Large Scale Parallel Document Mining for Machine Translation

Jakob Uszkoreit Jay M. Ponte Ashok C. Popat Moshe Dubiner

Google, Inc.
{uszkoreit,ponte,popat,moshe}@google.com

Abstract

A distributed system is described that reliably mines parallel text from large corpora. The approach can be regarded as cross-language near-duplicate detection, enabled by an initial, low-quality batch translation. In contrast to other approaches which require specialized metadata, the system uses only the textual content of the documents. Results are presented for a corpus of over two billion web pages and for a large collection of digitized public-domain books.

1 Introduction

While the World Wide Web provides an abundance of readily available monolingual text, parallel data is still a comparatively scarce resource, yet plays a crucially important role in training statistical machine translation systems.

We describe an approach to mining document-aligned parallel text to be used as training data for a statistical machine translation system. Previous approaches have focused on rather homogeneous corpora and relied on metadata such as publication dates (Munteanu and Marcu, 2005; Munteanu and Marcu, 2006; Udupa et al., 2009; Do et al., 2009; Abdul-Rauf and Schwenk, 2009) or information about document structure (Resnik and Smith, 2003; Chen and Nie, 2000). In large and unstructured collections of documents such as the Web, however, metadata is often sparse or unreliable. Our approach, in contrast, scales computationally to very large and diverse collections of documents and does not require metadata. It is based solely on the textual contents of the input documents.

Casting the problem as one of cross-language near duplicate detection, we use a baseline machine translation system to translate all input documents into a single language. However, the words and phrases that are most discriminatory for the purposes of information retrieval and duplicate detection are the relatively rare ones, precisely those that are less likely to be translated well by the baseline translation system.

Our approach to circumvent this problem and to avoid the prohibitive quadratic computational complexity of the naive approach of performing a comparison of every possible pair of input documents is similar to previous work in near duplicate detection (Broder, 2000; Henzinger, 2006; Manber, 1994) and noisy data retrieval (Harding et al., 1997).

We use shingles consisting of word n-grams to construct relatively rare features from more common, in-vocabulary words. For each input document, we identify a comparatively small set of candidate pairings with documents sharing at least a certain number of such features. We then perform a more expensive comparison between each document and all documents in its candidate set using lower order n-gram features that would typically be too frequent to be used efficiently in forming candidate pairings, but provide a higher coverage of the scored document pairs. Another important aspect of our approach is that it can be implemented in a highly parallel way, as we describe in the following section.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1101–1109,
Beijing, August 2010

2 System Description

The input is a set of documents from diverse sources such as web pages and digitized books. In a first stage, all documents are independently translated into English using a baseline statistical machine translation system.

We then extract two different sets of n-grams from the translated documents: matching n-grams that are used to construct the candidate sets as well as scoring n-grams used only in the computation of a score for a given pair of documents. This stage generates two indexes: a *forward index* listing all extracted scoring n-grams, indexed by document; and an *inverted index* referencing all documents from which we extracted a given matching n-gram, indexed by n-grams. The inverted index is also used to accumulate global information about scoring n-grams, such as their document frequency, yet for scoring n-grams we do not accumulate a posting list of all documents in which they occur.

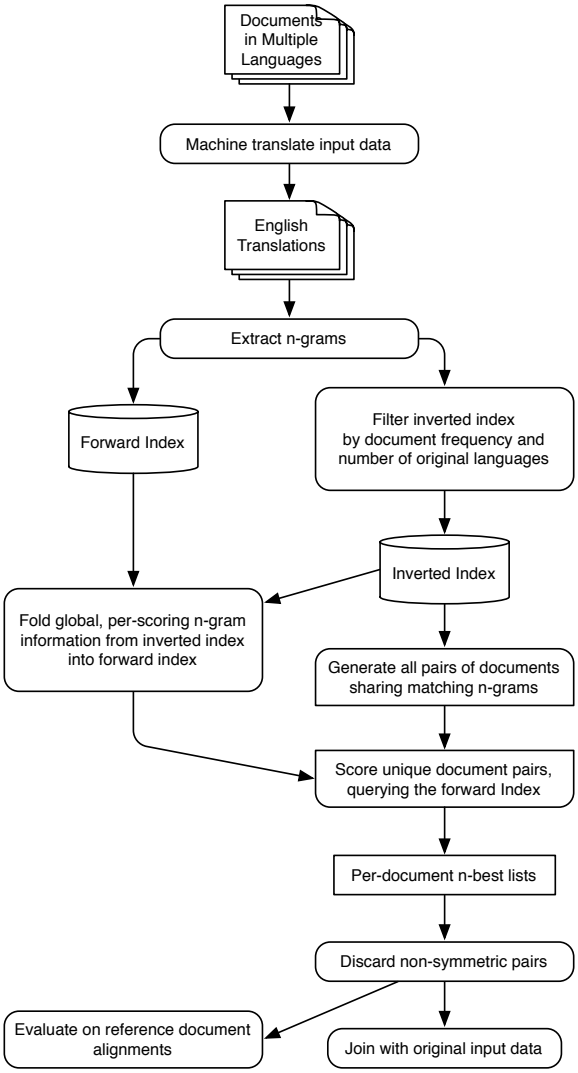

Figure 1: Architecture of the Parallel Text Mining System.

In the next step, the system generates all possible pairs of documents for each matching n-gram posting list in the inverted index. Since we keep only those pairs of documents that originated in different languages, we can discard posting lists from the inverted index that contain only a single document, i.e. those of singleton n-grams, or only documents in a single language.

Crucially, we further discard posting lists for matching n-grams whose frequency exceeds a certain threshold. When choosing a sufficiently large order for the matching n-grams, their long-tailed distribution causes only a small fraction of matching n-grams to be filtered out due to frequency, as we show empirically in Section 5. It is this filtering step that causes the overall runtime of the system to be linear in the size of the input data and allows the system to scale to very large document collections.

In parallel, global information about scoring n-grams accumulated in the inverted index that is required for pairwise scoring, such as their document frequency, is folded into the forward index by iterating over all forward index entries, requesting the respective per-feature quantities from the inverted index and storing them with each occurrence of a scoring n-gram in an updated forward index.

In the next stage, we compute pairwise scores for all candidate document pairs, accessing the forward index entry of each of the two scored documents to obtain the respective scoring n-grams. Document pairs with a score below a given threshold are discarded. For each input document, this results in one n-best list per language. In the last step we retain only those document pairs where each document is contained in the n-best list of the other document for its original language. Finally we perform a *join* of our identified translation pairs with the original text by making another

pass over the original, untranslated input data where the contents of document pairs with sufficiently high scores are then aggregated and output. Document pairings involving all languages are identified simultaneously. Each stage of the system fits well into the MapReduce programming model (Dean and Ghemawat, 2004). The general architecture is shown in Figure 1.

2.1 Pairwise Scoring

For scoring a pair of documents d and d', the forward index is queried for the entries for both documents. Let $F_d = \{f_1, f_2, ...f_n\}$ and $F_{d'} = \{f'_1, f'_2, ...f'_{n'}\}$ be the sets of scoring n-grams in the forward index entries of d and d', respectively. Let $\text{idf}(f) = \log \frac{|D|}{df(f)}$ be the inverse document frequency of a scoring n-gram f, where $|D|$ is the number of documents in the input corpus and $df(f)$ is the number documents from which we extracted the feature f. Interpreting F_d and $F_{d'}$ as incidence vectors in the vector space of n-grams and replacing each non-zero component f with $\text{idf}(f)$, we compute the score of the document pair as the inverse document frequency weighted cosine similarity of F_d and $F_{d'}$

$$\text{score}(d, d') = \frac{F_d \cdot F_{d'}}{||F_d|| \cdot ||F_{d'}||} \quad (1)$$

The per-document n-best lists are sorted according to this score and document pairs for which the score is below a threshold are discarded completely.

We do not use term frequency in the scoring metric. In preliminary experiments, incorporating the term frequency to yield basic *tf/idf* as well as using other information retrieval ranking functions incorporating term frequencies such as *BM25* (Robertson et al., 1995) resulted in a degradation of performance compared to the simpler scoring function described above. We believe this is due to the fact that, in contrast to the standard information retrieval setting, the overall length of our queries is on par with that of the documents in the collection.

The scoring is completely agnostic regarding the scoring n-grams' positions in the documents. Since especially for long documents such as books this may produce spurious matches, we apply an additional filter to remove document pairs for which the relative ordering of the matching scoring n-grams is very different. Together with each scoring n-gram we also extract its relative position in each document and store it in the forward index. When scoring a document pair, we compute the normalized permutation edit distance (Cormode et al., 2001) between the two sequences of overlapping n-grams sorted by their position in the respective document. If this distance exceeds a certain threshold, we discard the document pair.

2.2 Computational Complexity

By limiting the frequency of matching n-grams, the complexity becomes linear. Let the tunable parameter c be the maximum occurrence count for matching n-grams to be kept in the inverted index. Let m be the average number of matching n-grams extracted from a single document whose count is below c and D be the set of documents in the input corpus. Then the system generates up to $|D| \cdot m \cdot c$ candidate pairings. Scoring a given candidate document pair according to cosine similarity involves computing three dot-products between sparse vectors with one non-zero component per scoring n-gram extracted and not filtered from the respective document. Let s be the average number of such scoring n-grams per document, which is bounded by the average document length. Then the time complexity of the entire document alignment is in

$$O(|D| \cdot m \cdot c \cdot s) \quad (2)$$

and therefore linear in the number of input documents in the corpus and the average document size.

The space complexity is dominated by the size of the inverted and forward indexes, both of which are linear in the size of the input corpus.

2.3 Sentence-Level Alignment

Further filtering is performed on a per-sentence basis during per-document-pair sentence alignment of the mined text with a standard dynamic programming sentence alignment algorithm using sentence length and multilingual probabilistic dictionaries as features. Afterwards we crudely align

words within each pair of aligned source and target sentences. This crude alignment is used only to filter nonparallel sentences. Let S be the set of source words, T the set of target words and $S \times T$ the set of ordered pairs. Let the source sentence contain words $S_0 \subset S$ and the target sentence contain words $T_0 \subset T$. An alignment $A_0 \subset S_0 \times T_0$ will be scored by

$$\text{score}(A_0) = \sum_{(s,t) \in A_0} \ln \frac{p(s,t)}{p(s)\, p(t)} \qquad (3)$$

where the joint probabilities $p(s,t)$ and marginal probabilities $p(s)$, $p(t)$ are taken to be the respective empirical distributions (without smoothing) in an existing word aligned corpus. This is greedily maximized and the result is divided by its approximate expected value

$$\sum_{(s,t) \in S_0 \times T} \frac{p(s,t)}{p(s)} \ln \frac{p(s,t)}{p(s)\, p(t)} \qquad (4)$$

We discard sentence pairs for which the ratio between the actual and the expected score is less than $1/3$. We also drop sentence pairs for which both sides are identical, or a language detector declares them to be in the wrong language.

2.4 Baseline Translation System

To translate the input documents into English we use phrase-based statistical machine translation systems based on the log-linear formulation of the problem (Och and Ney, 2002).

We train the systems on the Europarl Corpus (Koehn, 2002), the DGT Multilingual Translation Memory (European Commission Directorate-General for Translation, 2007) and the United Nations ODS corpus (United Nations, 2006). Minimum error rate training (Macherey et al., 2008) under the BLEU criterion is used to optimize the feature function weights on development data consisting of the *nv-dev2007* and *news-dev2009* data sets provided by the organizers of the 2007 and 2009 WMT shared translation tasks[1]. We use a 4-gram language model trained on a variety of large monolingual corpora. The BLEU scores of our baseline translation system on the test sets from various WMT shared translation tasks are listed in Table 5. An empirical analysis of the impact of the baseline translation system quality on the data mining system is given in Section 6.3.

[1] available at `http://statmt.org`

3 Input Document Collections

We evaluate the parallel text mining system on two input data sets:

web A collection of 2.5 Billion general pages crawled from the Web, containing only pages in Czech, English, French, German, Hungarian and Spanish

books A collection of 1.5 Million public domain books digitized using an optical character recognition system. The collection consists primarily of English, French and fewer Spanish volumes

3.1 Reference Sets

We created reference sets of groups of documents in multiple languages which are true translations of one another for both the *web* and the *books* data set. Due to the presence of duplicates, each reference pairing can contain more than a single alternative translation per language. The *web* reference set was constructed by exploiting the systematic hyperlink structure of the web-site `http://america.gov/`, that links pages in one language to their respective translations into one or more other languages. The resulting reference set contains documents in Arabic, Chinese, English, French, Russian and Spanish, however, for most English pages there is only one translation into one of the other languages. Overall, the reference set contains 6,818 documents and 7,286 translation pairs.

The *books* reference set contains 30 manually aligned groups of translations covering a total of 103 volumes in English and French.

4 Evaluation Metrics

The fact that the system outputs pairs of documents and the presence of duplicate documents in the corpus motivate the use of modified versions of *precision* and *recall*.

Let C be a set of candidate parallel document pairs and let R be a possibly incomplete reference set of groups of parallel documents known to exist in the corpus. Consider the following two subsets of C:

- *Matching* pairs which are in some reference cluster.
- *Touching* pairs which are non-matching but have at least one document in some reference cluster.

We define

$$\text{Precision} = \frac{|C_{\text{Matching}}|}{|C_{\text{Matching}}| + |C_{\text{Touching}}|}$$

and

$$\text{Recall} = \frac{|C_{\text{Matching}}|}{|R|} \quad (5)$$

5 Parameter Selection

We conducted a series of small-scale experiments on only those documents contained in the *web* reference data set to empirically determine good settings for the tunable parameters of the text mining system. Among the most important parameters are the orders of the n-grams used for pairing documents as well as scoring them. Aside from the obvious impact on the quality of the output, these parameters have a very large influence on the overall computational performance of the system. The choice of the order of the extracted matching n-grams is mainly a trade-off between recall and efficiency. If the order is too large the system will miss valid pairs; if too small the the threshold on matching n-gram frequency will need to be increased.

Figure 2 shows the F1-scores obtained running only on the documents contained in the *web* reference set with different orders of matching and scoring n-grams. Figure 3 shows the corresponding number of pairwise comparisons made when using different orders of matching n-grams. While there is a drop of 0.01 in F1 score between using 2-grams and 5-grams as matching n-grams, this drop in quality seems to be well worth the 42-fold reduction in resulting pairwise comparisons.

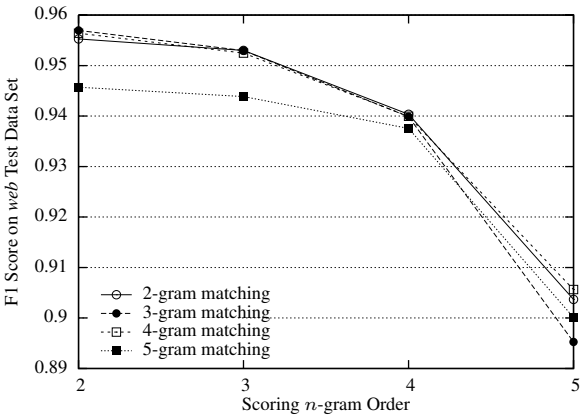

Figure 2: F1 scores on the *web* reference set for different scoring and matching n-gram orders.

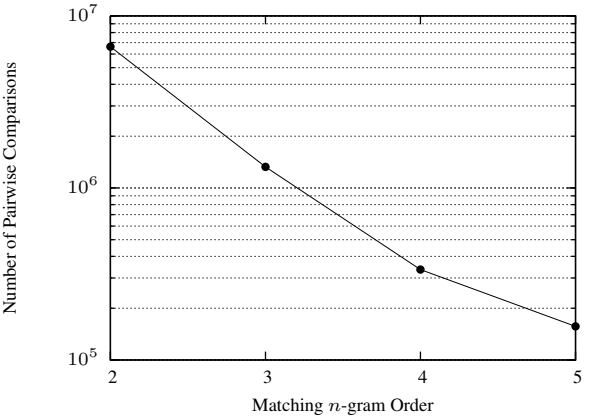

Figure 3: Number of pairwise comparisons made when using matching n-grams of different orders.

The largest portion of the loss in F1 score is incurred when increasing the matching n-gram order from 4 to 5, the reduction in pairwise comparisons, however, is still more than twofold.

Table 1 shows the precision and recall on the *web* reference set when running only on documents in the reference set using 5-grams as matching n-grams and bigrams for scoring for different values of the threshold on the cosine similarity score. In this setting as well as in large-scale experiments on both complete data sets described in section 6.1, a threshold of 0.1 yields the highest F1 score.

score threshold	0.06	0.10	0.12	0.16	0.20
precision	0.92	0.97	0.98	0.99	0.99
recall	0.91	0.91	0.90	0.89	0.83

Table 1: Precision and recall on the *web* reference set when running only on documents contained in the reference set.

6 Evaluation

We run the parallel text mining system on the *web* and *books* data sets using 5-grams for matching and bigrams for scoring. In both cases we discard matching n-grams which occurred in more than 50 documents and output only the highest scoring candidate for each document.

In case of the *web* data set, we extract every 5-gram as potential matching feature. For the *books* data set, however, we downsample the number of candidate matching 5-grams by extracting only those whose integer fingerprints under some hash function have four specific bits set, thus keeping on average only $1/16$ of the matching n-grams. Here, we also restrict the total number of matching n-grams extracted from any given document to 20,000. Scoring bigrams are dropped from the forward index if their document frequency exceeds 100,000, at which point their influence on the pairwise score would be negligible.

Running on the *web* data set, the system on average extracts 250 matching 5-grams per document, extracting a total of approximately 430 Billion distinct 5-grams. Of those, 78% are singletons and 21% only occur in a single language. Only approximately 0.8% of all matching n-grams are filtered due to having a document frequency higher than 50. The forward index initially contains more than 500 Billion bigram occurrences; after pruning out singletons and bigrams with a document frequency larger than 100,000, the number of indexed scoring feature occurrences is reduced to 40%. During scoring, approximately 50 Billion pairwise comparisons are performed.

In total the n-gram extraction, document scoring and subsequent filtering takes less than 24 hours on a cluster of 2,000 state-of-the-art CPUs.

The number of words after sentence-level filtering and alignment that the parallel text mining system extracted for the different languages from each dataset are listed in Table 2.

	baseline	*books*	*web*
Czech	27.5 M	0	271.9 M
French	479.8 M	228.5 M	4,914.3 M
German	54.2 M	0	3,787.6 M
Hungarian	26.9 M	0	198.9 M
Spanish	441.0 M	15.0 M	4,846.8 M

Table 2: The number of words per language in the baseline training corpora and extracted from the two different data sets.

score threshold	0.06	0.10	0.12	0.16	0.20
precision	0.88	0.93	0.95	0.97	0.97
recall	0.68	0.65	0.63	0.52	0.38

Table 3: Precision and recall on the reference set when running on the complete *web* data set with different score thresholds.

score threshold	0.06	0.10	0.12	0.16	0.20
precision	0.95	1.00	1.00	1.00	1.00
recall	0.71	0.71	0.71	0.48	0.38

Table 4: Precision and recall on the reference set when running on the complete *books* data set with different score thresholds.

6.1 Precision and Recall

Tables 3 and 4 show precision and recall on the respective reference sets for the *web* and the *books* input data sets. While the text mining system maintains a very high precision, recall drops significantly compared to running only on the documents in the reference set. One reason for this behavior is that the number of n-grams in the test data set which are sufficiently rare to be used as queries drops with increasing amounts of input data and in particular short documents which only share a small number of matching n-grams anyway, may happen to only share matching n-grams with a too high document frequency. Further analysis shows that another, more significant factor is the existence of multiple, possibly partial translations and near-duplicate documents which cause symmetrization to discard valid document pairs because each document in the pair is determined by the document pair score to be more similar to a different translation of a near-duplicate or sub-

Language Pair	Training Data	WMT 2007 news commentary	WMT 2008 news	WMT 2009 news
Czech English	baseline	21.59	14.59	16.46
	web	29.26 (+7.67)	20.16 (+5.57)	23.25 (+6.76)
German English	baseline	27.99	20.34	20.03
	web	32.35 (+4.36)	23.22 (+2.88)	23.35 (+3.32)
Hungarian English	baseline	-	10.21	11.02
	web	-	12.92 (+2.71)	14.68 (+3.66)
French English	baseline	34.26	22.14	26.39
	books	34.73 (+0.47)	22.39 (+0.25)	27.15 (+0.76)
	web	36.65 (+2.39)	23.22 (+1.08)	28.34 (+1.95)
Spanish English	baseline	43.67	24.15	26.88
	books	44.07 (+0.40)	24.32 (+0.17)	27.16 (+0.28)
	web	46.21 (+2.54)	25.52 (+1.37)	28.50 (+1.62)
English Czech	baseline	14.78	12.45	11.62
	web	20.65 (+5.86)	18.70 (+6.25)	16.60 (+4.98)
English German	baseline	19.89	14.67	14.31
	web	23.49 (+3.60)	16.78 (+2.11)	16.96 (+2.65)
English Hungarian	baseline	-	07.93	08.52
	web	-	10.16 (+2.23)	11.42 (+2.90)
English French	baseline	31.59	22.29	25.14
	books	31.92 (+0.33)	22.42 (+0.13)	25.46 (+0.32)
	web	34.35 (+2.76)	23.56 (+1.27)	27.05 (+1.91)
English Spanish	baseline	42.05	24.65	25.85
	books	42.05	24.79 (+0.14)	26.07 (+0.22)
	web	45.21 (+3.16)	26.46 (+1.81)	27.79 (+1.94)

Table 5: BLEU scores of the translation systems trained on the automatically mined parallel corpora and the baseline training data.

set of the document. This problem seems to affect news articles in particular where there are often multiple different translations of large subsets of the same or slightly changed versions of the article.

6.2 Translation Quality

Arabic English	NIST 2006	NIST 2008
Baseline (UN ODS)	44.31	42.79
Munteanu and Marcu	45.13	43.86
Present work	44.72	43.64
Chinese English	NIST 2006	NIST 2008
Baseline (UN ODS)	25.71	19.79
Munteanu and Marcu	28.11	21.69
Present work	28.08	22.02

Table 6: BLEU scores of the Chinese and Arabic to English translation systems trained on the baseline UN ODS corpus and after adding either the Munteanu and Marcu corpora or the training data mined using the presented approach.

We trained a phrase-based translation system on the mined parallel data sets and evaluated it on translation tasks for the language pairs Czech, French, German, Hungarian and Spanish to and from English, measuring translation quality with the BLEU score (Papineni et al., 2002). The translation tasks evaluated are the WMT 2007 news commentary test set as well the WMT 2008 and 2009 news test sets.

The parallel data for this experiment was mined using the general settings described in the previous section and a threshold of 0.1 on the pairwise score. We ensure that the test data is not included in the training data by filtering out all sentences from the training data that share more than 30% of their 6-grams with any sentence from one of the test corpora.

Table 5 shows the BLEU scores of the different translation systems. The consistent and significant improvements in BLEU score demonstrate the usefulness of the mined document pairs in training a translation system.

Even though the presented approach works on a less granular level than the sentence-level approach of Munteanu and Marcu (2005), we compare results on the same input data[2] used by those authors to automatically generate the

[2]LDC corpora LDC2005T12, LDC2005T14 and LDC2006T02, the second editions of the Arabic, Chinese and English Gigaword corpora.

Sampling Rate	WMT 2007 news commentary			WMT 2008 news			WMT 2009 news		
	degraded	Cz→En	En→Cz	degraded	Cz→En	En→Cz	degraded	Cz→En	En→Cz
1.0	21.59	29.26	20.65	14.59	20.16	18.70	16.46	23.25	16.60
0.5	20.12	29.16	20.55	13.65	20.16	18.71	15.44	23.16	16.56
0.25	18.59	29.09	20.61	12.79	20.09	18.58	14.35	23.18	16.50
0.125	16.69	29.10	20.39	11.87	20.07	18.48	13.05	23.06	16.53
0.0625	14.72	29.04	20.44	10.87	20.06	18.49	11.62	23.11	16.44
0.0312	12.60	28.75	20.28	09.71	19.97	18.45	10.43	23.04	16.41

Table 7: BLEU scores of the degraded Czech to English baseline systems used for translating Czech documents from the *web* data set as well as those of Czech to and from English systems trained on data mined using translations of varying quality created by sampling from the training data.

Arabic English and Chinese English sentence-aligned parallel LDC corpora LDC2007T08 and LDC2007T09. We trained Arabic and Chinese English baseline systems on the United Nations ODS corpus (United Nations, 2006); we also use these to translate the non-English portions of the input data to English. We then evaluate the effects of also training on either the LDC2007T08 and LDC2007T09 corpora or the parallel documents mined by our approach in addition to the United Nations ODS corpus on the NIST 2006 and 2008 MT evaluation test sets. The results are presented in Table 6.

The approach proposed in (Munteanu and Marcu, 2005) relies critically on the existence of publication dates in order to be computationally feasible, yet it still scales superlinearly in the amount of input data. It could therefore not easily be applied to much larger and less structured input data collections. While our approach neither uses metadata nor operates on the sentence level, in all but one of the tasks, the system trained on the data mined using our approach performs similarly or slightly better.

6.3 Impact of Baseline Translation Quality

In order to evaluate the impact of the translation quality of the baseline system on the quality of the mined document pairs, we trained artificially degraded Czech to English translation systems by sampling from the baseline training data at decreasing rates. We translate the Czech subset of the *web* document collection into English with each of the degraded systems and apply the parallel data mining system in the same configuration.

Table 7 shows the BLEU scores of the degraded baseline systems and those resulting from adding the different mined data sets to the non-degraded Czech English and English Czech systems. Degrading the input data translation quality by up to 8.9% BLEU results in a consistent but only comparatively small decrease of less than 0.6% BLEU in the scores obtained when training on the mined document pairs. This does not only show that the impact of variations of the baseline system quality on the data mining system is limited, but also that the data mining system will already work with a rather low quality baseline system.

7 Conclusion

We presented a scalable approach to mining parallel text from collections of billions of documents with high precision. The system makes few assumptions about the input documents. We demonstrated that it works well on different types of data: a large collection of web pages and a collection of digitized books. We further showed that the produced parallel corpora can significantly improve the quality of a state-of-the-art statistical machine translation system.

8 Acknowledgments

We thank the anonymous reviewers for their insightful comments.

References

Abdul-Rauf, Sadaf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *EACL*, pages 16–23.

Broder, Andrei Z. 2000. Identifying and filtering near-duplicate documents. In *COM '00: Proceedings of the 11th Annual Symposium on Combinatorial Pat-*

tern Matching, pages 1–10, London, UK. Springer-Verlag.

Chen, Jiang and Jian-Yun Nie. 2000. Parallel web text mining for cross-language IR. In *In In Proc. of RIAO*, pages 62–77.

Cormode, Graham, S. Muthukrishnan, and Süleyman Cenk Sahinalp. 2001. Permutation editing and matching via embeddings. In *ICALP '01: Proceedings of the 28th International Colloquium on Automata, Languages and Programming,*, pages 481–492, London, UK. Springer-Verlag.

Dean, Jeffrey and Sanjay Ghemawat. 2004. MapReduce: Simplified data processing on large clusters. In *Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI-04)*, San Francisco, CA, USA.

Do, Thi-Ngoc-Diep, Viet-Bac Le, Brigitte Bigi, Laurent Besacier Eric, and Castelli. 2009. Mining a comparable text corpus for a Vietnamese - French statistical machine translation system. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, pages 165–172, Athens, Greece, March.

European Commission Directorate-General for Translation. 2007. DGT-TM parallel corpus. `http://langtech.jrc.it/DGT-TM.html`.

Harding, Stephen M., W. Bruce Croft, and C. Weir. 1997. Probabilistic retrieval of OCR degraded text using n-grams. In *ECDL '97: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 345–359, London, UK. Springer-Verlag.

Henzinger, Monika. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 284–291, New York, NY, USA. ACM.

Koehn, Philipp. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Draft.

Macherey, Wolfgang, Franz Och, Ignacio Thayer, and Jakob Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 725–734, Honolulu, Hi, October. Association for Computational Linguistics.

Manber, Udi. 1994. Finding similar files in a large file system. In *Proceedings of the USENIX Winter 1994 Technical Conferenc*.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Comput. Linguist.*, 31(4):477–504.

Munteanu, Dragos Stefan and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *ACL*.

Och, Franz Josef and Hermann 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)*, pages 295–302, Philadelphia, PA, USA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing 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)*, pages 311–318, Philadelphia, PA, USA.

Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

Robertson, S E, S Walker, S Jones, M M Hancock-Beaulieu, and M Gatford. 1995. Okapi at TREC–3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*.

Udupa, Raghavendra, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *EACL*, pages 799–807.

United Nations. 2006. ODS UN parallel corpus. `http://ods.un.org/`.