Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment

Jason R. Smith[*]
Center for Lang. and Speech Processing
Johns Hopkins University
Baltimore, MD 21218
jsmith@cs.jhu.edu

Chris Quirk and Kristina Toutanova
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{chrisq,kristout}@microsoft.com

Abstract

The quality of a statistical machine translation (SMT) system is heavily dependent upon the amount of parallel sentences used in training. In recent years, there have been several approaches developed for obtaining parallel sentences from non-parallel, or comparable data, such as news articles published within the same time period (Munteanu and Marcu, 2005), or web pages with a similar structure (Resnik and Smith, 2003). One resource not yet thoroughly explored is Wikipedia, an online encyclopedia containing linked articles in many languages. We advance the state of the art in parallel sentence extraction by modeling the document level alignment, motivated by the observation that parallel sentence pairs are often found in close proximity. We also include features which make use of the additional annotation given by Wikipedia, and features using an automatically induced lexicon model. Results for both accuracy in sentence extraction and downstream improvement in an SMT system are presented.

1 Introduction

For any statistical machine translation system, the size of the parallel corpus used for training is a major factor in its performance. For some language pairs, such as Chinese-English and Arabic-English, large amounts of parallel data are readily available, but for most language pairs this is not the case. The domain of the parallel corpus also strongly influences the quality of translations produced. Many parallel corpora are taken from the news domain, or from parliamentary proceedings. Translation quality suffers when a system is not trained on any data from the domain it is tested on.

While parallel corpora may be scarce, comparable, or semi-parallel corpora are readily available in several domains and language pairs. These corpora consist of a set of documents in two languages containing similar information. (See Section 2.1 for a more detailed description of the types of non-parallel corpora.) In most previous work on extraction of parallel sentences from comparable corpora, some coarse document-level similarity is used to determine which document pairs contain parallel sentences. For identifying similar web pages, Resnik and Smith (2003) compare the HTML structure. Munteanu and Marcu (2005) use publication date and vector-based similarity (after projecting words through a bilingual dictionary) to identify similar news articles.

Once promising document pairs are identified, the next step is to extract parallel sentences. Usually, some seed parallel data is assumed to be available. This data is used to train a word alignment model, such as IBM Model 1 (Brown et al., 1993) or HMM-based word alignment (Vogel et al., 1996). Statistics from this word alignment model are used to train a classifier which identifies bilingual sentence pairs as parallel or not parallel. This classifier is applied to all sentence pairs in documents which were found to be similar. Typically, some pruning is done to reduce the number of sen-

[*] This research was conducted during the author's internship at Microsoft Research.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 403–411,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

tence pairs that need to be classified.

While these methods have been applied to news corpora and web pages, very little attention has been given to Wikipedia as a source of parallel sentences. This is surprising, given that Wikipedia contains annotated article alignments, and much work has been done on extracting bilingual lexicons on this dataset. Adafre and de Rijke (2006) extracted similar sentences from Wikipedia article pairs, but only evaluated precision on a small number of extracted sentences.

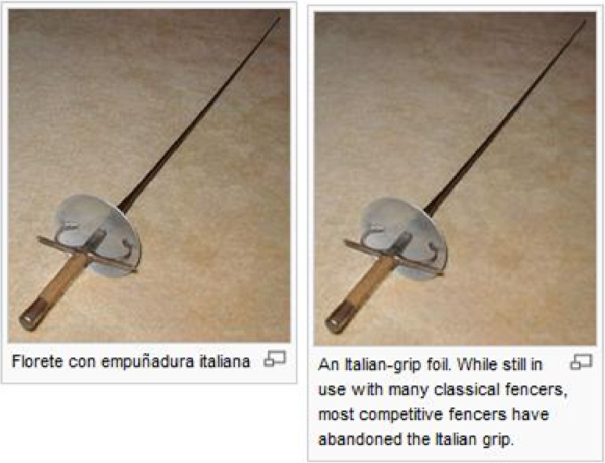
In this paper, we more thoroughly investigate Wikipedia's viability as a comparable corpus, and describe novel methods for parallel sentence extraction. Section 2 describes the multilingual resources available in Wikipedia. Section 3 gives further background on previous methods for parallel sentence extraction on comparable corpora, and describes our approach, which finds a global sentence alignment between two documents. In Section 4, we compare our approach with previous methods on datasets derived from Wikipedia for three language pairs (Spanish-English, German-English, and Bulgarian-English), and show improvements in downstream SMT performance by adding the parallel data we extracted.

2 Wikipedia as a Comparable Corpus

Wikipedia (Wikipedia, 2004) is an online collaborative encyclopedia available in a wide variety of languages. While the English Wikipedia is the largest, with over 3 million articles, there are 24 language editions with at least 100,000 articles.

Articles on the same topic in different languages are also connected via "interwiki" links, which are annotated by users. This is an extremely valuable resource when extracting parallel sentences, as the document alignment is already provided. Table 1 shows how many of these "interwiki" links are present between the English Wikipedia and the 16 largest non-English Wikipedias.

Wikipedia's markup contains other useful indicators for parallel sentence extraction. The many hyperlinks found in articles have previously been used as a valuable source of information. (Adafre and de Rijke, 2006) use matching hyperlinks to identify similar sentences. Two links match if the articles they refer to are connected by an "interwiki" link. Also, images in Wikipedia are often stored in a central source across different languages; this allows identification of captions which may be parallel (see Figure 1). Finally, there are other minor forms of markup which may be useful for finding similar content across languages, such as lists and section headings. In Section 3.3, we will explain how features are derived from this markup.



Florete con empuñadura italiana

An Italian-grip foil. While still in use with many classical fencers, most competitive fencers have abandoned the Italian grip.

Figure 1: Captions for an image of a foil in English and Spanish

2.1 Types of Non-Parallel Corpora

Fung and Cheung (2004) give a more fine-grained description of the types of non-parallel corpora, which we will briefly summarize. A *noisy parallel corpus* has documents which contain many parallel sentences in roughly the same order. *Comparable corpora* contain topic aligned documents which are not translations of each other. The corpora Fung and Cheung (2004) examine are *quasi-comparable*: they contain bilingual documents which are not necessarily on the same topic.

Wikipedia is a special case, since the aligned article pairs may range from being almost completely parallel (e.g., the Spanish and English entries for "Antiparticle") to containing almost no parallel sentences (the Spanish and English entries for "John Calvin"), despite being topic-aligned. It is best characterized as a mix of noisy parallel and comparable article pairs. Some Wikipedia authors will translate articles from another language; others

French	German	Polish	Italian	Dutch	Portuguese	Spanish	Japanese
496K	488K	384K	380K	357K	323K	311K	252K
Russian	Swedish	Finnish	Chinese	Norwegian	Volapük	Catalan	Czech
232K	197K	146K	142K	141K	106K	103K	87K

Table 1: Number of aligned bilingual articles in Wikipedia by language (paired with English).

write the content themselves. Furthermore, even articles created through translations may later diverge due to independent edits in either language.

3 Models for Parallel Sentence Extraction

In this section, we will focus on methods for extracting parallel sentences from aligned, comparable documents. The related problem of automatic document alignment in news and web corpora has been explored by a number of researchers, including Resnik and Smith (2003), Munteanu and Marcu (2005), Tillmann and Xu (2009), and Tillmann (2009). Since our corpus already contains document alignments, we sidestep this problem, and will not discuss further details of this issue. That said, we believe that our methods will be effective in corpora without document alignments when combined with one of the aforementioned algorithms.

3.1 Binary Classifiers and Rankers

Much of the previous work involves building a binary classifier for sentence pairs to determine whether or not they are parallel (Munteanu and Marcu, 2005; Tillmann, 2009). The training data usually comes from a standard parallel corpus. There is a substantial class imbalance ($O(n)$ positive examples, and $O(n^2)$ negative examples), and various heuristics are used to mitigate this problem. Munteanu and Marcu (2005) filter out negative examples with high length difference or low word overlap (based on a bilingual dictionary).

We propose an alternative approach: we learn a ranking model, which, for each sentence in the *source* document, selects either a sentence in the *target* document that it is parallel to, or "null". This formulation of the problem avoids the class imbalance issue of the binary classifier.

In both the binary classifier approach and the ranking approach, we use a Maximum Entropy classifier, following Munteanu and Marcu (2005).

3.2 Sequence Models

In Wikipedia article pairs, it is common for parallel sentences to occur in clusters. A global sentence alignment model is able to capture this phenomenon. For both parallel and comparable corpora, global sentence alignments have been used, though the alignments were monotonic (Gale and Church, 1991; Moore, 2002; Zhao and Vogel, 2002). Our model is a first order linear chain Conditional Random Field (CRF) (Lafferty et al., 2001). The set of source and target sentences are observed. For each *source* sentence, we have a hidden variable indicating the corresponding *target* sentence to which it is aligned (or null). The model is similar to the discriminative CRF-based word alignment model of (Blunsom and Cohn, 2006).

3.3 Features

Our features can be grouped into four categories.

Features derived from word alignments

We use a feature set inspired by (Munteanu and Marcu, 2005), who defined features primarily based on IBM Model 1 alignments (Brown et al., 1993). We also use HMM word alignments (Vogel et al., 1996) in both directions (*source* to *target* and *target* to *source*), and extract the following features based on these four alignments:[1]

1. Log probability of the alignment
2. Number of aligned/unaligned words
3. Longest aligned/unaligned sequence of words
4. Number of words with fertility 1, 2, and 3+

We also define two more features which are independent of word alignment models. One is a sentence length feature taken from (Moore, 2002),

[1] These are all derived from the one best alignment, and normalized by sentence length.

which models the length ratio between the *source* and *target* sentences with a Poisson distribution. The other feature is the difference in relative document position of the two sentences, capturing the idea that the aligned articles have a similar topic progression.

The above features are all defined on sentence pairs, and are included in the binary classifier and ranking model.

Distortion features

In the sequence model, we use additional distortion features, which only look at the difference between the position of the previous and current aligned sentences. One set of features bins these distances; another looks at the absolute difference between the expected position (one after the previous aligned sentence) and the actual position.

Features derived from Wikipedia markup

Three features are derived from Wikipedia's markup. The first is the number of matching links in the sentence pair. The links are weighted by their inverse frequency in the document, so a link that appears often does not contribute much to this feature's value. The image feature fires whenever two sentences are captions of the same image, and the list feature fires when two sentences are both items in a list. These last two indicator features fire with a negative value when the feature matches on one sentence and not the other.

None of the above features fire on a null alignment, in either the ranker or CRF. There is also a bias feature for these two models, which fires on all non-null alignments.

Word-level induced lexicon features

A common problem with approaches for parallel sentence classification, which rely heavily on alignment models trained from unrelated corpora, is low recall due to unknown words in the candidate sentence-pairs. One approach that begins to address this problem is the use of self-training, as in (Munteanu and Marcu, 2005). However, a self-trained sentence pair extraction system is only able to acquire new lexical items that occur in parallel sentences. Within Wikipedia, many linked article pairs do not contain any parallel sentences, yet contain many words and phrases that are good translations of each other.

In this paper we explore an alternative approach to lexicon acquisition for use in parallel sentence extraction. We build a lexicon model using an approach similar to ones developed for unsupervised lexicon induction from monolingual or comparable corpora (Rapp, 1999; Koehn and Knight, 2002; Haghighi et al., 2008). We briefly describe the lexicon model and its use in sentence-extraction.

The lexicon model is based on a probabilistic model $P(w_t|w_s, T, S)$ where w_t is a word in the target language, w_s is a word in the source language, and T and S are linked articles in the target and source languages, respectively.

We train this model similarly to the sentence-extraction ranking model, with the difference that we are aligning word pairs and not sentence pairs. The model is trained from a small set of annotated Wikipedia article pairs, where for some words in the source language we have marked one or more words as corresponding to the source word (in the context of the article pair), or have indicated that the source word does not have a corresponding translation in the target article. The word-level annotated articles are disjoint from the sentence-aligned articles described in Section 4. The following features are used in the lexicon model:

Translation probability. This is the translation probability $p(w_t|w_s)$ from the HMM word alignment model trained on the seed parallel data. We also use the probability in the other direction, as well as the log-probabilities in the two directions.

Position difference. This is the absolute value of the difference in relative position of words w_s and w_t in the articles S and T.

Orthographic similarity. This is a function of the edit distance between source and target words. The edit distance between words written in different alphabets is computed by first performing a deterministic phonetic translation of the words to a common alphabet. The translation is inexact and this is a promising area for improvement. A similar source of information has been used to create seed lexicons in (Koehn and Knight, 2002) and as part of the feature space in (Haghighi et al., 2008).

Context translation probability. This feature looks at all words occurring next to word w_s in the article S and next to w_t in the article T in a local context window (we used one word to the left and one word to the right), and computes several scoring functions measuring the translation correspondence between the contexts (using the IBM Model 1 trained from seed parallel data). This feature is similar to distributional similarity measures used in previous work, with the difference that it is limited to contexts of words within a linked article pair.

Distributional similarity. This feature corresponds more closely to context similarity measures used in previous work on lexicon induction. For each source headword w_s, we collect a distribution over context positions $o \in \{-2, -1, +1, +2\}$ and context words v_s in those positions based on a count of times a context word occurred at that offset from a headword: $P(o, v_s|w_s) \propto weight(o) \cdot C(w_s, o, v_s)$. Adjacent positions -1 and $+1$ have a weight of 2; other positions have a weight of 1. Likewise we gather a distribution over target words and contexts for each target headword $P(o, v_t|w_t)$. Using an IBM Model 1 word translation table $P(v_t|v_s)$ estimated on the seed parallel corpus, we estimate a cross-lingual context distribution as $P(o, v_t|w_s) = \sum_{v_s} P(v_t|v_s) \cdot P(o, v_s|w_s)$. We define the similarity of a words w_s and w_t as one minus the Jensen-Shannon divergence of the distributions over positions and target words.[2]

Given this small set of feature functions, we train the weights of a log-linear ranking model for $P(w_t|w_s, T, S)$, based on the word-level annotated Wikipedia article pairs. After a model is trained, we generate a new translation table $P_{lex}(t|s)$ which is defined as $P_{lex}(t|s) \propto \sum_{t \in T, s \in S} P(t|s, T, S)$. The summation is over occurrences of the source and target word in linked Wikipedia articles. This new translation table is used to define another HMM word-alignment model (together with distortion probabilities trained from parallel data) for use in the sentence extraction models. Two copies of each feature using the HMM word alignment model are generated: one using the seed data HMM model, and another using this new HMM model.

The training data for Bulgarian consisted of two partially annotated Wikipedia article pairs. For German and Spanish we used the feature weights of the model trained on Bulgarian, because we did not have word-level annotated Wikipedia articles.

4 Experiments

4.1 Data

We annotated twenty Wikipedia article pairs for three language pairs: Spanish-English, Bulgarian-English, and German-English. Each sentence in the *source* language was annotated with possible parallel sentences in the *target* language (the target language was English in all experiments). The pairs were annotated with a quality level: **1** if the sentences contained some parallel fragments, **2** if the sentences were mostly parallel with some missing words, and **3** if the sentences appeared to be direct translations. In all experiments, sentence pairs with quality **2** or **3** were taken as positive examples. The resulting datasets are available at http://research.microsoft.com/en-us/people/chrisq/wikidownload.aspx.

For our seed parallel data, we used the Europarl corpus (Koehn, 2005) for Spanish and German and the JRC-Aquis corpus for Bulgarian, plus the article titles for parallel Wikipedia documents, and translations available from Wiktionary entries.[3]

4.2 Intrinsic Evaluation

Using 5-fold cross-validation on the 20 document pairs for each language condition, we compared the binary classifier, ranker, and CRF models for parallel sentence extraction. To tune for precision/recall, we used minimum Bayes risk decoding. We define the loss $L(\tau, \mu)$ of picking target sentence τ when the correct target sentence is μ as 0 if $\tau = \mu$, λ if $\tau = \text{NULL}$ and $\mu \neq \text{NULL}$, and 1 otherwise. By modifying the null loss λ, the precision/recall trade-off can be adjusted. For the CRF model, we used posterior decoding to make the minimum risk decision rule tractable. As a summary measure of the performance of the models at different levels of recall we use average precision as defined in (Ido

[2] We restrict our attention to words with ten or more occurrences, since rare words have poorly estimated distributions. Also we discard the contribution from any context position and word pair that relates to more than 1,000 distinct source or target words, since it explodes the computational overhead and has little impact on the final similarity score.

[3] Wiktionary is an online collaborative dictionary, similar to Wikipedia.

Language Pair	Binary Classifier			Ranker			CRF		
	Avg Prec	R@90	R@80	Avg Prec	R@90	R@80	Avg Prec	R@90	R@80
English-Bulgarian	75.7	33.9	56.2	76.3	38.8	57.0	**80.6**	**52.9**	**59.5**
English-Spanish	90.4	81.3	87.6	93.4	81.0	84.5	**94.7**	**87.6**	**90.2**
English-German	61.8	9.4	27.5	66.4	25.7	42.4	**78.9**	**52.2**	**54.7**

Table 2: Average precision, recall at 90% precision, and recall at 80% precision for each model in all three language pairs. In these experiments, the Wikipedia features and lexicon features are omitted.

Setting	Ranker			CRF		
	Avg Prec	R@90	R@80	Avg Prec	R@90	R@80
English-Bulgarian						
One Direction	76.3	38.8	57.0	80.6	52.9	59.5
Intersected	78.2	47.9	60.3	79.9	38.8	57.0
Intersected +Wiki	80.8	39.7	68.6	82.1	53.7	62.8
Intersected +Wiki +Lex	89.3	64.4	79.3	**90.9**	**72.0**	**81.8**
English-Spanish						
One Direction	93.4	81.0	84.5	94.7	87.6	90.2
Intersected	94.3	82.4	89.0	95.4	88.5	91.8
Intersected +Wiki	94.5	82.4	89.0	95.6	89.2	92.7
Intersected +Wiki +Lex	95.8	87.4	91.1	**96.4**	**90.4**	**93.7**
English-German						
One Direction	66.4	25.7	42.4	78.9	52.2	54.7
Intersected	71.9	36.2	43.8	80.9	54.0	67.0
Intersected +Wiki	74.0	38.8	45.3	82.4	56.9	**71.0**
Intersected +Wiki +Lex	78.7	46.4	59.1	**83.9**	**58.7**	68.8

Table 3: Average precision, recall at 90% precision, and recall at 80% precision for the Ranker and CRF in all three language pairs. “+Wiki” indicates that Wikipedia features were used, and “+Lex” means the lexicon features were used.

et al., 2006). We also report recall at precision of 90 and 80 percent. Table 2 compares the different models in all three language pairs.

In our next set of experiments, we looked at the effects of the Wikipedia specific features. Since the ranker and CRF are asymmetric models, we also experimented with running the models in both directions and combining their outputs by intersection. These results are shown in Table 3.

Identifying the agreement between two asymmetric models is a commonly exploited trick elsewhere in machine translation. It is mostly effective here as well, improving all cases except for the Bulgarian-English CRF where the regression is slight. More successful are the Wikipedia features, which provide an auxiliary signal of potential parallelism.

The gains from adding the lexicon-based features can be dramatic as in the case of Bulgarian (the CRF model average precision increased by nearly 9 points). The lower gains on Spanish and German may be due in part to the lack of language-specific training data. These results are very promising and motivate further exploration. We also note that this is perhaps the first successful practical application of an automatically induced word translation lexicon.

4.3 SMT Evaluation

We also present results in the context of a full machine translation system to evaluate the potential utility of this data. A standard phrasal SMT system (Koehn et al., 2003) serves as our testbed, using a conventional set of models: phrasal models of source given target and target given source; lexical weighting models in both directions, language model, word count, phrase count, distortion penalty, and a lexicalized reordering model. Given that the extracted Wikipedia data takes the standard form of parallel sentences, it would be easy to exploit this same data in a number of systems.

For each language pair we explored two training conditions. The "Medium" data condition used easily downloadable corpora: Europarl for German-English and Spanish-English, and JRC/Acquis for Bulgarian-English. Additionally we included titles of all linked Wikipedia articles as parallel sentences in the medium data condition. The "Large" data condition includes all the medium data, and also includes using a broad range of available sources such as data scraped from the web (Resnik and Smith, 2003), data from the United Nations, phrase books, software documentation, and more.

In each condition, we explored the impact of including additional parallel sentences automatically extracted from Wikipedia in the system training data. For German-English and Spanish-English, we extracted data with the null loss adjusted to achieve an estimated precision of 95 percent, and for English-Bulgarian a precision of 90 percent. Table 4 summarizes the characteristics of these data sets. We were pleasantly surprised at the amount of parallel sentences extracted from such a varied comparable corpus. Apparently the average Wikipedia article contains at least a handful of parallel sentences, suggesting this is a very fertile ground for training MT systems.

The extracted Wikipedia data is likely to make the greatest impact on broad domain test sets – indeed, initial experimentation showed little BLEU gain on in-domain test sets such as Europarl, where out-of-domain training data is unlikely to provide appropriate phrasal translations. Therefore, we experimented with two broad domain test sets.

First, Bing Translator provided a sample of translation requests along with translations in German-English and Spanish-English, which acted our standard development and test set. Unfortunately no such tagged set was available in Bulgarian-English, so we held out a portion of the large system's training data to use for development and test. In each language pair, the test set was split into a development portion ("Dev A") used for minimum error rate training (Och, 2003) and a test set ("Test A") used for final evaluation.

Second, we created new test sets in each of the three language pairs by sampling parallel sentences from held out Wikipedia articles. To ensure that this test data was clean, we manually filtered the sentence pairs that were not truly parallel and edited them as necessary to improve adequacy. We called this "Wikitest". This test set is available at http://research.microsoft.com/en-us/people/chrisq/wikidownload.aspx. Characteristics of these test sets are summarized in Table 5.

We evaluated the resulting systems using BLEU-4 (Papineni et al., 2002); the results are presented in Table 6. First we note that the extracted Wikipedia data are very helpful in medium data conditions, significantly improving translation performance in all conditions. Furthermore we found that the extracted Wikipedia sentences substantially improved translation quality on held-out Wikipedia articles. In every case, training on medium data plus Wikipedia extracts led to equal or better translation quality than the large system alone. Furthermore, adding the Wikipedia data to the large data condition still made substantial improvements.

5 Conclusions

Our first substantial contribution is to demonstrate that Wikipedia is a useful resource for mining parallel data. The sheer volume of extracted parallel sentences within Wikipedia is a somewhat surprising result in the light of Wikipedia's construction. We are also releasing several valuable resources to the community to facilitate further research: manually aligned document pairs, and an edited test set. Hopefully this will encourage research into Wikipedia as a resource for machine translation.

Secondly, we improve on prior pairwise models by introducing a ranking approach for sentence pair extraction. This ranking approach sidesteps the problematic class imbalance issue, resulting in improved average precision while retaining simplicity and clarity in the models.

Also by modeling the sentence alignment of the articles globally, we were able to show a substantial improvement in task accuracy. Furthermore a

		German	English	Spanish	English	Bulgarian	English
Medium	sentences	924,416	924,416	957,884	957,884	413,514	413,514
	types	351,411	320,597	272,139	247,465	115,756	69,002
	tokens	11,556,988	11,751,138	18,229,085	17,184,070	10,207,565	10,422,415
Large	sentences	6,693,568	6,693,568	7,727,256	7,727,256	1,459,900	1,459,900
	types	1,050,832	875,041	1,024,793	952,161	239,076	137,227
	tokens	100,456,622	96,035,475	155,626,085	137,559,844	29,741,936	29,889,020
Wiki	sentences	1,694,595	1,694,595	1,914,978	1,914,978	146,465	146,465
	types	578,371	525,617	569,518	498,765	107,690	74,389
	tokens	21,991,377	23,290,765	29,859,332	28,270,223	1,455,458	1,516,231

Table 4: Statistics of the training data size in all three language pairs.

		German	English	Spanish	English	Bulgarian	English
Dev A	sentences	2,000	2,000	2,000	2,000	2,000	2,000
	tokens	16,367	16,903	24,571	21,493	39,796	40,503
Test A	sentences	5,000	5,000	5,000	5,000	2,473	2,473
	tokens	42,766	43,929	68,036	60,380	52,370	52,343
Wikitest	sentences	500	500	500	500	516	516
	tokens	8,235	9,176	10,446	9,701	7,300	7,701

Table 5: Statistics of the test data sets.

Language pair	Training data	Dev A	Test A	Wikitest
Spanish-English	Medium	32.6	30.5	33.0
	Medium+Wiki	36.7 (+4.1)	33.8 (+3.3)	39.1 (+6.1)
	Large	39.2	**37.4**	38.9
	Large+Wiki	**39.5** (+0.3)	37.3 (-0.1)	**41.1** (+2.2)
German-English	Medium	28.7	26.6	13.0
	Medium+Wiki	31.5 (+2.8)	29.6 (+3.0)	18.2 (+5.2)
	Large	**35.0**	33.7	17.1
	Large+Wiki	34.8 (-0.2)	**33.9** (+0.2)	**20.2** (+3.1)
Bulgarian-English	Medium	36.9	26.0	27.8
	Medium+Wiki	37.9 (+1.0)	27.6 (+1.6)	37.9 (+10.1)
	Large	**51.7**	**49.6**	36.0
	Large+Wiki	**51.7**(+0.0)	49.4 (-0.2)	**39.5**(+3.5)

Table 6: BLEU scores under various training and test conditions. The first column is from minimum error rate training; the next two columns are on held-out test sets. For training data conditions including extracted Wikipedia sentences, parenthesized values indicate absolute BLEU difference against the corresponding system without Wikipedia extracts.

small sample of annotated articles is sufficient to train these global level features, and the learned classifiers appear very portable across languages. It is difficult to say whether such improvement will carry over to other comparable corpora with less document structure and meta-data. We plan to address this question in future work.

Finally, initial investigations have shown that substantial gains can be achieved by using an induced word-level lexicon in combination with sentence extraction. This helps address modeling word pairs that are out-of-vocabulary with respect to the seed parallel lexicon, while avoiding some of the issues in bootstrapping.

References

S. F Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of EACL*, pages 62–69.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of ACL*.

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

P. Fung and P. Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1051.

W. A Gale and K. W Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, pages 152–157.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL*, pages 771–779.

Roy Bar-Haim Ido, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge.

P. Koehn and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 127–133, Edmonton, Canada, May.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

R. C Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. *Lecture Notes in Computer Science*, 2499:135–144.

D. S Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelpha, Pennsylvania, USA.

R. Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*.

P. Resnik and N. A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

C. Tillmann and J. Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of HLT/NAACL*, pages 93–96.

C. Tillmann. 2009. A Beam-Search extraction algorithm for comparable data. In *Proceedings of ACL*, pages 225–228.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841.

Wikipedia. 2004. Wikipedia, the free encyclopedia. [Online; accessed 20-November-2009].

B. Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 745. IEEE Computer Society.