The University of Maryland Statistical Machine Translation System for the Fourth Workshop on Machine Translation

Chris Dyer[*†]**, Hendra Setiawan**[†]**, Yuval Marton**[*†]**, and Philip Resnik**[*†]
[†]UMIACS Laboratory for Computational Linguistics and Information Processing
[*]Department of Linguistics
University of Maryland, College Park, MD 20742, USA
{redpony,hendra,ymarton,resnik} AT umd.edu

Abstract

This paper describes the techniques we explored to improve the translation of news text in the German-English and Hungarian-English tracks of the WMT09 shared translation task. Beginning with a convention hierarchical phrase-based system, we found benefits for using word segmentation lattices as input, explicit generation of beginning and end of sentence markers, minimum Bayes risk decoding, and incorporation of a feature scoring the alignment of function words in the hypothesized translation. We also explored the use of monolingual paraphrases to improve coverage, as well as co-training to improve the quality of the segmentation lattices used, but these did not lead to improvements.

1 Introduction

For the shared translation task of the Fourth Workshop on Machine Translation (WMT09), we focused on two tasks: German to English and Hungarian to English translation. Despite belonging to different language families, German and Hungarian have three features in common that complicate translation into English:

1. productive compounding (especially of nouns),

2. rich inflectional morphology,

3. widespread mid- to long-range word order differences with respect to English.

Since these phenomena are poorly addressed with conventional approaches to statistical machine translation, we chose to work primarily toward mitigating their negative effects when constructing our systems. This paper is structured as follows. In Section 2 we describe the baseline model, Section 3 describes the various strategies we employed to address the challenges just listed, and Section 4 summarizes the final translation system.

2 Baseline system

Our translation system makes use of a hierarchical phrase-based translation model (Chiang, 2007), which we argue is a strong baseline for these language pairs. First, such a system makes use of lexical information when modeling reordering (Lopez, 2008), which has previously been shown to be useful in German-to-English translation (Koehn et al., 2008). Additionally, since the decoder is based on a CKY parser, it can consider all licensed reorderings of the input in polynomial time, and German and Hungarian may require quite substantial reordering. Although such decoders and models have been common for several years, there have been no published results for these language pairs.

The baseline system translates lowercased and tokenized source sentences into lowercased target sentences. The features used were the rule translation relative frequency $P(\bar{e}|\bar{f})$, the "lexical" translation probabilities $P_{lex}(\bar{e}|\bar{f})$ and $P_{lex}(\bar{f}|\bar{e})$, a rule count, a target language word count, the target (English) language model $P(e_1^I)$, and a "pass-through" penalty for passing a source language word to the target side.[1] The rule feature values were computed online during decoding using the suffix array method described by Lopez (2007).

[1]The "pass-through" penalty was necessary since the English language modeling data contained a large amount of source-language text.

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 145–149,
Athens, Greece, 30 March – 31 March 2009.

2.1 Training and development data

To construct the translation suffix arrays used to compute the translation grammar, we used the parallel training data provided. The preprocessed training data was filtered for length and aligned using the GIZA++ implementation of IBM Model 4 (Och and Ney, 2003) in both directions and symmetrized using the `grow-diag-final-and` heuristic. We trained a 5-gram language model from the provided English monolingual training data and the non-Europarl portions of the parallel training data using modified Kneser-Ney smoothing as implemented in the SRI language modeling toolkit (Kneser and Ney, 1995; Stolcke, 2002). We divided the 2008 workshop "news test" sets into two halves of approximately 1000 sentences each and designated one the dev set and the other the dev-test set.

2.2 Automatic evaluation metric

Since the official evaluation criterion for WMT09 is human sentence ranking, we chose to minimize a linear combination of two common evaluation metrics, BLEU and TER (Papineni et al., 2002; Snover et al., 2006), during system development and tuning:

$$\frac{\text{TER} - \text{BLEU}}{2}$$

Although we are not aware of any work demonstrating that this combination of metrics correlates better than either individually in sentence ranking, Yaser Al-Onaizan (personal communication) reports that it correlates well with the human evaluation metric HTER. In this paper, we report uncased TER and BLEU individually.

2.3 Forest minimum error training

To tune the feature weights of our system, we used a variant of the minimum error training algorithm (Och, 2003) that computes the error statistics from the target sentences from the translation search space (represented by a *packed forest*) that are exactly those that are minimally discriminable by changing the feature weights along a single vector in the dimensions of the feature space (Macherey et al., 2008). The loss function we used was the linear combination of TER and BLEU described in the previous section.

3 Experimental variations

This section describes the experimental variants explored.

3.1 Word segmentation lattices

Both German and Hungarian have a large number of compound words that are created by concatenating several morphemes to form a single orthographic token. To deal with productive compounding, we employ *word segmentation lattices*, which are word lattices that encode alternative possible segmentations of compound words. Doing so enables us to use possibly inaccurate approaches to guess the segmentation of compound words, allowing the decoder to decide which to use during translation. This is a further development of our general source-lattice approach to decoding (Dyer et al., 2008).

To construct the segmentation lattices, we define a log-linear model of compound word segmentation inspired by Koehn and Knight (2003), making use of features including number of morphemes hypothesized, frequency of the segments as free-standing morphemes in a training corpus, and letters in each segment. To tune the model parameters, we selected a set of compound words from a subset of the German development set, manually created a linguistically plausible segmentation of these words, and used this to select the parameters of the log-linear model using a lattice minimum error training algorithm to minimize WER (Macherey et al., 2008). We reused the same features and weights to create the Hungarian lattices. For the test data, we created a lattice of every possible segmentation of any word 6 characters or longer and used forward-backward pruning to prune out low-probability segmentation paths (Sixtus and Ortmanns, 1999). We then concatenated the lattices in each sentence.

Source	Condition	BLEU	TER
German	baseline	20.8	60.7
	lattice	**21.3**	**59.9**
Hungarian	baseline	11.0	71.1
	lattice	**12.3**	**70.4**

Table 1: Impact of compound segmentation lattices.

To build the translation model for lattice system, we segmented the training data using the one-best split predicted by the segmentation model, and word aligned this with the English side. This variant version of the training data was then concatenated with the baseline system's training data.

3.1.1 Co-training of segmentation model

To avoid the necessity of manually creating segmentation examples to train the segmentation model, we attempted to generate sets of training examples by selecting the compound splits that were found along the path chosen by the decoder's one-best translation. Unfortunately, the segmentation system generated in this way performed slightly worse than the one-best baseline and so we continued to use the parameter settings derived from the manual segmentation.

3.2 Modeling sentence boundaries

Incorporating an n-gram language model probability into a CKY-based decoder is challenging. When a partial hypothesis (also called an "item") has been completed, it has not yet been determined what strings will eventually occur to the left of its first word, meaning that the exact computation must deferred, which makes pruning a challenge. In typical CKY decoders, the beginning and ends of the sentence (which often have special characteristics) are not conclusively determined until the whole sentence has been translated and the probabilities for the beginning and end sentence probabilities can be added. However, by this point it is often the case that a possibly better sentence beginning has been pruned away. To address this, we explicitly generate beginning and end sentence markers as part of the translation process, as suggested by Xiong et al. (2008). The results of doing this are shown in Table 2.

Source	Condition	BLEU	TER
German	baseline	21.3	**59.9**
	+boundary	**21.6**	60.1
Hungarian	baseline	12.3	70.4
	+boundary	**12.8**	**70.4**

Table 2: Impact of modeling sentence boundaries.

3.3 Source language paraphrases

In order to deal with the sparsity associated with a rich source language morphology and limited-size parallel corpora (bitexts), we experimented with a novel approach to paraphrasing out-of-vocabulary (OOV) source language phrases in our Hungarian-English system, using monolingual contextual similarity rather than phrase-table pivoting (Callison-Burch et al., 2006) or monolingual bitexts (Barzilay and McKeown, 2001; Dolan et al., 2004). Distributional profiles for source phrases were represented as context vectors over a sliding window of size 6, with vectors defined using log-likelihood ratios (cf. Rapp (1999), Dunning (1993)) but using cosine rather than city-block distance to measure profile similarity.

The 20 distributionally most similar source phrases were treated as paraphrases, considering candidate phrases up to a width of 6 tokens and filtering out paraphrase candidates with cosine similarity to the original of less than 0.6. The two most likely translations for each paraphrase were added to the grammar in order to provide mappings to English for OOV Hungarian phrases.

This attempt at monolingually-derived source-side paraphrasing did not yield improvements over baseline. Preliminary analysis suggests that the approach does well at identifying many content words in translating extracted paraphrases of OOV phrases (e.g., *a kommunista part vezetaje* ⇒ , *leader of the communist party* or *a ra tervezett* ⇒ *until the planned to*), but at the cost of more frequently omitting target words in the output.

3.4 Dominance feature

Although our baseline hierarchical system permits long-range reordering, it lacks a mechanism to identify the most appropriate reordering for a specific sentence translation. For example, when the most appropriate reordering is a long-range one, our baseline system often also has to consider shorter-range reorderings as well. In the worst case, a shorter-range reordering has a high probability, causing the wrong reordering to be chosen. Our baseline system lacks the capacity to address such cases because all the features it employs are independent of the phrases being moved; these are modeled only as an unlexicalized generic nonterminal symbol.

To address this challenge, we included what we call a *dominance feature* in the scoring of hypothesis translations. Briefly, the premise of this feature is that the function words in the sentence hold the key reordering information, and therefore function words are used to model the phrases being moved. The feature assesses the quality of a reordering by looking at the phrase alignment between pairs of function words. In our experiments, we treated the 128 most frequent words in the corpus as function words, similar to Setiawan et al. (2007). Due to space constraints, we will discuss the details in another publication. As Table 3 reports, the use of this feature yields positive results.

Source	Condition	BLEU	TER
German	baseline	21.6	60.1
	+dom	**22.2**	**59.8**
Hungarian	baseline	**12.8**	70.4
	+dom	12.6	**70.0**

Table 3: Impact of alignment dominance feature.

3.5 Minimum Bayes risk decoding

Although during minimum error training we assume a decoder that uses the maximum derivation decision rule, we find benefits to translating using a *minimum risk* decision rule on a test set (Kumar and Byrne, 2004). This seeks the translation E of the input lattice $\mathcal{F}$ that has the least *expected loss*, measured by some loss function L:

$$\hat{E} = \arg\min_{E'} \mathbb{E}_{P(E|\mathcal{F})}[L(E, E')] \quad (1)$$
$$= \arg\min_{E'} \sum_{E} P(E|\mathcal{F}) L(E, E') \quad (2)$$

We approximate the posterior distribution $P(E|\mathcal{F})$ and the set of possible candidate translations using the unique 500-best translations of a source lattice $\mathcal{F}$. If $H(E, \mathcal{F})$ is the decoder's path weight, this is:

$$P(E|\mathcal{F}) \propto \exp \alpha H(E, \mathcal{F})$$

The optimal value for the free parameter α must be experimentally determined and depends on the ranges of the feature functions and weights used in the model, as well as the amount and kind of pruning using during decoding.[2] For our submission, we used $\alpha = 1$. Since our goal is to minimize $\frac{\text{TER}-\text{BLEU}}{2}$ we used this as the loss function in (2). Table 4 shows the results on the dev-test set for MBR decoding.

[2]If the free parameter α lies in $(1, \infty)$ the distribution is sharpened, if it lies in $[0, 1)$, the distribution is flattened.

Source	Decoder	BLEU	TER
German	Max-D	22.2	59.8
	MBR	**22.6**	**59.4**
Hungarian	Max-D	12.6	70.0
	MBR	**12.8**	**69.8**

Table 4: Performance of maximum derivation vs. MBR decoders.

4 Conclusion

Table 5 summarizes the impact on the dev-test set of all features included in the University of Maryland system submission.

Condition	German		Hungarian	
	BLEU	TER	BLEU	TER
baseline	20.8	60.7	11.0	71.1
+lattices	21.3	59.9	12.3	70.4
+boundary	21.6	60.1	12.8	70.4
+dom	22.2	59.8	12.6	70.0
+MBR	22.6	59.4	12.8	69.8

Table 5: Summary of all features

Acknowledgments

This research was supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-001, and the Army Research Laboratory. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of the sponsors. Discussions with Chris Callison-Burch were helpful in carrying out the monolingual paraphrase work.

References

Regina Barzilay and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *In Proceedings of ACL-2001*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings NAACL-2006*.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

B. Dolan, C. Quirk, and C. Brockett. 2004. Unsupervised construction of large paraphrase corpora:

exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics of the Association for Computational Linguistics*, Geneva, Switzerland.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Chris Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, June.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of IEEE Internation Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the EACL 2003*.

Philipp Koehn, Abhishek Arun, and Hieu Hoang. 2008. Towards better machine translation quality for the German-English language pairs. In *ACL Workshop on Statistical Machine Translation*.

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Processings of HLT-NAACL*.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 976–985.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. In *Proceedings of COLING*, Manchester, UK.

Wolfgang Macherey, Franz Josef Och, Ignacio Thayer, and Jakob Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proceedings of EMNLP*, Honolulu, HI.

F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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

K. Papineni, 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 ACL*, pages 311–318.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics.*, pages 519–525.

Hendra Setiawan, Min-Yen Kan, and Haizhao Li. 2007. Ordering phrases with function words. In *Proceedings of ACL*.

S. Sixtus and S. Ortmanns. 1999. High quality word graphs using forward-backward pruning. In *Proceedings of ICASSP*, Phoenix, AZ.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Intl. Conf. on Spoken Language Processing*.

Deyi Xiong, Min Zhang, Ai Ti Aw, Haitao Mi, Qun Liu, and Shouxun Lin. 2008. Refinements in BTG-based statistical machine translation. In *Proceedings of IJCNLP 2008*.