Year: 2012

Towards a Wikipedia-extracted alpine corpus

Plamada, Magdalena ; Volk, Martin

Abstract: This paper describes a method for extracting parallel sentences from comparable texts. We present the main challenges in creating a German-French corpus for the Alpine domain. We demonstrate that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia, therefore we introduce an alternative information retrieval approach. Sentence alignment algorithms were used to identify semantically equivalent sentences across the Wikipedia articles. Using this approach, we create a corpus of sentence-aligned Alpine texts, which is evaluated both manually and automatically. Results show that even a small collection of extracted texts (approximately 10000 sentence pairs) can partially improve the performance of a state-of-the-art statistical machine translation system. Thus, the approach is worth pursuing on a larger scale, as well as for other language pairs and domains.

Originally published at:
Plamada, Magdalena; Volk, Martin (2012). Towards a Wikipedia-extracted alpine corpus. In: The Fifth Workshop on Building and Using Comparable Corpora, Istanbul, Turkey, 26 May 2012.

Towards a Wikipedia-extracted Alpine Corpus

Magdalena Plamada, Martin Volk

Institute of Computational Linguistics, University of Zurich
Binzmühlestrasse 14, 8050 Zürich
{plamada, volk}@cl.uzh.ch

Abstract

This paper describes a method for extracting parallel sentences from comparable texts. We present the main challenges in creating a German-French corpus for the Alpine domain. We demonstrate that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia, therefore we introduce an alternative information retrieval approach. Sentence alignment algorithms were used to identify semantically equivalent sentences across the Wikipedia articles. Using this approach, we create a corpus of sentence-aligned Alpine texts, which is evaluated both manually and automatically. Results show that even a small collection of extracted texts (approximately 10 000 sentence pairs) can partially improve the performance of a state-of-the-art statistical machine translation system. Thus, the approach is worth pursuing on a larger scale, as well as for other language pairs and domains.

Keywords: Comparable corpus, Alpine texts, Wikipedia, Information Retrieval, Sentence alignment, Statistical machine translation, French-German

1. Introduction

The performance of Statistical Machine Translation (SMT) systems depends strongly both on the quality and the quantity of the training data. A well-known problem of SMT systems for most language pairs is the limited amount of bilingual parallel training data. The existing parallel corpora cover a relatively small percentage of possible language pairs and very few domains, thus building new ones involves considerable efforts, both in terms of time and costs.

In the last decade, less expensive but very productive methods of creating such sentence-aligned bilingual corpora have been proposed, based on the extraction of parallel texts from comparable texts. Zhao and Vogel (2002) introduced an adaptive approach for mining parallel sentences from a bilingual news collection, which combines a sentence length model with an IBM Model 1 translation model. Fung and Cheung (2004) combine bootstrapping methods and an IBM Model 4 model in order to exploit "very-non-parallel corpora" consisting of news stories from different sources.

The availability of comparable corpora and their potential for creating parallel corpora have sparked the interest of the SMT community. Munteanu and Marcu (2005) propose a maximum entropy-based classifier for identifying parallel sentences in newspaper articles by referring to a bilingual dictionary. They evaluate the extracted corpus by using it as training data for an SMT system. A similar approach is presented in (Abdul Rauf and Schwenk, 2011), with the difference that the authors of the latter paper use automatic translations instead of bilingual dictionaries and the selection relies on other metrics, such as word or translation error rate (WER, TER).

The approaches mentioned up to this point have been tested only on news corpora, but the expansion of the Web has drawn the attention towards another fruitful resource: web corpora. Adafre and de Rijke (2006) describe an MT based approach to find corresponding sentences in Wikipedia based on sentence similarity, without investigating the improvements of their method for a specific task (e. g. SMT, information extraction). Alternatively, Fung et al. (2010) also crawl comparable web sites (in particular, Wikipedia) in order to extract potential parallel sentences. The authors mention the improvement of SMT systems as one of the main objectives, but do not report any results.

As previously discussed, work in this field has focused mainly on two types of corpora (news and web corpora), with the purpose of extracting good training material for SMT. Nevertheless, not all papers present their results in terms of SMT improvements. There is also no claim about the performance of these approaches for a different domain. This represents our motivation to develop an approach inspired by earlier work, with the aim to extract a parallel corpus of mountaineering texts from Wikipedia. Moreover, we are interested in investigating to what extent the extracted corpus improves the performance of a domain-specific SMT system.

Wikipedia is an important multilingual resource available for a variety of domains, in almost 300 languages. It is not a parallel corpus because its articles in different languages are edited independently by users and are not literal translations of each other. However, often an article in one language contains a number of sentences translated from its corresponding article in another language. We identify and extract the parallel sentences in the Wikipedia articles and, moreover, we reduce the search space to one specific domain: Alpine texts (i. e. mountaineering reports, hiking recommendations, popular science articles about the biology and the geology of mountainous regions).

In the project Domain-specific Statistical Machine Translation[1] we have developed an SMT system trained for the Alpine domain. The training data comes from the Text+Berg corpus[2], which contains the digitized publications of the Swiss Alpine Club (SAC) from 1864 until

[1] http://www.cl.uzh.ch/research_en.html

[2] See www.textberg.ch

2011. The most relevant part for SMT training is the parallel German-French one representing a sizable corpus of approx. 5 million words. We therefore have the expertise to use in-house developed tools for the purpose of this experiment.

This article describes our approach for exploiting Wikipedia in order to produce more parallel texts for the Alpine domain. In section 2. we describe the extraction workflow, and in the subsequent section we evaluate the resulting corpus. The last section discusses future experiments and further improvements of the extraction method.

2. Extraction Methods

The general architecture of our parallel sentence generation process is shown in Figure 1. The approach was applied only to the language pair German-French, as these are the main languages of the Text+Berg corpus. However, the procedure can be applied with little effort to any of the available Wikipedias and any other domain. In our case, the input consists of German and French Wikipedia dumps[3], which are available in a special XML format, called MediaWiki[4].

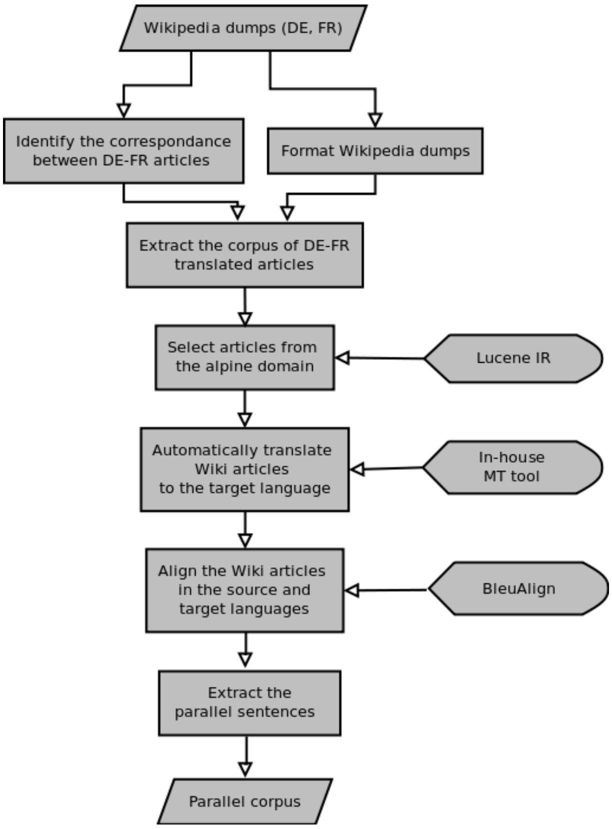
Wikipedia dumps (DE, FR)
Identify the correspondance between DE-FR articles
Format Wikipedia dumps
Extract the corpus of DE-FR translated articles
Select articles from the alpine domain
Lucene IR
Automatically translate Wiki articles to the target language
In-house MT tool
Align the Wiki articles in the source and target languages
BleuAlign
Extract the parallel sentences
Parallel corpus

Figure 1: The workflow of the extraction algorithm

In the first step, we identify Wikipedia articles available in both languages by using the procedure described in (Lopez and Otero, 2010). We had to adapt the configuration files for French and German, as the original tool was developed for English, Spanish and Portuguese. The relevant output for our task represents the mapping between the titles of the articles available in both languages. The simplified XML structure proposed by CorpusPedia (Lopez and Otero, 2010) cannot be validated by usual XML parsers (e.g. DOM, SAX, ElementTree), so we need an additional tool for converting MediaWiki to valid XML.

For this purpose we used WikiPrep[5], a preprocessing tool that transforms the Wikipedia dumps to a simple XML format (Gabrilovich and Markovitch, 2006). The content of Wikipedia pages is converted to plain text with XML markups for section headers and internal links. MediaWiki is localized for all the languages supported in Wikipedia. We therefore had to customize the configuration files for German and French, so that MediaWiki elements (namespaces, templates, date and number formats etc.) can be correctly identified. After updating the files, we run the tool over the two Wikipedia dumps and then filter the articles available in both German and French.

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. This procedure is detailed in the following subsection. Once we extract the Alpine comparable corpus, we proceed to the sentence alignment of the articles. The aim is to obtain a reasonably-sized set of sentence pairs that are likely to contain good data for our parallel corpus. This step is described in subsection 2.3.

2.1. Article classification in Wikipedia

In Wikipedia, articles are organized into topics and therefore they are assigned to one or more categories. This classification could allow us to extract articles on similar topics, in our case the topics of interest could be *Alpen*, *Berge* or *Ort*. However, many articles lack a category tag. This is the case with disambiguation articles, which distinguish between several contexts associated with an article title. For example, the article *Morgenstern* can refer to a planet, a medieval weapon, a magazine, a music band or a common last name for several personalities. Neither redirect articles, which automatically send the reader to another article, fall in any Wikipedia category. For instance, both pages *Wintersonnenwende* and *Sommersonnenwende* are linked to the more general article *Sonnenwende* (English: solstice). Apart from these cases, there are some arbitrary articles in our Wikipedia dump[3] that have no category tags. In fact, only 51.5% of the articles in the German Wikipedia have an assigned category. The remaining part consists of 33% redirect articles, 10% miscellaneous articles and 5.5% disambiguation articles. The percentages are similar in the French Wikipedia: 52.5% of the articles are categorised, 40% represent redirect articles, 4% mixed articles and 3.5% disambiguation articles.

Another interesting aspect is that articles are usually not placed in the most general category they logically belong to, if they are tagged as a subcategory thereof. For example, the article *Rosengartengruppe* is tagged with the following categories: *Bergmassiv (Dolomiten), Gebirge in*

[3] Accessed in September 2011

[4] http://www.mediawiki.org/wiki/MediaWiki

[5] http://sourceforge.net/projects/wikiprep/

Südtirol, Gebirge im Trentino, Dolomiten (English: massif in the Dolomites, mountains in South Tyrol, mountains in Trentino, Dolomites), but there is no reference to the Alps, although it is obvious that this mountain range belongs to the Alps. If we would like to use the Wikipedia classification as criterion for the extraction of domain-specific articles, we should come up with an extensive list of relevant categories. The categories in Wikipedia are sometimes very specific (e. g. *Berg im Kanton Appenzell Innerrhoden*), so compiling the list is not a trivial task. Besides, we would need an automatic classifier able to distinguish between relevant (e. g. *Bergführer*) and irrelevant categories (e. g. *Berg bei Neumarkt in der Oberpfalz*) for our corpus.

Another challenge for this task is that the categories assigned to the same article in different languages do not overlap. For example, the article *Trois Vallées* is tagged in German as *Wintersportgebiet in Frankreich, Alpen* (English: winter sports resort in France, Alps), whereas in French it belongs to the following categories: Tourisme en Savoie, Domaine skiable (English: tourism in Savoy, ski area). Identifying the semantic relationships between the German and the French categories is also not an easy task for a reasoner. One would therefore need to compile separate category lists for both German and French, as a simple translation of the categories from the other language would not help. This is not an isolated case in Wikipedia, but a general trend, as tables 1 and 2 show. They illustrate the distribution of the Wikipedia categories for the first 10 000 articles extracted with our approach (see section 2.2.). The German part contains 17 000 categories and the French one 16 000 categories, but more than 50% of them appear only once.

Category	Number of articles
Mann	1 278
Berg in Europa	325
Deutscher	279
Berg in den Alpen	210
Autor	190
Schweizer Gemeinde	157

Table 1: The most frequent categories in the top 10^4 German articles retrieved by Lucene

In the German Wikipedia, however, the leading category *Mann* (English: man, person) covers approximately 13% of the articles. As this category is rather general, we inspected the other categories assigned to these articles. We found that more than 90% of the articles were tagged with categories such as *Bergsteiger, Geograph, Entdecker, Extremsportler, Bergführer* (English: alpinist, geographer, explorer, extreme athlete, mountain guide). This proves that the retrieved articles are consistent with our domain of interest. Approximately 20% of them were also tagged with *Deutscher* (English: German), an expected percentage considering their corresponding values in table 1.

The next ranked categories cover significantly less articles (approx. $2-3\%$), but they obviously represent what we would expect in such a corpus (e. g. mountains in Europe and in the Alps, respectively). These results prove the accuracy of our extraction approach.

Category	Number of articles
Film américain	140
Ville de Bade-Wurtemberg	134
Ville de Rhénanie-du-Nord-Westphalie	121
Sommet des Alpes autrichiennes	98
Ville de Bavière	75
Sommet des Alpes suisses	64

Table 2: The most frequent categories in the top 10^4 French articles retrieved by Lucene

The French categories are much more diverse, therefore none of them covers a significant percentage of the articles. The leading category in the French Wikipedia, *Film américain* (English: American movie), is rather unexpected for this domain and belongs to the false positives in our results. However, the value is comparable to the following positions in the hierarchy, which are all relevant for our domain.

Taking all these aspects into consideration, we considered the extraction of domain-specific articles by means of the Wikipedia categorization time-consuming. We therefore decided to use an information retrieval-based approach, which will be detailed in section 2.2.

2.2. Extracting domain-specific articles

In order to extract the articles belonging to the Alpine domain, we have performed IR queries over the French and German Wikipedia. The input queries contained the 100 most frequent mountaineering keywords in the Text+Berg corpus (e. g. *Alp, Gipfel, Meter, Berg* in German and *montagne, sommet, mètre, cas* in French). The keyword lists are not translations of each other, as the term frequencies have been computed separately for German and French, respectively. However, they share common terms in the Alpine domain, such as *mountain, peak, meter*.

The extraction tool is based on the Lucene API[6], an open-source IR library. As Lucene does not have a module for morphological analysis, the reported results are based only on word-matching. We have decided to restrict the keywords to common nouns due to their limited inflectional variation. Lucene returns a list of the articles relevant to our query, ranked by their similarity score[7]. The score takes into consideration several factors such as term frequency, inverse document frequency, number of matched query terms etc.

Upon completion of this step our corpus was reduced to approx. 150 000 parallel articles. This value should be regarded with caution, as it stands for all articles that contain at least one occurrence of the top 100 Text+Berg keywords. Therefore in our experiments we use only articles that report a Lucene score above a certain threshold. The choice of the threshold depends highly on the targeted accuracy

[6] http://lucene.apache.org

[7] http://lucene.apache.org/core/old_versioned_docs/versions /3_0_0/scoring.html

and the task itself, as the similarity scores are sometimes misleading. It is possible that a short article about less important mountains (e. g. *Gurktaler Alpen*, similarity score: 0,010 97) receives a lower score than a long article about a collection of novels (e. g. *Die Arbeiten des Herkules*, similarity score: 0,034 29). Table 3 shows a selection of the articles with the highest scores in the German Wikipedia.

Title	Score
Reinhold Messner	0,089 43
Britische Mount-Everest-Expedition 1924	0,080 52
Hans Kammerlander	0,070 07
Ortler	0,069 66
Mount Everest	0,062 15
Mont Blanc	0,053 64

Table 3: The best ranked Alpine articles in the German Wikipedia according to Lucene

In contrast, table 4 presents the best articles in the French Wikipedia, sorted by their relevance according to Lucene. The French ranking differs from the German one firstly because the keyword lists partially contain different nouns. On the other hand, the content of the articles (including their structure and length) highly varies among the language variants of Wikipedia.

Title	Score
Lure	0,059 58
Parc national de Glacier	0,059 40
Mont Kenya	0,057 72
Nez-Percés	0,057 53
Mont Ventoux	0,057 15
Mont Blanc	0,057 09

Table 4: The best ranked Alpine articles in the French Wikipedia according to Lucene

However, the hit lists may also contain overlapping content, such as the article about *Mont Blanc* in the previous examples. An interesting finding is that the first hit for the French Wikipedia is an article about the city of Lure, which apparently does not have much in common with our topic, mountains. Taking a closer look at the whole article explains the score, as is contains thorough sections about the geology, the topography, the hydrology, and the climatology of the place, which are all areas closely related to mountaineering.

2.3. Extracting aligned sentence pairs

We use the Bleualign algorithm (Sennrich and Volk, 2010) for extracting parallel sentences from two Wikipedia articles. The aligned sentences (beads) are identified by means of an intermediary machine translation of the source. In our case, the translation is performed by our in-house SMT system trained on Alpine texts. Bleualign generates all possible sentence pairs between the automatic translation and the targeted article and computes for each of them the BLEU score (Papineni et al., 2002). Subsequently it reduces the search space by keeping only the 3 best-scoring alignment candidates for each sentence. Finally the algorithm returns the alignment pair which maximizes the BLEU score and respects the monotonic sentence order.

The algorithm can be applied in both directions. Translation direction does not matter in general, but we have decided to translate from French to German. We chose the French texts as the source texts because they are generally shorter. As the algorithm tries to align as much sentences as possible, this choice of the source texts allows us to maximize precision. In order to obtain a high-precision sentence alignment, Sennrich and Volk (2010) proposed computing the alignments in both directions, intersecting the results and then discarding all beads that differ between the two runs. For our purposes, however, we compute the alignments in a single direction (French-German).

In the end we filter the results once more by choosing only the 70% best-ranked alignments. The resulting set of alignment pairs represents a corpus containing semantically equivalent sentences.

As an example, the following sentence pair is a candidate for our parallel corpus which obtains the highest BLEU score.

FR: ainsi , la partie nord de l' himmelschrofenzug se compose de dolomite tandis que la partie sud se compose de roches du lias de la couche de l' allgäu

Automatic translation: damit ist der nördliche teil des himmelschrofenzug besteht aus dolomit , während der südliche teil besteht aus felsen des lias der schneedecke , das allgäu

DE Reference: so besteht der nördliche teil des himmelschrofenzugs aus hauptdolomit. der südliche teil besteht aus liasgesteinen der allgäudecke , die auf den hauptdolomit aufgeschoben worden sind

It is worth noting that the BLEU score is not computed between the source and target sentences, but between the automatic translation and the target sentence. This is how the BLEU values in Table 5 should be interpreted. Although the automatic translation is not perfectly correct, one notices that the word overlap between the translation and the target sentence is rather high. This explains why the extra tail in the German reference *die auf den hauptdolomit aufgeschoben worden sind* is not penalized by the BLEU score. Moreover, this example clearly shows that the algorithm deals not only with 1-to-1 alignments, but also with 1-to-n alignments.

3. Experiments and Results

3.1. Experimental setting

In this experiment we selected the top 4 000 ranked Wikipedia articles retrieved by Lucene. For this purpose we have merged the German and French lists and sorted the resulting list by the similarity score that Lucene provides. The articles have been sent to Bleualign for sentence alignment, using a customized configuration. We put great value on translations' fluency, so we measured the BLEU score on 3-grams, instead of 2-grams, as proposed by Sennrich and Volk (2010). In addition, we decided not to use any gap filling heuristics, because of the great variation of article structure between the Wikipedias.

French sentence	German sentence	BLEU Score
sur ce point , Andrée se démarque non seulement des explorateurs qui lui succéderont , mais aussi de bien de ceux qui l'ont précédé	darin unterschied sich Andrée nicht nur von den späteren sondern auch von vielen früheren Entdeckungsreisenden	0.5555
lors d' une conférence donnée en 1895 à l' académie royale des sciences de Suède, il fit grosse impression devant un public composé de géographes et météorologues	er hielt Vorlesungen bei der Königlichen Akademie der Wissenschaften und bei der schwedischen Gesellschaft für Anthropologie und Geologie und erhielt breite Zustimmung	0.6010
cinquante-sept personnes trouvèrent la mort et 200 habitations, 47 ponts, 24 km de chemin de fer et 300 km de routes furent détruits	in dem dünn besiedelten und zuvor evakuierten Gebiet verloren 57 Menschen ihr Leben und 200 Häuser, 47 Brücken, 24 km Eisenbahngleise sowie 300 km Highways wurden zerstört	0.4143
il est ainsi le premier homme à gravir trois sommets de plus de 8000 m en une même saison	mit dieser Besteigung war Messner der erste Mensch überhaupt , der mehr als zwei Achttausender bestiegen hatte	1.0
cette montagne est avec le plateau de Gottesack voisin l'attraction majeure du sous-groupe	dieser Berg ist zusammen mit dem benachbarten Gottesackerplateau auch die markanteste Erscheinung der Untergruppe	1.0

Table 5: Alignment pairs identified by Bleualign

Specifically, the dataset consists of 555 000 German and 290 000 French sentences. Bleualign identified 24 000 alignments out of them. For the evaluation, we manually check a set of 200 automatically aligned sentences and we report the precision of the algorithm for this dataset.

3.2. Results

Out of the 200 sentence pairs under consideration, 30% represent perfect translations, 45% contain only aligned segments (partial alignments) and 25% represent missalignments. We can therefore count on 75% precision of the alignment procedure. A large-scale automatic evaluation of the alignment quality could be indirectly performed by measuring the improvements of a SMT system trained with the aligned data.

Table 5 presents a selection of the alignment pairs identified by our approach, together with the BLEU score computed over the translation. An interesting finding is that a high BLEU score does not always correlate with a perfect translation. BLEU has been previously criticized as a measure of translation quality, and it is not considered reliable on sentence level (Callison-Burch et al., 2006). Take, for example, the fourth sentence in the table, whose automatic translation received the maximum alignment score. This is a perfect example of a comparable text, but not a translation. The topic is clearly the same: the first man ascending more than two (e. g. three) peaks, but the rest of the sentence modifies its meaning in different directions. This finding brings again in discussion the relativity of the BLEU scores and the central question whether this sort of alignments can be considered good training material for SMT.

On the other hand, the last sentence pair correctly receives the maximum BLEU score, as all the words in the French sentence have a correspondent in the German one. In fact, the French article that contains the sentence in question is a faithful translation of its German correspondent, performed by a human translator. This is not an unique case in Wikipedia, but part of the initiative *Projet:Traduction* aiming to enrich the French Wikipedia with translations from other Wikipedias. This information is marked up in the page source with *{{Traduction/Référence|de|Allgäuer Alpen|28915176|9 mars 2007}}*. For quality reasons, the translated articles are subject to double reviewing. These sentence pairs are therefore the ideal parallel data that we aim to find in Wikipedia.

Moreover, the first two sentence pairs also represent valid translations, but receive lower alignment scores due to the different construction types. For example, the French relative clause *des explorateurs qui lui succéderont* is replaced by a nominal phrase in German: *den späteren Entdeckungsreisenden*. And the passive voice *une conférence donnée* is expressed as active voice in the German sentence: *er hielt Vorlesungen*. However, as long as these results are in the upper part of the ranking, the differences between BLEU scores should not be a problem for our task.

Between the extremes we find example number three, situated in the second half of the BLEU ranking. In this case, the German sentence has one significant extra segment compared to the French one: the nominal phase *in dem dünn besiedelten und zuvor evakuierten Gebiet*. The rest of the sentence is perfectly translated into French, but the rather poor BLEU score can be ascribed to be a penalty for the length difference. This finding highlights the need of more fine-grained alignments, at sub-sentential level. Munteanu and Marcu (2006) proposed a method to extract these segments and demonstrated the relevance of the task by reporting improvements in SMT performance.

3.2.1. 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. The results discussed in this section refer only to the translation direction German-French. Our Baseline MT system is trained on the Text+Berg corpus (approx. 200 000 sentence pairs) and is the same used for the automatic translations required in the extraction step (see section 2.3.). We then train another MT system on the initial corpus plus 10 000 sentence pairs from the extracted corpus. In the following we will refer the latter one as ExtractedPlus. Both systems were tested on a test corpus of 1 000 sentences from the Text+Berg corpus. The translation performance was measured using the automatic BLEU evaluation metric on a single reference translation.

The system trained with the addition of the comparable texts has not achieved the expected improvements in performance, most probably because of the small amount of new training material (compared to the existing training data). Therefore we have manually inspected the performance of the two systems in terms of word coverage. The Baseline system failed to translate 700 words from the test corpus, whereas the ExtractedPlus system reports only 600 out-of-vocabulary words, most of them proper nouns and compounds.

An example is presented below. Both systems produce an imperfect output, following the same grammatical structure. The differences consist mainly in the choice of words. The baseline system leaves untranslated 3 words: *spitzenrouten, anziehungspunkte, bschüttigütti* and omits some words (e. g. *kletterer*). The ExtractedPlus system, however, can handle the domain specific terms like the ones mentioned before and translates them correctly: *voies extrêmes, points d' impact, grimpeurs*. Although it still cannot translate the proper noun *bschüttigütti*, the output sentence is still easier to understand and therefore the ExtractedPlus system can be considered better in this case.

DE: dasselbe gilt für die von den rein klettertechnischen schwierigkeiten her gesehenen spitzenrouten und anziehungspunkte für leistungsstarke kletterer bschüttigütti (10) und fusion (10 -).
Reference: cela vaut également pour bschüttigütti (10) et fusion (10 -) , voies extrêmes par leurs difficultés techniques , et objectifs de rêve pour de forts grimpeurs .
Baseline: il en est de même pour les difficultés purement techniques venant spitzenrouten anziehungspunkte par et pour doués bschüttigütti (10) et à s'être illustrée dans fusion (10 -) .
ExtractedPlus: il en est de même pour les difficultés purement techniques venant de la voies extrêmes et de points d' impact pour grimpeurs doués bschüttigütti (10) et de fusion (10 -) .

4. Conclusions and Outlook

We have presented our efforts in extracting a parallel corpus of Alpine texts from Wikipedia. Wikipedia, and, in general, comparable corpora are inherently heterogeneous collections of texts, where the same topic can be expanded in different ways. The differences can be found not only on the content level, but also on the formal level (i. e. MediaWiki syntax). One major problem of freely available resources like Wikipedia is that they can be edited independently by non-experts and there are no unification efforts. This makes it difficult, in the first place, to process the different Wikipedias in an uniform manner.

We demonstrated that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia. Our method proposes a IR approach in order to achieve a solution to this task. However, an interesting research direction for the future is to combine these two approaches, in order to increase the reliability of the extraction method.

We have identified semantically equivalent sentences from the German and French Wikipedia articles by computing alignments between them. The reported results support our claim that this approach is worth pursuing. The procedure can be refined by training a classifier based on the Bleualign algorithm to automatically distinguish between useful and less useful alignment pairs (without the need to manually set thresholds). Moreover, as shown in section 3.2., an important improvement step is to allow the alignment of subsentential segments.

After collecting a sizable collection of Alpine texts, we will investigate the contribution of the extracted corpus for SMT performance on a larger scale. Finally, the use of the improved SMT system in our extraction algorithm could allow us to compute new and better alignments in the next development cycle.

5. References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25:341–375. 10.1007/s10590-011-9114-9.

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.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *In Proceedings of EACL*, pages 249–256.

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.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using Wikipedia: enhancing text categorization with encyclopedic knowledge. In *proceedings of the 21st national conference on Artificial intelligence - Volume 2*, pages 1301–1306. AAAI Press.

Isaac Gonzalez Lopez and Pablo Gamallo Otero. 2010. Wikipedia as multilingual source of comparable corpora. In *Proceedings of the LREC 2010*, Malta.

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 and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010).*

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.