Fellbaum, editor. *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.
- Michael Gamon, Anthony Aue, and Martine Smets. Sentence-Level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*, pages 103–111, 2005.
- Jesús Giménez. 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>, 2007.
- Jesús Giménez and Enrique Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*, pages 685–690, 2006.
- Jesús Giménez and Lluís Màrquez. Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. In Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov, editors, *Recent Advances in Natural Language Processing III, Current Issues in Linguistic Theory (CILT)*, pages 153–162, Amsterdam, 2004a. John Benjamin Publishers. ISBN 90-272-4774-9.

- Jesús Giménez and Lluís Màrquez. Combining Linguistic Data Views for Phrase-based SMT. In *Proceedings of the Workshop on Building and Using Parallel Texts, ACL*, 2005.
- Jesús Giménez and Lluís Màrquez. The LDV-COMBO system for SMT. In *Proceedings of the NAACL Workshop on Statistical Machine Translation (WMT'06)*, 2006.
- Jesús Giménez and Lluís Màrquez. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 256–264, 2007a.
- Jesús Giménez and Lluís Màrquez. Context-aware Discriminative Phrase Selection for Statistical Machine Translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 159–166, 2007b.
- Jesús Giménez and Lluís Màrquez. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In *Proceedings of IJCNLP*, 2008.
- Jesús Giménez and Lluís Màrquez. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th LREC*, pages 43–46, 2004b.
- Edwin Thompson Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957.
- Peng Jin, Yunfang Wu, and Shiwen Yu. Semeval-2007 task 05: Multilingual chinese-english lexical sample. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 19–23, Prague, Czech Republic, June 2007. Association for Computational Linguistics.
- T. Joachims. 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, 1999.
- Hans Kamp. A Theory of Truth and Semantic Representation. In J.A.G. Groenendijk, T.M.V. Janssen, , and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematisch Centrum, address = Amsterdam, 1981.
- Philipp Koehn. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA*, 2004.
- Philipp Koehn and Hieu Hoang. Factored Translation Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 868–876, 2007.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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)*, 2003.
- Alex Kulesza and Stuart M. Shieber. A Learning Approach to Improving Sentence-level MT Evaluation. In *Proceedings of the 10th International Conference on*

- Theoretical and Methodological Issues in Machine Translation*, 2004.
- Patrik Lambert, Jesús Giménez, Marta R. Costa-jussà, Enrique Amigó, Rafael E. Banchs, Lluís Màrquez, and J.A. R. Fonollosa. Machine Translation System Development based on Human Likeness. In *Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology*, 2006.
- Yoong Keok Lee and Hwee Tou Ng. An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 41–48, 2002.
- Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 761–768, 2006.
- Chin-Yew Lin and Franz Josef Och. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 2004a.
- Chin-Yew Lin and Franz Josef Och. 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)*, 2004b.
- Ding Liu and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.
- Lluís Màrquez, Gerard Escudero, David Martínez, and German Rigau. Supervised Corpus-Based Methods for WSD. In Phil Edmonds and Eneko Agirre, editor, *Word Sense Disambiguation: Algorithms, Applications, and Trends*, NIPS, chapter 7. Kluwer, 2006.
- Dennis Mehay and Chris Brew. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 2007.
- I. Dan Melamed, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.
- Borja Navarro, Montserrat Civit, M. Antonia Martí, Raquel Marcos, and Belén Fernández. Syntactic, Semantic and Pragmatic Annotation in Cast3LB. In *Proceedings of SProLaC*, pages 59–68, 2003.

- Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.
- Franz Josef Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, RWTH Aachen, Germany, 2002.
- Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.
- Franz Josef Och and Hermann 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.
- Karolina Owczarzak, Josef van Genabith, and Andy Way. 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, 2007.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation, RC22176. Technical report, IBM T.J. Watson Research Center, 2001.
- John C. Platt. Probabilities for SV Machines. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, NIPS, chapter 5, pages 61–74. The MIT Press, 2000.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231, 2006.
- Lucia Specia, Mark Stevenson, and Maria das Graças Volpe Nunes. Learning Expressive Models for Word Sense Disambiguation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 41–48, 2007.
- Andreas Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP*, 2002.
- Nicolas Stroppa, Antal van den Bosch, and Andy Way. 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, 2007.
- C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*, 1997.
- Christoph Tillmann and Tong Zhang. A Discriminative Global Training Algorithm for Statistical MT. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 721–728, 2006.

- Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995. ISBN 0-387-98780-0.
- Sriram Venkatapathy and Srinivas Bangalore. Three models for discriminative machine translation using Global Lexical Selection and Sentence Reconstruction. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 152–159, 2007.
- 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.
- David Vilar, Jia Xu, Luis Fernando D’Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *Proceedings of the 5th LREC*, pages 697–702, 2006.



---

# Index

- Dedicated Word Selection, 3
- Discriminative Learning, 4
- Evaluation, 18
  - Heterogeneous Automatic MT Evaluation, 18
  - Human Acceptability, 20
  - Human Likeness, 20
  - Manual Evaluation, 22
  - The KING measure, 21
  - The  $A_{pt}$  measure, 22
  - The QUEEN measure, 21
- Maximum Entropy, 4
- Statistical Machine Translation, 1, 14
  - Adjustment of Parameters, 23
- Support Vector Machines, 4
  - Outcomes into probabilities, 15