Unsupervised Syntax-Based Machine Translation: The Contribution of Discontiguous Phrases

Rens Bod

School of Computer Science
University of St Andrews
St Andrews, Scotland
ILLC, University of Amsterdam
Amsterdam, Netherlands
rens@science.uva.nl

Abstract

We present a new unsupervised syntax-based MT system, termed U-DOT, which uses the unsupervised U-DOP model for learning paired trees, and which computes the most probable target sentence from the relative frequencies of paired subtrees. We test U-DOT on the German-English Europarl corpus, showing that it outperforms the state-of-the-art phrase-based Pharaoh system. We demonstrate that the inclusion of *noncontiguous* phrases significantly improves the translation accuracy. This paper presents the first translation results with the data-oriented translation (DOT) model on the Europarl corpus, to the best of our knowledge.

Introduction: Phrase-Based vs Syntax-Based Machine Translation

Phrase-based and syntax-based methods in MT have complementary strengths and shortcomings. While phrase-based methods have been highly successful (Koehn et al. 2003), it has often been noted that such methods are too constrained for translating *discontiguous* constructions like *take* SB *by surpise* (Chiang 2005; Nesson et al. 2006). Shieber (2007) gives evidence that more than half of the entries in the HarperCollins Italian College Dictionary can be subject to 'noncontiguity'. Yet, many syntax-based methods have achieved only small (or no) improvements over purely phrase-based methods. It has been noted that the disappointing contribution of syntactic methods may be due to the traditional notion of syntactic constituent which often harms rather than helps in finding a correct translation (see Chiang 2005). A well-known example is the German-English pair *Es gibt* and *There is*, that are both *non*-constituents. Purely linguistically syntax-based systems therefore often underperform phrase-based methods (e.g. Yamada and Knight 2001). What would be needed is a system that takes into account contiguous as well as discontiguous phrases, *regardless* whether they form linguistically motivated constituents.

In this paper we start an investigation into using a successful unsupervised parsing system, known as U-DOP (Bod 2006, 2007) for providing the tree structures for bilingual corpora such as the Europarl corpus. U-DOP induces a probabilistic tree-substitution grammar (PTSG) from raw data, and has achieved some of the best unsupervised parsing results in the literature (Klein and Manning 2002, 2004; Dennis 2005; Seginer 2007). We will use the structures induced by U-DOP for extending the so-called Data-Oriented Translation (DOT) system (Poutsma 2000; Hearne and Way 2003) towards unsupervised DOT, which we will term *U-DOT*. U-DOT starts by assigning all possible alignments between paired trees bootstrapped by U-DOP and uses the (smoothed) relative frequencies of the subtree pairs to compute the most probable target sentence given a source sentence. This leads to an MT model which takes into account all possible contiguous as well as discontiguous phrases.

Our model is riminiscent of the hierarchical phrase-based model of Chiang (2005) and the synchronic probabilistic tree-insertion grammar model of Nesson et al. (2006), but it differs also from these models in various ways. Firstly, we make use of a structure bootstrapping model, U-DOP, which computes the probability of each tree by summing up the probabilities of its derivations by means of Viterbi *n* best. This probability model, which is also used in computing the best translation, makes the model sensitive to both large and small subtrees. Secondly, our model only uses substitution as a combination operation between subtrees, while Shieber (2007) has shown the importance of including the *insertion* operation. DOP models with the insertion operation have been developed in Hoogweg (2003), and will be extended to MT in future research. Our model is congenial to Galley et al. (2006) who also use subtrees as productive units in a synchronous tree-substitution grammar for MT.

In the following, we will first briefly review the DOT model, and show how it can be generalized to unsupervised MT by extending it with U-DOP. We next discuss the algorithmic background of this new U-DOT model and present experiments involving machine

translation from German to English with the Europarl corpus. We end with a conclusion.

Data-Oriented Translation (DOT)

The Data-Oriented Translation model (DOT) uses the DOP model (Bod et al. 2003) as a basis for statistical MT (Poutsma 2000; Hearne and Way 2003, 2006; Groves et al. 2004). DOT starts with a bilingual treebank where each tree pair constitutes an example translation pair and where translationally equivalent constituents are linked, as e.g. in figure 1 for the English-French *Click Save – Cliquez sur Enregistrer* (taken from Groves et al. 2004).

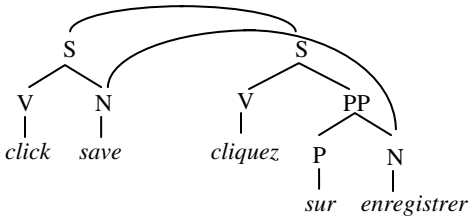


Figure 1. Aligned trees for the English-French pair *Click Save – Cliquez sur Enregistrer* as used in DOT

Like DOP, the DOT model then uses all linked subtree pairs from the bilingual treebank to form a probabilistic tree-substitution grammar (PTSG) where the productive units consist of linked subtrees which are used to compute the most probable translation of a target sentence given a source sentence. Linked subtrees from the tree pair in figure 1 are given in figure 2 (also the entire tree pair constitutes a linked subtree).

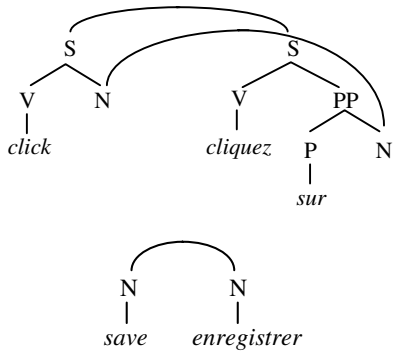


Figure 2. Linked subtrees from the translational tree pair in figure 1

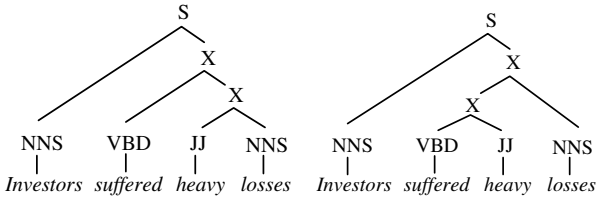
The probability model for this PTSG is similar to the DOP1 model (Bod et al. 2003) and is given in Poutsma (2000), Groves et al. (2004) and others: the probability of a target sentence given a source sentence is computed by summing up the probabilities of all derivations (which is in practice computed by Viterbi *n* best derivations). The probability of a derivation is the product of the probabilities of the subtree pairs involved in it, while the probability of a subtree pair is estimated by its (smoothed) relative frequency in the aligned treebank.

As to date, all experiments with DOT have been carried out on very small, manually annotated treebanks such as the HomeCentre corpus of 810 parsed and aligned sentence pairs (see Hearne and Way 2006). The extension of DOT to larger treebanks will run into formidable annotation tasks. While Groves et al. (2004) show how the alignment task can be partly automated, there is an additional issue in how far – if at all – DOT combines syntactic and phrase-based information. Since DOT is based on linguistically motivated syntax, the model equals the notion of phrase with the notion of syntactic constituent. However, it is well known by now (e.g. Koehn et al. 2003; Chiang 2005) that such an approach has difficulties in capturing phrase-pairs that go beyond constituents, such as the German-English pair *Ich möchte… - I would like to…* which are both non-constituents. On the other hand, purely phrase-based methods have difficulties in capturing *discontiguous* translation pairs such as the English-Italian *the nearest airport to Trento - l' aeroporto più vicino a Trento*, which reflects the dicontiguous translation pair *nearest* NP1 *to* NP2 - NP1 *più vicino a* NP2.

What would be needed is a DOT model which is not based on pre-annotated trees but a DOT model which allows for any substring, be it contiguous or discontiguous, to form a potential 'constituent'. This can be accomplished by using the unsupervised U-DOP model for learning trees, resulting in a new model which we will call *U-DOT*.

Unsupervised Data-Oriented Translation (U-DOT)

U-DOT is based on an extension of the DOP model to unsupervised parsing known as U-DOP (Bod 2006). U-DOP assigns all unlabeled binary trees to a set of given sentences (possibly tagged), and next takes (in principle) all subtrees from these binary trees to compute the most probable trees. For example, the tagged WSJ sentence *Investors suffered heavy losses* has a total of five different binary trees, as shown in figure 3 (where each root node is labeled with S and each internal node is labeled with X).



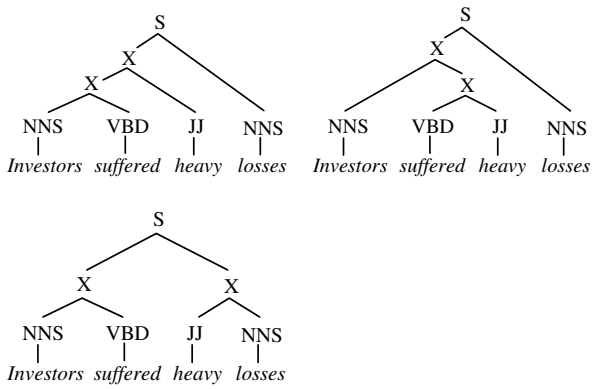


Figure 3. All binary trees for *Investors suffered heavy losses* as proposed by U-DOP

Although the number of binary trees for a sentence grows with the Catalan number, the total set of (unlabeled) binary trees can be stored efficiently by a packed parse forest.

The underlying idea of U-DOP is to use (the frequencies of) all subtrees from this binary tree set to compute the most probable tree for each sentence. Subtrees from the trees in figure 3 include for example the subtrees in figure 4.

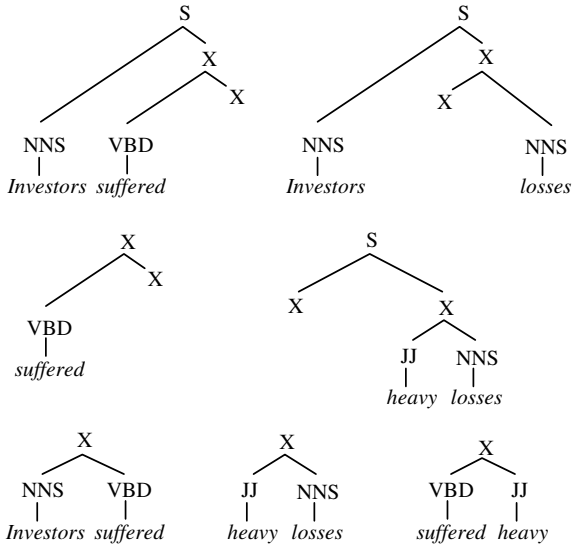


Figure 4. Some subtrees from the trees in figure 3

Thus U-DOP takes into account both contiguous substrings like *Investors suffered X* and non-contiguous substrings like *Investors X losses*. This property carries over to our unsupervised extension of DOT.

In Bod (2007) we have shown how a parse forest of binary trees can be converted into a compact PCFG in the vein of Goodman (2003), and which we will summarize in the next section. The PCFG reduction of parse forests allowed us to induce trees for very large corpora in Bod (2007), such as the four million sentences NANC corpus (Graff 1995). These large experiments could be accomplished also thanks to an efficient estimator known as DOP* (Zollmann and Sima'an 2005). While the resulting U-DOP model was called U-DOP* in Bod (2007), we will continue to refer to the model as U-DOP in the current paper as long as no confusion arises.

U-DOP was been evaluated on English, German and Chinese, obtaining some of the best unsupervised results in the literature (Bod 2007). However, compared to *supervised* parsers, U-DOP's results are considerably lower: where U-DOP obtains about 70% unlabeled f-score on the standard section 23 of the Wall St Journal, many supervised parsers obtain around 91% on the same set. Yet it should be kept in mind that the evaluation on hand-parsed data unreasonably favors supervised parsers. For instance, U-DOP learns constituents for word sequences such as *We would like to…* and *There are…*, which in the Penn Treebank are *non*-constituents. While U-DOP is thus punished for this 'incorrect' prediction if evaluated on the Penn Treebank, this property of U-DOP may be beneficial if evaluated in the context of machine translation using the Bleu score. Thus U-DOP can discover phrases which are typically neglected by linguistically motivated syntax-based translation models. At the same time, the model can also learn *discontiguous* dependencies that are typically neglected by phrase-based MT systems (Koehn et al. 2003).

The extension of DOT with U-DOP is now straightforward. Instead of starting from treebanks, we start from unlabeled corpora and use our new implementation of U-DOP in Bod (2007) to infer the 'best' binary trees directly for word strings from bilingual corpora. Next, we assign links between each two tree nodes for each sentence pair and compute the most probable translation for a held-out data set from the relative frequencies of the subtree pairs (see next section). We will refer to this unsupervised version of DOT as U-DOT. To give a simple illustration of U-DOT, consider the German-English pair *Es gibt viele Zeitungen* and *There are many newspapers* for which U-DOP induces respectively the structures in figure 5 (we leave the words untagged):

$$[[\text{Es gibt}]_X\ [\text{viele Zeitungen}]_X]_S$$

$$[[\text{There are}]_X\ [\text{many newspapers}]_X]_S$$

Figure 5. Two structures induced by U-DOP

Then, U-DOT assigns links between all subtree pairs, as in figure 6:

$[\text{Es gibt}]_X$ – $[\text{There are}]_X$

$[\text{Es gibt}]_X$ – $[\text{many newspapers}]_X$

$[[\text{Es gibt}]_X\ [\text{viele Zeitungen}]_X]_S$ – $[[\text{There are}]_X\ [\text{many newspapers}]_X]_S$

$[\text{viele Zeitungen}]_X$ – $[\text{There are}]_X$

$[\text{viele Zeitungen}]_X$ – $[\text{many newspapers}]_X$

Figure 6. Possible alignments between *Es gibt viele Zeitungen* and *There are many newspapers* according to U-DOT

Many of these alignments would result in incorrect translations. How does U-DOT rule out incorrect alignments such as *Es gibt - many newpapers* on the basis of frequency only? It is easy to see that we only need to observe one other sentence pair with *Es gibt*, for example *Es gibt keine Mitglieder … - There are no members …* to make the alignment *Es gibt – There are* more frequent than alternative alignments for *Es gibt*.[1]

Converting Parse Forests into PCFG Reductions

In principle we can use an $O(n^3)$ CKY-style parsing algorithm for (U-)DOT which first parses the source string, after which the target string is derived from it by following the links. The main computational problem is how to deal with the large number of subtrees. There already exists an efficient *supervised* algorithm that parses a sentence by means of all subtrees from a treebank. This algorithm was extensively described in Goodman (2003) and converts a DOP-based PTSG into a compact PCFG reduction that generates eight rules for each node in the treebank. Goodman's reduction is based on the following idea: every node in every tree is assigned a unique number which is called its address. The notation $A@k$ denotes the node at address k where A is the nonterminal labeling that node. A new nonterminal is created for each node in the training data. This nonterminal is called A_k. Let a_j represent the number of subtrees headed by the node $A@j$, and let a represent the number of subtrees headed by nodes with nonterminal A, that is $a = \Sigma_j\, a_j$. Then there is a PCFG with the following property: for every subtree in the training corpus headed by A, the grammar will generate an isomorphic subderivation. For example, for a node $(A@j\ (B@k, C@l))$, the following eight PCFG rules in figure 7 are generated, where the number following a rule is its weight.

[1] Of course, there is also the pair *Es gibt – There is*. Bod (forthcoming) shows that distinctions between singular and plural nouns can be learned by U-DOP.

$A_j \rightarrow BC$	$(1/a_j)$	$A \rightarrow BC$	$(1/a)$
$A_j \rightarrow B_kC$	(b_k/a_j)	$A \rightarrow B_kC$	(b_k/a)
$A_j \rightarrow BC_l$	(c_l/a_j)	$A \rightarrow BC_l$	(c_l/a)
$A_j \rightarrow B_kC_l$	(b_kc_l/a_j)	$A \rightarrow B_kC_l$	(b_kc_l/a)

Figure 7. PCFG reduction for supervised DOP

By simple induction it can be shown that this construction produces PCFG derivations isomorphic to DOP derivations (Goodman 2003: 130-133). The PCFG reduction is linear in the number of nodes in the corpus. In practice, we smooth the subtree numbers by a simple extension of the good-turing method (see Bod 2006).

While Goodman's reduction method was developed for supervised DOP, where each training sentence is annotated with exactly one tree, the method can be generalized to a corpus where sentences are annotated with all possible binary trees, as long as we represent the trees from the source-language by a shared parse forest. A shared parse forest can be obtained by adding pointers from each node in the chart (or tabular diagram) to the nodes that caused it to be placed in the chart. Such a forest can be represented in cubic space and time (Billot and Lang, 1989). Then, instead of assigning a unique address to each node in each tree, as done by the PCFG reduction for supervised DOP, we now assign a unique address to each node in each parse forest for each sentence. However, the same node may be part of more than one tree. A shared parse forest is an AND-OR graph where AND-nodes correspond to the usual parse tree nodes, while OR-nodes correspond to distinct subtrees occurring in the same context. The total number of nodes is cubic in sentence length n. This means that there are $O(n^3)$ many nodes that receive a unique address as described above, to which next our PCFG reduction is applied. This is a huge reduction compared to Bod (2006) where only ad hoc sampling could make U-DOP work.

Next, we compute the most probable target sentence from the 1,000 most probable derivations by means of Viterbi n-best (the exact computation of the most probable sentence from all derivations is NP hard – see Sima'an 1996). We incorporated the technique by Huang and Chiang (2005) into our implementation which allows for efficient Viterbi n-best parsing.

Experiments

We used the Europarl German-English corpus which consists of 750,000+ sentence pairs with roughly 15,3 million German words and 16,1 million English words. We evaluated the translation performance on a 2,000 sentence test set from a different part of the Europarl corpus. The BLEU score (Papineni et al. 2002) was used to measure translation accuracy, as calculated by the

NIST script (version 11a) with its default settings. The main reason to test U-DOT on German-English is that U-DOP has already been shown to obtain good results in learning structures for German and English sentences (resp. for the NEGRA corpus and the Wall St Journal corpus).

We computed for each test sentence the most probable translation as estimated from the 1,000 most probable derivations. We tested both the full U-DOT model, using all subtrees, and a restricted model, termed U-DOT−, which discards all subtrees with discontiguous yields. As a baseline, we compared our results against the publicly available state-of-the-art phrase-based system Pharaoh (Koehn et al. 2003), using the default feature set. Next, we also used human judgements of translation quality by randomly selecting 100 sentences from the test corpus. Three subjects evaluated the 100 translations produced by each system in random order against the gold standard reference translation using a 5 point fluency and adequacy scale. Table 1 shows the results, where U-DOT+ refers to the full U-DOT model containing subtrees with contiguous as well as noncontiguous yields, while U-DOT− uses only subtrees without discontiguous yields (that is, without any open node between lexicalized nodes).

System	BLEU	Fluency	Adequacy
Pharaoh	0.251	3.2	3.1
U-DOT+	0.280	3.4	3.3
U-DOT−	0.248	3.1	2.9

Table 1. Results of evaluating U-DOT with all subtrees (U-DOT+) and U-DOT without discontiguous subtrees (U-DOT−) against the phrase-based Pharaoh system.

The table shows that U-DOT+ outperforms both U-DOT− and the Pharaoh system, while the Pharaoh system outperforms U-DOT−. By using Zhang's significance tester (Zhang et al. 2004), which employs bootstrap resampling (Koehn 2004), we calculated that the difference in performance between U-DOT+ and Pharaoh is statistically significant ($p < 0.008$). Our system achieves an absolute improvement of 0.029 over the baseline. Also the difference between U-DOT+ and U-DOT− is statistically significant, but the difference between Pharaoh and U-DOT− is not. These experiments show that U-DOT+'s inclusion of discontiguous phrases significantly improves the translation accuracy for German-English. Also with respect to human judgments, U-DOT+ appears to perform better than the purely phrase-based Pharaoh system. It would be interesting to compare our system to the clause restructuring method by Collins et al. (2005) and to the hierarchical phrase-based model by Chiang (2005), but these systems are not (yet) publicly available, though we hope to include a comparison in future research.

We next wanted to compare the U-DOT system, which in Chiang (2005)'s terminology is only *formally* syntax-based, against the supervised DOT model which is *linguistically* syntax-based. Since there are no hand-annotated trees for the Europarl corpus, we employed the supervised DOP model from Bod (2003), which was trained on Penn's Wall St Journal corpus, to parse the trees from the English Europarl. We additionally used the unknown word model in Bod (2003) which uses statistics on word endings, hyphenation and capitalization. However, it is well known that DOP's f-score decreases if it is applied to another domain: for example, DOP's accuracy decreases from around 91% to 85.5% f-score if tested on the Brown corpus. Yet, this score is still considerably higher than the accuracy obtained by the unsupervised U-DOP model, which is 67.6% unlabeled f-score on unrestricted Brown sentences. Although in the absence of a gold standard, we cannot measure DOP's f-score on the Europarl, our use of DOP is motivated by the fact that we want to compare a supervised parser against an unsupervised one in the context of machine translation. For the German part of the Europarl, we trained the DOP parser on the Negra corpus, as Dubey and Keller (2003). Although the use of different training sets resulted in differently labeled trees for English and German under supervised DOP, we simply assigned all possible links between the nodes and let the statistics decide (using the PCFG reduction technique to compute the most probable translation from the 1,000 most probable derivations). Table 2 shows the results, where DOT+ refers to the full DOT model based on the supervised DOP parser, while DOT− refers to the DOT model after excluding subtrees containing discontiguous yields. For comparison, we also added the results of our fully unsupervised MT systems from table 1 again, i.e., U-DOT+, U-DOT− and Pharaoh.

System	BLEU
DOT+	0.221
DOT−	0.209
U-DOT+	0.280
U-DOT−	0.248
Pharaoh	0.251

Table 2. Results of evaluating the supervised DOT systems against the unsupervised U-DOT systems, compared to the phrase-based Pharaoh system.

The table shows that the unsupervised U-DOT+ model outperforms the supervised DOT+ model ($p < 0.001$). Surprisingly, the non-contiguous U-DOT− also

outperformed the DOT+ model ($p < 0.05$), even if DOT+ included both contiguous and discontiguous phrases.

Conclusions

We have shown that the inclusion of noncontiguous phrases in U-DOT significantly improves the translation accuracy for the German-English Europarl corpus, outperforming the state-of-the-art phrase-based Pharaoh system which is based on contiguous phrases only. Our experiments also indicated that an unsupervised syntax-based MT system outperforms a supervised syntax-based MT system. Our experiments need of course be extended to other languages and more complex test sets. In particular, we want to compare U-DOT to other syntax-based MT systems such as Chiang (2005).

References

Billot, S. and B. Lang (1989). The Structure of Shared Forests in Ambiguous Parsing. *Proceedings ACL 1989*.

Bod, R. (2003). An efficient implementation of a new DOP model. *Proceedings EACL 2003*, Budapest.

Bod, R. (2006). An All-Subtrees Approach to Unsupervised Parsing. *Proceedings ACL 2006*, Sydney.

Bod, R. (2007). Is the End of Supervised Parsing in Sight? *Proceedings ACL 2007*, Prague.

Bod, R. (forthcoming). From Exemplar to Grammar. Submitted for publication.

Bod R. R. Scha and K. Sima'an (eds.) (2003). *Data-Oriented Parsing*. CSLI Publications.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. *Proceedings ACL 2005*.

Collins, M., P. Koehn, and I. Kucerova (2005). Clause Restructuring for Statistical Machine Translation. *Proceedings ACL 2005*, Ann Arbor.

Dennis, S. (2005). An Exemplar-Based Approach to Unsupervised Parsing. *Proceedings CogSci 2005*.

Dubey, A. and F. Keller (2003). Parsing German with Sister-Head Dependencies. *Proceedings ACL 2003*, Sapporo.

Galley, M., J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer (2006). Scalable inference and training of context-rich syntactic translation models. In *Proc. ACL 2006*, Sydney.

Goodman, J. (2003). Efficient algorithms for the DOP model. In R. Bod, R. Scha and K. Sima'an (eds.). *Data-Oriented Parsing*, CSLI Publications, 125-146

Graff, D. (1995). *North American News Text Corpus*. Linguistic Data Consortium. LDC95T21.

Groves, D., M. Hearne and A. Way (2004). Robust Sub-Sentential Alignment of Phrase-Structure Trees. *Proceedings COLING 2004*, Geneva.

Hearne, M. and A. Way (2003). Seeing the Wood for the Trees: Data-Oriented Translation. *Proceedings of the Ninth Machine Translation Summit*, New Orleans.

Hearne, M and A. Way (2006). Disambiguation Strategies for Data-Oriented Translation. *Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo.

Hoogweg, L. (2003). Extending DOP with Insertion. In R. Bod, R. Scha and K. Sima'an (eds.). *Data-Oriented Parsing*, CSLI Publications, 317-335.

Huang, L. and D. Chiang (2005). Better *k*-best parsing. *Proceedings IWPT 2005*, Vancouver.

Klein, D. and C. Manning (2002). A general constituent-context model for improved grammar induction. *Proceedings ACL 2002*, Philadelphia.

Klein, D. and C. Manning (2004). Corpus-based induction of syntactic structure: models of dependency and constituency. *Proceedings ACL 2004*, Barcelona.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. *Proceedings of EMNLP 2004*.

Koehn, P., F. Och and D. Marcu (2003). Statistical phrase based translation. *Proceedings of HLT-NAACL 2003*.

McClosky, D., E. Charniak and M. Johnson (2006). Effective self-training for parsing. *Proceedings HLT-NAACL 2006*, New York.

Nesson, R., S. Shieber and A. Rush (2006). Induction of probabilistic synchronous tree-insertion grammars for machine translation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, Boston.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings ACL 2002*.

Poutsma, A. (2000). Data-Oriented Translation. *Proceedings COLING 2000*, Saarbruecken.

Seginer, Y. (2007). Fast Unsupervised Incremental Parsing. *Proceedings ACL 2007*, Prague.

Shieber, S. (2007). Probabilistic synchronous tree-adjoining grammars for machine translation: The argument from bilingual dictionaries. *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, Rochester, New York.

Sima'an, K. (1996). Computational complexity of probabilistic disambiguation by means of tree grammars. *Proceedings COLING 1996*, Copenhagen.

Yamada, K. and K. Knight (2001). A syntax-based statistical translation model. In *Proceedings ACL 2001*.

Zhang, Y., S. Vogel and A. Waibel (2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Zollmann, A. and K. Sima'an (2005). A consistent and efficient estimator for data-oriented parsing. *Journal of Automata, Languages and Combinatorics*, Vol. 10 (2005) Number 2/3, 367-388