Mining for Domain-specific Parallel Text from Wikipedia

Plamada, Magdalena ; Volk, Martin

Originally published at:
Plamada, Magdalena; Volk, Martin (2013). Mining for Domain-specific Parallel Text from Wikipedia. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora, Sofia, Bulgaria, August 2013 - August 2013, 112-120.

Mining for Domain-specific Parallel Text from Wikipedia

Magdalena Plamadă, Martin Volk
Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zurich
{plamada, volk}@cl.uzh.ch

Abstract

Previous attempts in extracting parallel data from Wikipedia were restricted by the monotonicity constraint of the alignment algorithm used for matching possible candidates. This paper proposes a method for exploiting Wikipedia articles without worrying about the position of the sentences in the text. The algorithm ranks the candidate sentence pairs by means of a customized metric, which combines different similarity criteria. Moreover, we limit the search space to a specific topical domain, since our final goal is to use the extracted data in a domain-specific Statistical Machine Translation (SMT) setting. The precision estimates show that the extracted sentence pairs are clearly semantically equivalent. The SMT experiments, however, show that the extracted data is not refined enough to improve a strong in-domain SMT system. Nevertheless, it is good enough to boost the performance of an out-of-domain system trained on sizable amounts of data.

1 Introduction

A high-quality Statistical Machine Translation (SMT) system can only be built with large quantities of parallel texts. Moreover, systems specialized in specific domains require in-domain training data. A well-known problem of SMT systems is that existing parallel corpora cover a small percentage of the possible language pairs and very few domains. We therefore need a language-independent approach for discovering parallel sentences in the available multilingual resources.

This idea was explored intensively in the last decade with different text sources, generically called comparable corpora, such as news feeds, encyclopedias or even the entire Web. The first approaches focused merely on news corpora and were either based on IBM alignment models (Zhao and Vogel, 2002; Fung and Cheung, 2004) or employing machine learning techniques (Munteanu and Marcu, 2005; Abdul Rauf and Schwenk, 2011).

The multilingual Wikipedia is another source of comparable texts, not yet thoroughly explored. Adafre and de Rijke (2006) describe two methods for identifying parallel sentences across it based on monolingual sentence similarity (MT and respectively, lexicon based). Fung et al. (2010) approach the problem by combining recall- and precision-oriented methods for sentence alignment, such as the DK-vec algorithm or algorithms based on cosine similarities. Both approaches have achieved good results in terms of precision and recall.

However, we are interested in real application scenarios, such as SMT systems. The following approaches report significant performance improvements when using the extracted data as training material for SMT: Smith et al. (2010) use a maximum entropy-based classifier with various feature functions (e.g. alignment coverage, word fertility, translation probability, distortion). Ştefănescu et al. (2012) propose an algorithm based on cross-lingual information retrieval, which also considers similarity features equivalent to the ones mentioned in the previous paper.

The presented approaches extract general purpose sentences, but we are interested in a specific topical domain. We have previously tackled the problem (Plamada and Volk, 2012) and encountered two major bottlenecks: the alignment algorithm for matching possible candidates and the similarity metric used to compare them. To our knowledge, existing sentence alignment algorithms (including the one we have employed in the first place) have a monotonic order constraint, meaning that crossing alignments are not

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 112–120,
Sofia, Bulgaria, August 8, 2013.

allowed. But this phenomenon occurs often in Wikipedia, because its articles in different languages are edited independently, without respecting any guidelines. Moreover, the string-based comparison metric proved to be unreliable for identifying parallel sentences.

In this paper we propose an improved approach for selecting parallel sentences in Wikipedia articles which considers all possible sentence pairs, regardless of their position in the text. The selection will be made by means of a more informed similarity metric, which rates different aspects concerning the correspondences between two sentences. Although the approach is language and domain-independent, the present paper reports results obtained through querying the German and French Wikipedia for Alpine texts (i.e. mountaineering reports, hiking recommendations, articles on the biology and the geology of mountainous regions). Moreover, we report preliminary results regarding the use of the extracted corpus for SMT training.

2 Finding candidate articles

The general architecture of our parallel sentence extraction process is shown in Figure 1. We applied the approach only to the language pair German-French, as these are the languages we have expertise in. In the project Domain-specific Statistical Machine Translation[1] we have built an SMT system for the Alpine domain and for this language pair. The training data comes from the Text+Berg corpus[2], which contains the digitized publications of the Swiss Alpine Club (SAC) from 1864 until 2011, in German and French. This SMT system will generate the automatic translations required in the extraction process.

The input consists of German and French Wikipedia dumps[3], available in the MediaWiki format [4]. Therefore our workflow requires a preprocessing step, where the MediaWiki contents are transformed to XML and then to raw text. Preprocessing is based on existing tools, such as WikiPrep[5], but further customization is needed in order to correctly convert localized MediaWiki elements (namespaces, templates, date and number formats etc.). We then identify Wikipedia articles

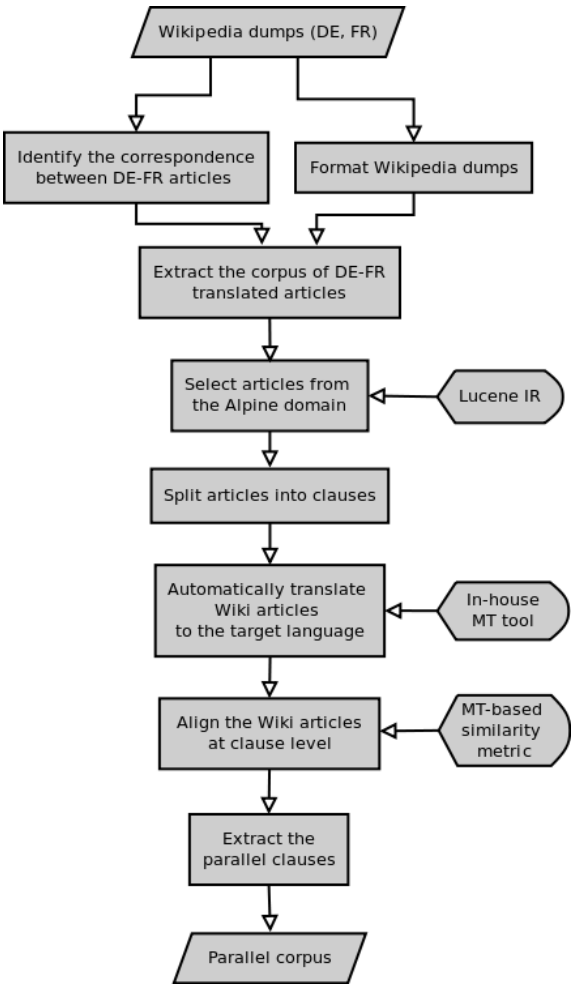

Figure 1: The extraction workflow

available in both languages by means of the interlanguage links provided by Wikipedia itself. This reliable information is a good basis for the extraction workflow, since we do not have to worry about the document alignment.

Upon completion of this step, we have extracted a bilingual corpus of approximately 400 000 articles per language. The corpus is subsequently used for information retrieval (IR) queries aiming to identify the articles belonging to the Alpine domain. The input queries contain the 100 most frequent mountaineering keywords in the Text+Berg corpus (e.g. *Alp, Gipfel, Berg, Route* in German and *montagne, sommet, voie, cabane* in French). This filter reduces the search space to 40 000 articles. Although we have refined our search terms by discarding the ones occurring frequently in other text types (e.g. *meter, day, year, end*), we were not able to avoid a small percentage of false positives. Once we extract the Alpine comparable corpus, we proceed to the extraction of

[1] http://www.cl.uzh.ch/research_en.html
[2] See www.textberg.ch
[3] Accessed in September 2011
[4] http://www.mediawiki.org/wiki/MediaWiki
[5] http://sourceforge.net/projects/wikiprep/

parallel sentences, which will be thoroughly discussed in the following section. See (Plamada and Volk, 2012) for more details about the extraction pipeline.

3 Finding parallel segments in Wikipedia articles

The analysis of our previous results brought into attention many "parallel" sentence pairs of different lengths, meaning that the shared translated content does not span over the whole sentence. As an example, consider the following sentences which have been retrieved by the extraction pipeline. Although they both contain information about the valleys connected by the Turini pass, the German sentence contains a fragment about its position, which has not been translated into French.

DE: Der Pass liegt in der äusseren, besiedelten Zone des Nationalpark Mercantour und stellt den Übergang zwischen dem Tal der Bévéra und dem Tal der Vésubie dar.

FR: Le col de Turini relie la vallée de la Vésubie à la vallée de la Bévéra.

If this sentence pair would be used for MT training, it would most probably confuse the system, because noisy word alignments are to be expected. Our solution to this problem is to split the sentences into smaller entities (e.g. clauses) and then to find the alignments on this granularity level. The clause boundary detection is performed independently for German and French, respectively, following the approach developed by Volk (2001). The general idea is to split the sentences into clauses containing a single full verb.

Our alignment algorithm, unlike previous approaches, ignores the position of the clauses in the texts. Although Wikipedia articles are divided into sections, their structure is not synchronized across the language variants, since articles are edited independently. We have encountered, for example, cases where one section in the French article was included in the general introduction of the German article. If we would have considered sections boundaries as anchor points, we would have missed useful clause pairs. We therefore decided to use an exhaustive matching algorithm, in order to cover all possible combinations.

For the sake of simplicity, we reduce the problem to a monolingual alignment task by using an intermediary machine translation of the source article. We decided that German articles should always be considered the source because we expect a better automatic translation quality from German to French. The translation is performed by our in-house SMT system trained on Alpine texts. The algorithm generates all possible clause pairs between the automatic translation and the targeted article and computes for each of them a similarity score. Subsequently it reduces the search space by keeping only the 3 best-scoring alignment candidates for each clause. Finally the algorithm returns the alignment pair which maximizes the similarity score and complies with the injectivity constraint. In the end we filter the results by allowing only clause pairs above the set threshold.

We defined our similarity measure as a weighted sum of feature functions, which returns values in the range [0,1]. The similarity score models two comparison criteria:

- **METEOR score**
  We used the METEOR similarity metric because, unlike other string-based metrics (e.g. BLEU (Papineni et al., 2002)), it considers not only exact matches, but also word stems, synonyms, and paraphrases (Denkowski and Lavie, 2011). Suppose that we compute the similarity between the following sentences in French: *j' aimerais bien vous voir* and *je voudrais vous voir* (both meaning *I would like to see you*). BLEU, which is a string-based metric, would assign a similarity score of 52.5. This value could hardly be considered reliable, given that the sentence *ta voiture vous voir* (paired with the first sentence) would get the same BLEU score, although the latter sentence (EN: your car see you) is obviously nonsense. On the other hand, METEOR would return a score of 90.3 for the original sentence pair, since it can appreciate that the two pronouns (*je* and *j'*) are both variations of the first person singular in French and that the predicates convey the same meaning.

- **Number of aligned content words**
  However, METEOR scores can also be misleading, since they rely on automatic word alignments. Two sentences are likely to receive a high similarity score when they share many aligned words. However, the alignments are not always reliable. We often saw

sentence pairs with a decent Meteor score where only some determiners, punctuation signs or simple word collocations (e.g. *de la montagne* (EN: of the mountain)) were matched. As an illustration, consider the following sentence pair and its corresponding alignment:

Hyp: les armoiries , le désir de la ville de breslau par ferdinand i. le 12 mars 1530 a

Ref: le 19 juin 1990 , le conseil municipal rétablit le blason original de la ville

```
2-4 3-5 5-12 6-13 7-14 13-0
```

Although the sentences are obviously not semantically equivalent (a fact also suggested by the sparse word alignments), the pair receives a METEOR score of 0.23. We decided to compensate for this by counting only the aligned pairs which link content words and dividing them by the total number of words in the longest sentence from the considered pair. In the example above, only one valid alignment (7-14) can be identified, therefore the sentence pair will get a partial score of 1/18. In this manner we can ensure the decrease of the initial similarity score.

Additionally, we have defined a token ratio feature to penalize the sentence length differences. Although a length penalty is already included in the METEOR score, we still found false candidate pairs with exceedingly different lengths. Therefore we decided to use this criterion as a selection filter rather than including it in the similarity function, in order to increase the chances of other candidates with similar length. Even if no other candidate will pass all the filters, at least we expect the precision to increase, since we will have one false positive less.

The final formula for the similarity score between two clauses *src* in the source language and, respectively *trg* in the target language is:

$$score(src, trg) = w_1 * s_1 + (1 - w_1) * s_2 \quad (1)$$

where s_1 represents the METEOR score and s_2 the alignment score.

The weights, as well as the final threshold are tuned to maximize the correlation with human judgments. We modeled the task as a minimization problem, where the function value increases by 1 for each correctly selected clause pair and decreases by 1 for each wrong pair. The solution (consisting of the individual weights and the threshold) is found using a brute force approach, for which we employed the `scipy.optimize` package from Python. The training set consists of an article with 1300 clause pairs, 25 of which are parallel and the rest non-parallel. We chose this distribution of the useful/not useful clauses because this corresponds to the real distribution observed in Wikipedia articles. In the best configuration, we retrieve 23 good clause pairs and 1 wrong. This corresponds to a precision of 95% and a recall of 92% on this small test set.

However, we can influence the quantity of extracted parallel clauses by manually adjusting the final filter threshold. Figure 2 depicts the size variations of the resulting corpus at different thresholds, where the relative frequency is represented on a logarithmic scale. We notice that the rate of decrease is linear in the log scale of the number of extracted clause pairs. We start at a similarity score of 0.2 because the pairs below this threshold are too noisy. The data between 0.2 and 0.3 is already mixed, as it will be shown in the following sections. However, since this data segment contains approximately twice as much data as the summed superior ones, we decided to include it in the corpus.

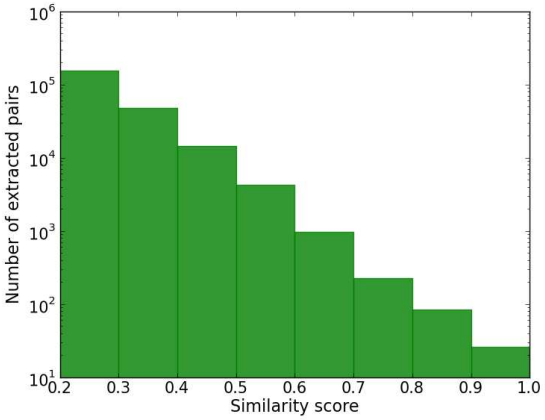

Figure 2: The distribution of the extracted clause pairs at different thresholds

Table 1 presents German-French clause pairs with their corresponding similarity scores. On the top we can find rather short clauses (up to 10 words) with perfectly aligned words. One expects that the decrease of the values implies that

Nr.	French clause	German clause	Score
1	mcnish écrit dans son journal :	mcnish schrieb in sein tagebuch :	1.0
2	son journal n' a pas été retrouvé	sein tagebuch wurde nie gefunden	0.950
3	elle travailla pendant plusieurs semaines avec lui	während mehrerer wochen arbeitete sie mit ihm zusammen	0.840
4	en 1783, il fait une première tentative infructueuse avec marc théodore bourrit	paccard startete 1783 zusammen mit marc theodore bourrit einen ersten, erfolglosen besteigungsversuch	0.717
5	en 1962, les bavarois toni kinshofer, siegfried löw et anderl mannhardt réussirent pour la première fois l' ascension par la face du diamir	1962 durchstiegen die bayern toni kinshofer, siegfried löw und anderl mannhardt erstmals die diamir-flanke	0.623
6	le 19 août 1828 il tenta, avec les deux guides jakob leuthold et johann wahren l' ascension du finsteraarhorn	august 1828 versuchte er zusammen mit den beiden bergführern jakob leuthold und johann währen das finsteraarhorn zu besteigen	0.519
7	le parc protège le mont robson, le plus haut sommet des rocheuses canadiennes	das 2248 km^2 große schutzgebiet erstreckt sich um den 3954 m hohen mount robson, dem höchsten berg der kanadischen rocky mountains	0.470
8	la plupart des édifices volcaniques du haut eifel sont des dômes isolés plus ou moins aplatis	die meisten der vulkanbauten der hocheifel sind als isolierte kuppen vereinzelt oder in reihen der mehr oder minder flachen hochfläche aufgesetzt	0.379
9	le site, candidat au patrimoine mondial, se compose d' esplanades-autels faits de pierres	die stätte, ein kandidat für das unesco-welterbe, besteht aus altarplattformen aus steinen und erde, gestützt auf einer unterirdischen konstruktion aus bemalten ton-pfeilern	0.259
10	qu' un cas mineur ayant un effet limité sur la santé	wie sich diese substanzen auf die gesundheit auswirken,	0.200

Table 1: Examples of extracted clause pairs

the clauses contain less or even no translated fragments. A manual inspection of the extracted pairs showed that this is not always the case. We have found clause pairs with almost perfect 1-1 word correspondences and a similarity score of only 0.51. The "low" score is due to the fact that we are comparing human language to automatic translations, which are not perfect.

On the other hand, a comparable score can be achieved by a pair in which one of the clauses contains some extra information (e.g. pair number 7). The extra parts in the German variant (*2248 km^2 große* - EN: with an area of 2248 km^2; *3954 m hohen* - EN: 3954 m high) cannot be separated by means of clause boundary detection, since they don't contain any verbs. This finding would motivate the idea of splitting the phrases into subsentential segments (linguistically motivated or not) and aligning the segments, similar to what Munteanu (2006) proposed. Nevertheless, we consider this pair a good candidate for the parallel corpus.

Pair number 8 has the same coordinates (i.e. an extra tail in the German variant), yet it receives a lower score, which might disqualify it for the final list, if we only look at the numbers. In this case, the low score is caused by the German compounds (*Vulkanbauten, Hocheifel*), which are unknown to the SMT system, therefore they are left untranslated and cannot be aligned. However, we argue that this clause pair should also be part of the extracted corpus.

Score range	Average sentence length	
	German	French
$[0.9 - 1.0]$	4	4.26
$[0.8 - 0.9)$	4.87	5.04
$[0.7 - 0.8)$	6.47	6.65
$[0.6 - 0.7)$	10.78	10.71
$[0.5 - 0.6)$	12.09	11.51
$[0.4 - 0.5)$	11.91	11.80
$[0.3 - 0.4)$	11.28	11.22
$[0.2 - 0.3)$	11.22	11.01

Table 2: The average sentence length for different score ranges

The last pair is definitely a bad candidate for a parallel corpus, since the clauses do not convey the same meaning, although they share many words (*avoir un effet - auswirken, sur la santé - auf die Gesundheit*). A subsentential approach would allow us to extract the useful segments in this case, as well. There are, of course, pairs with similar scores and poorer quality, therefore 0.2 is the lowest threshold which can provide useful candidate pairs. At the other end of the scale, we consider pairs above 0.4 as parallel and everything below as comparable. As a general rule, a high threshold ensures a high accuracy of the extraction pipeline.

Table 2 presents the average length (number of tokens) of the extracted clauses for different ranges of the similarity score. We notice that the best ranked clauses tend to be very short, whereas the last ranked are longer, as the examples in table 1 confirm. However, the average length over the whole extracted corpus is below 10 words, a small value compared to the results reported on Wikipedia articles by Ştefănescu and Ion (2013). This finding is due to the fact that we are aligning clauses instead of whole sentences.

We expected the German sentences to be usually shorter than the French ones (or at least have a similar number of words), since they are more likely to contain compounds. This fact is confirmed by the first part of the table. A turnaround occurs in the range (0.5,0.6), where the German sentences become slightly longer than the French ones, since they tend to contain extra information (see also table 1).

4 Experiments and Results

The conducted experiments have focused only on the extraction of parallel clauses and their use in a SMT scenario. For this purpose, we have used as input the articles selected and preprocessed in the previous development phase (Plamada and Volk, 2012). Specifically, the data set consists of 39 000 parallel articles with approximately 6 million German clauses and 2.7 million French ones. We were able to extract 225 000 parallel clause pairs out of them, by setting the final filter threshold to 0.2. This means that roughly 4% of the German clauses have an French equivalent (and 8% when reporting to the French clauses), figures comparable to our previous results on a different sized data set. However, the quality of the extracted data is higher than in our previous approaches.

To evaluate the quality of the parallel data extracted, we manually checked a set of 200 automatically aligned clauses with similarity scores above 0.25. For this test set, 39% of the extracted data represent perfect translations, 26% are translations with an extra segment (e.g. a noun phrase) on one side and 35% represent misalignments. However, given the high degree of parallelism between the clauses from the middle class, we consider them as true positives, achieving a precision of 65%. Furthermore, 40% of the false positives have been introduced by matching proper names, 32% contain matching subsentential segments (word sequences longer than 3 words) and 27% represent failures in the alignment process.

4.1 SMT Experiments

In addition to the manual evaluation discussed in the previous subsection, we have run preliminary investigations with regard to the usefulness of the extracted corpus for SMT. In this evaluation scenario, we use only pairs with a similarity score above 0.35. The results discussed in this section refer only to the translation direction German-French. The SMT systems are trained with the Moses toolkit (Koehn et al., 2007), according to the WMT 2011 guidelines[6]. The translation performance was measured using the BLEU evaluation metric on a single reference translation. We also report statistical significance scores, in order to indicate the validity of the comparisons between the MT systems (Riezler and Maxwell, 2005). We consider the BLEU score difference significant if the computed p-value is below 0.05.

We compare two baseline MT systems and several systems with different model mixtures (trans-

[6]http://www.statmt.org/wmt11/baseline.html

lation models, language models or both). The first baseline system is an in-domain one, trained on the Text+Berg corpus and is the same used for the automatic translations required in the extraction step (see section 3). The second system is purely out-of-domain and it is trained on Europarl, a collection of parliamentary proceedings (Koehn, 2005). The development set and the test set contain in-domain data, held out from the Text+Berg corpus. Table 3 lists the sizes of the data sets used for the SMT experiments.

Data set	Sentences	DE Words	FR Words
SAC	220 000	4 200 000	4 700 000
Europarl	1 680 000	37 000 000	43 000 000
Wikipedia	120 000	1 000 000	1 000 000
Dev set	1424	30 000	33 000
Test set	991	19 000	21 000

Table 3: The size of the German-French data sets

Our first intuition was to add the extracted sentences to the existing in-domain training corpus and to evaluate the performance of the system. In the second scenario, we added the extracted data to an SMT system for which no in-domain parallel data was available. For this purpose, we experimented with different combinations of the models involved in the translation process, namely the German-French translation model (responsible for the translation variants) and the French language model (ensures the fluency of the output). Besides of the models trained on the parallel data available in each of the data sets, we also built combined models with optimized weights for each of the involved data sets. The optimization was performed with the tools provided by Sennrich (2012) as part of the Moses toolkit. We also want to compare several language models, some trained on the individual data sets, others obtained by linearly interpolating different data sets, all optimized for minimal perplexity on the in-domain development set. The results are summarized in table 4.

A first remark is that an out-of-domain language model (LM) adapted with in-domain data (extracted from Wikipedia and/or SAC data) significantly improves on top of a baseline system trained with out-of-domain texts (Europarl, EP) with up to 1.7 BLEU points. And this improvement can be achieved with only a small quantity of additional data compared to the size of the original training data (120k or 220k versus 1680k sentence pairs). When replacing the out-of-domain

Translation model	Language model	BLEU score
Europarl TM	EP LM	9.45
Europarl TM	EP+Wiki LM	10.39
EP+Wiki TM	EP+Wiki LM	10.37
Europarl TM	EP+Wiki+SAC LM	11.22
EP+Wiki TM	EP+Wiki+SAC LM	11.74
EP+WMix TM	EP+Wiki+SAC LM	10.40
SAC TM	SAC LM	16.71
SAC+Wiki TM	SAC+Wiki LM	16.51
SAC+WMix TM	SAC+Wiki LM	16.37

Table 4: SMT results for German-French

translation model with a combined one (including the Wikipedia data set) and keeping only the adapted language models, we can observe two tendencies. In the first case (using a combination of out-of-domain and Wikipedia-data for the language model), the BLEU score remains approximately at the same level (10.37-10.39), the difference not being statistically significant (p-value = 0.387).

The addition of quality in-domain data for the LM from the previous configuration brings an improvement of 0.5 BLEU points on top of the best Europarl system (11.22 BLEU points). Given that all other factors are kept constant, this improvement can be attributed to the additional translation model (TM) trained on Wikipedia data. Moreover, the statistical significance tests confirm that the improved system performs better than the previous one (p-value = 0.005). To demonstrate that these results are not accidental, we replaced the Wikipedia extracted sentences with a random combination thereof (referred to as WMix) and retrained the system. Under these circumstances, the performance of the system dropped to 10.40 BLEU points. These findings demonstrate the effect of a small in-domain data set on the performance of an out-of-domain system trained on big amounts of data. If the data is of good quality, it can improve the performance of the system, otherwise it significantly deteriorates it.

We notice that the performance of a strong in-domain baseline system (SAC) cannot be heavily influenced (either positively or negatively) by translation and language model mixtures combining existing in-domain data with Wikipedia data. In terms of BLEU points, the mixture models trained with "good" Wikipedia data cause a perfor-

mance drop of 0.2, but the significance test shows that the difference is not statistically significant (p-value = 0.08). On the other hand, the TM including shuffled Wikipedia sentences causes a performance drop of 0.34 BLEU points, which is statistically significant (p-value = 0.013). We can conclude that the quantity of the data is not the decisive factor for the performance change, but rather the quality of the data. The Wikipedia extracted data set maintains the good performance, whereas a random mixture of the Wikipedia data set causes a performance decrease. Therefore the focus of future work should be on obtaining high quality data, regardless of its amount.

5 Conclusions and Outlook

In this paper we presented a method for extracting domain-specific parallel data from Wikipedia articles. Based on previous experiments, we focus on clause level alignments rather than on full-sentence extraction methods. Moreover, the ranking of the candidates is based on a metric combining different similarity criteria, which we defined ourselves. The precision estimates show that the extracted sentence pairs are clearly semantically equivalent. The SMT experiments, however, show that the extracted data is not refined enough to improve a strong in-domain SMT system. Nevertheless, it is good enough to overtake an out-of-domain system trained on 10 times bigger amounts of data.

Since our extraction system is merely a prototype, there are several ways to improve its performance, including better filtering for in-domain articles, finer grained alignment and more sophisticated similarity metrics. For example, the selection of domain-specific articles can be improved by means of an additional filter based on Wikipedia categories. The accuracy of the extraction procedure can be improved by means of a more informed similarity metric, weighting more feature functions. Moreover, we can bypass the manual choice of thresholds by employing a classifier (e.g. SVM^{light} (Joachims, 2002)). Additionally, we could try to align even shorter sentence fragments (not necessarily linguistically motivated).

We are confident that Wikipedia can be seen as a useful resource for SMT, but further investigation is needed in order to find the best method to exploit the extracted data in a SMT scenario. For this purpose, quality data should be preferred over sizable data. We would therefore like to experiment with different ratio combinations of the data sets (Wikipedia extracted and in-domain data) until we find a combination which outperforms our in-domain baseline system.

References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25:341–375.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of EMNLP*.

Pascale Fung, Emmanuel Prochasson, and Simon Shi. 2010. Trillions of comparable documents. In *Proceedings of the the 3rd workshop on Building and Using Comparable Corpora (BUCC'10)*, Malta.

Thorsten Joachims. 2002. *Learning to classify text using Support Vector Machines*. Kluwer Academic Publishers.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit X*, pages 79–86.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504, December.

Dragos Stefan Munteanu. 2006. *Exploiting comparable corpora*. Ph.D. thesis, University Of Southern California.

Kishore Papineni, 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Magdalena Plamada and Martin Volk. 2012. Towards a Wikipedia-extracted alpine corpus. In *Proceedings of the Fifth Workshop on Building and Using Comparable Corpora*, Istanbul, May.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France. Association For Computational Linguistics.

Jason Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Ştefănescu and Radu Ion. 2013. Parallel-Wiki: A collection of parallel sentences extracted from Wikipedia. In *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*.

Dan Ştefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In Mauro Cettolo, Marcello Federico, Lucia Specia, and AndyEditors Way, editors, *Proceedings of the 16th Conference of the European Association for Machine Translation EAMT 2012*, pages 137–144.

Martin Volk. 2001. *The automatic resolution of prepositional phrase - attachment ambiguities in German*. Habilitation thesis, University of Zurich.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–748, Washington, DC, USA. IEEE Computer Society.