Context-aware Discriminative Phrase Selection for Statistical Machine Translation

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

Abstract

In this work we revise the application of discriminative learning to the problem of phrase selection in Statistical Machine Translation. Inspired by common techniques used in Word Sense Disambiguation, we train classifiers based on local context to predict possible phrase translations. Our work extends that of Vickrey et al. (2005) in two main aspects. First, we move from word translation to phrase translation. Second, we move from the *'blank-filling'* task to the *'full translation'* task. We report results on a set of highly frequent source phrases, obtaining a significant improvement, specially with respect to adequacy, according to a rigorous process of manual evaluation.

1 Introduction

Translations tables in Phrase-based Statistical Machine Translation (SMT) are often built on the basis of Maximum-likelihood Estimation (MLE), being one of the major limitations of this approach that the source sentence context in which phrases occur is completely ignored (Koehn et al., 2003).

In this work, inspired by state-of-the-art Word Sense Disambiguation (WSD) techniques, we suggest using Discriminative Phrase Translation (DPT) models which take into account a wider feature context. Following the approach by Vickrey et al. (2005), we deal with the *'phrase translation'* problem as a classification problem. We use Support Vector Machines (SVMs) to predict phrase translations in the context of the whole source sentence. We extend the work by Vickrey et al. (2005) in two main aspects. First, we move from *'word translation'* to *'phrase translation'*. Second, we move from the *'blank-filling'* task to the *'full translation'* task.

Our approach is fully described in Section 2. We apply it to the Spanish-to-English translation of European Parliament Proceedings. In Section 3, prior to considering the 'full translation' task, we analyze the impact of using DPT models for the isolated 'phrase translation' task. In spite of working on a very specific domain, a large room for improvement, coherent with WSD performance, and results by Vickrey et al. (2005), is predicted. Then, in Section 4, we tackle the full translation task. DPT models are integrated in a 'soft' manner, by making them available to the decoder so they can fully interact with other models. Results using a reduced set of highly frequent source phrases show a significant improvement, according to several automatic evaluation metrics. Interestingly, the BLEU metric (Papineni et al., 2001) is not able to reflect this improvement. Through a rigorous process of manual evaluation we have verified the gain. We have also observed that it is mainly related to adequacy. These results confirm that better phrase translation probabilities may be helpful for the full translation task. However, the fact that no gain in fluency is reported indicates that the integration of these probabilities into the statistical framework requires further study.

2 Discriminative Phrase Translation

In this section we describe the phrase-based SMT baseline system and how DPT models are built and integrated into this system in a 'soft' manner.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 159–166,
Prague, June 2007.

2.1 Baseline System

The baseline system is a phrase-based SMT system (Koehn et al., 2003), built almost entirely using freely available components. We use the *SRI Language Modeling Toolkit* (Stolcke, 2002) for language modeling. We build trigram language models applying linear interpolation and Kneser-Ney discounting for smoothing. Translation models are built on top of word-aligned parallel corpora linguistically annotated at the level of shallow syntax (i.e., lemma, part-of-speech, and base phrase chunks) as described by Giménez and Màrquez (2005). Text is automatically annotated, using the *SVMTool* (Giménez and Màrquez, 2004), *Freeling* (Carreras et al., 2004), and *Phreco* (Carreras et al., 2005) packages. We used the *GIZA++ SMT Toolkit*[1] (Och and Ney, 2003) to generate word alignments. We apply the phrase-extract algorithm, as described by Och (2002), on the Viterbi alignments output by GIZA++ following the *'global phrase extraction'* strategy described by Giménez and Màrquez (2005) (i.e., a single phrase translation table is built on top of the union of alignments corresponding to different linguistic data views). We work with the union of source-to-target and target-to-source alignments, with no heuristic refinement. 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 unsmoothed relative frequency (i.e., MLE). Regarding the argmax search, we used the *Pharaoh* beam search decoder (Koehn, 2004), which naturally fits with the previous tools.

2.2 DPT for SMT

Instead of relying on MLE estimation to score the phrase pairs (f_i, e_j) in the translation table, we suggest considering the translation of every source phrase f_i as a multi-class classification problem, where every possible translation of f_i is a class.

We use *local linear SVMs* [2]. Since SVMs are binary classifiers, the problem must be binarized. We have applied a simple *one-vs-all* binarization, i.e., a SVM is trained for every possible translation candidate e_j. Training examples are extracted from the same training data as in the case of MLE models, i.e., an aligned parallel corpus, obtained as described in Section 2.1. We use each sentence pair in which the source phrase f_i occurs to generate a positive example for the classifier corresponding to the actual translation of f_i in that sentence, according to the automatic alignment. This will be as well a negative example for the classifiers corresponding to the rest of possible translations of f_i.

2.2.1 Feature Set

We consider different kinds of information, always from the source sentence, based on standard WSD methods (Yarowsky et al., 2001). As to the local context, inside the source phrase to disambiguate, and 5 tokens to the left and to the right, we use n-grams ($n \in \{1, 2, 3\}$) of: words, parts-of-speech, lemmas and base phrase chunking IOB labels. As to the global context, we collect topical information by considering the source sentence as a bag of lemmas.

2.2.2 Decoding. A Trick.

At translation time, we consider every instance of f_i as a separate case. In each case, for all possible translations of f_i, we collect the SVM score, according to the SVM classification rule. We are in fact modeling $P(e_j|f_i)$. However, these scores are not probabilities. We transform them into probabilities by applying the *softmax function* described by Bishop (1995). We do not constrain the decoder to use the translation e_j with highest probability. Instead, we make all predictions available and let the decoder choose. We have avoided implementing a new decoder by pre-computing all the SVM predictions for all possible translations for all source phrases appearing in the test set. We input this information onto the decoder by replicating the entries in the translation table. In other words, each distinct occurrence of every single source phrase has a distinct list of phrase translation candidates with their corresponding scores. Accordingly, the source sentence is transformed into a sequence of identifiers,

[1] `http://www.fjoch.com/GIZA++.html`

[2] We use the SVM^{light} package, which is freely available at `http://svmlight.joachims.org` (Joachims, 1999).

in our case a sequence of (w, i) pairs[3], which allow us to uniquely identify every distinct instance of every word in the test set during decoding, and to retrieve DPT predictions in the translation table. For that purpose, source phrases in the translation table must comply with the same format.

This imaginative trick[4] saved us in the short run a gigantic amount of work. However, it imposes a severe limitation on the kind of features which 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.

3 Phrase Translation

Analogously to the *'word translation'* 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, but provides an insight to the gain prospectives.

We used the data from the *Openlab 2006* Initiative[5] promoted by the TC-STAR Consortium[6]. This test suite is entirely based on European Parliament Proceedings. We have focused on the Spanish-to-English task. The training set consists of 1,281,427 parallel sentences. Performing phrase extraction over the training data, as described in Section 2.1, we obtained translation candidates for 1,729,191 source phrases. We built classifiers for *all* the source phrases with more than one possible translation and more than 10 occurrences. 241,234 source phrases fulfilled this requirement. For each source phrase, we used 80% of the instances for training, 10% for development, and 10% for test.

Table 1 shows "phrase translation" results over the test set. We compare the performance, in terms of accuracy, of DPT models and the "most frequent translation" baseline ('MFT'). The MFT baseline 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. The 'all' set includes results for the 241,234 phrases, whereas the 'frequent' set includes results for a selection of 41 very frequent phrases ocurring more than 50,000 times.

phrase set	model	macro	micro
all	MFT	0.66	0.70
	DPT	0.68	0.76
frequent	MFT	0.76	0.75
	DPT	0.86	0.86

Table 1: "Phrase Translation" Accuracy (test set).

A priori, DPT models seem to offer a significant room for potential improvement. Although phrase translation differs from WSD in a number of aspects, the increase with respect to the MFT baseline is comparable. Results are also coherent with those attained by Vickrey et al. (2005).

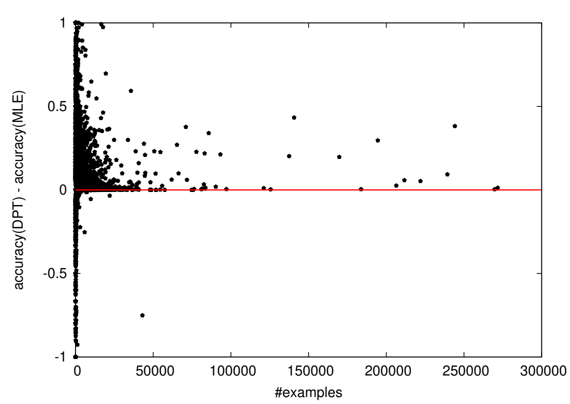

Figure 1: Analysis of "Phrase Translation" Results on the development set (Spanish-to-English).

Figure 1 shows the relationship between the accuracy[7] gain and the number of training examples. In general, with a sufficient number of examples (over 10,000), DPT outperforms the MFT baseline.

[3] w is a word and i corresponds to the number of instances of word w seen in the test set before the current instance.

[4] We have checked that results following this type of decoding when translation tables are estimated on the basis of MLE are identical to regular decoding results.

[5] `http://tc-star.itc.it/openlab2006/`

[6] `http://www.tc-star.org/`

[7] We focus on micro-averaged accuracy.

4 Full Translation

In the "phrase translation" task the predicted phrase does not interact with the rest of the target sentence. In this section we analyze the impact of DPT models when the goal is to translate the whole sentence.

For evaluation purposes we count on a set of 1,008 sentences. Three human references per sentence are available. We randomly split this set in two halves, and use them for development and test, respectively.

4.1 Evaluation

Evaluating the effects of using DPT predictions, directed towards a better word selection, 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 3. 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, Vilar et al. (2006) found that only around 25% of the errors are 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. For instance, metrics such as BLEU (Papineni et al., 2001) tend to favour longer n-gram matchings, and are, thus, biased towards word ordering. We might find better suited metrics, such as METEOR (Banerjee and Lavie, 2005), which is oriented towards word selection[8]. However, a new problem arises. Because different metrics are biased towards different aspects of quality, scores conferred by different metrics are often controversial.

In order to cope with evaluation difficulties we have applied several complementary actions:

1. Based on the results from Section 3, we focus on a reduced set of 41 very promising phrases trained on more than 50,000 examples. This set covers 25.8% of the words in the test set, and exhibits a potential absolute accuracy gain around 11% (See Table 1).
2. With the purpose of evaluating the changes related only to this small set of very promising phrases, we introduce a new measure, $\mathbf{A}_{pt}$, which computes "phrase translation" accuracy for a given list of source phrases. For every test case, A_{pt} counts the proportion of phrases from the list appearing in the source sentence which have a valid[9] translation both in the target sentence and in any of the reference translations. In fact, because in general source-to-target alignments are not known, A_{pt} calculates an approximate[10] solution.
3. We evaluate overall MT quality on the basis of 'Human Likeness'. In particular, we use the QUEEN[11] meta-measure from the QARLA Framework (Amigó et al., 2005). QUEEN operates under the assumption that a good translation must be similar to all human references according to all metrics. Given a set of automatic translations A, a set of similarity metrics X, and a set of human references R, 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 reference r than two other references r' and r'' to each other. Formally:

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

QUEEN captures the features that are common to all human references, rewarding those automatic translations which share them, and penalizing those which do not. Thus, QUEEN provides a robust means of combining several metrics into a single measure of quality. Following the methodology described by Giménez and Amigó (2006), we compute the QUEEN measure over the metric combination with highest KING, i.e., discriminative power. We have considered all the lexical metrics[12] provided by

[8] METEOR works at the unigram level, may consider word stemming and, for the case of English is also able to perform a lookup for synonymy in WordNet (Fellbaum, 1998).

[9] Valid translations are provided by the translation table.

[10] Current A_{pt} implementation searches phrases from left to right in decreasing length order.

[11] QUEEN is available inside the IQMT package for MT Evaluation based on 'Human Likeness' (Giménez and Amigó, 2006). `http://www.lsi.upc.edu/~nlp/IQMT`

[12] Consult the IQMT Technical Manual v1.3 for a detailed description of the metric set. `http://www.lsi.upc.edu/~nlp/IQMT/IQMT.v1.3.pdf`

	QUEEN	A_{pt}	BLEU	METEOR	ROUGE
$P(e) + P_{MLE}(f\|e)$	0.43	0.86	0.59	0.77	0.42
$P(e) + P_{MLE}(e\|f)$	0.45	0.87	0.62	0.77	0.43
$P(e) + P_{DPT}(e\|f)$	**0.47**	**0.89**	0.62	0.78	0.44

Table 2: Automatic evaluation of the 'full translation' results on the test set.

IQMT. The optimal set is:

$\{$ METEOR$_{wnsyn}$, ROUGE$_{w_1.2}$ $\}$

which includes variants of METEOR, and ROUGE (Lin and Och, 2004).

4.2 Adjustment of Parameters

Models are combined in a log-linear fashion:

$$logP(e|f) \propto \lambda_{lm} logP(e) + \lambda_g logP_{MLE}(f|e) + \lambda_d logP_{MLE}(e|f) + \lambda_{DPT} logP_{DPT}(e|f)$$

$P(e)$ is the language model probability. $P_{MLE}(f|e)$ corresponds to the MLE-based generative translation model, whereas $P_{MLE}(e|f)$ corresponds to the analogous discriminative model. $P_{DPT}(e|f)$ corresponds to the DPT model which uses SVM-based predictions in a wider feature context. In order to perform fair comparisons, model weights must be adjusted.

Because we have focused on a reduced set of frequent phrases, in order to translate the whole test set we must provide alternative translation probabilities for all the source phrases in the vocabulary which do not have a DPT prediction. We have used MLE predictions to complete the model. 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 phrases f_j overlapping with or containing f_i (See Section 4.3). 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; and separately adjusting their weights.

Counting on a reliable automatic measure of quality is a crucial issue for system development. Optimal configurations may vary very significantly depending on the metric governing the optimization process. We optimize the system parameters over the QUEEN measure, which has proved to lead to more robust system configurations than BLEU (Lambert et al., 2006). We exhaustively try all possible parameter configurations, at a resolution of 0.1, over the development set and select the best one. In order to keep the optimization process feasible, in terms of time, the search space is pruned[13] during decoding.

4.3 Results

We compare the systems using the generative and discriminative MLE-based translation models to the discriminative translation model which uses DPT predictions for the set of 41 very 'frequent' source phrases. Table 2 shows automatic evaluation results on the test set, according to several metrics. Phrase translation accuracy (over the 'frequent' set of phrases) and MT quality are evaluated by means of the A_{pt} and QUEEN measures, respectively. For the sake of informativeness, BLEU, METEOR$_{wnsyn}$ and ROUGE$_{w_1.2}$ scores are provided as well.

Interestingly, discriminative models outperform the (noisy-channel) default generative model. Improvement in A_{pt} measure also reveals that DPT predictions provide a better translation for the set of 'frequent' phrases than the MLE models. This improvement remains when measuring overall translation quality via QUEEN. If we take into account that DPT predictions are available for only 25% of the words in the test set, we can say that the gain reported by the QUEEN and A_{pt} measures is consistent with the accuracy prospectives predicted in Table 1. METEOR$_{wnsyn}$ and ROUGE$_{w_1.2}$ reflect a slight improvement as well. However, according to BLEU there is no difference between both systems. We suspect that BLEU is unable to accurately reflect the possible gains attained by a better 'phrase selection' over a small set of phrases because of its tendency

[13]For each phrase only the 30 top-scoring translations are used. At all times, only the 100 top-scoring solutions are kept. We also disabled distortion and word penalty models. Therefore, translations are monotonic, and source and target tend to have the same number of words (that is not mandatory).

to reward long n-gram matchings. In order to clarify this scenario a rigorous process of manual evaluation has been conducted. We have selected a subset of sentences based on the following criteria:

- sentence length between 10 and 30 words.
- at least 5 words have a DPT prediction.
- DPT and MLE outputs differ.

A total of 114 sentences fulfill these requirements. In each translation case, assessors must judge whether the output by the discriminative 'MLE' system is better, equal to or worse than the output by the 'DPT' system, with respect to adequacy, fluency, and overall quality. In order to avoid any bias in the evaluation, we have randomized the respective position in the display of the sentences corresponding to each system. Four judges participated in the evaluation. Each judge evaluated only half of the cases. Each case was evaluated by two different judges. Therefore, we count on 228 human assessments.

Table 3 shows the results of the manual system comparison. Statistical significance has been determined using the sign-test (Siegel, 1956). According to human assessors, the 'DPT' system outperforms the 'MLE' system very significantly with respect to adequacy, whereas for fluency there is a slight advantage in favor of the 'MLE' system. Overall, there is a slight but significant advantage in favor of the 'DPT' system. Manual evaluation confirms our suspicion that the BLEU metric is less sensitive than QUEEN to improvements related to adequacy.

Error Analysis

Guided by the QUEEN measure, we carefully inspect particular cases. We start, in Table 4, by showing a positive case. The three phrases highlighted in the source sentence (*'tiene'*, *'señora'* and *'una cuestión'*) find a better translation with the help of the DPT models: *'tiene'* translates into *'has'* instead of *'i give'*, *'señora'* into *'mrs'* instead of *'lady'*, and *'una cuestión'* into *'a point'* instead of *'a ... motion'*.

In contrast, Table 5 shows a negative case. The translation of the Spanish word *'señora'* as *'mrs'* is acceptable. However, it influences very negatively the translation of the following word *'diputada'*, whereas the 'MLE' system translates the phrase *'señora diputada'*, which does not have a DPT prediction, as a whole. Similarly, the translation of *'cuestión'* as *'matter'*, although acceptable, is breaking the phrase *'cuestión de orden'* of high cohesion, which is commonly translated as *'point of order'*. The cause underlying these problems is that DPT predictions are available only for a subset of phrases. Thus, during decoding, for these cases our DPT models may be in disadvantage.

	Adequacy	**Fluency**	**Overall**
MLE > DPT	39	**84**	83
MLE = DPT	100	76	46
MLE < DPT	**89**	68	**99**

Table 3: Manual evaluation of the 'full translation' results on the test set. Counts on the number of translation cases for which the 'MLE' system is better than (>), equal to (=), or worse than (<) the 'DPT' system, with respect to adequacy, fluency, and overall MT quality, are presented.

5 Related Work

Recently, there is a growing interest in the application of WSD technology to MT. For instance, Carpuat and Wu (2005b) suggested integrating WSD predictions into a SMT system in a *'hard'* manner, either for decoding, by constraining the set of acceptable translation candidates for each given source word, or for post-processing the SMT system output, by directly replacing the translation of each selected word with the WSD system prediction. 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 (2005a) analyzed the converse question, i.e. they measured the WSD performance of SMT models. They showed that dedicated WSD models significantly outperform current state-of-the-art SMT models. Consequently, SMT should benefit from WSD predictions.

Simultaneously, Vickrey et al. (2005) studied the

Source	**tiene** la palabra la **señora** mussolini para **una cuestión** de orden .
Ref 1	**mrs** mussolini **has** the floor for **a point** of order .
Ref 2	you have the floor , **missus** mussolini , for **a question** of order .
Ref 3	**ms** mussolini **has** now the floor for **a point** of order .
$P(e) + P_{MLE}(e\|f)$	**i give** the floor to the **lady** mussolini for **a procedural motion** .
$P(e) + P_{DPT}(e\|f)$	**has** the floor the **mrs** mussolini on **a point** of order .

Table 4: Case of Analysis of sentence #422. DPT models help.

Source	**señora** diputada , ésta **no es una cuestión** de orden .
Ref 1	**mrs mussolini** , that is **not a point** of order .
Ref 2	**honourable member** , this is **not a question** of order .
Ref 3	**my honourable friend** , this is **not a point** of order .
$P(e) + P_{MLE}(e\|f)$	**honourable member** , this is **not a point** of order .
$P(e) + P_{DPT}(e\|f)$	**mrs karamanou** , this is **not a matter** of order .

Table 5: Case of Analysis of sentence #434. DPT models fail.

application of discriminative models based on WSD technology 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 they approached the task in a *'soft'* way, i.e., allowing their WSD models to interact with other models during decoding. Similarly, our DPT models are, as described in Section 2.2, *softly* integrated in the decoding step, and thus do not suffer from the detrimental "language model effect" either, in the context of the "full translation" task. Besides, DPT models enforce phrasal cohesion by considering disambiguation at the level of phrases.

6 Conclusions and Further Work

Despite the fact that measuring improvements in word selection is a very delicate issue, we have showed that dedicated discriminative translation models considering a wider feature context provide a useful mechanism in order to improve the quality of current phrase-based SMT systems, specially with regard to adequacy. However, the fact that no gain in fluency is reported indicates that the integration of these probabilities into the statistical framework requires further study.

Moreover, there are several open issues. First, for practical reasons, we have limited to a reduced set of 'frequent' phrases, and we have disabled reordering and word penalty models. We are currently studying the impact of a larger set of phrases, covering over 99% of the words in the test set. Experiments with enabled reordering and word penalty models should be conducted as well. Second, automatic evaluation of the results revealed a low agreement between BLEU and other metrics. For system comparison, we solved this through a process of manual evaluation. However, this is impractical for the adjustment of parameters, where hundreds of different configurations are tried. In this work we have relied on automatic evaluation based on 'Human Likeness' which allows for metric combinations and provides a stable and robust criterion for the metric set selection. Other alternatives could be tried. The crucial issue, in our opinion, is that the metric guiding the optimization is able to capture the changes.

Finally, we argue that, if DPT models considered features from the target side, and from the correspondence between source and target, results could further improve. However, at the short term, the incorporation of these type of features will force us to either build a new decoder or extend an existing one, or to move to a new MT architecture, for instance, in the fashion of the architectures suggested by Tillmann and Zhang (2006) or Liang et al. (2006).

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, projects OpenMT (TIN2006-15307-C03-02) and TRAN-

GRAM (TIN2004-07925-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.

References

Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*.

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

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

Marine Carpuat and Dekai Wu. 2005a. Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation. In *Proceedings of IJCNLP*.

Marine Carpuat and Dekai Wu. 2005b. Word Sense Disambiguation vs. Statistical Machine Translation. In *Proceedings of ACL*.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th LREC*.

Xavier Carreras, Lluís Márquez, and Jorge Castro. 2005. Filtering-ranking perceptron learning for partial parsing. *Machine Learning*, 59:1–31.

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

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

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

Jesús Giménez and Lluís Màrquez. 2005. Combining Linguistic Data Views for Phrase-based SMT. In *Proceedings of the Workshop on Building and Using Parallel Texts, ACL*.

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA*.

Patrik Lambert, Jesús Giménez, Marta R. Costa-jussá, Enrique Amigó, Rafael E. Banchs, Lluís Márquez, and J.A. R. Fonollosa. 2006. Machine Translation System Development based on Human Likeness. In *Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology*.

Percy Liang, Alexandre Bouchard-Côté, Dan Klein, , and Ben Taskar. 2006. An End-to-End Discriminative Approach to Machine Translation. In *Proceedings of COLING-ACL06*.

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

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

Franz Josef Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen, Germany.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, rc22176. Technical report, IBM T.J. Watson Research Center.

Sidney Siegel. 1956. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

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

Christoph Tillmann and Tong Zhang. 2006. A Discriminative Global Training Algorithm for Statistical MT. In *Proceedings of COLING-ACL06*.

D. Vickrey, L. Biewald, M. Teyssier, and D. Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of HLT/EMNLP*.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *Proceedings of the 5th LREC*.

David Yarowsky, Silviu Cucerzan, Radu Florian, Charles Schafer, and Richard Wicentowski. 2001. The Johns Hopkins Senseval2 System Descriptions. In *Proceedings of Senseval-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*.