Dependency Relations as Source Context in Phrase-Based SMT

Rejwanul Haque[a], Sudip Kumar Naskar[a], Antal van den Bosch[b], and Andy Way[a]

[a] CNGL, School of Computing,
Dublin City University, Dublin 9, Ireland
{rhaque, snaskar, away}@computing.dcu.ie
[b] ILK Research Group, Tilburg centre for Creative Computing,
Tilburg University, Tilburg, The Netherlands
{Antal.vdnBosch}@uvt.nl

Abstract. The Phrase-Based Statistical Machine Translation (PB-SMT) model has recently begun to include source context modeling, under the assumption that the proper lexical choice of an ambiguous word can be determined from the context in which it appears. Various types of lexical and syntactic features such as words, parts-of-speech, and supertags have been explored as effective source context in SMT. In this paper, we show that position-independent syntactic dependency relations of the head of a source phrase can be modeled as useful source context to improve target phrase selection and thereby improve overall performance of PB-SMT. On a Dutch—English translation task, by combining dependency relations and syntactic contextual features (part-of-speech), we achieved a 1.0 BLEU (Papineni *et al.*, 2002) point improvement (3.1% relative) over the baseline.

Keywords: phrase-based SMT, syntactic dependencies, memory-based learning

1 Introduction

In log-linear phrase-based SMT (Koehn *et al.*, 2003), the probability $P(e^I_1|f^J_1)$ of a target phrase e^I_1 given a source phrase f^J_1 is modelled as a log-linear combination of features which typically consist of a finite set of translation features, and a language model (Och and Ney, 2002). The usual translation features involved in those models express dependencies between the source and target phrases, but not among the phrases in the source language themselves. Stroppa *et al.* (2007) were the first to show that incorporating source-language context using neighbouring words and part-of-speech tags had the potential to improve translation quality. Due to a strand of related work, source context modeling has now been shown to offer a new dimension to PB-SMT. By making use of similarity in the contexts of source phrases, information can be added that can positively influence the weighting and selection of target phrases.

Approaches to include source context for proper selection of target phrases have been inspired by methods for word sense disambiguation (WSD), that employ rich context-sensitive features to determine the contextually most likely sense of a polysemous word. These contextual features may include lexical features of words appearing in the context and bearing sense-discriminatory information, position-specific neighbouring words (Giménez and Márquez, 2007; Stroppa *et al.*, 2007), shallow and deep syntactic features of the sentential context (Gimpel and Smith, 2008) and full sentential context (Carpuat and Wu, 2007). Most of the work on syntactic features has made use of part-of-speech taggers (Stroppa *et al.,* 2007), supertaggers (Haque *et al.*, 2009) and shallow and deep syntactic parsers (Gimpel and Smith, 2008). In the present work, we explore how the local sentential context information from a dependency parse can be modeled as source context features to be integrated into a PB-SMT model.

The remainder of the paper is organized as follows. In Section 2 we discuss related work. Section 3 provides a brief overview of PB-SMT. In Section 4 we describe how we model dependency information as context-informed features in our baseline log-linear PB-SMT system. Section 5 describes the memory-based classification approach. In Section 6 we describe the features used in the experiments, and the pre-processing required. Section 7 presents the results obtained, and offers some analysis. In Section 8 we formulate our conclusions, and offer some avenues for further work.

2 Related Work

Brown *et al.* (1991) were the first to propose the use of dedicated WSD models in word-based SMT systems. Results were limited to the case of binary disambiguation, i.e., deciding between only two possible translation candidates, and to a reduced set of common words. A significant improvement in translation was reported according to manual evaluation. Berger *et al.* (1996) suggested context-sensitive modeling of word translations in order to integrate local contextual information into their IBM translation models using a Maximum Entropy (MaxEnt) model, but the work is not supported by any significant evaluation results.

García Varea *et al.* (2001) present a MaxEnt approach to integrate contextual dependencies into the EM algorithm of the statistical alignment model to develop a refined context-dependent lexicon model. Using such a model on the German—English Verbmobil corpus, they obtained better alignment quality in terms of improved alignment error rate (AER). However, since alignment is not an end task in itself and is most often used as an intermediate task to generate phrase pairs for the t-tables in PB-SMT systems, improved AER scores do not necessarily result in improved translation quality, as noted by a number of researchers.

Vickrey *et al.* (2005) built classifiers inspired by those used in WSD to fill in any blanks in a partially completed translation. Giménez and Màrquez (2007) extended this work by considering the more general case of frequent phrases and moved to full translation rather than blank-filling on the target side. Attempts to embed context-rich approaches from WSD methods into SMT systems to enhance lexical selection did not lead to any improvement in translation quality (Carpuat and Wu, 2005). However, more recent approaches of integrating state-of-the-art WSD methods into SMT to improve the overall translation quality have met with more success (Carpuat and Wu, 2007; Chan *et al.*, 2007; Giménez and Màrquez, 2007, 2009).

Recently, Bangalore *et al.* (2008) employed an SMT architecture based on stochastic finite-state transducers that addresses global lexical selection, i.e. dedicated word selection. Specia *et al.* (2008) use dedicated predictions for the re-ranking of *n*-best translations, limited to a small set of words from different grammatical categories. Significant BLEU improvements were reported in both approaches. Hasan *et al.* (2008) present target context modeling into SMT using a triplet lexicon model that captures long-distance (global) dependencies. Their approach is evaluated in a re-ranking framework; slight improvements are observed over IBM model 1 in terms of BLEU and TER (Snover et al., 2006).

Target-language models arguably play the most significant role in today's PB-SMT systems. However, for some time now people have believed that some incorporation of source language information into SMT systems was bound to help. Stroppa *et al.* (2007) added source-side contextual features to a state-of-the-art log-linear PB-SMT system by incorporating context-dependent phrasal translation probabilities learned using decision trees. They considered up to two words and/or POS tags on either side of the source focus word as contextual features. In order to overcome problems of estimation of such features, they used a decision-tree classifier (Daelemans *et al.*, 2005) that implicitly smoothes the probability estimates. Significant improvements over a baseline state-of-the-art PB-SMT system were obtained on Italian—English and Chinese—English IWSLT tasks.

Several proposals have recently been made to fully exploit the accuracy and the flexibility of discriminative learning (Cowan *et al.*, 2006; Liang *et al.*, 2006). Work of this type generally requires a redefinition of the training procedure; in contrast, our approach introduces new features while retaining the strength of existing state-of-the-art systems.

Like the work of (Max *et al.*, 2008), the present work is directly motivated by and is an extension of the approach of (Stroppa *et al.*, 2007). The work of both (Max *et al.*, 2008) and (Gimpel and Smith, 2008) focuses on language pairs where the target is not English. While (Gimpel and Smith, 2008) are unable to show any improvements for English-to-German, (Max *et al.*, 2008) conduct experiments from English-to-French. Using the same sorts of local contextual features as (Stroppa *et al.*, 2007), as well as using broader context in addition to grammatical dependency information, (Max *et al.*, 2008) show modest gains over a PB-SMT baseline model according to manual evaluation. Inspired by the supertag-based target-language modeling (Hassan *et al.*, 2008), Haque *et al.* (2009) extended the work of Stroppa *et al.* (2007) on the IWSLT'06 Chinese—English data and showed that supertag-based source context modeling significantly improves the translation quality.

Discriminative lexical selection in PB-SMT can be broadly divided into two categories: (i) hard interaction such as (Carpuat and Wu, 2005), and (ii) soft interaction such as (Stroppa *et al.*, 2007; Carpuat and Wu, 2007; Chan *et al.*, 2007; Giménez and Màrquez, 2007, 2009; Haque *et al.*, 2009). In the first group, WSD-like predictions are used during pre-processing or post-processing. In the second group, predictions are allowed to interact with other models (e.g., language, distortion, additional translation models etc.) during decoding time. The present work falls into the second type of interaction methods.

3 Log-Linear PB-SMT

Translation is modelled in PB-SMT as a decision process, in which the translation $e_1^I = \mathrm{e}_1 \ldots \mathrm{e}_\mathrm{I}$ of a source sentence $f_1^J = \mathrm{f}_1 \ldots \mathrm{f}_\mathrm{J}$ is chosen to maximize (1):

$$\underset{I,e_1^I}{\arg\max} P(e_1^I \mid f_1^J) = \underset{I,e_1^I}{\arg\max} P(f_1^J \mid e_1^I) P(e_1^I) \quad (1)$$

where $P(f_1^J \mid e_1^I)$ and $P(e_1^I)$ denote respectively the translation model and the target-language model (Brown *et al.*, 1993). In log-linear phrase-based SMT, the posterior probability $P(f_1^I \mid e_1^J)$ is directly modelled as a (log-linear) combination of features (Och and Ney, 2002), that usually comprise M translational features, and the language model, as in (2):

$$\log P(e_1^I \mid f_1^J) = \sum_{m=1}^{M} \lambda_m h_m(f_1^J, e_1^I, s_1^K) + \lambda_{LM} \log P(e_1^I) \quad (2)$$

where $s_1^K = s_1 \ldots s_k$ denotes a segmentation of the source and target sentences respectively into the sequences of phrases $(\hat{e}_1, \ldots, \hat{e}_k)$ and $(\hat{f}_1, \ldots, \hat{f}_k)$ such that (we set $i_0 = 0$) (3):

$$\forall 1 \le k \le K,\ s_k = (i_k; b_k, j_k),\quad \hat{e}_k = e_{i_{k-1}+1} .. e_{i_k},\ \hat{f}_k = f_{b_k} \ldots f_{j_k} \quad (3)$$

The translational features depend only on pairs of source/target phrases and do not take into account any context of these phrases, i.e. each feature h_m in (2) can be rewritten as in (4):

$$h_m(f_1^J, e_1^I, s_1^K) = \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) \quad (4)$$

where $\hat{h}_m$ is a feature that applies to a single phrase-pair. Thus (2) can be rewritten as:

$$\sum_{m=1}^{m} \lambda_m \sum_{k=1}^{K} \hat{h}_m(\hat{f}_k, \hat{e}_k, s_k) = \sum_{k=1}^{K} \hat{h}(\hat{f}_k, \hat{e}_k, s_k) \quad (5)$$

where $\hat{h} = \sum_{m=1}^{m} \lambda_m \hat{h}_m$. In this context, the translation process amounts to: (i) choosing a segmentation of the source sentence, (ii) translating each source phrase, and (iii) reordering the target segments obtained.

4 Dependency Features in PB-SMT

In addition to using local words and POS-tags as features, as in (Stroppa *et al.*, 2007), Haque *et al.* (2009) introduced supertags as a syntactic source context feature type in the log-linear model of PB-SMT. The context of a source phrase $\hat{f}_k$ is defined as the sequence of features immediately before and after the focus phrase $\hat{f}_k$. This contextual information is local, position-specific, and does not carry grammatical dependency information relating to other words in the sentence outside the focus phrase. In this paper, we experiment with incorporating position-independent source-side context information related to the focus phrase, namely the grammatical dependencies linking from and to the head word of the focus phrase $\hat{f}_k$ with words occurring elsewhere in the sentence. Following (Stroppa *et al.,* 2007) and (Haque *et al.*, 2009), we compare this with incorporating words and part-of-speech tags as context, in order to observe the relative effects of position-independent and position-dependent features; we also combine the two types of features.

The identification of the head word of a phrase is not trivial, as SMT phrases are not restricted to linguistic phrases; simple linguistic rules of thumb cannot be used. Therefore, head-words of SMT phrases in a sentence are identified from the dependency tree generated for that sentence. For all words in a given source phrase, the word that occupies the hierarchically highest position in the dependency graph is chosen as the head word. We consider the following dependency features, drawing on the syntactic dependencies emanating from or pointing to the head word of the source focus phrase:

(A) For the head word of the focus phrase, we extract a list of the zero or more relations with other words of which the head word is the parent (i.e. the dependency type labels on all modifying dependency relations). The list of relations is concatenated and sorted uniquely and alphabetically into a single feature. This feature is denoted as IR (incoming relations).

(B) For the head word of the focus phrase we extract the relation it has with its parent; it will always have one and only one. If the head word is a verb, then frame sub-categorization information is extracted and used as this feature. This feature is denoted as PR (parent relation).

(C) Extending (B), we encode the identity of the single parent word of the head word of the focus phrase. This feature is denoted as PW (parent word).

Together we refer to these dependency features as the grammatical *dependency information* (DI) of the focus phrase $\hat{f}_k$, DI ($\hat{f}_k$). They are expressed as the conditional probability of the target phrase given the source phrase $\hat{f}_k$ and its grammatical dependency information DI ($\hat{f}_k$), as in (6):

$$\hat{h}_m (\hat{f}_k , \mathrm{DI} (\hat{f}_k), \hat{e}_k , s_k) = \log P (\hat{e}_k \mid \hat{f}_k , \mathrm{DI}(\hat{f}_k)) \tag{6}$$

5 Memory-Based Classification

As (Stroppa *et al.,* 2007) point out, directly estimating context-dependent phrase translation probabilities using relative frequencies is problematic. Indeed, Zens and Ney (2004) showed that the estimation of $P(\hat{e}_k \mid \hat{f}_k)$ using relative frequencies results in the overestimation of the probabilities of long phrases. In the case of grammatical dependency-informed features, which include the identity of the parent word of the focus phrase, this estimation problem can only become worse.

As an alternative, in this work we make use of memory-based machine learning classifiers that are able to estimate P ($\hat{e}_k \mid \hat{f}_k$, DI ($\hat{f}_k$)) by similarity-based reasoning over memorized nearest-neighbour examples of source—target phrase translations to a new source phrase to be translated. Memory-based classification uses a distance function operating on $\{ \hat{f}_k$, DI ($\hat{f}_k$)$\}$, producing a numeric distance between a source focus phrase to be translated, against all memorized examples of source phrases in the training set. In this study we adopt the commonly used Overlap metric (Daelemans and Van den Bosch, 2005).

A parameter k determines the k-closest radii of distances around the source phrase that encompass the nearest neighbours; then, the distribution of target phrases associated with these nearest neighbours is taken as the output of the classification step. The contribution of a single nearest neighbour in this set can be weighted by its distance to the source phrase to be translated, e.g. by assigning higher weights to closer neighbours. We set k=3 in our experiments, and use exponential decay for the distance-weighted class voting (Daelemans and Van den Bosch, 2005).

As the search for nearest neighbours can be slow when there are large amounts of training examples, heuristic methods exist that produce approximate nearest neighbour search. We employ one such approximate memory-based classifier: TRIBL[1] (Daelemans *et al.,* 1997). The TRIBL approximation performs an initial decision-tree split of the database of training examples on the n most informative features (we set n=1). Feature importance is estimated by computing the gain ratio of all features. After this sub-selection of training examples matching on the most informative feature, the nearest-neighbour distance function is applied to the remaining features (weighted by their gain ratio) to arrive at the set of nearest neighbours. When predicting a target phrase given a source phrase and its dependency information, the identity of the source phrase is (also intuitively) the feature with the highest prediction power. This implies that nearest neighbours always match on the source phrase, and are most similar (preferably, identical) with respect to their contextual dependency features.

6 Experimental Set-Up

6.1 Features Used

The result of memory-based classification is a set of weighted class labels, representing the possible target phrases $\hat{e}_k$ given a source phrase and its dependency information. Once normalized, these weights can be seen as the posterior probabilities of the target phrases $\hat{e}_k$, which thus give access to $P(\hat{e}_k \mid \hat{f}_k, DI(\hat{f}_k))$. Therefore, the expected feature is derived as in (7):

$$\hat{h}_{mbl} = \log P(\hat{e}_k \mid \hat{f}_k, DI(\hat{f}_k)) \tag{7}$$

In addition to the above feature, we derived a simple binary feature $\hat{h}_{best}$. The feature $\hat{h}_{best}$ is defined as in (8):

$$\hat{h}_{best} = \begin{cases} 1 & \text{if } \hat{e}_k \text{ maximizes } P(\hat{e}_k \mid \hat{f}_k, CI(\hat{f}_k)) \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

We performed experiments by integrating these two features $\hat{h}_{mbl}$ and $\hat{h}_{best}$ directly into the log-linear model. Their weights are optimized using minimum error-rate training (Och, 2003) on a held-out development set for each of the experiments.

Our approach in terms of experimental set-up and classification of a source phrase along with contextual dependency features differs from Stroppa *et al.*, (2009) and Haque *et al.,* (2009) in the following respects:

(i) Stroppa *et al.* (2007) and Haque *et al.*, (2009) integrate local, position-specific contextual features into the log-linear framework. Here we integrate a feature encoding position-independent dependency information;

(ii) Haque *et al.* (2009) interpolate the context-dependent phrase translation probability with the forward phrase translation probability; the interpolation weight is tuned manually. Here we directly integrate our derived features into the PB-SMT log-linear framework.

[1] TRIBL is available as part of the TiMBL software package, which can be downloaded from http://ilk.uvt.nl/timbl.

Two more papers are closely related to the present work. Carpuat and Wu (2007) mention in passing that their WSD system uses basic dependency relations, but the nature of this information is not further described, and neither is the effect. Max *et al.*, (2008) exploit grammatical dependency information, in addition to information extracted from the immediate context of a source phrase. Our approach differs with Max *et al.*, (2008) in three respects:

(i) Max *et al.*, (2008) select the set of 16 most informative dependency relations for their experiments. Dependencies are considered that link any of the tokens of the given source phrase to tokens outside this phrase. Each dependency type is represented in the vector by the outside word it involves, or by the symbol 'nil', which indicates that this type of dependency does not occur in the phrase under consideration. In contrast to this approach, we used all 26 dependency relations in our experiments, and we only generate features from the head-words of the SMT phrases, identified from the dependency graph generated for the source sentence (as described earlier in Section 4).

(ii) They filter out phrases from phrase table entries for which $P(\hat{e}_k \mid \hat{f}_k) < 0.0002$. In contrast, we keep all phrase pairs for more discrimination.

(iii) Their experimental data contains 95K English-to-French training pairs, while we trained our models on about three times as many (286K) Dutch-to-English translation pairs, a less explored direction.

6.2 Pre-processing

As (Stroppa *et al.*, 2007) point out, PB-SMT decoders such as Moses (Koehn *et al.*, 2007) rely on a static phrase table, represented as a list of aligned phrases accompanied by several estimated metrics. Since these features do not express the context information in which those phrases occur, no dependency information is kept in the phrase table, and there is no way to recover this information from the phrase table.

In order to take into account the dependency information features within such decoders, the test text to be translated is pre-processed. Each word appearing in the test set (and, during development, the development set) is assigned a unique identifier. First we prepare the phrase table using the training data. Subsequently, we generate all possible phrases from the testset. These phrases are then looked up in the phrase table, and when found, the phrase along with its dependency information is given to TRIBL for classification. As stated above, TRIBL produces target phrase distributions according to the training examples found within the *k*-nearest distance radii around the source phrase to be classified. We derive target phrase probabilities from this distribution and temporarily insert them instead of the original phrase table estimates of the found target phrases, to directly take our feature functions ($\hat{h}_{mbl}$ and $\hat{h}_{best}$) into account in the log-linear model. Thus we create a dynamic phrase table.

A lexicalized reordering model is used for all the experiments undertaken on development and test texts. The source phrase in the reordering table is replaced by the sequence of unique identifiers when the new phrase table is created. After replacing all words by their unique identifyers, we perform MERT using our new phrase table to optimize the feature weights.

7 Results and Analysis

The experiments were carried out on the Dutch-to-English Open Subtitles corpus,[2] which is collected as part of the Opus collection of freely available parallel corpora (Tiedemann and Nygaard, 2004). The corpus contains user-contributed translations of movie subtitles. The training text contains 286,160 sentences; the development set and test set each contain 1,000 sentences. Dutch sentences were parsed using Tadpole[3], a morphosyntactic analyzer and dependency parser (Van den Bosch *et al.*, 2007).

[2] http://urd.let.rug.nl/tiedeman/OPUS/OpenSubtitles.php

[3] http://ilk.uvt.nl/tadpole/

Table 1: Experiments with words and part-of-speech.

Experiments	BLEU	NIST	METEOR	TER	WER	PER
+ Word±2	33.05	6.11	56.02	50.62	49.82	43.68
+ POS±2	33.30	6.09	**56.57**	50.52	50.17	43.81
+ POS±2*	**33.39**	6.11	56.3	50.43	50.34	43.54

Table 2: Experiments with dependency relations.

Experiments	BLEU	NIST	METEOR	TER	WER	PER
PR	32.69	6.08	55.08	50.48	50.11	43.58
IR	32.61	6.00	55.53	52.40	51.56	45.09
PR + PW	32.74	6.06	**55.98**	51.15	50.75	43.61
PR + IR	**33.06**	**6.20**	55.70	**49.45**	**48.83**	**42.44**
PR + IR+ PW	32.79	6.18	55.37	49.51	49.03	42.43

Table 3: Experiments combining dependency relations, words and part-of-speech.

Experiments	BLEU	NIST	METEOR	TER	WER	PER
Baseline	*32.39*	*6.11*	*55.39*	*50.15*	*49.67*	*43.12*
Word±2	32.48	6.11	55.72	50.40	50.43	42.91
POS±2	33.07	6.13	56.17	50.07	49.38	42.85
POS±2*	**33.29**	**6.17**	**55.72**	**49.56**	**48.91**	**42.77**
Word±2+POS±2	32.59	6.09	55.36	50.11	49.63	43.10

We performed three series of experiments. In the first series, words, part-of-speech, and their combination are added as contextual information, respectively denoted by Word±2, POS±2 and Word±2+POS±2 (Stroppa *et al.*, 2007). The experimental results are reported in Table 1. In all cases, the size of the left and right contexts is 2. An additional experiment was performed in which the parts-of-speech of the focus phrases were ignored (Haque *et al.*, 2009), identified in Table 1 as POS±*. As can be observed from Table 1, the POS±* experiment produces the best improvements over the baseline of 0.90 BLEU points.

A second series of experiments was performed involving dependency relations as source context. Results are shown in Table 2. Five different experiments were performed combining dependency features (IR, PR and PW). The combination of PR and IR produces the best results in terms of BLEU and NIST (Doddington, 2002): we observe a 0.67 point improvement in BLEU. In the third series we combined the position-independent PR+IR dependency features with the position-dependent word and part-of-speech features. The combined experimental results are reported in Table 3. We observe that combining POS±2* with PR+IR yields the highest BLEU improvements (1.0 BLEU point; 3.08% relative) over the baseline. The best METEOR (Banerjee and Lavie, 2005) score (an improvement of 1.18 over the baseline) is obtained when PR+IR is combined with POS±2.

As an additional analysis, Figure 1 displays the distribution of distances (number of tokens) between the source phrase boundary and words outside the phrase linked through a dependency relation. There are about two times as many incoming modifier dependency relations linking to modifier words outside the focus phrase than to phrase-internal modifiers. About half of the phrases have the root of the dependency graph as the parent, i.e., they are the main verb. For the remainder of the phrases, by definition the parent of the head-word is a phrase-external word.

8 Conclusion and Future Work

From the distance distribution statistics we find that the average distance of head-modifying words to the phrase boundary is only 0.75 when including phrase-internal relations, indicating that modifiers of the phrase are usually not far away. In contrast, parent words of the phrase's head word are found relatively further away, at an average distance of 1.69 tokens outside the phrase boundary. By incorporating dependency features such relatively distant relations can be incorporated as contextual information, in addition to the local information explored in earlier studies. With the present study we have demonstrated that considering dependency relations as source context in PB-SMT, the system yields gains with all evaluation metrics (with a best 2.08% relative gain in BLEU). Nevertheless, considering only local contextual features (words and part-of-speech tags), the system produces better gains (such as a 2.78% relative gain in terms of BLEU) over the baseline. Fortunately we can gain from combining the two information sources: when we test the combination, we observe the best improvement over the baseline both in terms of BLEU (an improvement of 1.0 point, a relative gain of 3.08%) and METEOR (an improvement of 1.18 points, a relative gain of 2.13%).

Our experiments have focused on the Dutch-to-English Open Subtitles dataset, using a dependency parser for Dutch. We intend to further validate our conclusions by scaling up to larger datasets. We also intend to perform experiments in the reverse direction, English-to-Dutch, which would enable us to test the combination of dependency features and supertags.

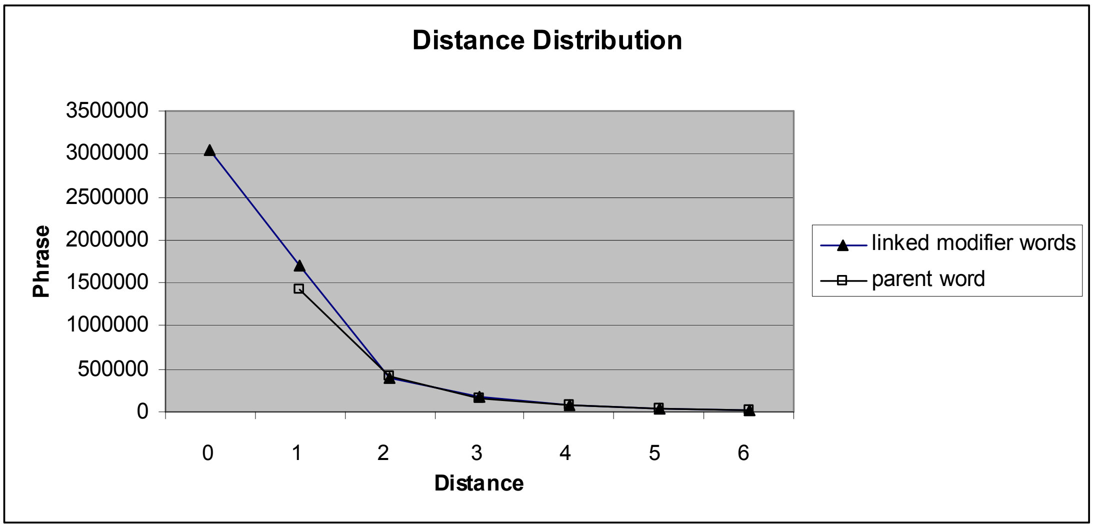

Figure 1: Distances found between phrase boundaries with linked modifier words and parent word.

References

Banerjee S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics (ACL-05)*, University of MI, Ann Arbor, MI, pp.65–72.

Bangalore S., P. Haffner and S. Kanthak. 2007. Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL-07)*, Prague, Czech Republic, pp.152–159.

Berger A., S.A. Della Pietra and V.J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–68.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263—311.

Brown P.F., S.A. Della Pietra, V.J. Della Pietra and R.L. Mercer. 1991. A statistical approach to sense disambiguation in machine translation. In *Proceedings of HLT '91: Workshop on Speech and Natural Language*, Morristown, NJ, USA, pp.146–151.

Carpuat M. and D. Wu. 2005. Evaluating the word sense disambiguation performance of statistical machine translation. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, Jeju Island, Republic of Korea, pp.120-125.

Carpuat M. and D. Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation, *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp.73-80.

Chan Y.S., H.T. Ng and D. Chang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL-2007),* Prague, Czech Republic. pp.33-40.

Cowan B., I. Kucerov`a and M. Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, Sydney, Australia, pp.232–241.

Daelemans W., A. van den Bosch, and J. Zavrel. 1997. A feature-relevance heuristic for indexing and compressing large case bases. In M. van Someren and G. Widmer (Eds.), *9th European Conference on Machine Learning - Poster Papers*. Prague: Laboratory of Intelligent Systems, pp.29-38.

Daelemans W. and A. van den Bosch. 2005. Memory-based language processing, Cambridge University Press, Cambridge, UK.

Doddington G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT-2002)*, San Diego, California, ed. Mitchell Marcus [San Francisco, CA: Morgan Kaufmann for DARPA], pp. 138-145.

García-Varea I., F.J. Och, H. Ney and F. Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. *Proceedings of the 39th Annual meeting [of the Association for Computational Linguistics] and 10th Conference of the European Chapter [of ACL] (ACL-EACL 2001),* Toulouse, France, pp.204-211.

Giménez J. and L. Màrquez. 2007. Context-aware discriminative phrase selection for statistical machine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL-2007): 2nd Workshop on Statistical Machine Translation,* Prague, Czech Republic, pp.159-166.

Giménez J. and L. Márquez. 2008. Discriminative Phrase Selection for Statistical Machine Translation. In C. Goutte, N. Cancedda, M. Dymetman and G. Foster (eds.*) Learning Machine Translation. NIPS Workshop Series. MIT Press*.

Gimpel K. and N.A. Smith. 2008. Rich source-side context for statistical machine translation. *ACL-08: HLT. Proceedings of the Third Workshop on Statistical Machine Translation*, The Ohio State University, OH, pp.9-17.

Haque R., S.K. Naskar, Y. Ma and A. Way. 2009. Using Supertags as Source Language Context in SMT. In *Proceedings of the 13th EAMT Conference*, Barcelona, Spain, pp.234-241.

Hasan S., J. Ganitkevitch, H. Ney and J. Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-*2008), Honolulu, Hawaii, USA, pp.372-381.

Hassan H., K. Sima'an and A. Way. 2008. Syntactically Lexicalized Phrase-Based SMT. *IEEE Transactions on Audio, Speech and Language Processing* 6(7):1260-1273.

Koehn P., F.J. Och and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series (HLT-NAACL 2003)*, Edmonton, Canada, pp.48-54.

Koehn P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical ma-chine translation. In *Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007): Proceedings of demo and poster sessions*, Prague, Czech Republic, pp.177-180.

Liang P., A. Bouchard-Côté, D. Klein and B. Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006*) , Sydney, Australia, pp.761-768.

Max A., R. Makhloufi and P. Langlais. 2008. Explorations in Using Grammatical Dependencies for Contextual Phrase Translation Disambiguation. In *Proceedings of the 12th EAMT Conference,* Hamburg, Germany, pp.112-117.

Och F. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, pp.160-167.

Och F. and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002),* Philadelphia, PA, pp.295-302.

Papineni K., S. Roukos, T. Ward and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, pp.311-318.

Snover M., B. Dorr, R. Scwartz, J. Makhoul and L. Micciula. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, MA, pp. 223-231.

Specia L., B. Sankaran and M.G.V. Nunes. 2008. n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation, In *Computational Linguistics and Intelligent Text Processing*, Springer Berlin/Heidelberg, pp.399-410.

Stroppa N., A. van den Bosch and A. Way. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2007)*, Skövde, Sweden, pp.231-240.

Tiedemann J. and L. Nygaard. 2004. The OPUS corpus - parallel & free. In *Proceedings of the 4th International Conference on language resources and evaluation (LREC-2004).* Lisbon, Portugal, pp.1183-1186.

Van den Bosch A., G.J. Busser, S. Canisius and W. Daelemans. 2007. An efficient memory-based morpho-syntactic tagger and parser for Dutch. In *Proceedings of Computational Linguistics in the Netherlands: Selected Papers from the Seventeenth CLIN Meeting*, Leuven, Belgium, pp. 99-114.

Vickrey D., L. Biewald, M. Teyssier and D. Koller. 2005. Word-sense disambiguation for machine translation. In *Proceedings of Human Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, Vancouver, BC, Canada, pp.771-778.

Zens R. and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proceedings of Human Language Technology conference and North American Chapter of the Association for Computational Linguistics annual meeting (HLT/NAACL 2004)*, Boston, MA, pp.257-264.