Exploiting Comparable Corpora for Domain-specific Statistical Machine Translation

Thesis presented to the
Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of Doctor of Philosophy

by

Magdalena Plamadă

Accepted in the fall semester 2017
on the recommendation of the doctoral committee:

Prof. Dr. Martin Volk (Main supervisor)
Prof. Dr. Jörg Tiedemann
Prof. Dr. Mark Fishel
PD Dr. Gerold Schneider

Zurich, 2018

Abstract

The availability of parallel texts in various language combinations is the main bottleneck for Statistical Machine Translation (SMT) systems, as their performance is strongly influenced by the amount and the quality of the training texts. To overcome this bottleneck the present thesis investigates several methods for extracting parallel text segments from comparable text collections (e.g. Wikipedia, the Common Crawl). Moreover, the extraction is focused on a specific topical domain, whereas the segments have different granularities: clauses and sentences, respectively. The approaches have been tested for the language pair German-French and for the Alpine domain, which includes texts about mountaineering expeditions, the geology or the culture of the Alps.

The main contributions of this thesis are as follows:

- We suggest a three step approach for extracting domain-specific parallel segments from the considered comparable corpora. First we select the relevant documents in accordance with the following criteria: availability in the languages of interest and similarity to the chosen domain. The selected documents are aligned by means of a similarity metric making use of an intermediate automatic translation. Finally document pairs below a given similarity threshold are filtered out.
- We introduce a metric for measuring the similarity between two candidate segments based on metrics from Machine Translation evaluation (e.g. BLEU and METEOR) and on the percentage of aligned content words. The features have different weights and they are determined automatically on a training set. The metric is part of the segment alignment procedure.
- We evaluate the extracted pairs in a SMT scenario with respect to an out-of-domain baseline and an in-domain one. For this purpose we apply different domain adaptation techniques in order to make the most use of the extracted texts. We measure the performance on the same test set withheld from the Text+Berg corpus by means of state of the art evaluation metrics (BLEU, METEOR, OOV rate). The results revealed that the extracted texts bring significant improvements to an out-of-domain SMT system, but only marginal improvements to an in-domain system (i.e. in terms of lexical coverage).

Abstract (German)

Die grösste Herausforderung für statistische maschinelle Übersetzungssysteme (SMÜ) besteht in der eingeschränkten Verfügbarkeit von parallelen Texten in verschiedenen Sprachkombinationen, denn ihre Leistung hängt sowohl von der Menge, als auch von der Qualität der Texte ab. Um diesen Engpass zu überwinden, werden in dieser Arbeit unterschiedliche Methoden zur Extraktion von parallelen Textsegmenten aus vergleichbaren Korpora (z.B. Wikipedia, Common Crawl) untersucht. Die Experimente auf eine bestimmte Domäne beschränkt, wobei mit unterschiedlichen Textgranularitäten bei der Extraktion getestet wird. Die Methoden wurden für das Sprachpaar Deutsch-Französisch und für die Alpinismus-Domäne getestet.

Die Hauptbeiträge dieser Dissertation sind Folgende:

- Wir stellen einen dreistufigen Ansatz zur Extraktion von domänen-spezifischen Textsegmenten aus den ausgewählten vergleichbaren Korpora vor. Als erstes wählen wir die relevanten Dokumente gemäss folgender Kriterien aus: die Verfügbarkeit in den gewünschten Sprachen und die Ähnlichkeit zur Domäne. Diese Dokumente werden anhand einer Ähnlichkeitsmetrik aligniert, die auf einer automatischen Übersetzung aufbaut. Abschliessend werden Dokumentpaare mit niedriegem Ähnlichkeitswert entfernt.

- Wir schlagen eine Ähnlichkeitsmass vor, die einerseits auf Evaluationsmassen aus der maschinellen Übersetzung aufbaut und andererseits auf dem Anteil der alignierten Inhaltswörter. Diese Merkmale haben unterschiedliche Gewichte, die anhand eines Trainingsets bestimmt werden. Das Ähnlichkeitsmass ist in dem Alignierungsverfahren integriert.

- Wir evaluieren die extrahierten Segmente in einer SMÜ-Umgebung hinsichtlich einer Fremddomänen-Baseline und einer Zieldomänen-Baseline. Wir verwenden unterschiedliche Techniken zur Domänen-Anpassung, um die extrahierten Texte optimal auszunutzen. Wir berechnen die Übersetzungsqualität anhand eines Testsatzes aus dem Text+Berg Korpus, und mittels Standard Evaluationsmassen, wie z.B. BLEU, METEOR, OOV Rate). Die Resultate zeigen, dass die extrahierten Texte eine deutliche Verbesserung des Fremddomäne-Systems erzielen, während sich die Qualität des Zieldomäne-Systems nur geringfügig verbessert (z.B. in Bezug auf die lexikalische Abdeckung)

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Martin Volk, for the opportunity to work in his research team. His guidance, his steady support and encouragements were very much appreciated. I would further like to thank him for being a valuable source of inspiration in both professional and personal respect.

I would also like to thank the three other members of my doctoral committee, Jörg Tiedemann, Mark Fishel and Gerold Schneider for taking the time to read my manuscript and for their useful comments.

I am also grateful for the valuable support and objective feedback of my former and present colleagues at the Institute of Computational Linguistics, especially Rico Sennrich, Anne Göhring, Alexandra Bünzli and Annette Rios. And of course, for the inspiring coffee breaks that we shared.

Last but not least, I would like to thank my friends and family for their constant encouragement, especially my parents and my husband Andrei, for keeping me motivated throughout this time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Statistical Machine Translation (SMT) systems make use of statistical models for mapping the cross-lingual correspondences between words and phrases, respectively. Until a few years back, the underlying models were phrase-based, but they tend to be replaced more and more by models employing neural networks. The latter proved to perform significantly better by considering richer word contexts and conveying cohesive translations. Although the coverage of neural MT systems is currently limited due to the high computational costs of training such models, these systems are rapidly gaining popularity, as they outperform the systems of the previous generation. As the developments in the NMT field are following at a rapid pace, such systems are expected to spread out rapidly both in research and language technology industry.

This performance would not be possible without the usage of parallel texts, from where the systems learn the most probable translation of words and phrases in a given context. Unfortunately such texts are only available for a limited number of language pairs. Moreover, SMT systems optimized for translating specific texts (travel reports, parliamentary proceedings, news wires, etc.) profit from parallel texts from the same domain, restricting their availability even more. Therefore many efforts are being invested in the development of translation systems under low resource conditions.

A straightforward solution to this problem is to generate synthetic parallel corpora by means of automatic translations. This technique is unsafe because most SMT systems are far from perfect and would therefore generate erroneous translations. When the automatically translated texts are used to train a new phrase-based system, the existing errors propagate to the output. Another option is the usage of pivot languages for generating either a new parallel corpus or a translation model by aligning the phrase translations. This technique performs well for closely related languages (e.g. French, Italian, Spanish), but the performance declines for distant language pairs (French, Greek).

A neat solution to the problem of lacking parallel data is the exploitation of other multilingual resources (such as comparable corpora) with the purpose of identifying parallel text segments. Significant efforts have been devoted to mining parallel texts from comparable ones, but previous approaches concentrated mainly on news corpora for a handful of well-resourced languages. With the expansion of the Web, approaches have also been applied for Web corpora. This work follows the second trend and introduces language- and domain-independent methods for extracting parallel texts from freely available comparable corpora from the Web (specifically Wikipedia and the Common Crawl).

Moreover, we apply our methods to the narrow domain of Alpine texts (texts about mountains and mountaineering expeditions, articles about the geology or the culture of the Alps). We choose to work with this domain because we can exploit parallel texts of good quality previously collected in the Text+Berg project (see Section 1.6). This will allow us to optimize the SMT systems for this specific domain and to achieve better translations. The extraction approach is not conditioned by preexistent in-domain parallel texts, but we expect a better performance of the systems optimized with in-domain texts.

Further motivation to work on this domain is the challenge to work with other types of texts than the ones usually exploited in the literature, such as news or parliamentary texts. Our texts often use a descriptive language, which is characterized by a high number of adjectives or adverbs, sometimes embedded linguistic structures and figures of speech. Non-linear sentences like the following *Die folgenden Tage kämpften wir den langwierigen, aber zum Schluss erfolgreichen Papierkrieg, der scheinbar unvermeidlich als erstes Hindernis von jeder Expedition überwunden werden muss.* require great efforts in order to reproduce their semantics in a different language. We are interested in investigating how a domain-specific system handles these kind of texts.

Since the phrase-based approach was considered state-of-the-art at the time we started this project, our experiments focus on phrase-based MT, denoted throughout this work SMT. This work is nevertheless of general interest for the MT community because it provides insight into state-of-the-art domain-specific SMT for a narrow domain. Since the developed approach is to a great extent domain- and language-independent, cross-domain comparisons can be easily performed in order to check the relevance of different domain adaptation techniques. The findings will be particularly relevant when building resources for non-mainstream language pairs and domains.

1.2 Comparable Corpora

Although there is no formal definition of comparable corpora in the literature, the generally accepted reading of the term is: a collection of texts in different languages selected by their similarity in terms of a set of dimensions (topic, genre, time period, etc.) (Sharoff et al., 2013). On the other hand, Fung and Cheung (2004) provides a more fine grained distinction between comparable corpora in terms of their degree of comparability:

- **Parallel corpora** can be seen as particular cases of comparable corpora in which documents represent bilingual translations of each other and are thus aligned at sentence level. Typical examples are corpora used for training SMT systems, such as Europarl [1] or United Nations corpus [2].
- **Noisy parallel corpora** are collections of texts not strictly aligned at sentence level, but which contain translations of each other to a great extent, such as the Proceedings of the European Parliament, from which the Europarl corpus has been extracted.
- **Quasi-comparable corpora** are collections of text on the same topic, which do not represent strict translations of each other and have thus a different structure. Examples are Parliamentary debates on the tax system in the Canadian Hansard and the European Parliament, respectively, or Wikipedia articles.
- **Very-non-parallel corpora** are collections of texts which might (but don't have to) be topic-related, with clearly different structure. An example is the TDT3 corpus, which contains news stories from the period 1998-2000 in different languages[3] or the texts available on the Web.

News feeds, which represent an accessible resource for Natural Language Processing (NLP) applications, can fall in any of the mentioned categories. They can report on the same topic and be released simultaneously by the press agency (in which case they are likely to contain translated content), but they can also be edited independently (in which case the content may vary significantly). We will exemplify this phenomenon with pairs

[1] http://www.statmt.org/europarl/

[2] https://conferences.unite.un.org/uncorpus

[3] https://catalog.ldc.upenn.edu/LDC2001T58

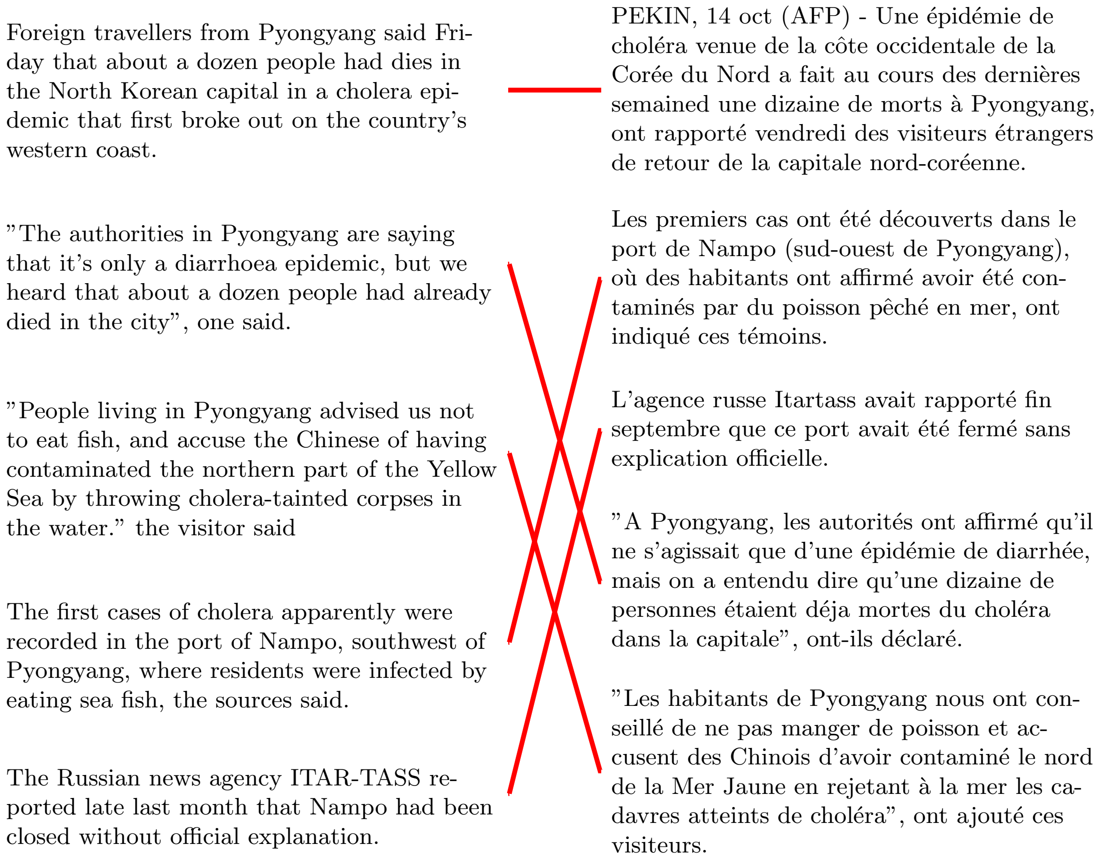
of English-French news released by the Agence France Presse. In the following examples, the correspondence between paragraphs will be marked with red lines.

For example, the news articles in Figure 1.1 can be considered parallel, as each paragraph on the English side has a correspondent on the French side and they occur in the same order. Moreover, they contain accurate translations of each other and could as well be used as training material for SMT.

ALEPPO, Syria, November 10 – I came across Shuruq by chance one day, while walking in a playground in Aleppo. The nine-year-old little girl was playing with her brother, two sisters and mother. Since she has no legs, her big brother was pushing her on the swing.	ALEP (Syrie), 10 novembre 2014 – Je suis tombé sur Shuruq par hasard en me promenant dans un square d'Alep. Cette petite fille de 9 ans jouait en compagnie de son frère, de ses deux sœurs et de sa mère. Comme elle n'a pas de jambes, c'était son grand frère qui poussait la balançoire pour elle.
Aleppo was once Syria's economic capital. The town has been ravaged by more than two years of merciless fighting between rebels and Syrian government forces. Daily bombings by forces loyal to President Bashar al-Assad have left thousands of people dead and caused mass destruction. Shuruq's mother says she lost her legs when a bomb destroyed her home.	Alep était autrefois la capitale économique de la Syrie. La ville est ravagée depuis deux ans par les combats sans merci entre la rébellion et les forces gouvernementales. Les bombardements quotidiens par l'armée loyale au président Bachar Al-Assad y ont fait des milliers de morts et provoqué d'énormes destructions. Selon sa mère, Shuruq a perdu ses deux jambes à cause d'une bombe qui a détruit sa maison.
....	...
Shuruq is a smart, friendly little girl, full of energy. She learned to walk on her hands and moves around nimbly. I didn't speak to her much – she was far more interested in playing than being photographed.	Shuruq est une petite fille intelligente, amicale et pleine d'énergie. Elle a appris à marcher sur ses mains et se déplace avec agilité. Je ne lui ai pas beaucoup parlé : elle était plus intéressée de jouer dans le square que de se faire prendre en photo.
Here in Aleppo, there is a school that has decided to take in all the handicapped children and help them lead a more or less normal life – whether or not their handicap is linked to the war. Life here is very tough for everyone. Even more so for these mutilated children you often see selling sweets on the street to survive.	A Alep, nous avons une école qui a pris l'initiative d'accueillir tous les enfants handicapés et qui les aide à mener une existence à peu près normale, que leur handicap soit dû à la guerre ou non. La vie est très dure ici pour tout le monde, elle l'est encore plus pour ces enfants mutilés que l'on voit souvent vendre des sucreries dans la rue pour survivre.

Figure 1.1: Example of parallel news feeds.

On the other hand, the news articles in Figure 1.2 are representative for the category *noisy parallel corpora*. We notice that most paragraphs are aligned, although they do not occur in the same order in both languages. However, the aligned paragraphs actually contain parallel text segments, which can be extracted and further exploited in various NLP applications.



Foreign travellers from Pyongyang said Friday that about a dozen people had dies in the North Korean capital in a cholera epidemic that first broke out on the country's western coast.

"The authorities in Pyongyang are saying that it's only a diarrhoea epidemic, but we heard that about a dozen people had already died in the city", one said.

"People living in Pyongyang advised us not to eat fish, and accuse the Chinese of having contaminated the northern part of the Yellow Sea by throwing cholera-tainted corpses in the water." the visitor said

The first cases of cholera apparently were recorded in the port of Nampo, southwest of Pyongyang, where residents were infected by eating sea fish, the sources said.

The Russian news agency ITAR-TASS reported late last month that Nampo had been closed without official explanation.

PEKIN, 14 oct (AFP) - Une épidémie de choléra venue de la côte occidentale de la Corée du Nord a fait au cours des dernières semained une dizaine de morts à Pyongyang, ont rapporté vendredi des visiteurs étrangers de retour de la capitale nord-coréenne.

Les premiers cas ont été découverts dans le port de Nampo (sud-ouest de Pyongyang), où des habitants ont affirmé avoir été contaminés par du poisson pêché en mer, ont indiqué ces témoins.

L'agence russe Itartass avait rapporté fin septembre que ce port avait été fermé sans explication officielle.

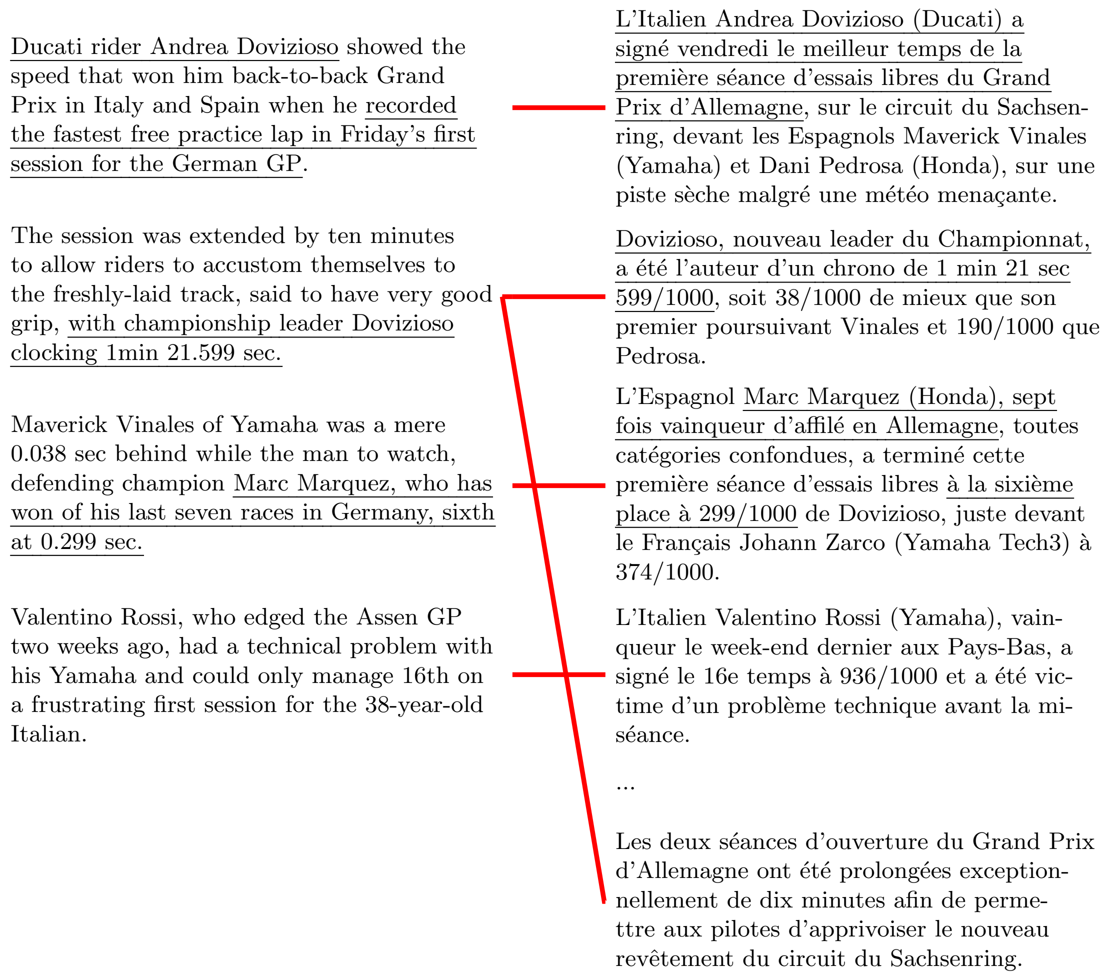
"A Pyongyang, les autorités ont affirmé qu'il ne s'agissait que d'une épidémie de diarrhée, mais on a entendu dire qu'une dizaine de personnes étaient déja mortes du choléra dans la capitale", ont-ils déclaré.

"Les habitants de Pyongyang nous ont conseillé de ne pas manger de poisson et accusent des Chinois d'avoir contaminé le nord de la Mer Jaune en rejetant à la mer les cadavres atteints de choléra", ont ajouté ces visiteurs.

FIGURE 1.2: Example of "noisy non-parallel" comparable corpora (Munteanu and Marcu, 2005).

Unfortunately, many news feeds fall into the third category of quasi-comparable corpora, illustrated in Figure 1.3. In this case, the articles report about the same event, but focus on different aspects and thus provide different information. Corresponding paragraphs are less frequent than in the previous cases and even so, they contain paraphrases or additions on either side (e.g. the first paragraph). Moreover, a paragraph on the English side can be scattered over two French paragraphs and supplemented with new information, not found anywhere on the English side, such as in the case of the second English paragraph. Finally, there are also paragraphs which occur only in one language and do not have a correspondent in the other language.

The last category subsumes completely unrelated news, which sometimes share the same topic and other times diverge from each others topics. Such texts can still be useful for comparative linguistic research because the texts have been written independently from each other and are yet topically related. Moreover, their availability is not limited to a handful of languages. From the perspective of SMT, such texts cannot be used to build SMT systems alone, but can help improve the language models, which are responsible for the fluency of the MT output.



Ducati rider Andrea Dovizioso showed the speed that won him back-to-back Grand Prix in Italy and Spain when he recorded the fastest free practice lap in Friday's first session for the German GP.

The session was extended by ten minutes to allow riders to accustom themselves to the freshly-laid track, said to have very good grip, with championship leader Dovizioso clocking 1min 21.599 sec.

Maverick Vinales of Yamaha was a mere 0.038 sec behind while the man to watch, defending champion Marc Marquez, who has won of his last seven races in Germany, sixth at 0.299 sec.

Valentino Rossi, who edged the Assen GP two weeks ago, had a technical problem with his Yamaha and could only manage 16th on a frustrating first session for the 38-year-old Italian.

L'Italien Andrea Dovizioso (Ducati) a signé vendredi le meilleur temps de la première séance d'essais libres du Grand Prix d'Allemagne, sur le circuit du Sachsenring, devant les Espagnols Maverick Vinales (Yamaha) et Dani Pedrosa (Honda), sur une piste sèche malgré une météo menaçante.

Dovizioso, nouveau leader du Championnat, a été l'auteur d'un chrono de 1 min 21 sec 599/1000, soit 38/1000 de mieux que son premier poursuivant Vinales et 190/1000 que Pedrosa.

L'Espagnol Marc Marquez (Honda), sept fois vainqueur d'affilé en Allemagne, toutes catégories confondues, a terminé cette première séance d'essais libres à la sixième place à 299/1000 de Dovizioso, juste devant le Français Johann Zarco (Yamaha Tech3) à 374/1000.

L'Italien Valentino Rossi (Yamaha), vainqueur le week-end dernier aux Pays-Bas, a signé le 16e temps à 936/1000 et a été victime d'un problème technique avant la mi-séance.

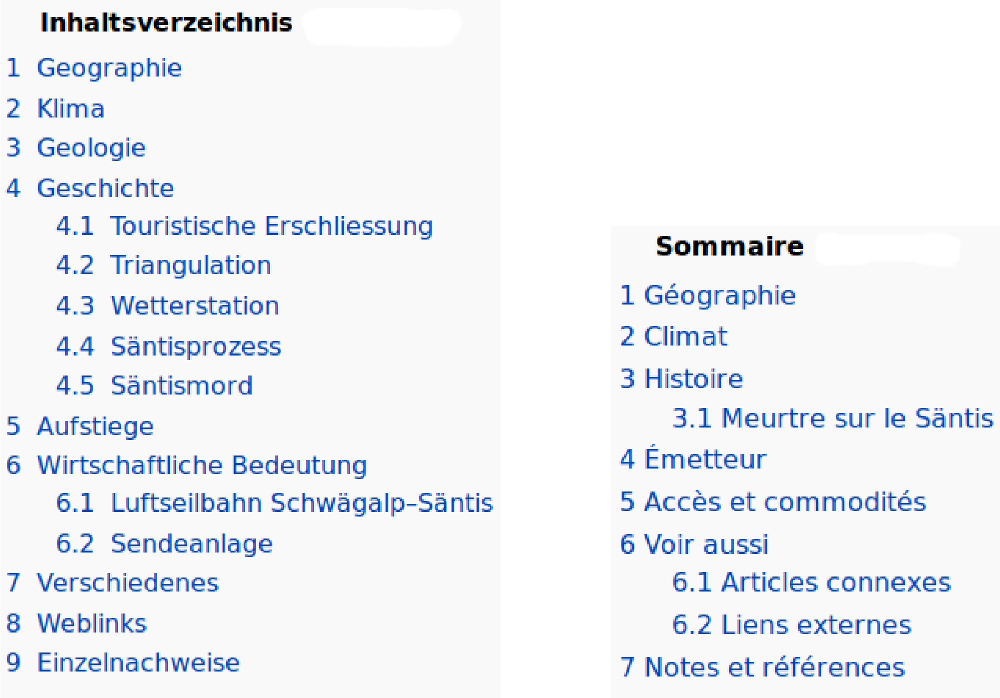
...

Les deux séances d'ouverture du Grand Prix d'Allemagne ont été prolongées exceptionnellement de dix minutes afin de permettre aux pilotes d'apprivoiser le nouveau revêtement du circuit du Sachsenring.

Figure 1.3: Example of quasi-comparable corpora

Similar phenomena are frequently encountered in Wikipedia as well, but the advantage of Wikipedia is that its articles are topically-related. To this point, no study has investigated the comparability degree of Wikipedia articles, but the general assumption is that most articles share pieces of information, although they are written independently from each other. We will give a first intuition based on the table of contents of the German and French articles about the mountain *Säntis*, depicted in Figure 1.4.

We notice that the structure of the German article is more detailed than its French correspondent and that the order of the sections differs (similar to the example in Figure 1.3). Moreover, if we look at the contents of the corresponding sections (e.g. *Klima* and *Climat*), we notice that the German side contains much more information than the French counterpart. Moreover, if we compare the length of the articles, the German article is 2.7 times longer than the French one (1630 words vs. 600 words). The length ratio of Wikipedia articles across languages at word level can be even higher, ranging from 1:219 in favor of the German version to 75:1 in favor of the French version. We will investigate if, even in such extreme cases, articles share parallel text segments.



Inhaltsverzeichnis

1 Geographie
2 Klima
3 Geologie
4 Geschichte
4.1 Touristische Erschliessung
4.2 Triangulation
4.3 Wetterstation
4.4 Säntisprozess
4.5 Säntismord
5 Aufstiege
6 Wirtschaftliche Bedeutung
6.1 Luftseilbahn Schwägalp-Säntis
6.2 Sendeanlage
7 Verschiedenes
8 Weblinks
9 Einzelnachweise

(A) German ToC

Sommaire

1 Géographie
2 Climat
3 Histoire
3.1 Meurtre sur le Säntis
4 Émetteur
5 Accès et commodités
6 Voir aussi
6.1 Articles connexes
6.2 Liens externes
7 Notes et références

(B) French ToC

Figure 1.4: The table of contents (ToC) of the German and French versions of the same Wikipedia article.

In this work, we will focus on two types of comparable corpora: Wikipedia and the Web. We will investigate whether they contain parallel segments and how their structure facilitates their extraction. The next section will first describe the particularities of Wikipedia which are relevant for NLP applications, in general, and the extraction of parallel segments, in particular.

1.3 Wikipedia

Wikipedia is an important multilingual resource available for a variety of domains, in almost 300 languages. Its particularity is that the resource is under ongoing development, i.e. it is constantly being updated and extended. According to official statistics[4], over the past 6 years Wikipedia has been growing yearly with approximately 20% (in terms of overall number of articles). This aspect poses difficulties in replicating experiments, since results are likely to change over time. On the other hand, it also opens the possibility to improve results when more texts become available.

[4] http://stats.wikimedia.org/EN/TablesArticlesTotal.htm

Wikipedia can be seen as a valuable linguistic resource due to its structured content, such as the topical organization of articles, the internal link system referring to both internal articles and external resources and its multilingualism. We next list the structural elements of Wikipedia that are relevant for NLP, in general, and partially also for our purpose.

Articles are the basic information units in Wikipedia and they correspond to encyclopedic concepts. Although their structure is not fixed, articles are required to start with a definition of the concept. Articles consist of text (with specific markups), images, tables or stand-alone lists and can include references to other Wikipedia articles.

Inter-language links make the connections between Wikipedia articles on the same topic in multiple languages. Although this information is displayed in the online version of Wikipedia, neither the source code of the page nor the Wikipedia dumps encode it. This information is dynamically aggregated to the page from a different database, Wikidata[5]. By the time we started our extraction experiments, inter-language links were part of the Wikipedia dumps, therefore our approach fully exploited this feature.

The following snippet illustrates how the inter-language links are marked in the 2011 Wikipedia dumps. The markup consists of pairs of language codes and article titles in the corresponding languages. In this example, extracted from the German Wikipedia, the links point to articles about the mountain pasture in several languages, such as English, Spanish, Finish etc.

```
[[en:Alpine meadow]]
[[es:Pradera alpina]]
[[fi:Alppiniitty]]
[[fr:Alpage]]
[[io:Alpo]]
```

(Internal) hyperlinks represent references to Wikipedia articles in the same language which denote relevant stand-alone concepts. The Wikipedia guidelines recommend a moderate usage of the hyperlinks, in order to ensure the article readability. The pairing of the anchor text and the title of the referred article can be used for synonym extraction or word sense disambiguation. The links are coloured in blue in the web version and are marked with square brackets in the underlying source code (see the following example).

[5] www.wikidata.org

```
Abgeleitet ist er vom früh-rätoromanischen Eigennamen Sambatinus
(der am Samstag Geborene), womit zunächst wohl eine am Berghang
gelegene Alp bezeichnet wurde.

Abgeleitet ist er vom früh-[[Rätoromanische Sprache|
rätoromanischen]] [[Eigenname]]n ''Sambatinus'' (der am Samstag
Geborene), womit zunächst wohl eine am Berghang gelegene
[[Alm (Bergweide)|Alp]] bezeichnet wurde.
```

In this example, the link associated with the word *rätoromanischen* redirects the reader to the article about Rhaeto-Romance languages (*Rätoromanische Sprache*). Similarly, the word *Alp* contains a hyperlink to the disambiguated page *Alm (Bergweide)* (EN: mountain pasture). Since the word *Alm* can refer to several concepts in German (e.g. pasture or lime precipitation), the correct meaning in the context is specified in brackets.

Redirect pages are pages which only contain a link to a different Wikipedia article or to a specific section thereof on the same topic, in the same language. This feature ensures the minimization of duplicate content and can also be used for synonym extraction.

In the following example we list the source code of the page *Rätoromanisch*. The actual content of the page is subsumed in the `<text>` tag, all other tags being relevant for identification. In this case, the page only contains a link to the Wikipedia page *Bündnerromanisch*. This term can be used alternatively with *Rätoromanisch* to denote the language spoken in the canton of Graubünden in Switzerland.

```
<page>
  <title>Rätoromanisch</title>
  <id>8</id>
  <redirect />
  <revision>
    <id>38028819</id>
    <timestamp>2010-09-3T09:33:05Z</timestamp>
    <contributor>
      <id>38244</id>
    </contributor>
    <text xml:space="preserve">#REDIRECT [[Bündnerromanisch]]
    </text>
  </revision>
</page>
```

Disambiguation pages represent pages detailing possible meanings of a term, each of them being associated with a brief explanation and a link to the corresponding Wikipedia article. This can be seen in the following excerpt from the Wikipedia page *Kiefer*, which contains several references related to the different readings of the term: as tree and as jaw. The page also includes reference to pages about persons whose last name is Kiefer (not listed here).

```
Kiefer steht für:
* Kieferngewächse (Pinaceae), eine Familie der Pflanzen
* Kiefern oder Föhren (Pinus), eine Gattung der Nadelholzge-
    wächse mit zirka 115 Arten
* Waldkiefer (Pinus sylvestris), Gemeine Kiefer, Rotföhre, eine
    Art der Kiefern
* Schwarzkiefer (Pinus nigra) als vorherrschende Kiefernart in
    Südostösterreich
* Kiefer (Anatomie), ein dem Kauen dienender Knochen der Wirbel-
    tiere
* Kiefer (Insekt), Teile der Mundwerkzeuge von Insekten
```

Categories represent "abstract" concepts that subsume the articles in the Wikipedia. This feature provides the means to organize Wikipedia articles by grouping together articles on similar topics. The categories are organized in a hierarchic structure, although the relation between them is not always *ISA*.

This is exemplified in Figure 1.5, which depicts the Wikipedia category structure from the root to the category *Alpinismus* (EN: Alpinism). In this view, only categories subsuming Alpinism are depicted, but the category structure is more expanded in breadth. For example, the category *Sport* subsumes more than 20 categories apart from the one listed here. We notice there is an *ISA* relation between the end category (Alpinism) and its antecedents *Sport nach Sportart* (EN: type of sport) and *Sport*. This is not the case for its other direct antecedent category *Umwelt und Natur* (EN: environment and nature).

As shown before, Wikipedia is a rich resource with several layers of structured information. This feature facilitates the extraction of various information relevant in Language Technology research, from cross-lingual lists of named entities (useful for named entity recognition) to comprehensive lists of possible readings for ambiguous terms (useful for word sense disambiguation). In this work we will focus on exploiting Wikipedia (and comparable corpora, in general) for SMT, therefore the next subsection briefly introduces a few usage scenarios.

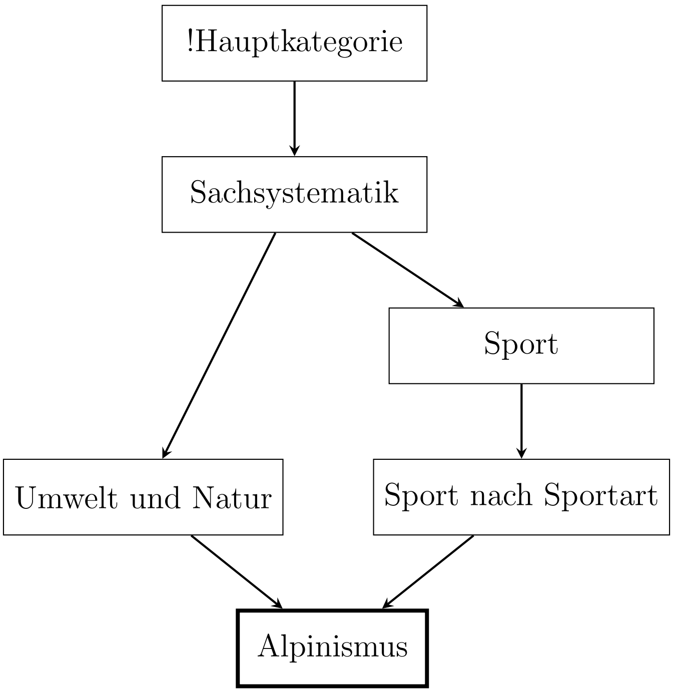


FIGURE 1.5: The Wikipedia category structure subsuming the category *Alpinismus* (EN: Alpinism).

1.4 Exploiting Comparable Corpora for SMT

Parallel data (i.e. bilingual texts representing translations of each other) is crucial for many NLP applications, including SMT. The main usage of comparable corpora for SMT is mining for parallel sentences. The extracted texts can either be used as the main training corpus or as additional training corpora (in different model mixtures). The latter applies especially when translating texts from a specific domain, for which only small amounts of parallel texts are available. The combinations are usually applied to the two core models involved in the translation process, namely the translation model (responsible for the translation variants) and the language model (which ensures the fluency of the output).

The extraction process is equivalent to a multi-level alignment procedure, starting at document level and ending at sentence, phrase or word level, depending on the intended usage. If the extracted data will be used for phrase-based SMT, the final alignment granularity will be at sentence or phrase level. If the extracted data is needed for a bilingual terminology database, the final alignment will be performed at word level.

Document alignment is a challenge especially when working with news collections, since neither the release date nor the title are always reliable anchor points. Moreover, some news are only published in one language. Most approaches proposed in the literature suggest limiting the search space to a time span of a couple of days and then matching the documents by means of word overlap. Other text sources, instead, contain meta-information which simplify the task of document alignment. For example, web pages can be aligned at document level by means of URL matching, whereas in Wikipedia the task is trivial, since the cross-lingual equivalences are available by default.

Comparable corpora are also used for extracting domain-specific bilingual terminology, multi-word expressions or named entities. The most frequent approach is to compute the similarity between the context vectors of different word pairs and to extract only the pairs with a high similarity. In order to compare the context vectors, a seed bilingual dictionary is required. Some approaches bypass the need for a dictionary by relying on identical words or cognates, which are frequent in case of similar languages, such as Swedish and Danish.

Such resources can be easily plugged-in to a SMT system afterwards and are especially beneficial when only small amounts of parallel data are available for the considered language pair and domain. The resulting hybrid systems usually outperform the baseline SMT systems due to the enhanced terminological resources. Chapter 2 will provide a more detailed overview of approaches exploiting comparable corpora for SMT.

1.5 Domain Adaptation for SMT

Domain adaptation subsumes approaches that exploit extensively texts similar to the ones to be translated (also called in-domain texts) in order to improve the performance of a general SMT system. The challenge of this task lies in the limited availability of in-domain texts, as compared to the general domain texts (also referred to as out-of-domain texts). One might think that simply adding in-domain texts to the existing training texts can solve the problem, but, in practice, we have no control on the preferred translation in case an input word has different translations in the in- and out-of-domain corpora, respectively. Therefore we need more precise adaptation methods, which can give preference to the in-domain texts.

The following example illustrates the need for domain adaptation in machine translation. We translated a sentence from the Alpine domain with several commercial translation systems and with an in-house SMT system trained with in-domain data. The challenge lied in conveying the correct translation in context of the German noun *Pass*, which can mean *passport, passage or (mountain) pass.* General domain MT systems, such as Google Translate, Personal Translator (PT 18) or Systran, provide the most frequent translations of the word: *passeport or passage* (EN: passport, passage). The system trained with in-domain data, instead, generates the appropriate translation in this context: *col* (EN: mountain pass). Domain adaptation can help MT systems deal with such ambiguity problems by adjusting the translation probabilities in favour of the domain-specific translations.

Source:	Der nächste Pass, der Trescolmen (2161 m), war schlimmer.
Reference:	Le deuxième col (Trescolmen 2161 m) fut pire.
Google:	Le passage suivant, le Trescolmen (2161 m) était pire.
PT 18:	Le prochain passeport, le m Trescolmen (2161), a été pire.
Systran:	Le prochain passeport, le Trescolmen (2161 m), était plus mauvais.
Our system:	Le prochain col, le Trescolmen (2161 m), fut pire encore .
Gloss:	The next mountain pass, the Trescolmen (2161 m), was worse.

Typical domain adaptation approaches used in the literature are filtering of domain-like training data or weighting the models trained on different training corpora. These approaches can be used both independently (especially weighted models) or in conjunction. For example, Moore and Lewis (2010) propose a perplexity-based method to select in-domain-like monolingual texts and use the selected texts for language modelling. Axelrod et al. (2011) propose a similar method, but tailored for the selection of parallel texts. The resulting corpus is then used for training translation models, both in isolation and combined with an in-domain translation model.

By far the most frequent technique used in practice is mixture modelling, that is the weighted combination of the core models used by the SMT system (the language and the translation models). The component models are trained on the individual corpora and then mixed through a linear or log-linear combination (Foster and Kuhn, 2007). The translation models can also be combined through instance weighting, i.e. by assigning weights on sentence or phrase level (Matsoukas et al., 2009, Foster et al., 2010).

In our experiments, we follow the method introduced by Sennrich (2012) for translation model combination. The author suggests a weighted combination of the translation models with the objective of minimizing perplexity, which can be applied to any number of models. The supported combination methods are linear interpolation and instance weighting (in case there is a small in-domain development test set). Language models are combined by means of weighted linear interpolation with weights computed on an in-domain development set. We also withheld a part of the in-domain parallel corpus (which will be described in the following section) for optimization purposes.

1.6 The Alpine Domain

The notion of domain is not a clearly defined concept in Computational Linguistics. In the literature, the term has been used to refer to news stories, parliamentary texts, medical texts or movie subtitles as well. It is however, an open question, how the distinction between domains should be made. For example, in which category should a news article about a recent discovery in the treatment of cancer be placed (news vs. medical)? Throughout this work, we refer to domain as to the subject of the texts, in a broader sense. We therefore consider that texts from the same domain use a common vocabulary and are make similar lexical choices. If they share the same writing style, we can also talk about a common genre. In reality, texts sharing the same domain can have different functions and cover different writing styles, therefore also cover different genres.

Our experiments are carried out for the Alpine domain, which covers texts about mountains altogether, mountaineering expeditions etc. We choose this domain because we can exploit the outcomes of a previous project, the Text+Berg project[6]. Its purpose is to digitize Alpine texts from different sources: the yearbooks of the Swiss Alpine Club from 1864 until today, the French journal *Echo des Alpes* from 1872 until 1924 and the British Alpine Journal from 1969 until 2008. The resulting corpus contains texts which focus on conquering and understanding the mountains and covers a wide variety of text genres such as expedition reports, (popular) scientific papers, book reviews, etc. The corpus is multilingual and contains articles in German (some also in Swiss German), French, English, Italian and even Romansh.

Whilst the Echo des Alpes yearbooks and the Alpine Journal are exclusively in French and English, respectively, the Swiss yearbooks represent a language mix. Initially, the yearbooks contained mostly German articles and few in French. Since 1957 the books appeared in parallel German and French versions and from 2012 in trilingual versions German, French and Italian. This sums up to 57 parallel editions German-French, a bit more than 4200 parallel articles and almost 300,000 parallel sentences, aligned with Bleualign (Sennrich and Volk, 2010). In contrast, the German-French-Italian parallel corpus only contains 600 articles and estimated 26,000 parallel sentences. For our experiments with domain-specific SMT we use the German-French parallel corpus, which represents a valuable in-domain corpus for the Alpine domain.

The corpus is available in XML format and is annotated with both structural (article boundaries, title and author information, footnotes and captions) and linguistic information (sentence boundaries, language identification, Part-of-Speech tags and lemmas),

[6] www.textberg.ch

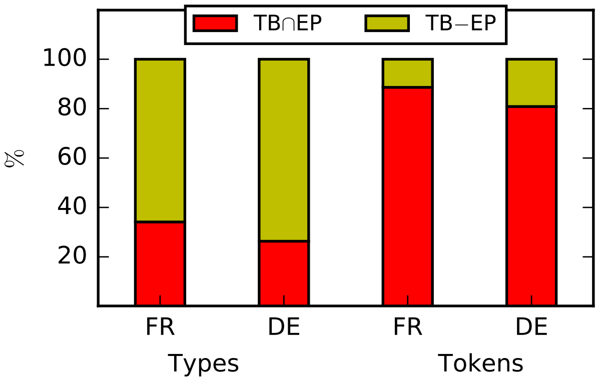


FIGURE 1.6: Domain Overlap between Text+Berg (TB) and Europarl (EP) with respect to the TB word frequencies.

as well as stand-off annotations of toponyms, person names and temporal expressions. Since we are working with phrase-based SMT, we only extract the textual information (word forms) from the corpus.

To illustrate the distinction between domains, we compute the overlap between our Alpine corpus and a corpus from a totally different domain, Europarl, which contains parliamentary proceedings. Here we compare only the German-French parallel sections thereof, which we also use in the SMT experiments in Chapter 6. We compute the overlap on token and type level, both on the French and on the German side. The results are depicted in Figures 1.6 and 1.7.

The statistics only consider content words occurring more than once in the respective corpora. We deliberately discard function words such as determiners, prepositions, conjunctions, pronouns and auxiliary verbs since they occur most frequently, but are not representative for any corpus. Our function words lists contain approximately 300 words in German and French, respectively. They have been extracted from word frequency lists compiled in the project *Deutscher Wortschatz* at the University of Leipzig [7]. A manual inspection was required since the lists also contained content words.

On type level (number of distinct words), the figures are similar for both corpora. The domain-specific vocabulary (i.e. words occurring only in one of the corpora), be it Alpine or parliamentary texts, covers 60-65% of the French vocabulary and 74-78% of the German one. The rest of the vocabulary is shared by the two corpora. The difference between the German and the French figures is caused by the different frequencies of the domain-specific words. For example, the most frequent German word specific to the Text+Berg vocabulary occurs 2.3 times more often than the most frequent French word.

[7] http://wortschatz.uni-leipzig.de/html/wliste.html

In Europarl, the ratio between the absolute frequency of the most frequent domain-specific word in German and French, respectively, is 4.2.

On token level, the trends are reversed. The domain-specific words cover only a small part of the words in the considered corpora, namely 11.4% on the French side and 20.3% on the German side of the Text+Berg corpus. For Europarl, the figures are even lower: 4% on the French side and, respectively 12% on the German side. This means that the domain-specific words, although higher in number, occur less frequently than common words. On the other hand, we can conclude that the domain-specific vocabulary has a greater weight in the case of the Alpine corpus. This feature makes the corpus more useful for training domain-adapted SMT systems.

Table 1.1 shows a selection of words specific to the Text+Berg corpus, that do not occur in Europarl. One notices that the lists contain mainly domain-specific words, such as mountaineering related named entities (e.g. *SAC* [Swiss Alpine Club]) of parts thereof (e.g. *Col* as in Col du Midi, *Piz* as in Piz Bernina). A second category refers to generic mountaineering keywords, such as Besteigung (EN: ascent), Alpinisten (EN: alpinists) in German or grimpeurs (EN: climbers), cordée (EN: rope) in French. Another identified category represents specific terminological words referring to alpinism, such as Steigeisen (EN: crampons) or bivouac (EN: bivouac). The lists partially overlap, although the frequency of the respective words is different in German and French, respectively.

The first category (named entities) are unproblematic for an automatic translation system if it has not seen them before, since the system would simply transfer them to the target language and the reader would still be able to understand what they refer to. In order to be able to translate correctly the other two word categories, only a domain-specific dictionary can help. Leaving these words untranslated or generating wrong translation can be fatal to the understanding of the texts, since they contain the

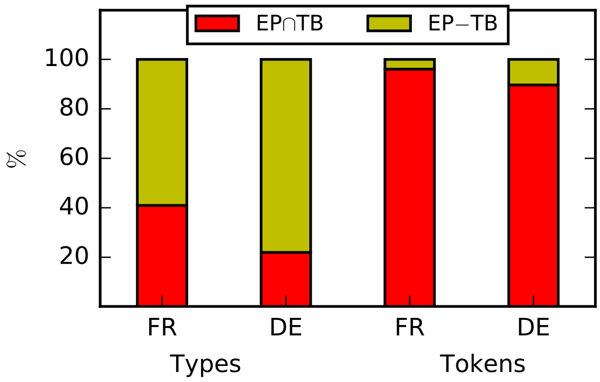


Figure 1.7: Domain Overlap between Text+Berg (TB) and Europarl (EP) with respect to the EP word frequencies.

German		**French**	
Word	**Freq.**	**Word**	**Freq.**
SAC (*Swiss Alpine Club*)	4569	grimpeurs (*climbers*)	1957
Besteigung (*ascent*)	1487	bivouac (*bivouac*)	1113
Alpinisten (*alpinists*)	1128	piz (*peak*)	1105
Piz (*peak*)	1079	cordée (*rope*)	995
Kletterer (*climber*)	1065	pitons (*pitons*)	946
Bergführer (*mountain guide*)	1014	assurage (*security*)	895
Col (*mountain pass*)	991	crampons (*climbing irons*)	840
Kletterei (*climb*)	941	éboulis (*scree*)	792
Alpinismus (*alpinism*)	885	ascensions (*ascents*)	790
Steigeisen (*climbing iron*)	759	névé (*névé*)	703

Table 1.1: Selection of the most frequent words in the Text+Berg corpus which do not occur in Europarl.

essence of the texts. These findings demonstrate once more the importance of using in-domain texts for training SMT systems.

1.7 The Project "Domain specific SMT"

In the project Domain specific Statistical Machine Translation (SMT) [8] we have investigated different ways of incorporating domain-specific texts into the workflow of a SMT system. Specifically, we worked on the Alpine domain and with the Text+Berg corpus [3] described in the previous section. Part of the experiments in this dissertation have been conducted within the frame of this project.

In order to be able to fully exploit the domain-specific texts for SMT, accurate sentence alignments had to be generated. Classical sentence alignment algorithms, such as (Gale and Church, 1993), did not perform reliably for this kind of texts, due to the following factors: few anchor points, missing sentences between corresponding blocks of text, as well as the relatively high number of many-to-many alignments. Therefore Sennrich and Volk (2010) developed an algorithm more robust to the noise caused by the differing positions of the text, also known as Bleualign. The algorithm implements a distance-search to find sentence pairs with the highest similarity and which also comply with the monotonicity constraint. Moreover, this algorithm outperformed existing alignment approaches.

The parallel texts aligned with Bleualign, as well as the monolingual texts from the Text+Berg corpus have been used for several SMT experiments aiming to improve the

[8] http://www.cl.uzh.ch/en/research/machine-translation/domainspecificsmt.html

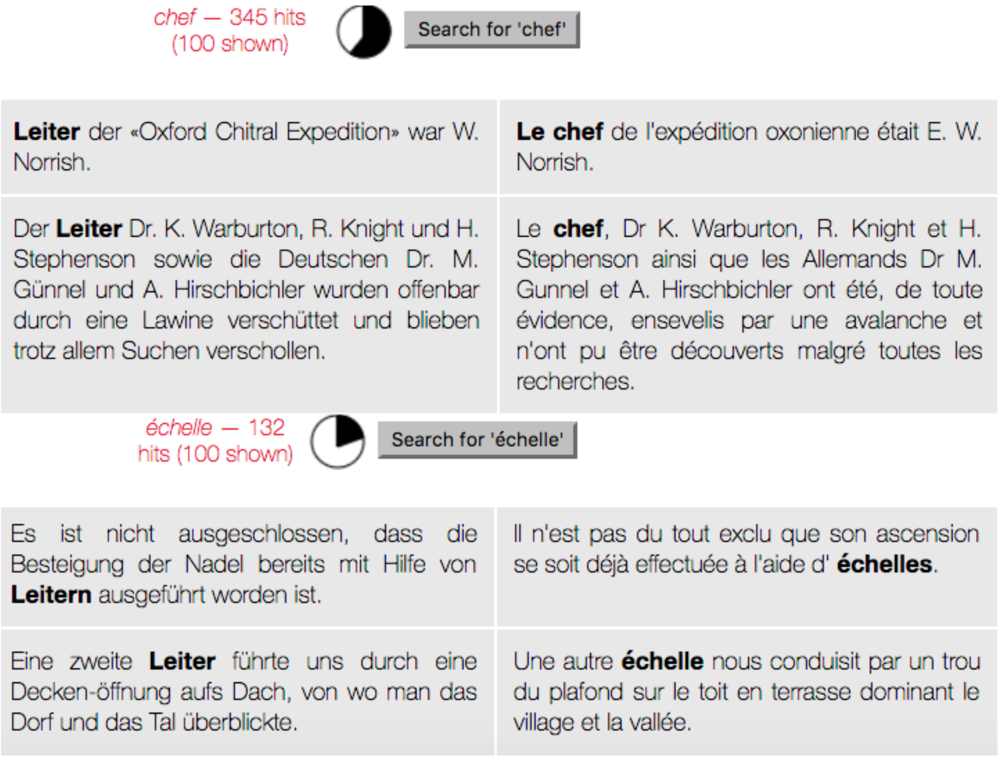
performance of our in-domain system. These include the combination of the out-of-domain and the in-domain models used by the SMT system (language model and translation model) optimized on an in-domain development set. Another experiment aimed at incorporating translations from external systems (such as Google Translate or Personal Translator) with the purpose of filling data gaps. These brought significant performance improvements to the in-domain baseline.

A useful by-product of the project was a bilingual concordancing tool called Bilingwis[9]. The tool identified possible translations of a German term in the French side of the Text+Berg corpus and illustrated them with corresponding examples from the corpus. Apart from being a show-case for the developed alignment algorithm, the tool provides the translations of the words in context. For example, one can check the usage of mountaineering terms such as *Steigeisen or Biwak* or check translation variants of ambiguous words such as *Leiter*.

The following example illustrates the case of the word *Leiter*, which can be translated into French as *chef* (EN: head) or *échelle* (EN: ladder), sometimes also as *moniteur* (EN: monitor), *directeur*(EN: director) or *leader* (EN: leader). Figure 1.8 depicts the suppressed Bilingwis output for the two most frequent translations, *chef* and *échelle*. However, all translation variants can be identified in the Text+Berg corpus, which means that a translation model trained on this corpus will have to consider all these possibilities when translating a new sentence where this word occurs. An advantage in this case is the fact that some readings of the term can be used alternatively (e.g. head, leader or director), thus reducing the translation error rate.

Another research topic of this project was the exploitation of comparable corpora in view of expanding our domain-specific corpus. First we considered aligning the Echo des Alpes collection with the SAC yearbooks, specifically the articles published during an overlapping time period (1872-1924). Taken together, these texts form a noisy comparable corpus, since they cover similar topics and probably report about the same events, given the specific geographical area (the Alps), but are written independently. The difficulty consisted in identifying potentially parallel documents, since the time stamp was not as relevant as in the case of news articles and we therefore had to rely on word matches. Since only a few matches could be identified, we decided not to pursue this direction any further. Instead, we focused on other comparable corpora, in which the document alignment was more likely (such as Wikipedia or the Web). These experiments will be discussed thoroughly in Chapters 3 - 5.

[9] https://pub.cl.uzh.ch/projects/bilingwis



Leiter der «Oxford Chitral Expedition» war W. Norrish.	**Le chef** de l'expédition oxonienne était E. W. Norrish.
Der **Leiter** Dr. K. Warburton, R. Knight und H. Stephenson sowie die Deutschen Dr. M. Günnel und A. Hirschbichler wurden offenbar durch eine Lawine verschüttet und blieben trotz allem Suchen verschollen.	Le **chef**, Dr K. Warburton, R. Knight et H. Stephenson ainsi que les Allemands Dr M. Gunnel et A. Hirschbichler ont été, de toute évidence, ensevelis par une avalanche et n'ont pu être découverts malgré toutes les recherches.

Es ist nicht ausgeschlossen, dass die Besteigung der Nadel bereits mit Hilfe von **Leitern** ausgeführt worden ist.	Il n'est pas du tout exclu que son ascension se soit déjà effectuée à l'aide d' **échelles**.
Eine zweite **Leiter** führte uns durch eine Decken-öffnung aufs Dach, von wo man das Dorf und das Tal überblickte.	Une autre **échelle** nous conduisit par un trou du plafond sur le toit en terrasse dominant le village et la vallée.

FIGURE 1.8: Screenshot of the Bilingwis concordancing tool for the search term *Leiter*.

1.8 Research Questions

This thesis will address a series of questions related to the extraction of parallel texts from comparable corpora and their usage in a SMT scenario.

1. To what extent do comparable corpora contain cross-lingual overlapping information (i.e. translated pieces of text)?
2. How can comparable corpora be exploited for mining parallel pieces of text (e.g. sentences, clauses, fragments)?
3. How can we measure the similarity between two candidate sentences from different languages?
4. How can we evaluate the quality of the extracted sentence pairs?
5. Will the extracted corpus improve the performance of an SMT system?

1.9 Thesis Outline

In this chapter we briefly introduced the key resources of the current work: comparable corpora, in general, and Wikipedia, in particular. We also demonstrated the importance of exploiting such corpora for (domain-specific) Statistical Machine Translation and gave a short overview of domain adaptation approaches. Finally we described the domain we focus on in this thesis: Alpine texts and stated the research questions that the current work will clarify. The rest of this thesis is structured as follows:

Chapter 2 *Comparable Corpora for Statistical Machine Translation* gives an overview of previous approaches for extracting parallel sentences from comparable corpora. We also pinpoint the similarities between existing methods and the methods presented in this work.

Chapter 3 *Extracting Parallel Sentences from Wikipedia* describes our approach for extracting parallel sentences from Wikipedia.

Chapter 4 *Extracting Parallel Sub-Sentential Segments from Wikipedia* illustrates a refined approach for extracting parallel text segments from Wikipedia.

Chapter 5 *Extracting Parallel Sentences from the Web* describes our approach for extracting parallel segments from the Common Crawl, a public crawl of the Web hosted on Amazon's Elastic Cloud.

Chapter 6 *SMT experiments* presents our experiments with the extracted data for domain-specific SMT and discusses our main findings.

Chapter 7 *Conclusions* summarizes the main contributions and findings of this thesis and provides an outlook for future research on exploiting comparable corpora for SMT.

Chapter 2

Comparable Corpora for Statistical Machine Translation

Comparable corpora represent a useful resource for different Natural Language Processing tasks, such as word sense disambiguation, text mining and Statistical Machine Translation (SMT), as they promise to bypass the lack of parallel corpora. This chapter reviews some of the approaches exploiting comparable corpora in respect to SMT. Moreover, we discuss how they relate to the extraction approaches described in the current work.

SMT is a data-driven approach, therefore it requires large amounts of bilingual texts to identify and assess regularities in the data (e.g. equivalences between words/phrases, word order). An SMT system learns from the translations it has seen during training and assigns probabilities for each possible translation of a word sequence in a given context. When translating a new sentence, the system recombines known text fragments, in order to yield the best possible output (i.e. maximize the total probability).

There are several criteria which play a significant role in the development of SMT systems and thus influence their quality. First, the *availability* of parallel texts for the desired language pair and/or domain. The available parallel corpora cover a limited number of language pairs. Moreover, considering that SMT systems specialized in translating specific texts (e.g. travel reports, parliamentary proceedings) profit from parallel texts from the same domain, their availability is restricted even more. Secondly, the *size* of the available parallel corpora strongly influences the quality of the SMT output. Experiments indicate that a corpus of 10 million words is a good starting point to build a state-of-the-art SMT system (Hardmeier and Volk, 2009). However, such an amount of training data is rarely at hand for most language pairs, even without restricting the search space to a specific domain. Last but not least, the *quality* of the training data is crucial to the performance of the SMT system. Training datasets have to consist of mutual translations, otherwise the system will not be able to learn accurate word and phrase alignments.

Consequently, the efforts of building parallel corpora are high, both in terms of time and costs. In recent years, researchers have sought to exploit readily available resources, such as comparable corpora, to alleviate this bottleneck. The proposed approaches focused on extracting possible translations (on word, phrase or sentence level) and using them as additional training material in SMT. The efforts concentrated mainly on news corpora and recently also on web corpora.

2.1 Approaches for News Corpora

One of the most influential works is that of Munteanu and Marcu (2005), who proposed a maximum entropy classifier for identifying parallel sentences in newspaper articles. Articles stemmed from two monolingual, independent collections, therefore no predefined correspondence between the articles existed. A bilingual dictionary (learned from external parallel corpora) was used to identify similar articles, and then to filter the candidate sentence pairs. The extracted corpus was evaluated as training material for an out-of-domain SMT system, achieving significant performance improvements.

To distinguish between parallel and non-parallel sentences, the classifier used a set of features based on sentence lengths and word alignments. Some of them, such as the sentence length ratio or the number of unaligned content words, are also part of the similarity metric we proposed to tackle this problem. Since most features are based on word alignments, their accuracy is of utmost importance. The authors computed the alignments in five different manners (variations of the IBM Model 1), whereas we rely on the internal alignment generated by METEOR. It is noteworthy that the authors used bilingual alignments, whereas we work with monolingual alignments.

The extracted data was then used as additional training data for several phrase-based SMT systems, some of them out-of-domain and others mixed (in- and out-of-domain). Significant BLEU score improvements have been reported when translating from Arabic and Chinese, respectively, into English (1.2-10 BLEU for Arabic-English and 1-4.5 BLEU for Chinese-English). In our SMT experiments, we compare the improvements both against an out-of-domain and a purely in-domain baseline system. Moreover, we do not simply concatenate the training data sets (baseline and extracted), but combine them using mixture modelling. We also conduct experiments with different splits of the extracted data in order to analyze the performance improvement in terms of data size, but since our data sets are smaller than the ones reported in the paper, the improvements are not always visible.

Tillmann and Xu (2009) also proposed an approach for extracting parallel sentences from pairs of monolingual news corpora. They adopted an exhaustive approach to compare all possible sentence pairings between a source document and a collection of possibly equivalent documents within a 7 days window, focusing on four methods to optimize the computation of the scoring function. The scoring function was used to describe the "parallelism" of a sentence pair and was based on lexical weights/IBM Model 1.

Our approach is similar in the sense that we also consider all possible sentence pairs, but we have the advantage that the article alignment is given (for Wikipedia articles), therefore the search space is smaller. Likewise, we use a scoring function to rank the candidate sentence pairs, which shares comparison criteria with the function used in (Tillmann and Xu, 2009), such as the sentence length ratio or the percentage of translated words (in our case, aligned words).

The authors reported significant improvements of 3.2-3.3 BLEU when adding the extracted data to the training corpus of an SMT system. The experiments concerned open-domain translation for Spanish-English and Portuguese-English. Unlike in our experiments, no distinction was made between in-domain and out-of-domain training data. Another important finding of this work was the reduction of the computation time of the extraction pipeline, as an effect of the optimized implementation.

Another approach to mine comparable news corpora is presented in (Abdul Rauf and Schwenk, 2011). Here, the authors automatically translated one side of the comparable corpus (source-side) into the second language of the corpus (target) and used the translations as information retrieval (IR) queries on the target-side corpus. Specifically, the search space was restricted to news articles within a window of ±5 days from the publication date of the source article. Candidate sentence pairs were then filtered by means of word or translation error rate (WER, TER, TERp) filters. Additionally, they investigated the benefits of sentence tail removal in case the reference sentence had an extra tail which was not included in the MT query.

In our approach, we also perform a monolingual comparison between an automatic translation of the source language article into the target language and the target article. Since in our case the correspondence between articles is uniquely defined and foreknown, we simply generate all sentence pairings between the articles and subsequently rank them according to a similarity criterion. We also use an automatic evaluation metric inspired from SMT as a component of the scoring function ranking the candidate sentence pairs, but none of the ones used in this paper.

The extracted data was then used in different SMT experiments for the language pairs Arabic-English and French-English. The authors built various SMT systems by adding sentences selected with different filter thresholds to the baseline system. We adopt the same approach in order to identify the correlation between the similarity thresholds and the SMT performance. In both our experiments and the ones in (Abdul Rauf and Schwenk, 2011), the BLEU score trends are not constantly ascending with respect to the addition of training data. Nevertheless, the authors reported significant BLEU score improvements when the extracted data was added to an in-domain baseline (1.5-2 BLEU for French-English and up to 1.4 BLEU for Arabic-English). The same applied when non-matching tails were removed from the extracted sentences.

2.2 Approaches for Web Corpora

The Web is also an extensive source of inherent parallel texts, but only few large-scale attempts to extract them are known. One of the early works in mining parallel data from the Web belongs to Resnik and Smith (2003), who performed a structural comparison of web pages (both in terms of URLs and page content) in order to identify candidate parallel documents (i.e. web pages). First they identified possibly parallel web pages by matching URLs generated by manual substitution rules. For each pair of candidate web pages, their underlying HTML structures were linearized and aligned, considering several alignment criteria based on the aligned non-markup text chunks or on the non-shared content.

Later, the authors enriched the system with a translational similarity metric, responsible for the content matching between the web pages. For the computation of the similarity score, a word lexicon (either hand-crafted or automatically generated) was required. This aspect alone achieved results comparable to the structural matching, whereas the combination of the two methods outperformed their individual performances. For a test set consisting of nearly 300 web document pairs, the authors reported an average precision of 97.4% and an average recall of 98%.

In a different experiment, the authors applied their method to the documents in the Internet Archive, a project aiming to archive the entire publicly available web. This project is similar to the Common Crawl, which we have used in our experiments to mine the Web, but is more complex, since it also indexes several types of textual, audio and video resources. Our extraction approach follows the method described in (Resnik and Smith, 2003) up to the content matching, therefore extracting parallel documents. We use a classical sentence alignment algorithm to identify parallel sentences instead.

More recently, Fung et al. (2010) proposed a method to continuously crawl web sites (irrespective of domain, URL or publication date) and to extract potential parallel sentences from them. The crawled web pages have first been indexed with respect to their content, but also to several external features, such as the URL structure, document length, HTML structure etc.. The pages have been then translated into the target language and used for IR queries (via a search engine) aiming to retrieve a set of candidate matching documents. The documents were aligned at sentence level by means of the DK-vec algorithm, which bootstraps a bilingual lexicon from the candidate document pairs and then uses its entries as anchor points for sentence alignment. Finally candidate parallel sentences were filtered by means of Inversion Transduction Grammar constraints (Wu and Fung, 2005).

The authors reported an experiment on a sample of 1000 Wikipedia articles in English and French, for which there was a clear 1-1 document alignment due to the matching article titles. A word overlap measure (based on an existing bilingual dictionary) was then used to identify candidate parallel sentences. Since the approach was developed for general web pages, the authors did not make use of the Wikipedia-specific structure (e.g. inter-language links, article structure), as we do. Therefore, more than 85% of their retrieved phrases represented “boilerplate” texts (identical proper names, dates, menu items). We overcome this problem by using only the text content of Wikipedia articles for parallel sentence mining. The remaining sentences represent parallel sentences, but also partially aligned sentences or mismatches. These findings are in agreement with ours.

At the same time, Uszkoreit et al. (2010) introduced a method to extract parallel texts from large collections of Web documents without making use of the associated metadata. Their method used monolingual near duplicate detection to identify parallel documents, therefore, in order to facilitate the comparison, documents had to be automatically translated into a common language. The document alignment was performed in two steps: first document matching based on rare n-gram sequences, secondly document matching based on lower-order n-grams overlap. The pairs of parallel documents were then aligned at sentence level by means of a dynamic algorithm based on sentence length and multilingual probabilistic dictionaries. The authors evaluated their approach on two datasets (web documents and digitized books) and achieved 99% precision and 83% recall when using the most restrictive selection threshold. In another experiment, they used the extracted data as additional training data for several open domain SMT systems translating between English and one of the following languages: Czech, German, Hungarian, French and Spanish. They reported improvements ranging from 0.3 BLEU for English-French to 7.7 BLEU for Czech-English, the biggest increase being achieved for baseline systems with a performance below 20 BLEU (e.g. Czech-English and Hungarian-English).

It is difficult to compare the extraction method above with ours, since they are so different. However, they both rely on a selection threshold to filter the candidate sentence pairs and show the same trends when analyzing the accuracy of the extracted sentence pairs. The more restrictive the threshold is, the less false positives are retrieved, therefore the more accurate are the extracted pairs. In this paper, the authors emphasized the extraction approach rather than the SMT experiments with the extracted data. Their experiments were conducted on concatenated datasets and concerned open domain translation, whereas we are interested in domain-specific translations and have experimented with different model mixtures based on the available datasets. A similar trend in the reported results is that the performance improvements are rather small for "resource richer" language pairs such as English-French or English-Spanish and the same applies for German-French, which we analyzed in our experiments.

In the literature, there was also a considerable amount of extraction approaches, which had been applied to a domain-specific subset of Web, such as news wires (Zhao and Vogel, 2002) or automotive texts (Ştefănescu et al., 2012). In these cases, the document alignment was tailored to the particularities of the respective corpora and could not be applied to any web crawled corpus. Since the parallel sentences extraction part is general, we decide to discuss some of these approaches in this section.

Ştefănescu et al. (2012) proposed a parallel sentence mining algorithm based on cross-lingual information retrieval (CLIR). First, target language sentences were reduced to their stemmed content words and then indexed for searching, together with additional information about the length of the sentence and the source document. Then, for each source language sentence, a IR query was generated and run, which contained a set of possible translations of the subsumed words into the target language, as well as information about the sentence length and the target document. The retrieved sentences were further filtered by means of a "viability score", which modeled, amongst others, the sentence length, the number of aligned words and the position of the aligned words. Finally the remaining sentence pairs were ranked by means of a translation similarity measure, which considered, amongst other features, the percentage of translated content/function words, the number of word alignments or the occurrence of aligned words in the beginning and in the end of the considered phrases.

There are a number of similarities, but also of dissimilarities between this approach and our extraction approach. We also use IR queries during extraction, but instead of selecting candidate sentence pairs, we use them to select in-domain articles. Therefore we apply the queries on document level and not on sentence level. Moreover, we also rank the candidate sentence pairs by means of a similarity metric, but our metric compares texts in the same language, whereas theirs compares bilingual texts and it is also more complex than ours. Our approach also includes a filtering step based on sentence length differences, but it is simpler than the one in the paper.

The authors also reported significant improvements in translation performance (up to 6.5 BLEU points) when the extracted data is added to a baseline out-of-domain SMT system. The experiments are performed on a comparable corpus of automotive texts in English-German collected from the Web. They showed that the extracted data is equivalent (in terms of BLEU scores) to a three times smaller amount of "clean" parallel sentences.

Jehl et al. (2012) proposed an approach to extract translations from a particular type of Web resource, namely independent microblog posts (i.e. Twitter posts). Their approach was also based on CLIR, but unlike Munteanu and Marcu (2005), who had used the standard IR framework, the authors here used a probabilistic, translation-based IR framework. They explored two extraction methods, one based on heuristic phrase extraction by means of an external dictionary and another one employing unsupervised word alignment on the query-document pairs.

Finally they used the extracted sentence pairs for domain-specific SMT, where in-domain data consisted of Twitter (microblog) posts and out-of-domain data consisted of parallel texts used in the NIST evaluation campaign. They tested three domain adaptation methods: optimization on an in-domain development set generated by means of crowdsourcing, the usage of an in-domain language model (LM) trained on monolingual Twitter messages and the usage of synthetic training data generated by automatically translating a subcorpus of the Arabic monolingual data into English by means of a SMT system employing the previous two adaptation methods.

The first two methods brought significant improvements in terms of BLEU scores, uni- and bigram precision and OOV rate compared to a baseline SMT system trained on the NIST data. Since their BLEU scores were below 20% and could therefore be regarded as unreliable, the authors put great value on the out-of-vocabulary (OOV) rates. In our SMT experiments, we also applied these methods and achieved similar results (increase of the BLEU scores and decrease of the OOV rates). Unlike them, we did not test the effect of the in-domain language model in isolation, but in combination with a mixed translation model using both out-of-domain and in-domain data.

The acquisition of domain-specific parallel data from the Web was a topic of interest for several research projects running almost simultaneously in the period 2010-2012, such as

PANACEA[1], ACCURAT[2] and TTC[3]. Moreover, these projects evaluated the extracted data in a Machine Translation setting, by this confirming the severe bottleneck of lacking parallel texts, illustrated in the previous chapter.

The PANACEA project (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies) was set to develop automatic tools for the acquisition/generation and maintenance of language resources required for NLP applications and beyond this, to suggest extraction workflows appropriate to the end applications. For example, they generated both monolingual and bilingual data collections for SMT, as well as bilingual dictionaries and transfer grammars for Rule-based MT. They exemplified their methods on the language pairs English–French and English–Greek and the textual domains *environment* and *labour legislation*.

The domain-specific texts were extracted by means of a focused crawler, which extracted specialized documents from the web complying with a user-defined set of keywords or URLs (Papavassiliou et al., 2013). The extraction of bilingual texts built on the monolingual crawls and imposed the restriction that the websites should be multilingual (with respect to the requested languages). Then Bitextor[4] (Esplà-Gomis and Forcada, 2010) was used to identify potentially parallel pages (i.e. pages that could be translations of each other) and pairs of parallel paragraphs within them. The aligned paragraphs were then split into sentences and aligned with Hunalign (Varga et al., 2005), which assigned a match score for the considered sentence pairs. This score served as a final filter in the extraction workflow. The intrinsic evaluation of the extracted pairs indicated a strict precision of 53-72% and a lax precision estimate of 87.5-94%.

The above approach is particularly relevant for our purposes because it also addresses domain adaptation in SMT. By using a manual definition of the domain (in terms of seed websites), the selection of in-domain documents becomes very accurate. We decided to save the manual efforts and to perform the in-domain selection automatically, nevertheless with the cost of retrieving false positives. The authors also model the extraction of parallel sentences as an alignment problem, but they work with bitexts, whereas we work with monolingual texts. These reported accuracy figures are also in the same range with the ones we obtained in our experiments, as we will show later on.

The SMT experiments conducted in this project can be split into two categories. First, the extracted texts have been used for training ”plain” phrase-based models. Then they have been enriched with linguistic annotations (lemmas and POS tags) and have

[1] http://panacea-lr.eu/en/
[2] http://www.accurat-project.eu/
[3] http://www.ttc-project.eu/
[4] http://bitextor.sourceforge.net/

been used to train factored models. In the first experiments, the authors tested several domain adaptation techniques on top of an out-of-domain baseline system. The methods included the optimization with an in-domain development set, training a LM on the monolingual corpus in the target language, concatenation of in- and out-of-domain bilingual training data. These methods outperformed the baseline system and in most cases also the previous system configurations. Specifically, the average relative improvement of the BLEU score for these experiments was 49.5%. The only exception occurred when in-domain monolingual data was added to a system trained with in-domain parallel data, in which case no significant improvement could be observed.

This case is somehow similar to our experiments on top of an in-domain baseline. In both cases, the additional in-domain data does not change the BLEU scores dramatically. In our experiments with an out-of-domain baseline, we tested a single system combination, in which we used bilingual in- and out-of-domain texts for both the LM and the translation model (TM) training and an in-domain development set for optimization. Moreover, our models represented a weighted combination of the different data sets, whereas in these experiments, the data sets are simply concatenated. Nevertheless, as expected, we obtained a significant performance boost over the out-of-domain baseline.

In the experiments with factored SMT, the BLEU score improvements are relatively modest: 1 BLEU for English-Greek for the best system configuration and practically no improvement for English-French. The results suggest that factored models are more effective when dealing with highly inflected languages, such as Greek.

More recently, Pecina et al. (2015) extended the experiments in the PANACEA project with different domain adaptation techniques, such as LM and TM interpolation, as well as other evaluation metrics, such as OOV rates of the test sentences, the perplexity of the reference translations given the target-side language models or the average phrase length in the test set translations. The evaluation setting did not change neither in terms of language pairs (English–French and English–Greek in both directions) nor of the textual domains (environment and labour legislation). The authors reported significant SMT improvements over an out-of-domain baseline: up to 5 BLEU on average when employing either LM or TM interpolation. Moreover, the OOV rate dropped by 30% on average when the adapted TM is employed, whereas the perplexity of the reference translations given the target language models dropped with both LM and TM adaptation.

The domain adaptation experiments in this paper are similar to the ones described in Chapter 6, but we apply the methods conjointly, i.e. we do not evaluate the effect of LM and TM interpolation or tuning with in-domain texts in isolation. Moreover, the authors evaluated the effect of the extracted data only on top of a purely out-of-domain baseline, whereas we compare the performance against both an out-of-domain and an in-domain baseline system, both tuned with in-domain texts. We observed similar trends of the SMT performance when we applied the domain adaptation techniques and we will discuss them in detail in Chapter 6.

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) was set to acquire bilingual terminologies from comparable corpora and to investigate their usage for computer-assisted translation and machine translation (both rule-based and statistical). The project focused on five European languages (English, French, German, Spanish and Latvian) and two non-European ones (Chinese and Russian), as well as two main topical domains: renewable energy and computer science. The SMT experiments were conducted for English-Latvian and German-French, the latter pair being of particular interest for comparison with the present work.

The goal of the SMT experiments was to integrate domain adaptation methods and domain-specific data in order to improve the translation performance. The domain-specific data comprises of monolingual texts, on the one hand, and of bilingual terminologies, on the other. The monolingual texts have been crawled from the Web with Babouk (de Groc, 2011), a focused crawler. They have been primarily used to train an in-domain language model. The bilingual terminologies also originate from the extracted monolingual domain-specific texts, but they are first generated independently for each language and then aligned at term level (i.e. a term can consist of more than one word). The bilingual terminologies have been used in two ways in the SMT setting. In the exclusive mode, the choice of the translation is forced if a translation is provided in the terminology, whereas in the inclusive mode, the available translation competes with the one from the phrase table.

The closest experiments to our research are the domain adaptation experiments conducted for the language pair German-French (in both translation directions) and for the narrow domain wind energy. The following system configurations have been compared: out-of-domain baseline, language model enriched with in-domain texts, translation models enriched with in-domain parallel terms (with the distinction between single and multi-word terms). For German-French, the combined LM brings an improvement of 0.9 BLEU, whereas the best TM configuration (using only multi-word terms in the inclusive mode) brings additional 0.9 BLEU. For the opposite translation direction, similar trends can be observed (0.8 BLEU improvement with the combined LM and additional 0.6 BLEU with the mixed TM).

In another experiment for German-French, the authors tried to overcome the data sparseness caused by the rich morphology of the target language, in this case French. They adapted an architecture which proved to be successful for translating into German

(Fraser et al., 2012). Specifically, they trained the SMT systems with stemmed texts on the target language and subsequently generated the inflected forms on the French output. In this context, however, no improvements could be seen, most probably due to the insufficient quality of the morphological resources on the French side.

2.3 Approaches for Wikipedia

The multilingual Wikipedia is another particular Web corpus and therefore a potential source of comparable texts. Adafre and de Rijke (2006) described two methods for identifying parallel sentences across it based on monolingual sentence similarity. In the first approach, the source article was automatically translated into the target language and then split into sentences. Then all possible sentence pairings between the translation and the reference article in the target language were generated and ranked by their Jaccard similarity scores. Finally the candidate sentence pairs were filtered so that each sentence had an unique correspondent in the other language (if possible, the one with the highest similarity score).

In their second approach, the authors generated a dictionary from Wikipedia article titles (based on the inter-language links) and represented sentences in terms of the dictionary concepts. Then, as described before, all sentence pairs were ranked by the Jaccard similarity and filtered in order to obtain 1-1 alignments. Our approach is very similar to the first method presented above, but we use a more informed similarity metric to rank candidate sentence pairs.

In the manual evaluation of 30 English-Dutch articles, the first method achieved 26% precision, whereas the second one 45% precision. Our MT-based extraction method, instead, obtains a much higher precision (57% strict precision and 93% lax precision), which demonstrates the power of our informed similarity metric. Since the purpose of the presented approaches was limited to the extraction of parallel sentences, there was no discussion about their usage in other applications, such as SMT.

Smith et al. (2010) also demonstrated that Wikipedia is a useful resource for parallel sentence extraction. They proposed two mining methods tailored for this task: a maximum entropy-based classifier and a conditional random field (CRF) model. All models used approximately the same set of feature functions, which were grouped into three big classes: features based on word alignments (e.g. number of aligned/unaligned words, word fertility), features derived from the Wikipedia markup (e.g. number of matching links, captions of the same image) and word-level induced lexicon features (e.g. translation probability, position difference). The CRF model additionally used distortion features, such as the position of the previous and current aligned sentences.

The authors reported significant performance improvements when using the extracted data as training material for open-domain SMT (up to 6.1 BLEU for Spanish-English, up to 5.2 BLEU for German-English and up to 10.1 BLEU for Bulgarian-English). Although their extraction methods were different, some of the SMT experiments findings were similar to ours. Despite the parallelism of Wikipedia-extracted data, its effect on in-domain test sets was not as substantial as the one achieved on open-domain test sets. Therefore they reported BLEU scores for three different test sets, all of them open-domain. The best results were achieved for a test set consisting of Wikipedia articles, most probably due to the simplicity of the language.

Tufiş et al. (2014) applied the extraction approach in (Ştefănescu et al., 2012) on Wikipedia articles as well, with the difference that they apply it in two steps. First, the extraction procedure was run using a dictionary compiled from out-of-domain parallel data. SMT systems were built with the data extracted at different similarity thresholds. Then, a new dictionary has been compiled from the extracted dataset which maximized the SMT performance in the first step. The extraction procedure was run again with the improved dictionary consisting of the initial dictionary merged with the extracted one. This approach improved the BLEU scores with 2.5-3 points for Romanian-English and German-English, but slightly decreased them (-1.3 BLEU) for Spanish-English. The decrease in the latter case was probably due to the fact that duplicates had been removed in the second extraction round.

The experiments discussed so far have been carried out by individual research groups, mainly for a handful of language pairs. The usage of comparable corpora for SMT has also been exploited on a large scale in the ACCURAT project (Analysis and evaluation of comparable corpora for under resourced areas of machine translation), which ran between January 2010 and June 2012. In this project, *general usage* comparable corpora for under-resourced languages and comparable corpora for a variety of *narrow domains* (e.g. renewable energy, sports, software) have been collected from the Web (e.g. Google News or Wikipedia). The languages of interest were Greek, Estonian, Croatian, Latvian, Lithuanian, Romanian and Slovenian, English and German.

They also evaluated the extracted data in terms of usefulness for domain-specific SMT, whereby they made a distinction between different source corpora (e.g. Google News vs. Wikipedia). The domain adaptation was achieved through several methods, such as LM interpolation, TM mixture, factored models or the addition of domain-specific terminologies. While most of the mentioned methods were applied in the standard way, the TM mixture approach was particular. First the word alignments were computed on the concatenated datasets, then the TMs were trained separately on each dataset, sorted and merged, at the same time avoiding duplicate entries. A couple of new features denoting the origin of the phrases (i.e. which individual TM) were added to the combined TM. The experiments were performed for 12 language pairs, from English into Greek, Estonian, Croatian, Latvian, Lithuanian, Romanian and Slovenian, from German, Greek and Lithuanian into Romanian, German-English and Latvian-Lithuanian.

The most successful adaptation methods were LM interpolation and TM mixture, which generally achieved improvements in the range of 0.3-8.8 BLEU (most of them lower than 1). For some language pairs, these approaches even lead to a decrease of the BLEU scores. The factored models and the bilingual terminologies have not achieved any improvements over the baseline, regardless of the language pair. On the other hand, the OOV rate generally decreased compared to the baseline, implying that the additional data had a positive impact on the SMT systems. The same trends were visible in their experiments performed with increasing amounts of data extracted from comparable corpora.

The described experimental setting is similar to the one used in our experiments, although the topical domains differ. We also worked with interpolated language models and mixed translation models, but we used a different technique to mix them. Moreover, we also report the results obtained with increasing amounts of additional training data extracted from comparable corpora. The results are similar to the ones we achieved in our experiments. As the final project report concludes, the changes generated by the systems using additional extracted data were restricted to a few lexical or word order differences. This statement explains why no spectacular BLEU score improvements could be expected.

2.4 Summary

This chapter has provided an overview of the existing approaches exploiting comparable corpora for parallel sentence extraction. We divided them into three categories, depending on the exploited resource: approaches for news corpora, approaches for general Web corpora and approaches for Wikipedia. We allow a separate category for Wikipedia-based approaches due to Wikipedia's structural particularities, which distinguish it from general Web corpora. Moreover, many studies exploit this resource in particular, thus establishing a standalone research direction.

To wrap up this section, Table 2.1 gives an overview of the discussed extraction approaches. Although the applied methods are manifold, they follow the same generic extraction steps, on which we also rely in this comparison. Specifically, we considered the following criteria: the document alignment method, the sentence alignment method, sentence similarity features and the granularity of the extracted data. We notice that the approaches can be grouped by various dimensions (e.g. the corpus type, specific methods for sentence alignment). For example, several approaches applied on news corpora use a search space limited to articles published during a window of n-days. IR queries are employed in various approaches for both document and sentence alignment, on both monolingual and bilingual texts. There are also sentence similarity features common to several distinct alignment approaches, such as sentence length or features based on the word alignments. Regarding the granularity of the extracted texts, most approaches retrieve parallel sentences. The ones which also retrieve parallel sub-sentential fragments are also based on the first category. We deliberately choose approaches working on this granularity level, in order to facilitate the comparison with our own approaches.

Another interesting aspect is the language distribution amongst the discussed approaches. We encounter a "Zipfian" distribution of the language pairs across approaches, which implies that most language pairs only occur once. Moreover, we notice that most language pairs include English (either on the source or on the target side). Figure 2.1 depicts the distribution for language pairs occurring more than once, also grouped by the corpus type. These language pairs are Arabic, German, Spanish, French and Chinese paired with English. This possibly indicates that English (and other "big" languages) are a safe choice for obtaining sizable corpora. For these prominent language pairs, the most used corpus was the Web (9 pairs), followed by the news corpora (6 pairs) and Wikipedia (4 pairs).

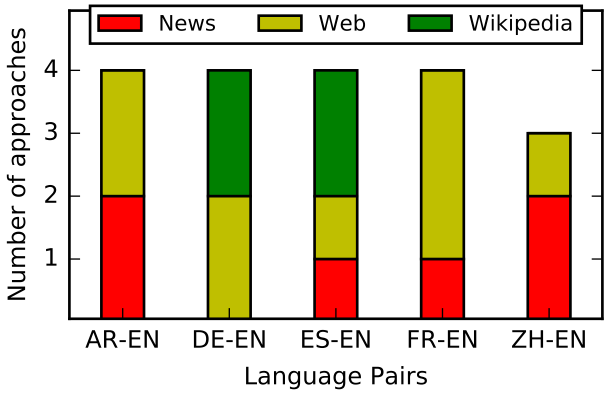


FIGURE 2.1: The distribution of the extraction approaches grouped by language pair and corpus type.

This chapter has presented several methods to extract parallel sentences from different types of corpora, starting with news corpora, continuing with Web corpora and finally with a particular Web resource, Wikipedia. The most predominant approaches are classifier-based or Information Retrieval-based, often making use of a bilingual dictionary. While the list of selected approaches is not exhaustive, it contains approaches which follow an extraction workflow similar to the one proposed in this thesis. However, since we worked on a specific domain-language pair combination, our results are not directly comparable with any of the described approaches. We think that the exploitation of comparable corpora is still a growing research field, since it offers the possibility to generate valuable linguistic resources for many different language pairs and practical applications. This fact is also evident from the latest Proceedings of the BUCC Workshop Series, which include approaches tested on less frequent language pairs, such as Russian or Portuguese paired with English and on various topical domains, such as biomedical texts or dubbed subtitles (Zweigenbaum et al., 2017).

Approach	Corpus type	Document alignment	Sentence alignment	Similarity	Granularity
Munteanu & Marcu, 2005	News stories	n-day window search Bilingual dictionary	ME classifier	f1-f5	Sentence and phrase level
Tillman & Xu, 2009		None	Exhaustive search	f5	Sentence level
Abdul Rauf & Schwenk, 2011		n-day window search	IR (monolingual)	f6	Sentence and phrase level
Resnik & Smith, 2003	Web pages	URL and HTML structure	Bilingual dictionary	None	Sentence level
Fung & al., 2010		IR (monolingual)			
Uszkoreit & al., 2010		matching n-grams	Dynamic alignment	f1, f5	
Ştefanescu & al., 2012		None	IR (crosslingual)	f7-f9	
Jehl & al, 2012		IR (crosslingual)	Bilingual dictionary	None	
Pecina & al, 2015		focused bilingual crawler	Hunalign	None	
Current work		URL and HTML structure	Gale & Church	f1	
Adafre & de Rijke, 2006	Wikipedia	A-priori	Exhaustive search	f10	Sentence level
Smith & al., 2010			ME classifier CRF model	f2-f3, f5 f11-f12	
Tufiş & al., 2014			IR (crosslingual)	f7-f9	
Current work			Exhaustive search	f2,f7	Clause level

Table 2.1: Overview of the extraction approaches from comparable corpora. Features used for sentence similarity include:

- f1: sentence length
- f2: percentage of aligned /translated words
- f3: fertility
- f4: longest contiguous span
- f5: word alignment score based on IBM Model-1 probabilities
- f6: translation error rate (WER, TER, TERp)
- f7: content / function words translation strength
- f8: alignment obliqueness
- f9: strong translation sentinels
- f10: Jaccard similarity
- f11: number of matching Wikipedia links
- f12: shared Wikipedia markup (e.g. captions, list items)

Chapter 3

Extracting Parallel Sentences from Wikipedia

This chapter describes our domain-specific approach for extracting parallel sentences from Wikipedia articles. Particularly, we work with German and French texts covering the Alpine domain (e.g. hiking recommendations, texts about the biology and the geology of mountainous regions). These choices are motivated by our ultimate goal of using the extracted corpus for SMT experiments in conjunction with the Text+Berg corpus, which covers the same domain and includes an extensive German-French parallel part.[1]

3.1 Wikipedia: Corpus Profile

We use Wikipedia as our starting corpus because its articles cover a variety of topical domains and they are usually available in multiple languages. Many of the Wikipedia articles also match our domain of interest, Alpine texts, thus representing a good resource for our extraction approach. However, Wikipedia is not a parallel corpus because its articles are not available in all the supported languages and, more importantly, because the articles do not represent translation of each other. Figure 3.1 shows the distribution of the articles in the best represented Wikipedia language variants[2]. The units on the Y-axis represent a million articles. The red bars represent the total number of monolingual articles and, from this point of view, the English Wikipedia has by far the largest coverage, followed by the Swedish (SV) and the Dutch (NL) one. The yellow bars represent the number of bilingual articles, i.e. articles from the previously considered monolingual Wikipedias which have an equivalent in the German (DE) Wikipedia as

[1] This section builds upon the work described in (Plamada and Volk, 2012, 2013).

[2] Accessed May 2016

well. In this case, the figures drop considerably, as most Wikipedia versions count less than 1 million bilingual articles.

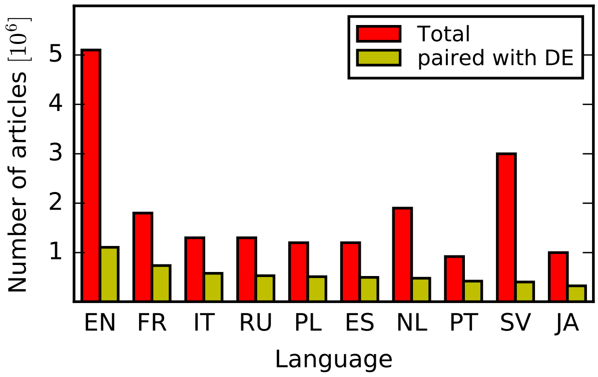


FIGURE 3.1: Article counts in the multilingual Wikipedia in May 2016.

Moreover, the Wikipedia articles in different languages are edited independently by users and are not translations of each other. They often have different structures and therefore different lengths. To demonstrate this, Figure 3.2 illustrates the sentence length of a random selection of French articles compared to their German correspondents. As we notice, in most cases there are considerable discrepancies between the number of sentences in each language.

The article length discrepancy is strongly correlated with the article structure. This can be clearly seen in the particular case of the Wikipedia article about the *Säntis* mountain (see Figure 1.4). The French article has a simpler structure and contains 3 times less sentences than its German counterpart. However, many sections have an equivalent in the German article (e.g. Histoire ↔ Geschichte, Émetteur ↔ Sendeanlage,

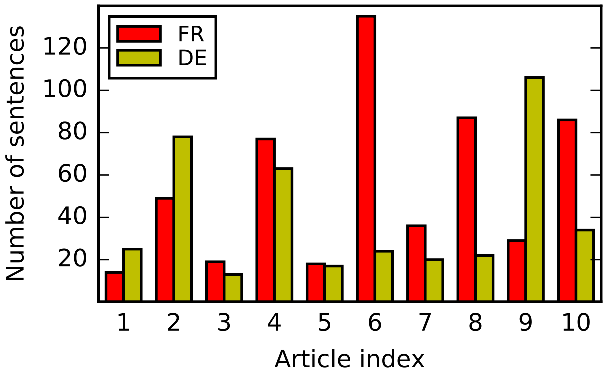


FIGURE 3.2: Compared article length between the German and the French Wikipedia.

Accès et commodités ↔ Wirtschaftliche Bedeutung), hence content overlap is likely to occur between these articles. In other words, we assume that an article in one language contains a number of sentences translated from its corresponding article in another language. These sentences are the target of our extraction approach.

3.2 The Extraction Workflow

We propose the following workflow for identifying domain-specific parallel sentences in Wikipedia articles. The general architecture of the workflow is shown in Figure 3.3. The approach is applied particularly to the language pair German-French and the Alpine domain, but can be applied to any of the available Wikipedias and any other domain. The input consists of German and French Wikipedia dumps[3], available in the MediaWiki format[4]. Altough the first experiments date back to 2011, the structure of the Wikipedia dumps has changed only minimally since then and not in a way that influences our extraction workflow. Specifically, the inter-language links are no longer included in the dumps, but are stored separately. However, the results and the examples in this dissertation refer to the 2011 Wikipedia dumps.

Since the MediaWiki markup of an article comprises much more information than its textual content and we are only interested in the latter one, our workflow requires a preprocessing step, in which the irrelevant markup is stripped off. The MediaWiki format cannot be directly parsed with a regular XML parser, therefore we first transform the MediaWiki files to XML files and then extract the raw text. At this point we also identify the Wikipedia articles available in both languages by means of the inter-language links. The extraction workflow can be divided into three main steps:

1. **In-domain article selection**: We select in-domain articles by running information retrieval (IR) queries over the German and the French Wikipedia, respectively. The queries contain domain-specific terms, such as *mountain, peak, climb*. For this particular set of query terms, this step reduces the original set of 400,000 DE-FR parallel articles to solely 20,000 articles.

2. **Sentence alignment**: For each pair of articles (of which one is considered the source and the other one the target), we divide the article texts into smaller segments (sentences or clauses) and perform the segment level alignment. We have experimented with different alignment algorithms based on segment similarity and different segment granularities (sentences and clauses, respectively).

[3] Accessed in September 2011

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

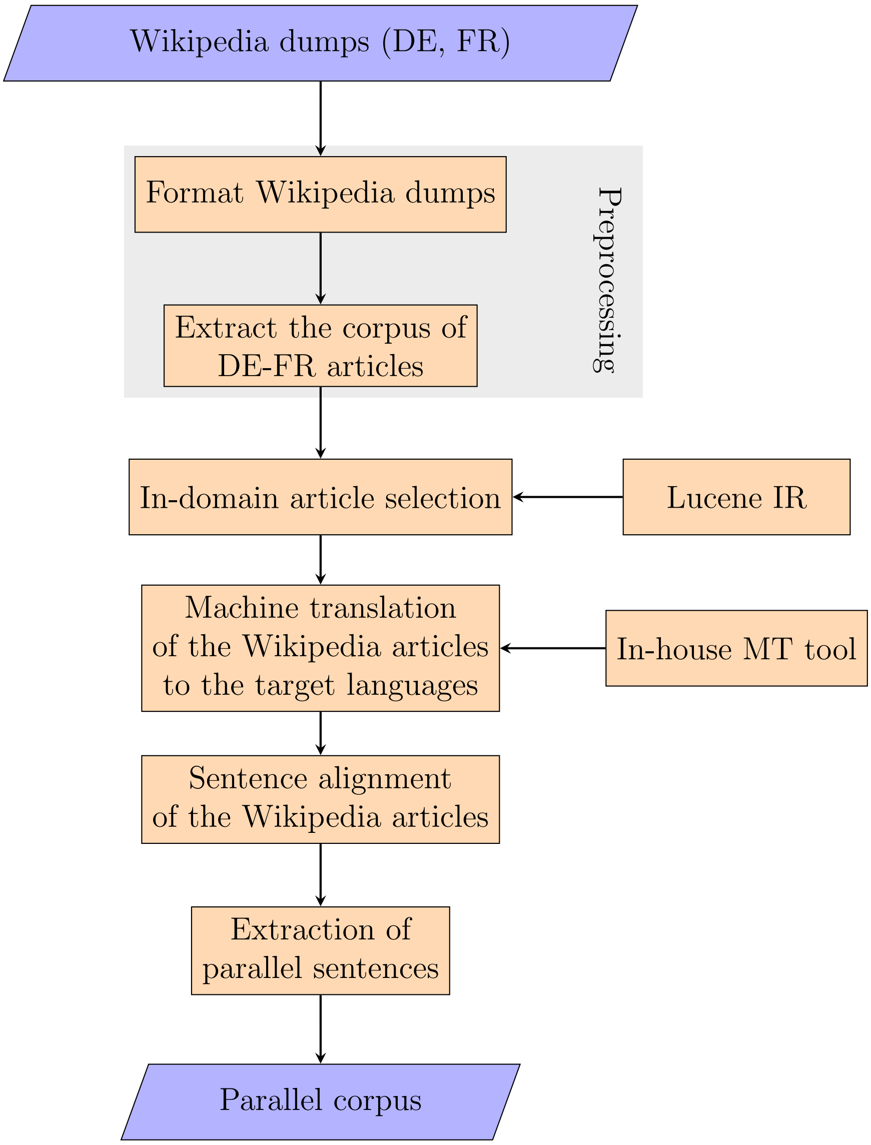

Wikipedia dumps (DE, FR)
Format Wikipedia dumps
Extract the corpus of DE-FR articles
Preprocessing
In-domain article selection
Lucene IR
Machine translation of the Wikipedia articles to the target languages
In-house MT tool
Sentence alignment of the Wikipedia articles
Extraction of parallel sentences
Parallel corpus

FIGURE 3.3: The extraction workflow.

3. **Filtering**: Finally we filter out segments with low similarity scores, assumed to be misalignments. The similarity is computed between segments in the same language, namely between an automatic translation of the source article into the target language and the target article. The target language is chosen so that we obtain a better machine translation performance, for this particular language pair it is always French.

3.3 Preprocessing

Preprocessing is an important step in our extraction pipeline, as it is in any cascaded pipeline, since the errors in this step will propagate in the output. The goal of this step was to extract the textual information from the corresponding Wikipedia dumps and to store it in valid XML files. This task was performed by an external tool, WikiPrep[5], but it could not correctly extract the information encoded in all types of MediaWiki elements. This was particularly problematic for localized elements such as namespaces and templates (storing different date and number formats). We therefore extended the tool with a template interpreter for the German and the French Wikipedia.

The following example illustrates the problem for a German Wikipedia template. The original sentence in MediaWiki format is depicted under the label *Original*. We notice that in the WikiPrep output (labeled *Before*) the altitude information is missing. With our additional template preprocessing we are able to extract this information, as it is shown in *After*.

Original: `Der '''Säntis''' ist mit {{Höhe|2501.9|CH|link=true}} der höchste [[Berg]] im [[Alpstein]] ([[Ostschweiz]]).`

Before: Der Säntis ist mit der höchste Berg im Alpstein (Ostschweiz).

After: Der Säntis ist mit 2501.9 m der höchste Berg im Alpstein (Ostschweiz).

There are several issues that make template processing difficult. We first need a comprehensive list of possible templates and a standardized usage thereof[6]. This means that we first have to compile a list of template definitions and a related set of procedures for extracting the textual information from the templates. Since the templates are language specific, we identified a number of 575,000 different templates in the German Wikipedia and 600,000 templates in the French Wikipedia dump. We further process all templates enclosed by curly brackets except for the ones belonging to tables.

The following example illustrates how the same information can be marked up differently in the French Wikipedia. Such temporal expressions (i.e. dates) are very frequent in Wikipedia (they cover approximately 38% of the detected templates) and are of utmost importance for the correct understanding of the text. We therefore have to identify all possible usages thereof in order to correctly extract the encoded information. On the other hand, the German Wikipedia does not use templates to mark dates and includes them verbatim in the text of the articles.

```
{{Date|25|juillet|2006|en informatique}}
{{Date de naissance|25| 7|2006}}
{{date|25|Juillet|2006|au Québec}}
```

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

[6] It is only recently that several initiatives aiming to give an overview of the existing templates and remove the obsolete templates arose.

```
{{Date|25|juillet|2006}}
{{date|25|juillet|2006}}
{{Date||juillet|2006}}
```

We chose to “implement” only the templates that are most likely to appear in the types of texts that we are interested in, such as transcriptions of named entities in different languages, geographical information (altitude, length), dates or numerical transformations. The transformations are done directly in the MediaWiki file, which is then sent to WikiPrep for the XML conversion. Upon completion of this step, we obtain a XML-formated, document-aligned comparable corpus in German and French.

3.4 The Selection of In-domain Articles

3.4.1 Article Classification in Wikipedia

In Wikipedia, articles are mapped to a hierarchical structure of topics (i.e. Wikipedia categories) and can be assigned to one or more categories. However, only 51.5% of the articles in our German Wikipedia dump have an assigned category. Out of the remaining ones, 33% represent 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% miscellaneous articles and 3.5% disambiguation articles. This organization could allow us to extract articles on topics relevant for our topical domain, such as *Alps*, *mountains* or *alpinism*. However, the extraction of in-domain articles by exploiting the Wikipedia category structure poses a number of challenges.

To start with, 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 Südtirol, Gebirge im Trentino, Dolomiten* (EN: massif in the Dolomites, mountains in South Tyrol, mountains in Trentino, Dolomites), but there is no explicit reference to the Alps, although it is obvious that this mountain range belongs to the Alps. If we would like to use the Wikipedia category system for the extraction of domain-specific articles, we should come up with an extensive list of relevant categories. Since the categories in Wikipedia are sometimes very specific (e. g. *Berg im Kanton Appenzell Innerrhoden* - EN: mountain in the canton Appenzell Innerrhoden), compiling the list is not a trivial task and involves considerable manual efforts.

Another challenge for this task is the mismatch between the categories assigned to the same article in different languages. For example, the article *Trois Vallées* is tagged in German as *Wintersportgebiet in Frankreich, Alpen* (EN: winter sports resort in France, Alps), whereas in French it belongs to the following categories: *Tourisme en Savoie, Domaine skiable* (EN: tourism in Savoy, ski area). One would therefore need to compile separate category lists for 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 Table 3.1 shows. Here we illustrate the distribution of the Wikipedia categories assigned to a sample of 10,000 articles from the Alpine domain retrieved with our IR-based approach (see section 3.4.2). The figures in the second column denote the number of articles labeled with the respective category.

Category	**No.**
Mann	794
Berg in Europa	673
Schweizer Gemeinde	516
Berg in den Alpen	414
Ort in Rhône-Alpes	383
Gebirge in Europa	184

Category	**No.**
Commune de la Haute-Savoie	122
Sommet des Alpes suisses	69
Point culminant d'un pays	61
Station française de sports d'hiver	44
Station autrichienne de sports d'hiver	38
Montagne du canton du Valais	38

Table 3.1: The most frequent categories in a sample of 10^4 Wikipedia articles from the Alpine domain (left: German, right: French). The abbreviation *No.* stands for the *number of articles.*

One notices that there is no overlap between the most frequent French and German categories, respectively. Moreover, French categories tend to be finer grained than the German ones and this might be the cause of the mismatch. For example, the French categories *Sommet des Alpes suisses* (EN: peak of the Swiss Alps) and *Montagne du canton du Valais* (EN: mountain from the Valais canton) could be subsumed under the German category *Berg in den Alpen* (EN: mountain in the Alps).

To date, the above discussed problems still exist, although the article categorization has slightly changed over the past years in some places. For example, the categories list of the German article *Trois Vallées* has been extended with the category *Albertville* denoting a village in the region, whereas the existing category *Alpen* has been changed to *Tourismus (Alpen).* At the same time, its French equivalent article has changed its category *Tourisme en Savoie* to *Station de sports d'hiver en Savoie* (EN: winter sport resort in Savoy) and added a new category *Tarentaise* denoting a valley in the region. In the end, the category lists of the corresponding articles continue to diverge and no improvement in terms of standardization can be observed over the years. On the contrary, the categories added in this case are in our opinion improper, as they denote entities more specific than the article itself and cannot thus characterize a more general concept.

We considered that the Wikipedia categorization was not consistent enough across languages in order to be used for the automatic extraction of domain-specific articles. At least for the language pair German-French, the categories on one side did not necessarily have an equivalent on the other side, so the extraction based on a brief set of bilingual category pairs would not have been able to find all possible articles. A manual correction of the categories was also out of question, given the size of the corpus. We therefore decided to use instead an information retrieval-based approach, which will be described in the following section.

3.4.2 IR-based Extraction

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 nouns in the Text+Berg corpus (e. g. *Alp, Gipfel, Berg, Route* in German and *montagne, sommet, voie, cabane* in French). We think that a more extended list of keywords, such the 1000 most frequent nouns in the corpus, would not have contributed to retrieving more relevant articles, as many of the keywords would either be generic words occurring in other types of texts as well, or very specific ones, which would unlikely occur in an encyclopedia for a broad audience, such as Wikipedia.

Moreover, we have filtered the keyword list by removing words that also occurred frequently in other types of texts (e.g. *meter, day, year, end*). The general domain texts come from the 2000 archives of the Swiss newspaper *Tagesanzeiger* (in German) and the *Le Monde* newspaper (in French), respectively. 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 frequently encountered terms in the Alpine domain, such as *mountain, peak, route.*

The extraction tool is based on the Lucene API[7], an open-source IR library. The tool first indexes the full text of the articles and then queries them by means of plain text queries. Beyond, Lucene supports several types of queries (e.g. wildcards or proximity queries). In our case, the query consists of the above mentioned keywords, connected by the logical operator OR. As Lucene does not have a module for morphological analysis, the reported results are based only on lemma matches. 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[8]. The score is

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

[8] http://lucene.apache.org/core/3_0_3/scoring.html

DE Title	Score
Reinhold Messner	0.1006
Hans Kammerlander	0.0684
Alpinstil	0.0676
Mount Everest	0.0674
Eiger-Nordwand	0.0658
Ortler	0.0638
Mont Blanc	0.0462

FR Title	Score
Alpes d'Allgäu	0.0391
Massif du Vercors	0.0351
Mont Kenya	0.0348
Piz Bernina	0.0320
Aneto	0.0315
Puncak Jaya	0.0308
Heinrich Harrer	0.0287

TABLE 3.2: The best ranked Alpine articles in Wikipedia according to Lucene.

computed using the following formula:

$$\begin{gathered} score(q,d) = coord(q,d) \times queryNorm(q) \times \\ \sum_{\text{t in q}} (tf(t\ in\ d) \times idf(t)^2 \times t.getBoost() \times norm(t,d)) \end{gathered} \tag{3.1}$$

where *coord(q,d)* is a score factor based on the number of query terms found in the specified document,
queryNorm(q) is a normalizing factor,
tf(t in d) is the frequency of the term *t* in document *d*,
idf(t) is the inverse document frequency of the term *t*,
t.getBoost is the weight of the term *t* in the query *q* and
norm(t,d) represents the product between several indexing time weights and length factors.

Upon completion of this step our corpus was reduced to approximately 20,000 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. Although we have refined our search terms by discarding the ones occurring frequently in other text types, we still could not avoid a small percentage of false positives. 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 level 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: 0.017,45) receives a lower score than a longer article about a glacial lake (e. g. *Weißensee*, similarity score: 0.026,98).

Table 3.2 shows a selection of the articles with the highest scores in the German and French Wikipedia, respectively, 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 and it is likely that a different number of keywords were identified in the French and the corresponding German articles. This is how we can explain the considerably lower scores obtained by French articles.

Nevertheless, the top retrieved articles are undoubtedly relevant for the topical domain of Alpine texts, since they refer to mountains (e.g. Mount Everest, Piz Bernina, Aneto) or famous alpinists (e.g. Reinhold Messner, Heinrich Harrer). Therefore, so far, we do not need to worry that the lower French scores could introduce out-of-domain articles. There is a small number of false positives in both the German and the French results lists, but they can be filtered out by setting a higher selection threshold for the Lucene scores.

In the previous section, we illustrated the distribution of the Wikipedia categories for a sample of articles (see Table 3.1). Those articles represent the first 10,000 articles retrieved by the method described above, using the German keyword list. Since the corresponding categories denote terms strongly correlated with the Alpine domain (e.g. mountains in Europe, winter sports resorts), we are convinced that our IR-based approach for selecting in-domain articles is well-founded.

Consequently, we ran the same experiment on the first 10^4 articles extracted from the French Wikipedia by means of the French keyword list. Our findings (see Table 3.3) imply that the French articles have a looser connection to the Alpine domain. Some refer to cities (e.g. *Ancien chef-lieu de district* (EN: former county seats)), another groups articles about flora. There are also categories which make more sense for this domain, such as *Sommet des Alpes suisses* (EN: peak of the Swiss Alps) or *Point culminant d'un pays* (EN: the highest point of a country). This is most probably a cummulative effect of the chosen keywords, on the one hand, and of the lower Lucene scores, on the other hand.

Despite of the cross-validation of the search terms, we still found terms which can occur in different contexts other than Alpine texts (sometimes even with slighty different meanings). For example, the word *section* refers, in the Text+Berg corpus, to a division of the Swiss Alpine Club, but in other texts refers to a portion of an object (a book, a fruit, etc.). This might lead to wrongly tagging out-of-domain articles as in-domain ones. In this case, we noticed a considerable amount of articles about biology, where the word *section* was used to classify plants or organisms.

Since the articles retrieved using the German keywords seem to be closer to our expectations, we decided to rely on the German keywords list for the selection of in-domain

Category	**Number of articles**
Ancien chef-lieu de district	147
Flore (nom vernaculaire)	70
Camino frances	64
Commune du Bas-Rhin	56
Sommet des Alpes suisses	55
Point culminant d'un pays	55

Table 3.3: The most frequent categories in the top 10^4 French articles.

articles and furthermore on the inter-language links for retrieving the corresponding French articles.

3.5 The Extraction of Parallel Sentences

In this step, we mine for parallel (i.e. semantically equivalent) sentences in the previously selected in-domain Wikipedia articles. We model the extraction from a pair of Wikipedia articles as a sentence alignment task. Moreover, we overcome the cross-lingual similarity task by using an intermediary machine translation of the source text, which means that the task is reduced to a monolingual alignment. The translations are generated with our in-house SMT system trained on Alpine texts, whereas the alignment is performed using the Bleualign algorithm (Sennrich and Volk, 2010).

Bleualign generates all possible sentence pairs between the automatic translation of the source article and the target article and computes for each of them the BLEU score (Papineni et al., 2002). The algorithm subsequently reduces the search space by keeping only the 3 best-scoring alignment candidates for each sentence and outputs the alignment pair which maximizes the BLEU score and respects the monotonic sentence order.

To serve our purposes, we run the Bleualign algorithm with some modified settings, namely the n-grams considered for the BLEU score computation and the gap filling heuristics. We choose to compute the BLEU score up to 3-grams in order to give preference to fluent translations. Bleualign considers by default only unigrams and bigrams for the computation of the BLEU score. Moreover, we do not use any heuristics to fill the alignment gaps (i.e. blocks of unaligned sentences) between sparsely aligned sentence pairs due to the different structure of the source and the target Wikipedia articles. The resulting set of alignment pairs represents a corpus containing semantically equivalent sentences. We call them *semantically equivalent* because they do not always represent literal translations of each other, as they sometimes contain paraphrases or extra segments.

The following example illustrates the case where one of the sentences contains an extra tail. Despite this, the sentence pair is one of the best ranked candidates for our parallel corpus (i.e. it obtains the highest BLEU score). It is worth noting that the BLEU score is not computed between the source (*FR*) and the target (*DE*) sentences, but between the automatic translation (*MT*) and the target sentence (both tokenized and lowercased).

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

MT: Damit ist der nördliche Teil des Himmelschrofenzug besteht aus Dolomit, während der südliche Teil besteht aus Felsen des lias der Schneedecke, das Allgäu

DE: So besteht der nördliche Teil des Himmelschrofenzugs aus Hauptdolomit. Der südliche Teil besteht aus Liasgesteinen der Allgäudecke, die auf den Hauptdolomit aufgeschoben worden sind

Although the automatic translation is not perfect, one notices that the word overlap between the translation and the target sentence is rather high. Since the source sentence (and also its translation) are shorter than the German reference, the extra tail in the reference *die auf den hauptdolomit aufgeschoben worden sind* is not penalized by the BLEU score. On the other hand, if we were to remove the extra tail in the German sentence, the remaining part would be a perfect equivalent of the French sentence.

3.6 Extraction Results

In this section we illustrate the outcome of the proposed extraction method, in particular by means of examples which caused challenging decisions during the extraction process. Table 3.4 shows a sample of extracted sentences and the intermediate similarity score that led to their selection. The similarity scores (based on the BLEU evaluation metric) have been computed between an automatic translation of the French sentence into German and the German target sentence. We chose this translation direction in order to avoid the brevity penalty associated with the BLEU score, as the French sentences (and therefore also their translations) are longer than the German ones.

The first sentence pair illustrates the issue of paraphrases, which often make difficult the decision regarding the parallelism of a given sentence pair. Although the meaning of the sentences is approximately the same, different wordings are used to express it. The French phrase *à gravir trois sommets de plus de 8000 m en une même saison* (EN: to climb three peaks above 8000 m in the same season) corresponds semantically to the German relative clause *der mehr als zwei Achttausender bestiegen hatte* (EN: who climbed more than two eight-tausenders), but there are several different nuances between them (three vs. more than two, eight-tausender vs. peaks above 8000 m).

On the other hand, the sentences in the second example convey the same meaning by using the same words (i.e. translations of each other). This sentence pair obtains the same similarity score as the one in the previous example, although we could argue that the second one is more entitled to obtain a high score. Since the BLEU-based similarity metric sometimes fails to make a clear cut distinction between parallel and comparable sentences, the choice of the confidence threshold is a difficult task.

No.	**French sentence**	**German sentence**	**Score**
1	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
2	Cette montagne est avec le plateau de Gottesack voisin l'attraction majeure du sous-groupe	Dieser Berg ist zusammen mit dem benachbarten Gottesacker-plateau auch die markanteste Erscheinung der Untergruppe	1.0
3	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
4	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
5	La voie normale (AD- en utilisant le câble, D sans), dont l'approche se fait depuis le refuge Turin au col du Géant, est équipée de grosses cordes fixes sur la partie difficile, des dalles Burgener à la pointe Sella.	Der Gipfel ist am besten von der Turiner Hütte (ital. Rifugio Torino) von der italienischen Seite zu erreichen.	0.4448
6	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

Table 3.4: Aligned pairs identified by Bleualign.

Another support for this claim is provided by the third pair, which also does not contain parallel sentences and yet obtains a relatively high similarity score. Although the sentences do not convey the same meaning, they contain many overlapping words (8 content words and several function words). However, close similarity scores can also correspond to sentences which are semantically equivalent. For example, the next sentence pair (4) represents a valid translation, but receives a lower score due to the different types of linguistic constructions used to express the same meaning. In this case, the French relative clauses are expressed as adjectives in German (e.g. *qui lui succéderont-späteren*).

The last two pairs in the table (5-6) illustrate cases where one of the sentences contains extra information. In the first case, the German sentence corresponds only to the relative clause of the French version, whereas the rest of the French sentence lacks any equivalent translation. In the second case, the German sentence contains an extended nominal phrase (*in dem dünn besiedelten und zuvor evakuierten Gebiet*), which is not translated into French. Since they have the same drawbacks, the sentence pairs obtain similar scores, which allow us to take consistent decisions regarding their selection/ discard.

We think that sentence pairs such as number (4)-(6) can only be useful for MT if we are able to separate the parallel segments, as they are prerequisite for obtaining good word alignments. For this purpose, a finer-grained partition of the input sentences (e.g. into clauses or text chunks) is required. The following chapter discusses our approach to extract parallel sub-sentential segments from comparable texts.

Additionally, we performed a different qualitative evaluation of the extracted sentences, aiming to answer the question: How much of the extracted data actually consists of parallel sentences? To measure this, we conducted a manual analysis of 200 randomly selected sentence pairs. Since some of the extracted pairs contained only partial alignments, we distinguished between the following categories: good, bad and partial. We therefore reported both *strict* and *lax* precision estimates. For a strict true positive, the German and the French sentence have to be reciprocal translations (e.g. sentence (2) in Table 3.4), whereas lax true positives include sentence pairs with partial alignments. Partial alignments can cover n-word sequences, with n higher than the number of n-grams used for the ui computation (e.g. sentence (6) in Table 3.4). For the considered test set, the *strict precision* estimate was 35% and the *lax precision* estimate was 55%.

Misalignments are traced back to sentence pairs with overrated BLEU scores, caused by overlapping n-grams acting as false friends (e.g. proper names, dates, frequent preposition + determiner sequences). We therefore investigate the effect of varying n-grams considered for the BLEU score computation. The choice of this parameter is a difficult decision since we have to consider the fact that the algorithm compares machine translations to human references, which often differ. In order to minimize the translation errors, the translation direction in this experiment is from German into French and the comparison language is therefore French.

We analyze the effect of 2-, 3- and 4-grams (considered for the BLEU score computation) on the extraction precision. The test set consists of 10 random articles from our in-domain corpus of Wikipedia articles, each of them at least 20 sentences long. The highest number of alignment pairs is generated when considering up to 2-grams. By taking into consideration 1- up to 3-grams the number of alignments drops with 25%. For the considered test set, the removed alignments were all misalignments, therefore the precision estimates improved. If we additionally consider 4-grams, the number of alignments continues to drop (in our case with 20%). However, this time the removed alignments are not only misalignments, but also good alignments with no/little overlap on 4-grams level.

This performance can be improved if one computes the alignments in both directions, i.e. from German into French and from French into German. This allows us to suppress some misalignments. In the 3-gram BLEU configuration from the previous example, this option reduces the number of alignment pairs with 75%, but significantly improves the alignment quality. In this case, the strict precision estimate is 76% and the lax precision estimate is 97%. However, this improvement triggers decreasing recall estimates, since some of the good alignments obtained in the initial setting are now left out.

Considering these findings, we use the following settings for the experiments on the whole Wikipedia: We compute the BLEU score on up to 3-grams, in order to maximize the number of true positives and to minimize the number of false positives. We furthermore compute the alignments in a single direction, firstly because we want to obtain as much parallel material as possible and secondly because we prefer to use only the most reliable translation direction (German-French).

In addition to the manual evaluation of precision, we also evaluated the extracted data in a SMT scenario. The results will be presented and discussed in Chapter 6.

Chapter 4

Extracting Parallel Sub-sentential Segments from Wikipedia

This chapter presents an improved approach for extracting parallel text segments from Wikipedia. In these experiments, we set the segment granularity at sub-sentential level (i.e. clauses) and use a different alignment algorithm. First we explain our rationale for these changes and then discuss the modifications to the initial extraction workflow. Finally we compare the results obtained with the sentence- and clause-based extraction approaches, respectively.[1]

4.1 Motivation

The analysis of the results presented in the previous section brought into attention many "parallel" sentence pairs of different lengths. By this we mean that the shared translated content does not always span the whole sentence. As an example, consider the following sentences which have been retrieved by the initial extraction pipeline.

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.
The pass is situated in the external, populated area of the Mercantour national park and represents the transition between the Bévéra and the Vésubie valleys.

FR Le col de Turini relie la vallée de la Vésubie à la vallée de la Bévéra.
The Turini pass connects the Vésubie and the Bévéra valleys.

[1] Parts of this chapter have been published in (Plamada and Volk, 2013).

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. If this sentence pair would be used for SMT training, it would most probably confuse the system, because noisy alignments are to be expected.

One solution to this problem is to split the sentences into smaller entities (e.g. clauses) and provide them as input for the extraction workflow in Chapter 3. Our claim is that we can increase the reliability of the proposed algorithm by matching shorter sentence fragments. Moreover, the selection of the candidates will be simplified because we will only have to consider 1-1 sentence alignments.

Another bottleneck of the previous approach is the alignment algorithm for matching possible candidates. 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 allowed. But this phenomenon occurs often in Wikipedia, because its articles in different languages are edited independently, as shown in Section 1.2. We therefore think that an alignment algorithm without position constraints is more appropriate for Wikipedia texts.

Moreover, the string-based comparison in Bleualign proved to be unreliable for our purpose, allowing many false positives. The following example illustrates one of the frequent cases: sentence pairs with similar length and structure, which only partially overlap (see the text marked in bold). The last parts of the sentences look similar *à 300 kilomètres au nord - 325 km südlich entfernt*, but convey different meanings and make these sentences not parallel. We therefore need a more powerful similarity metric, that can rule out such sentence pairs.

FR orig Le Mont Méru se trouve à 75 kilomètres au sud-ouest et **le Mont Kenya, deuxième sommet d'Afrique par l'altitude**, à 300 kilomètres au nord.

Mount Méru is situated 75 km southwest and Mount Kenya, the second highest mountain in Africa, 300 km to the north.

DE orig Vom Batian, dem im **Mount-Kenya-Massiv befindlichen zweithöchsten Berg des Kontinents**, ist der Kibo 325 km südlich entfernt.

Kibo lies 325 km south of Batian, which is part of Mount Kenya, the second highest mountain of the continent.

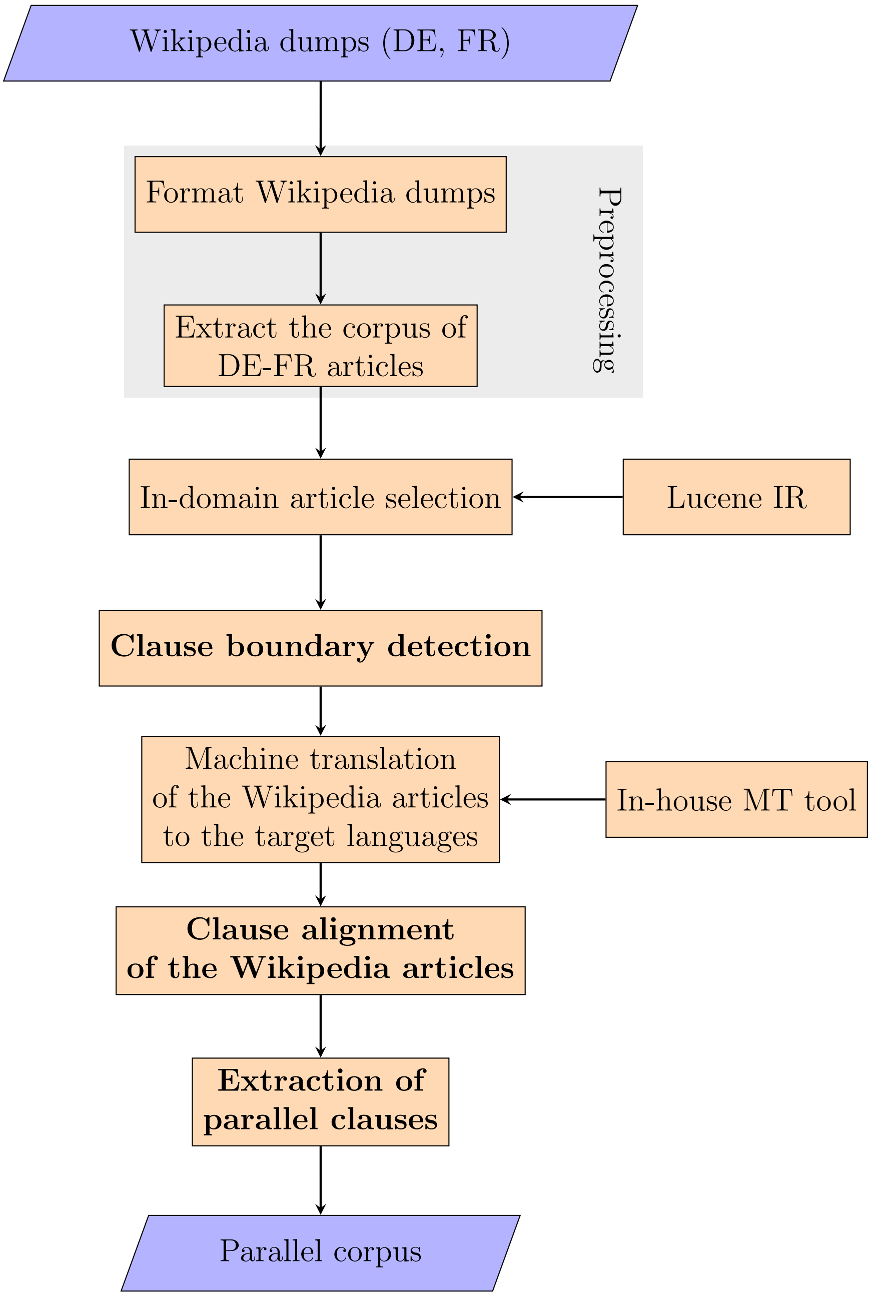

Wikipedia dumps (DE, FR)
Format Wikipedia dumps
Extract the corpus of DE-FR articles
Preprocessing
In-domain article selection
Lucene IR
Clause boundary detection
Machine translation of the Wikipedia articles to the target languages
In-house MT tool
Clause alignment of the Wikipedia articles
Extraction of parallel clauses
Parallel corpus

Figure 4.1: The modified extraction workflow.

4.2 The Modified Extraction Worflow

Based on the previous considerations, we adapt the extraction workflow from Chapter 3, in order to obtain parallel clauses. Figure 4.1 depicts the modified extraction workflow (the additional steps are marked in bold). Since the first part of the pipeline stays the same, we start with the in-domain articles selected with the previous approach. We could not reuse the machine translation of the source language articles due to the different partition of the text (into sentences and clauses, respectively). The modifications thus concern the splitting of the Wikipedia articles into clauses (further referenced as *clause* *boundary detection*) and the improved alignment algorithm at clause level. They will be discussed in detail in the following sections.

4.3 Clause Boundary Detection

The identification of clauses in long sentences was a hot topic in NLP research in the late 1990s early 2000s, as it was considered an important preliminary step for NLP applications, such as discourse analysis, bitext alignment or Text-to-Speech. Before discussing a few suggested solutions for this problem, we provide a definition and respectively, a possible classification of clauses.

We see clauses as the minimal standalone pieces of text which comprise a message. They are usually centered around a verb, although sometimes the verb can be omitted without changing the meaning of the clause. Quirk et al. (1985) define three main structural types of clauses:

Finite clauses: clauses containing a finite verb, such as *reads, has taken, can see*

Nonfinite clauses: clauses containing a non-finite verb, such as *to see, walking*

Verbless clauses: a clause which can be analyzed although no verb is present

The following example illustrates how these clause types can be combined in German. In this example, as well as in the following ones, $\langle CB \rangle$ stands for clause boundary.

> Unter seinem Nachfolger Andrew Scott Waugh wurde der $\langle CB \rangle$ zunächst als "Peak b" bezeichnete $\langle CB \rangle$ Gipfel 1848 erstmals von Indien aus vermessen, $\langle CB \rangle$ da Nepal den Zugang zu seinem Territorium verweigerte.
>
> *The summit, initially referred to as "Peak b", was first measured from India in 1848 under his successor Andrew Scott Waugh, as Nepal refused access to its territory.*

The sentence consists of two finite clauses, a main clause (*Unter seinem Nachfolger [...] wurde der Gipfel 1848 erstmals [...] vermessen*) and a subordinate one (*da Nepal den Zugang [...] verweigerte*), and a nonfinite one, which is embedded in the main clause (*zunächst als "Peak b" bezeichnete*).

The following example illustrates the third group of clauses, where the verb is omitted.

> Der Mann fuhr in den Bergen, $\langle CB \rangle$ die Frau an den Strand.
> *The man went to the mountains, $\langle CB \rangle$ the woman to the beach.*

The above sentence consists of two assertions with a similar structure, yet only the first one contains a predicate. The sentence can be completed by repeating the verb *fuhr* (EN: went) in the second assertion without changing its meaning, in which case the sentence would be undoubtedly split into two clauses. It thus makes sense to split the original sentence into a finite and a verbless clause.

Existing solutions for clause identification range from rule-based approaches to machine learning ones. In 2001 there was a Shared Task on using machine learning (ML) for splitting English texts into clauses, where six research teams participated. The best performing system was proposed by Carreras and Màrquez (2001) and it was based on binary decisions trees and a boosting algorithm. They outperformed the competing teams in all the tasks, namely identifying clause starts and clause ends, but also extracting complete clauses. They achieved F-scores between 78.6% and 91.7%, the lowest corresponding to the identification of whole clauses and the highest corresponding to clause start identification.

Much of the existing literature on clause identification focused on English, thus there are relatively few approaches for other languages. For example, Puşcaşu (2004) introduced a multilingual approach consisting of a language-independent machine learning component and a rule-based, language-specific component. The ML module was responsible for identifying clause boundaries represented by coordinating conjunctions or punctuation marks (considered the most frequent delimiters), whereas the rule-based module identified clause boundaries introduced by subordinators and reanalyzed clauses which contained several finite verbs. At this stage, the module also employed a language-specific list of unambiguous subordinators. The approach was tested for Romanian and English and achieved F-scores of 95% and, respectively, 92% for the specific task of clause start identification.

Our approach is similar to the rule-based module described above, as they both relied on POS tags. Moreover, the rules focused on the same anchor points, such as verbs, coordinating and subordinating conjunctions or punctuation marks. Unlike them, we did not include explicit word lists (e.g. subordinating conjunctions) in the rules.

For French, Afantenos et al. (2010) proposed an approach for identifying elementary discourse units, which is a more general task, since it does not only identify verbal clauses, but also prepositional phrases and adjuncts. Their approach was based on a maximum entropy classifier, which used both linguistic features (e.g. POS tags, dependency relations) and positional features (e.g. distance from sentence boundaries, context n-grams). Finally they applied a set of heuristic rules resulting in adding/deleting boundaries in order to generate well-formed sentence segments. The approach was evaluated on a small section of the Annodis corpus [2] consisting of Wikipedia and newspaper articles and achieved F-scores of 87.8% for the task of clause start identification and 73.3% for full clause identification. The lower figures compared to English are possibly a consequence of the fact that the French sentences contain more embedded segments and its boundaries are more difficult to identify.

Clause identification is not a standalone topic neither in the research conducted for German, but it can be encountered as a subtask of discourse segmentation. A reference work for this language is represented by Lüngen et al. (2006), who identified discourse segments according to three levels of information: the logical document structure, the punctuation and the linguistic information. In their interpretation, discourse segments were not only clauses (in the sense defined in the beginning of this section), but also prepositional phrases or appositions. Their approach was rule-based and did not require annotated training data. The approach was evaluated on a corpus consisting of four scientific and two web-published articles and achieved an average F-score of 75.57% for the identification of sentence-internal boundaries.

These approaches are relevant for our task of clause boundary identification from several perspectives. First, they both have a rule-based component for language-specific split decisions as a consequence of the different word order or of the different usage of punctuation marks. For example, the full stop is often used in German for other purposes than final sentence boundaries (e.g. after ordinal numbers denoting dates) and one should make sure that no false boundary is introduced in these case. Secondly, the defined rules relied on similar information, such as POS tags, punctuation or verb forms. The differences consisted in the definition of the elementary units: clauses in our case, discourse units in the mentioned approaches. We considered that clauses should contain a single verb, but discourse segmentation allowed verbless segments.

In the present work, we used a rule-based approach to tackle the problem of clause boundary identification. The rules are based on POS tags, therefore our main anchor points are verbs (as the main content bearers), conjunctions (which connect similar sentence structures, in our case clauses) and punctuation marks, especially commas and semicolons (which introduce a new idea, mainly corresponding to a new clause). By this means we are able to distinguish reliably the clauses containing a verb, be it finite or nonfinite.

The rules are language-specific due to the fact that the tag sets are different for German and French, respectively. Since many rules are valid for both languages, defining a new set of rules only requires to replace the corresponding POS tags, in case they exist. Our approach builds on the approach used by Volk (2001) for German. For French we have

[2] http://redac.univ-tlse2.fr/corpus/annodis/

developed a similar approach based on a smaller set of rules covering the most frequent patterns of main + subordinate /main + main clauses. The German set of rules consists of 30 rules, whereas the French one consists of merely 10 rules.

We exemplify the rule matching process by means of the the German sentence in Section 4.1. Below we list the sentence together with its POS annotation generated by the TreeTagger. We manually replaced the NN-tags assigned to named entities with NE-tags due to consistency reasons. The sentence consists of two finite clauses (containing a finite verb labeled with *VVFIN*) connected by the conjunction *und/KON* (EN: and).

> Der/ART Pass/NN **liegt/VVFIN** in/APPR der/ ART äusseren/ADJA ,/$, besiedelten/ADJA Zone/NN des/ART Nationalpark/NE Mercantour/NE **und/KON** **stellt/VVFIN** den/ART Übergang/NN zwischen/ APPR dem/ART Tal/NN der/ART Bévéra/NE und/KON dem/ART Tal/NN der/ART Vésubie/NE dar/PTKVZ ./$.

This sentence triggers the following rule, which implies that a sentence containing two finite verbs and a conjunction between them can be split into two clauses. In this case, we add a clause boundary in front of the coordinating conjunction:

```
V[AMV]FIN * KON * V[AMV]FIN * → V[AMV]FIN * <CB> KON * V[AMV]FIN *
```

The rule generates the following output:

> Der Pass liegt in der äusseren, besiedelten Zone des Nationalpark Mercantour ⟨*CB*⟩ und stellt den Übergang zwischen dem Tal der Bévéra und dem Tal der Vésubie dar.

Since the German tag set is richer than the French one, rules have to be redefined for the French tag set. The equivalent French rule for coordinated main clauses is thus:

```
V * C_C * V * → V * <CB> C_C * V *
```

We used the TreeTagger[3] to POS-tag the input sentences, for German with the standard parameter file and for French with a customized parameter file. The French parameter file was trained on the Le Monde Treebank (Abeillé et al., 2003), which contains general domain texts. The tag set is a bit more extensive than the standard one used by the TreeTager for French (40 tags instead of 30 tags), but still less complex than the German one (50 tags). For example, our tag set uses a unique tag for verbs, whereas the German

[3] www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

tagset has 12 tags to describe different verb forms. This is also the main reason why the French set of rules is smaller than the German one.

An improvement of our clause detection algorithm over the original one is that it merges retroactively parts of clauses separated by a subordinate clause (so-called nested clauses). By this means we can avoid generating verbless clauses, as illustrated in the following example:

> Ein Informationszentrum, $\langle CB \rangle$ das von der Parkverwaltung unterhalten wird, $\langle CB \rangle$ bietet Informationen über die Gesteinsschichten im Park.

Here the algorithm generates three clause chunks: the first chunk consists only of a nominal phrase, the second one of a relative clause, and the third one is the main clause (without its subject). The first and the third clause chunk build up together the main clause, whereby the first part represents the grammatical subject *Ein Informationszentrum* and the latter one contains the verb and its arguments. These chunks are glued together in order to obtain meaningful pieces of texs (i.e. clauses). In this way we can avoid aligning subjectless or objectless clauses to whole clauses.

We evaluated the clause boundary detector over a set of 100 sentences per language, which contained 125 intra-sentential clause boundaries. For this purpose, the clause boundaries were first determined manually. We did not count end of sentence boundaries (e.g. period), since they are trivial to identify. We noticed that most of the clause boundaries occured between coordinated clauses or between a main clause and a relative one. There were also cases of boundaries between a main and a subordinate clause (other than relative clauses), but they were less frequent than expected.

For German, the precision estimate was 89.3%, whereas the recall is 80.6%, similar to the figures reported by Volk (2001). For French, the figures were slightly lower, namely 76.5% precision and 77.1% recall. The drop in precision was due to a higher number of false positives (false clause boundaries), probably a side effect of the French PoS tag set, which was smaller than the German one and therefore did not allow very specific rule definitions. The figures for French are also comparable to the ones reported by Afantenos et al. (2010), though for a different test set.

4.4 Clause Alignment Algorithm

Having discussed how to split the text into clauses, we will now describe how to identify the cross-lingual equivalences between them. We model this step as a monolingual alignment task based on an intermediary machine translation of the source article. We consider German articles as the source because we expect a better automatic translation quality from German into French. The translation is performed by an in-house SMT system trained on Alpine texts, the same previously used to translate the Wikipedia articles.

As argued before, a position-independent alignment algorithm is more suitable for Wikipedia texts. Therefore our alignment 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 (i.e. enforcing 1-1 alignments). In the end we filter the results by allowing only clause pairs above a customizable threshold.

The similarity measure for comparing clauses is a key factor in the extraction process, as it directly influences its accuracy. In the first iteration of the extraction workflow we have used a string-based similarity metric (i.e. BLEU (Papineni et al., 2002)), which is often used for evaluating machine translated output (see Section 3). This lead to many misalignments: sentence pairs with overlapping words, but totally different meaning. We therefore want to avoid these cases by using a more informed similarity metric.

We define 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 use the METEOR similarity metric because, unlike other string-based metrics (e.g. BLEU), it considers not only exact matches, but also word stems, synonyms, and paraphrases (Denkowski and Lavie, 2011).

Number of aligned content words: Although the METEOR score exploits the word alignments, it makes no distinctions between content and function words. This feature is meant to give preference to sentence pairs with many aligned content words.

The rationale for choosing these criteria is detailed below. Suppose that we compute the similarity between the following tokenized 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 can hardly be considered reliable, given that the sentence *ta voiture vous voir*(EN: your car see you), paired with the first sentence, would get the same BLEU score, although the latter sentence is obviously nonsense. On the other hand, METEOR would return a score of 90.3 for the original sentence pair, since it can tell 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. The deliberately false French sentence achieves a score of 34.4, allowing a clear cut beween good and bad candidates for equivalent (parallel) sentences.

However, METEOR scores can also be misleading, since their computation is based on automatic word alignments. This means that two sentences are likely to receive a high similarity score when they share many aligned words, regardless of their lexical meaning in the sentence (content vs. function words). We often encounter sentence pairs with a decent METEOR score where only some determiners, punctuation marks or simple word collocations (e.g. *de la montagne* (EN: of the mountain)) 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

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 23.0. This score can be considered valid in case one pursues an extensive search targeted at obtaining as much data as possible. 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.

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

$$score(src, trg) = w_1 * s_1 + (1 - w_1) * s_2 \tag{4.1}$$

where s_1 represents the METEOR score and s_2 the alignment score and w_1 $/(1 - w_1)$ represent the weights of the associated scores.

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

No.	French clause	German clause	Sim. Score
1	McNish écrit dans son journal:	McNish schrieb in sein Tagebuch:	1.0
2	Elle travailla pendant plusieurs semaines avec lui	Während mehrerer Wochen arbeitete sie mit ihm zusammen	0.840
3	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
4	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
5	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
6	qu' un cas mineur ayant un effet limité sur la santé	wie sich diese Substanzen auf die Gesundheit auswirken,	0.200

Table 4.1: Aligned clause pairs extracted from Wikipedia. The abbreviation *No.* stands for the *number of articles* and *Sim.* stands for *Similarity.*

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 and 1 wrong clause pairs. This corresponds to a precision of 95% and a recall of 92% on this small test set.

Additionally, we define 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 decide 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.

4.5 Extraction Results

Table 4.1 provides examples of automatically extracted clause pairs and their corresponding similarity score. As mentioned before, the scores are computed between the original French clause and a machine translation of the German clause. We notice that the shorter segments (such as number 1, 2 and 6) resulted from splitting longer sentences, as suggested by punctuation marks or by the word order, whereas the others are single-clause sentences. The split phrases are also more likely to contain contiguous word alignments and therefore achieve high similarity scores (see sentences number 1 and 2). It is not the case for the last clause pair, which, despite sharing many words (*avoir un effet - auswirken, sur la santé - auf die Gesundheit*), does not convey the same meaning in both languages. Since the alignments are very fragmented, the similarity score penalizes this clause pair with a low value.

However, apparently perfect 1-1 word correspondences are not a guarantee of high similarity scores. Sentence number (3) illustrates this by achieving a score of only 0.51, in contrast to the human assessment which would consider them parallel. The "low" score is most likely an effect of the comparison between natural language texts and automatic translations. A similar score is achieved by the following clause pair (number 4), where the German counterpart contains additional information wrapped in appositions, such as *das 2248 km^2 grosse Schutzgebiet* (EN: the protected area with a surface of 2248 km^2). In French, the same information would be expressed as a relative clause, but several relative clauses in the same sentence would easily make it cumbersome. Since appositions are not clauses on their own, our approach cannot identify them as extra tails and therefore also retrieves such sentence pairs.

To estimate the quality of the extracted parallel data in numbers, 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. In terms of the strict and lax true positives definition from Section 3.6, the strict precision estimate is 39% and the lax precision estimate is 65%. These figures are higher than the ones obtained when aligning whole sentences, thus supporting the choice of clause-level alignment.

Moreover, we compute the same estimates for sentence pairs above a more restrictive threshold (i.e. 0.4). We notice that the precision (therefore also the quality) of the extraction process increases with the threshold. Specifically, we measure a strict precision estimate of 43% and a lax precision estimate of 82%. The difference between the strict and the lax precision estimates is an indicator of the different phrase structures in German and French, which don't allow a total content overlap.

We used the same test set to investigate the incidence of clauses which are not standalone sentences (and thus could not have been retrieved otherwise) in the extracted pairs. We found that 40% of the extracted pairs contain at least one segment belonging to longer sentences, thus demonstrating the usefulness of applying the extraction at this granularity level. Some multi-clause, perfectly aligned sentences will generate several parallel clause pairs, but this will not bias the results by any means.

Finally, we performed estimations regarding the amount of parallel texts available in Wikipedia based on the domain-specific subset used in the previous experiments. Out of the comparable in-domain corpus consisting of 4.5 Million German clauses and the 2.3 Million French clauses, our approach extracts up to 222,000 parallel clauses. These represent 5% and 10%, respectively of the initial corpus. It is likely to obtain considerably more parallel segments if one would consider a broader topic, such as geography. On the other hand, it is probable to obtain even less parallel segments when choosing a very specific domain, such as tidal energy. A topical model of the Wikipedia articles would be a good indicator of the most representative domains/topics in Wikipedia, which ideally contain sizable amounts of parallel texts.

4.6 Comparison to the Sentence-based Extraction

To allow a fair comparison between the approach from Chapter 3 and the one described in the current chapter, we apply them on the same test corpus and compare the results. The corpus consists of 10 random Wikipedia article pairs with at least 20 sentences. For the sentence-based extraction we compute 3-gram BLEU scores in a single direction.

The first thing we notice is that the clause-based approach generates 30% less alignments than the sentence-based one. Moreover, the quantitative evaluation shows that the clause-based approach achieves a significant drop of the false positives and a moderate increase of the true positives. We cannot observe a clear trend regarding the number of partial alignments. Table 4.2 details the precision estimates for the considered extraction approaches on this particular test set.

Approach	**Strict Precision**	**Lax Precision**
Sentence	43%	54%
Clause	57%	93%

Table 4.2: Precision of the different approaches to extract parallel data.

We notice that the figures for the clause-based approach (both strict and lax precision) are much better than the ones for the competing approach, as expected from the drop of false positives. This finding motivates our choice to work on this granularity level.

However, there is still a relatively high number of partial alignments despite the finer granularity of the extraction approach. We already illustrated this aspect in Table 4.1, but these new figures allow us to determine the extent of this phenomenon. Generally

we observed two types of partial alignments:
a. when one segment contains information which could not have been expressed with similar grammatical structures in the second language (e.g. examples 4-5 in Table 4.1)
b. when one segment contains additional information which could have been expressed in a similar way in the other language (such as in the following example)

DE Reinhold Andreas Messner (* 17. September 1944 in Brixen) ist ein Extrembergsteiger, Abenteurer, Buch- und Filmautor (u.a. über seine Expeditionen) und Politiker aus Südtirol, Italien

FR Reinhold Messner, né le 17 septembre 1944 dans le Tyrol du Sud est un alpiniste italien,

In this case, the French clause is part of a longer sentence (given that it ends with a comma), but the remainder of the French sentence (not displayed here) does not match the German clause. The German clause comprises an extensive enumeration (*ein Extrembergsteiger, Abenteurer, Buch- und Filmautor (u.a. über seine Expeditionen) und Politiker*), which is neither present in the French counterpart, nor can it be separated any further in terms of clauses.

For this kind of phrases, only a finer-grained partition thereof (e.g. in grammatically motivated chunks) could enable the extraction of more precise alignment pairs. This approach would require a grammar-aware chunker with customizable chunk length, as well as the optimization of the alignment algorithm, since the search space would increase significantly at chunk level. We choose to stop the investigation at this level due to time constraints.

So far this thesis has focused on methods to mine Wikipedia for parallel texts. The presented approaches have a common ground, but differ in the alignment method and in the level of granularity of the extracted texts. These changes are clearly reflected by the reported accuracy values. The following chapter will discuss an approach developed for another type of comparable corpus, the Web. The approach follows roughly the same general steps, but is adapted to the particularities of the exploited Web resource, the Common Crawl.

Chapter 5

Extracting Parallel Sentences from the Web

This chapter describes an approach for extracting parallel text segments from the Web (excluding Wikipedia). The experiments are carried out on the Common Crawl corpus, a public crawl of the Web hosted on Amazon's Elastic Cloud[1]. For this purpose, we follow the general extraction procedure described in (Smith et al., 2013) for the language pair German-French and then apply a domain-specific filter for Alpine texts.

5.1 Common Crawl: Corpus Profile

The Common Crawl corpus contains pages crawled from the Web during several years. For these experiments, we use the 2009-2010 version of the crawl, consisting of 32.3 terabytes of data, corresponding to approximately 2.5 billion URLs (web pages). The figures are growing with every release, so that the current order of magnitude of the corpus is petabytes. The corpus is stored on Amazon's Simple Storage System and can be easily accessed from Amazon's Elastic MapReduce services. The 2010 corpus consists of compressed ARC files, which contain raw web documents and their corresponding metadata headers.

To have a better idea about the content of the corpus, we analyze the distribution of the web domains in the corpus. For this purpose, we choose a representative share thereof representing the top 10,000 domains ranked by the number of URLs[2]. We then group the web domains by their top level domain (the suffix of a web domain) and compute the

[1] http://commoncrawl.org

[2] http://webdatacommons.org/structureddata/2010-09/stats/top_domains_by_urls_with_triples.html

absolute and the relative frequency of the resulting groups. Table 5.1 lists the top level domains (TLD) with a relative frequency above 1%. As expected, generic TLDs (*.com, .net, .org*) are among the most frequent, covering almost 85% of the URLs retrieved in the corpus. The remaining ones contain country codes, such as *.uk* for the United Kingdom, *.nl* for the Netherlands. Domain names starting with a country/language code and ending with a generic TLD (e.g. **nl**.tripadvisor.com) are counted as part of the category corresponding to the country code.

TLD	**Absolute frequency**	**Relative frequency**
.com	26,236,181	0.768,30
.uk	1,718,308	0.050,32
.net	1,374,106	0.040,24
.org	882,510	0.025,84
.en	522,258	0.015,29
.nl	476,104	0.013,94
.es	387,473	0.011,35
.de	385,808	0.011,30

Table 5.1: The distribution of the top level domains (TLD) in the 2009/10 Common Crawl corpus.

We believe that these figures are also representative for the language distribution in the corpus, although this fact cannot be directly entailed. The main reason is that the language code of a web page usually occurs after the TLD (e.g. www.myswitzerland.com/**de-ch**/home.html). Judging by the frequency of the non-generic TLDs, we can infer that the most prominent language in the corpus is English, followed by Dutch, Spanish and German. Based on these figures, the size of the comparable corpus German-French, which will be used for mining parallel segments, can be approximated to less than 1% of the initial corpus size.

In a similar initiative, Baroni et al. (2009) released a set of very large collections of web crawled texts (more than 1 billion words) for a handful of languages, such as English, German, Italian and French. The texts are selected from webpages with explicit TLDs (e.g. *.en, .de, .it., .fr*) by means of language-specific keyword lists. Unlike the Common Crawl, these text collections (grouped under the label *Wacky corpus*) were curated prior to their release, which also included linguistic annotations. The Common Crawl texts, instead, were crawled blindly from the Web and stored on the Amazon Cloud. To our knowledge, the Wacky corpus contains a static crawl of the web (i.e. it has not been regularly updated), whereas the Common Crawl is updated monthly. This might have to do with the fact that the Wacky corpus requires linguistic preprocessing, which is a resource consuming task.

	German		**French**	
	Total	≥ 20	Total	≥ 20
Text+Berg	254,596	16,201	141,657	15,513
Common Crawl	212,934	11,040	103,659	6,627
Wikipedia	209,785	9,328	136,779	8,373

Table 5.2: Word type counts in the corpora referred in this work (Text+Berg, Common Crawl and Wikipedia)

From a linguistic point of view, the texts contained in the Common Crawl corpus are different from the in-domain corpora collected so far, either from Wikipedia or from the SAC yearbooks. We will pinpoint the differences between these corpora from three different perspectives. First, we compute the number of different word types occurring in these corpora, depicted in Table 5.2. For this comparison we used subsets of the Common Crawl and Wikipedia, respectively, containing domain-specific texts similar to the Text+Berg corpus (selected by means of the filter described in the previous Chapter). Moreover, the subsets where chosen such that they are of a similar size to the reference in-domain corpus, Text+Berg. We report both absolute frequencies and partial ones (i.e. frequency of types that occur at least 20 times in the corpus). We follow Baroni et al. (2009) in choosing this threshold as a measure of representativeness of a word in the corpus.

From this table we can see that the distribution of the word types in all the considered corpora follows a Zipfian distribution, as the number of representative types represents merely 10% of the total word types. The trends are similar in German and French, respectively, whereby the percentages for German are systematically lower than the French ones. The figures in the case of the Common Crawl are poorer than the ones obtained for the in-domain corpus, but still higher than the ones computed for a corpus extracted from Wikipedia. These figures suggest that the corpora extracted from the Web are less variate than a similarly sized in-domain corpus, as they contain less word types than the reference corpus. This trend is a consequence of the fact that Web pages are more likely to use repetitive language, such as copyright statements or navigation menu items.

However, whilst this holds for Common Crawl texts, this observation is surprising for the corpus extracted from Wikipedia, since this collection only contains the article texts, without navigation menus. We therefore conducted a detailed investigation aiming to identify the source of the repetitive phrases in the in-domain fractions of the Common Crawl and Wikipedia, respectively. For this purpose, we identified similar lines in the corpus (i.e. with a fuzzy match value above 75%) and then grouped them by the overlapping patterns occurring in the beginning of the line. Table 5.3 depicts the most frequent opening phrases from the Common Crawl-extracted corpus together with their absolute frequencies.

German		**French**	
Phrase	**Frequency**	**Phrase**	**Frequency**
Sport vom [date]	99	sport du [date entry]	95
Datum	83	date [date entry]	83
Zeitraum [date-date]	35	period [date-date]	38
Öffnungszeiten und preise	24	heures d'ouverture et prix	23
[height] meter	18	[height] mètres	35

Table 5.3: Frequent opening phrases in the in-domain fraction of the Common Crawl corpus.

The most frequent patterns represent lines containing dates, sometimes accompanied by nouns. For example, the phrase *Sport vom DD.MM.YYYY* represents the title of a sports report on a given date, which might have been selected in the corpus due to the ski-related contents. The following two phrases concerning time periods, opening times and prices, are frequently encountered on travel websites. This is a consequence of the fact that the Common Crawl contains many web pages of this kind (see Table 4 in (Smith et al., 2013)). The last row in Table 5.3 contains altitude indications, which are also frequently encountered in an Alpine corpus. Less frequent is the position of the phrase, as in most cases, the lines would start with a proper name and then an altitude indication would follow. In this case, although not very frequent, the order is reversed.

An interesting finding is the fact that the frequency values are very close in German and French, respectively. This is an indicator that the selected texts are close translations of each other. In the Wikipedia-extracted corpus, however, the repetitive phrases have different absolute frequencies in German and French, respectively, although they also represent translations of each other. The figures for the latter corpus are illustrated in Table 5.4.

German		**French**	
Phrase	**Frequency**	**Phrase**	**Frequency**
offizielle website offizielle homepage	508 110	site officiel	1232
webseite de[rs]	106	site d[eu]	1003
liste de[r]	449	liste d[eu]	689
die ortschaft liegt	53	située dans le	1840

Table 5.4: Frequent opening phrases in the Wikipedia-extracted corpus from Chapter 4.

We notice the number of repetitive opening phrases in Wikipedia, both in the German and in the French versions, are much higher than in the Common Crawl. If we would compute the language distribution summarized in this table, we would have a handful of n-grams occurring very often (*e.g. webseite, liste, offizielle webseite*) and a large collection of n-grams denoting named entities which occur once (*e.g. Département Haute-Savoie, Kanton Graubünden*). This is a clear indicator of the sparsity of the language distribution in the Wikipedia-extracted corpus.

In the second perspective, we compute the overlap between the domain-specific fraction of the Common Crawl and the Text+Berg corpus on token and type level, both for French and German. The statistics omit function words and content words occurring only once in the corpus. For this experiment, we use the stop words lists described in Section 1.6. The results are depicted in Figure 5.1.

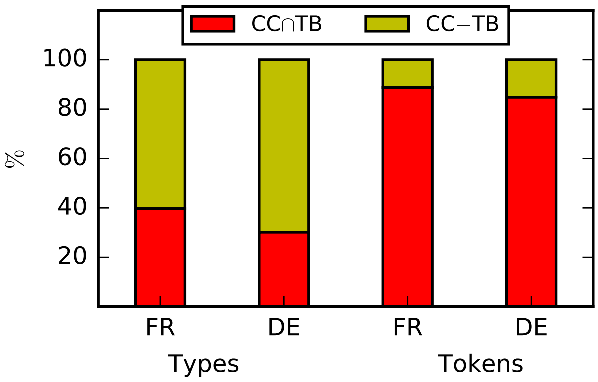


Figure 5.1: Domain Overlap between Common Crawl (CC) and Text+Berg(TB) with respect to the CC word frequencies.

The trends in the graph are similar to the ones in Figures 1.6 and 1.7, in the sense that the overlap ratio is higher on word level compared to the overlap on type level. In this case, the overlap between the Common Crawl and Text+Berg reaches similar values in French and German and these values are higher than in the case of Europarl and Text+Berg. This implies that the Common Crawl is more similar to Text+Berg (the reference in-domain corpus) than Europarl (considered an out-of-domain corpus). This finding is a confirmation of the fact that the in-domain filter applied to the Common Crawl is working properly.

Finally, we take a closer look at the vocabulary exclusive for the Common Crawl, i.e. the most frequent words in the German-French Common Crawl that do not occur in the Text+Berg corpus. The comparison is performed purely on lexical level and takes into consideration spelling variations (e.g. the use of the character *β* in standard German instead of *ss* in Swiss German). The statistics have been computed on a subcorpus of the Common Crawl consisting of approximately 240,000 in-domain parallel segments, selected in descending order of their similarity scores. Table 5.5 lists the most frequent words that do not occur in the Text+Berg corpus.

German		**French**	
Word	**Frequency**	**Word**	**Frequency**
Gästebewertungen	520	resort	603
Pattaya	506	Pattaya	509
County	385	Bali	374
Apartments	370	views	373
Bali	366	cliquez	358
Views	363	souhaitez-vous	267
Location	324	États-unis	235
Apartment	229	Beret	196
Infoboxen	201	appartements	176
Nichtraucherzimmer	195	residence	164

Table 5.5: Selection of the most frequent words that do not occur in the Text+Berg corpus.

The vocabulary specific to the "in-domain" section of the Common Crawl consists of terms related to tourism, such as *Gästebewertungen (EN: customer reviews), Location, Nichtraucherzimmer (EN: non-smoking room)* in German or *resort, appartements, residence* in French. Many of them are loan words from English, such as *County, Views, Location, resort, residence.* Since they also use the English spelling instead of the "localized" one (e.g. Apartment instead of Appartement), we assume some pitfalls in the language identification. Besides, the selection includes proper names denoting holiday destinations: *Pattaya, Bali, États-unis, Beret*, some of them common to both the German and the French vocabulary.

The list of most frequent German terms consists only of nouns, whereas in French verbs rank among the most frequent words, such as *cliquez* (EN: [you] click) or *souhaitez-vous* (EN: do you wish). Another particularity of the corpus is that the French verbs are in the imperative mood (polite form), a characteristic of advertisement texts which address their customers directly. A thorough analysis of the Common Crawl specific vocabulary (compared to the German side of the Text+Berg corpus) also reveals a couple of verbs (*auschecken*(EN: check out), *einrasten*(EN: engage)) or adjectives (*familiengeführte* (EN: family run), *behindertenfreundliche* (EN: disabled friendly)), typical for hotel descriptions/reviews.

If we compare the Common Crawl corpus with the Wikipedia extracted corpus instead, the list of frequent words specific to the Common Crawl includes considerably more adjectives, apart from the previously discussed hotel-specific nouns. For example, words like *gemütliche, konfortable* (EN: comfortable), *atemberaubende*(EN: breathtaking) or *unvergesslichen* (EN: unforgettable) do not occur (at all or more than 10 times) in Wikipedia texts. This is a strong indicator that Wikipedia texts are written in an impersonal manner, avoiding emotional statements. On the other hand, such adjectives are frequent in the SAC articles describing mountain expeditions.

The most obvious finding of this qualitative analysis is the linguistic diversity of the considered corpora, despite the fact that they cover similar topics. The in-domain Common Crawl corpus contains mostly descriptions of touristic places with advertising purposes, particular to travel websites. The Text+Berg corpus contains descriptive or narrative pieces of writing with an informative purpose, but written in a subjective manner. Wikipedia texts also have an informative purpose, but they are written in an encyclopedic, objective manner (so-called expository writings). The collected corpora thus represent a heterogeneous collection of texts, which should not be treated as a whole for further applications (such as word alignment or language model training). For example, in our SMT experiments we will first train the models separately on each corpus and then combine them in order to obtain the best performance on the test corpus.

5.2 The Extraction Workflow

Similar to our approach to mine parallel texts from Wikipedia, described in the previous chapter, the approach to mine the Common Crawl is language and domain independent. It follows roughly the same steps, but in a different order, as imposed by the general extraction workflow and by the structure of the initial corpus. The algorithm is *language independent* because it only requires the language codes corresponding to the languages of interest and does not imply any language-specific processing. It is *domain independent* because the domain-specific filtering is keyword-based and it occurs in the very end of the extraction workflow. This means that one can easily extract domain-specific subcorpora by just plugging in a set of in-domain keywords. This approach, developed together with the University of Edinburgh and Johns Hopkins University, was originally designed to extract general purpose parallel texts from the Common Crawl (Smith et al., 2013). We extend the approach with domain-specific filters to meet the purpose of our research questions.

The adapted workflow (depicted in Figure 5.2) can be thus divided into three main steps:

1. **Document alignment**: We first identify candidate document pairs (webpages available in German and French) by means of URL matching. For this purpose, we assume that the language of a webpage appears in its URL either as a ISO-639 language code or as spelled in English. For example, if the pages

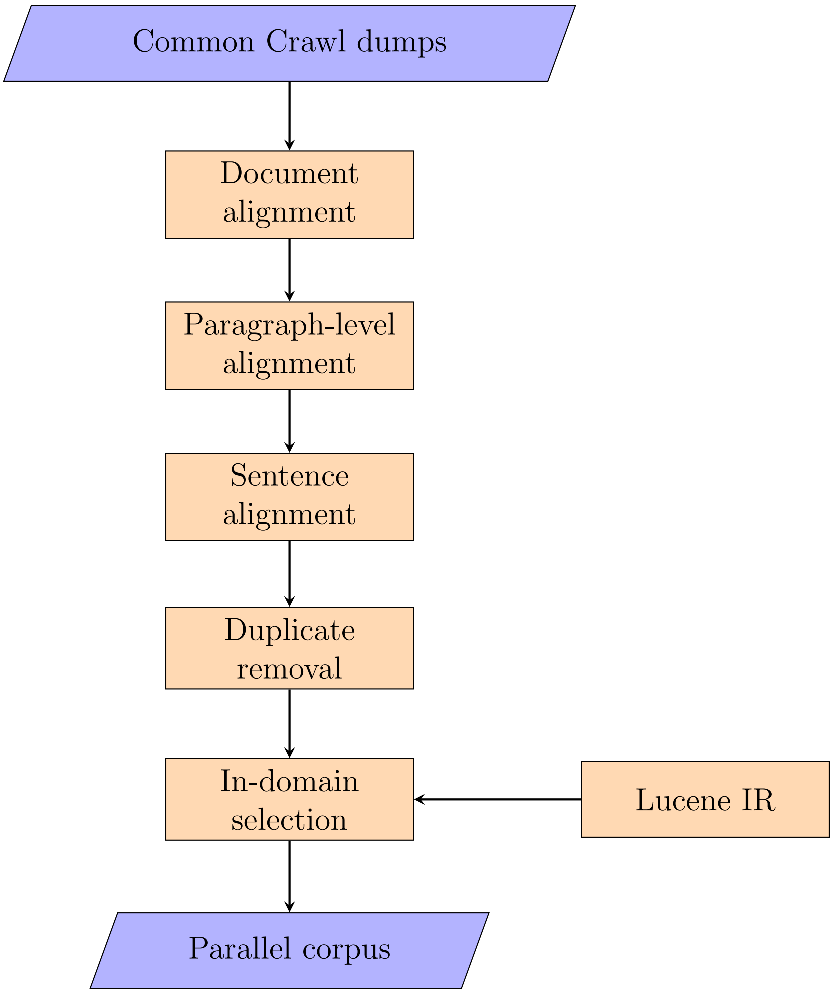


FIGURE 5.2: The workflow for extracting in-domain parallel segments from the Common Crawl.

`www.website.com/fr/` and `www.website.com/de/` can be found in the corpus, they represent a valid candidate pair.

2. **Sentence alignment**: The alignment of the document pairs is performed at two granularity levels. First we do a "coarse" section alignment based on the HTML structure of the web pages. The matching sections (text blocks) are then split into sentences and aligned at sentence level. Moreover, duplicate segments (identical source and target texts) are removed from the output.

3. **In-domain filtering**: Finally we apply a domain-specific filter (based on IR queries, as described in section 3.4.2) in order to identify the webpages containing texts from the Alpine domain. The resulting corpus contains only sentence pairs extracted from these webpages.

The size of the source corpus (32 T of data) requires an optimized processing architecture, since the usual cascaded pipelines would rapidly become computationally expensive. We therefore use Amazon's Elastic Map Reduce architecture[3] to process the input data in a distributed way, in the same time reducing the search space considerably. Specifically, we use the Map Reduce architecture to identify the correspondences between the URLs, or, in other words, to identify candidate document pairs.

In this setting, the Mapper goes through the corpus and scans the URLs of the contained web pages in search for language codes. If such a code is identified, the URL and the content of the corresponding webpage are mapped to a generic URL, where the language code is replaced with a wildcard. For example, the original URL `www.website.com/de/` will be mapped to `www.website.com/*/`. The Reducer retrieves all the webpages mapped to the same generic URL and outputs the associated values, in case matches exist for all the languages of interest (in our case, German and French).

The resulting document pairs are then downloaded to a local cluster, so that the remaining steps can be performed locally. The local processing includes further filtering of the candidates by means of the HTML structure of the webpages and the alignment of the enclosed text blocks. For this purpose, the HTML code of the webpages is first linearized, keeping only structural tags and chunks of raw text. The matching text blocks are further split into sentences and words, respectively, by means of the NLTK Punkt tokenizer (Bird et al., 2009). Sentence alignment is performed using the classical algorithm of Gale and Church (1993), since the same information is likely to appear at similar positions in the source and target texts.

As we expect web pages to contain many repetitive texts, we include a cleanup step in which such phrases are filtered out. Particularly we remove duplicate sentences (identical source and target texts) and boilerplate segments (e.g. sentences consisting only of named entities and numbers/dates, repetitive phrases such as copyright statements or web links). The extracted (and cleaned) corpus consists of over 500,000 webpage pairs and approximately 8 million sentence pairs (segments).

Finally we apply a domain-specific filter (based on IR queries, as described in section 3.4.2) in order to identify webpages containing texts related to the Alpine domain. This filter reduces the search space to 70,000 webpages, most of them related to travel. The advantage of running this step at last is that one can apply different in-domain filters on the general-domain data, without having to go through the whole costly workflow. Table 5.6 shows a selection of the webpages[4] with the highest similarity to Alpine texts

[3] http://aws.amazon.com/elasticmapreduce/

[4] Note that the URLs may no longer be valid, since the crawl used in these experiments dates from 2009-2010.

(in terms of Lucene scores) in the Common Crawl corpus. The rankings are computed independently for German and French, respectively.

The URLs are sometimes self explanatory (e.g. number 2,3,6,7), so we can be convinced that these pages contain texts about mountains or hiking expeditions judging by the enclosed keywords (e.g. *rock climb, hike, mountain*). Other links contain named entities denoting or related to mountains, such as Pico Bolivar (the highest mountain in Venezuela) in link number 4, Innsbruck in link number 5, Switzerland in link number 8 or the Chabod refuge (a mountain cabin in the Italian Alps) in link number 10, so the relation to the Alpine domain can be easily intuited. For the remaining links, only a look at the content of the web page can tell whether they represent true or false positives. In the first case (1), we deal with the hiking recommendation page of a hotel in the Aosta Valley (in the Italian Alps). The second last entry in the table (9) links to a blog-like webpage where someone describes his/her trip to Mount Cook (the highest mountain in New Zealand).

No.	**Webpage**	**Score**
1	`http://www.hostellerieduparadis.it/deu/attivita_dett.asp?id=1`	0.0985
2	`http://www.hottnez.com/de/the-15-most-spectacular-rock-climbs`	0.0890
3	`http://sierra-nevada.costasur.com/de/hiking.html`	0.0863
4	`http://www.guamanchi.com/german/trpicobolivar.html`	0.0811
5	`http://www.innsbruck-multimedia.at/index.php?lang=de`	0.0647
6	`http://www.iralto.com/fr/iran-mountains.htm`	0.0277
7	`http://www.hottnez.com/fr/the-15-most-spectacular-rock-climbs`	0.0267
8	`http://www.myswitzerland.com/fr/offer-Activities_Sports*.html`	0.0195
9	`http://www.workabroadprograms.net/fr/mount-cook/`	0.0154
10	`http://www.rifugiochabod.com/html/fra/index.php`	0.0144

Table 5.6: The best ranked in-domain webpages from the Common Crawl corpus and their corresponding Lucene scores. The scores are computed independently for German and French, respectively.

We notice from the table that the German websites achieve significantly higher scores than the French ones. For example, links number (2) and (7) refer to the different language variants of the same webpage, but the score for the French page is 70% lower than the German one. This probably happens because less keywords could be identified in the French page, since we are using slightly different keywords lists in German and French, respectively. Nevertheless, we decided to use the absolute values (Lucene scores) for merging the two rankings, since they are decisive for the selection of in-domain web pages/ texts.

An important remark is that Wikipedia pages are not included in the 2009-2010 version of the Common Crawl corpus, or at least not in the parallel German-French subcorpus. Specifically, there are no pages in the corpus that stem from Wikipedia URLs (e.g. `de.wikipedia.org`). This means that there is no overlap with the Wikipedia texts extracted in the previous experiments (see chapter 4). We can therefore fairly compare the influence of each of the extracted data sets (Wikipedia vs. Common Crawl) on the SMT performance.

5.3 Extraction Results

Unlike in the case of Wikipedia, where the amount of extracted parallel data depends on the similarity between the bilingual texts, the data extracted from the Web is selected by its similarity to the topical domain. This change is imposed by the modified extraction workflow, which first extracts generic parallel sentences and then restricts the search space to a specific domain.

Table 5.7 presents a sample of in-domain parallel sentences from the German-French Common Crawl corpus. Among them are both verbless phrases and complex sentences containing several subordinate clauses. The average sentence length ranges between 12 and 13 words in German and up to 14.5 words in French. The observed increase in length (compared to Wikipedia-extracted data) can be attributed to the fact that the alignment is performed at sentence level and not at clause level.

Nr.	French sentences	German sentences
1	Les zones touristiques de l'Aveyron	Die touristischen Zonen des Aveyron
2	Une piste de randonnée a été aménagée pour permettre de voir l'ensemble des cascades.	Ein Wanderpfad führt an den Wasserfällen vorbei.
3	Du plateau volcanique de l'Aubrac qui profile ses horizons à l'infini, au vaste plateau du Carladez qui hésite entre Rouergue et Auvergne, c'est le pays de l'authentique.	Die vulkanischen Hochfläche des Aubrac mit seinen unendlichen Weiten und die Ebene des Carladez zwischen Rouergue und Auvergne sind an Ursprünglichkeit kaum zu übertreffen.
4	Des sommets enneigés, des villages idylliques, des lacs de montagne d' un bleu intense lors de randonnées estivales, du soleil et du ski, une promenade au moyen-âge, des flâneries et du shopping.	Schneebedeckte Gipfel, verträumte Dörfer, tiefblaue Bergseen auf sommerlichen Wanderungen, Sonne + Ski, ein Spaziergang im Mittelalter, flanieren und shoppen .
5	Le pays étire 2090 kilomètres (1299 milles) de nord aux sud.	Das Land dehnt 2090 Kilometer (1299 Meilen) vom Norden bis zum Süden.

Table 5.7: Examples of parallel sentence pairs extracted from the Common Crawl

A core aspect of SMT is the quality of the word alignments generated during training, which depends on the parallelism of the input sentences. Pairs 1, 4 and 5 are examples of well aligned sentences, which are prerequisite for extracting accurate word alignments.

Nevertheless, the data set also contains sentences with slightly different meanings, but with some word overlap (such as the second example in the table). Such sentence pairs will hopefully generate some good alignments for words and phrases occurring at similar positions in the text, but misalignments are also to be expected.

A similar situation is observed in the third example, in which the two phrases convey the same meaning using different phrase structures. The French sentence includes two nested relative clauses, whereas its equivalent German sentence consists of a single complex clause. Although the word overlap is relatively high, the different length of the considered sentences will most likely pose difficulties for the word alignment. One would think that splitting the sentences into smaller units (e.g. clauses) and aligning them at the sub-sentential level could solve the problem. Alignment at clause level, however, is not beneficial in this case because only the French sentence can be split into shorter clauses, therefore the segments to align would be of considerably different lengths, similar to the initial situation.

To demonstrate this hypothesis, we compared the word alignments generated in each of the described situations: sentence-sentence and clause-sentence alignment, respectively. The word alignments are generated in a standard SMT training environment. To facilitate the understanding, we display the sentences once more below and mark the clause boundaries with $\langle CB \rangle$.

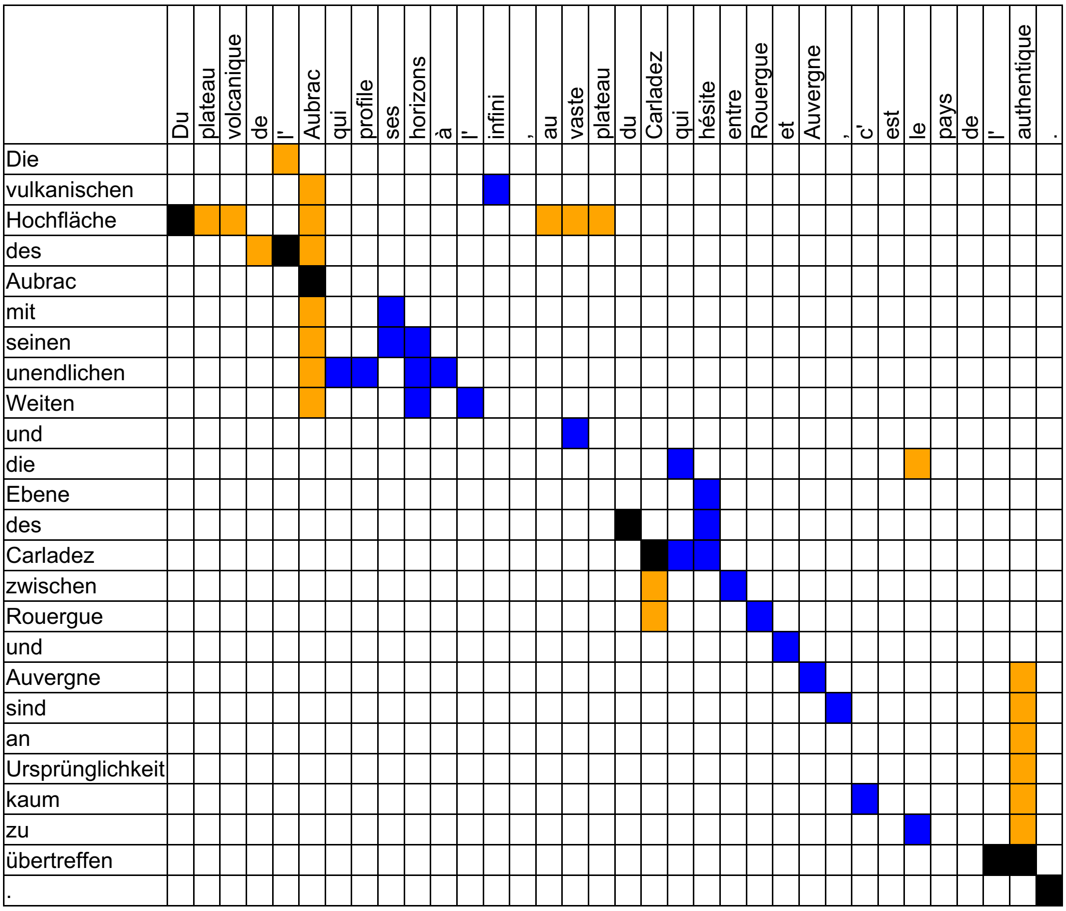
FR Du plateau volcanique de l'Aubrac $\langle CB \rangle$ qui profile ses horizons à l'infini, $\langle CB \rangle$ au vaste plateau du Carladez $\langle CB \rangle$ qui hésite entre Rouergue et Auvergne, $\langle CB \rangle$ c'est le pays de l'authentique.

From the Aubrac volcanic plateau with its horizons extending towards the infinite to the vast plateau of Carladez, which links Rouergue and Auvergne, it's the land of authenticity.

DE Die vulkanischen Hochfläche des Aubrac mit seinen unendlichen Weiten und die Ebene des Carladez zwischen Rouergue und Auvergne sind an Ursprünglichkeit kaum zu übertreffen.

The Aubrac volcanic plateau with its infinite horizons and the Carladez plateau between Rouergue and Auvergne can hardly be surpassed in authenticity.

Figure 5.3 depicts the overlaid word alignments matrix generated by aligning a. the pair of original sentences (blue squares) and b. the pair consisting of the original German sentence and the French main clause (orange squares). The rows in the matrix correspond to the German words, whereas the columns contain the French words. The filled points (squares) in the matrix represent word alignments between the words in the corresponding row and column, respectively. The black squares represent word alignments which coincide in both alignment settings. We consider two sentences to be perfectly



	Du	plateau	volcanique	de	l'	Aubrac	qui	profile	ses	horizons	à	l'	infini	,	au	vaste	plateau	du	Carladez	qui	hésite	entre	Rouergue	et	Auvergne	,	c'	est	le	pays	de	l'	authentique	.
Die					🟧																													
vulkanischen						🟧							🟦																					
Hochfläche	⬛	🟧	🟧			🟧									🟧	🟧	🟧																	
des				🟧	⬛	🟧																												
Aubrac						⬛																												
mit						🟧			🟦																									
seinen						🟧			🟦	🟦																								
unendlichen						🟧	🟦	🟦		🟦	🟦																							
Weiten						🟧				🟦		🟦																						
und																🟦																		
die																				🟦									🟧					
Ebene																					🟦													
des																		⬛			🟦													
Carladez																			⬛	🟦	🟦													
zwischen																			🟧			🟦												
Rouergue																			🟧				🟦											
und																								🟦										
Auvergne																									🟦								🟧	
sind																										🟦							🟧	
an																																	🟧	
Ursprünglichkeit																																	🟧	
kaum																											🟦						🟧	
zu																													🟦				🟧	
übertreffen																																⬛	⬛	
.																																		⬛

FIGURE 5.3: The word alignments computed for different partitions (sentences in blue versus clauses in orange) of the following sentences:
Du plateau volcanique de l'Aubrac qui profile ses horizons à l'infini, au vaste plateau du Carladez qui hésite entre Rouergue et Auvergne, c'est le pays de l'authentique. / Die vulkanischen Hochfläche des Aubrac mit seinen unendlichen Weiten und die Ebene des Carladez zwischen Rouergue und Auvergne sind an Ursprünglichkeit kaum zu übertreffen.

aligned if the filled points of its word alignment matrix are grouped around the main diagonal.

This trend can be observed partially in the original alignment setting, where we aligned whole sentences (blue squares). The clause level alignment, instead, has many gaps and several 1-to-n alignments, where $n \geq 7$ (orange squares). This is an indicator that something went wrong with the word alignment. This hypothesis is confirmed by the number of misalignments, which becomes twice as high as in the first case. We can therefore conclude that, in this case, it is sensible to align whole sentences directly despite their different length, as the generated word alignments are more accurate.

Another interesting finding is that some extracted sentences are machine translated, such as sentence pair number 5 from Table 5.7. The clue for this conclusion are the lexical choices of the verbs, which should stand for the English *stretch* (to extend between limits). This intransitive usage of the verb *stretch* requires in both French and German a reflexive verb, *s'étendre* and *sich erstrecken*, respectively. Instead, both versions use the transitive correspondents of the verb and this leads to the assumption that both sentences have been translated automatically from English. Since the current approach does not have a filter for machine translated texts, we cannot avoid such cases.

In addition to this qualitative evaluation, we also perform a quantitative one aiming to answer the question: How much of the extracted data represents parallel sentences? For this, we manually evaluate a sample of 100 sentence pairs randomly selected from the 10,000 pairs most similar to the Alpine domain. As in the previous evaluation scenarios, we use a 3-fold evaluation scheme: good, partial and bad alignments (misalignments). Sentences presumably containing automatic translations are considered true positives, although their quality and usefulness is debatable.

In this evaluation scenario, 68 of the extracted pairs are indeed parallel, 12 are only partially aligned, whereas 20 pairs represent misalignments. This translates into a 68% strict precision estimate and a 80% lax precision estimate. These figures are comparable with the ones obtained on clause level for the Wikipedia-extracted data, slightly better in terms of strict precision, but poorer in terms of lax precision. This means that the approach applied on Web data is able to retrieve more sentence pairs with similar wordings (and therefore more "parallel").

Finally, in reply to the main research questions, we estimated the amount of domain-specific parallel texts that could be extracted from the considered version of the Common Crawl. Our initial German-French corpus consisted of 8 Million aligned segments. After applying the in-domain filters, the corpus has been reduced to 242,000 segments, representing merely 3% of the initial size. If we would instead mine for sentences from a more general topical domain, such as tourism, we expect to extract a considerably bigger domain-specific corpus. On the other hand, if we would search for sentences from the political domain, for example, it is likely to obtain even less parallel segments. These assumptions are based on personal observations and on the topical analysis of the Common Crawl corpus in (Smith et al., 2013).

This chapter concludes the part of this thesis concerned with the methods for extracting parallel segments from multilingual comparable corpora. In the next chapter we will describe our experiments with the extracted data for phrase-based SMT. We will compare the parallel texts extracted with the different methods extrinsically by evaluating the performance of the SMT systems using them.

Chapter 6

SMT Experiments

In the previous chapters, we presented several approaches for extracting parallel segments from different comparable corpora (Wikipedia, Common Crawl) and evaluated the results intrinsically. In this chapter we evaluate the usefulness of the extracted data for an external application, namely SMT. The upcoming experiments use the extracted data as additional training material for Statistical Machine Translation (SMT) systems translating Alpine texts from German into French. The purpose of these experiments is to test whether adding training data guarantees a performance boost. We assess this by gradually adding data extracted from the above mentioned corpora to several baseline systems. We measure the performance using different automatic measures and we complete the analysis with human judgments.

6.1 Experimental Data and System Configurations

The key strategy in SMT is to make the most out of the available data. Although our target is to translate in-domain texts, in these experiments we use both in-domain and out-of-domain corpora. We start by giving an overview of the used corpora in Table 6.1. Note that the corpus size has been computed after the tokenization and clean-up steps (standard procedures for SMT training).

Europarl is a collection of parliamentary proceedings, which we use as out-of-domain corpus (Koehn, 2005).

Text+Berg (Release 149) is our in-domain corpus containing the publications of the Swiss Alpine Club. The development and the test data (**Dev set** and **Test set**) are also withheld from the in-domain corpus.

Wikipedia_bleu is a sample of the best ranked parallel sentences extracted from Wikipedia by means of the sentence-level approach.

Wikipedia_meteor is a sample of the best ranked parallel clauses extracted from Wikipedia by means of the clause-level approach.

Wikipedia Share{1-4} consist of clauses extracted from Wikipedia at different similarity thresholds (in terms of parallelism, as defined in Chapter 4). Share1 contains the best ranked segments, whereas the other ones include gradually less similar segments (with lower similarity scores). Moreover, the shares with high numbers include the ones with lower numbers (e.g. Share1 $\subset$ Share2 $\subset$ Share3).

Web Share{1-4} contain parallel segments extracted from the Common Crawl corpus at different domain-similarity thresholds (as described in Chapter 5). The definition of the shares follows the principles used for the Wikipedia shares.

Data set	**Segments**	**DE Words**	**FR Words**
Europarl (EP)	1,680,000	37,000,000	43,000,000
Text+Berg (TB)	280,000	4,850,000	5,500,000
Wikipedia_bleu	10,000	194,500	234,000
Wikipedia_meteor	20,000	231,000	227,000
Wikipedia Share 1	10,500	120,000	117,000
Wikipedia Share 2	68,000	768,500	762,000
Wikipedia Share 3	123,000	1,375,500	1,363,000
Wikipedia Share 4	222,000	2,495,000	2,456,000
Web Share 1	12,000	151,000	172,500
Web Share 2	36,000	432,000	480,000
Web Share 3	93,000	1,202,500	1,346,000
Web Share 4	242,000	3,060,500	3,429,000
Dev set	1424	30,000	33,000
Test set	991	19,000	21,000

Table 6.1: The size of the German-French data sets.

The SMT systems are trained with the Moses toolkit (Koehn et al., 2007), following the guidelines on the official website [1]. For comparability reasons, the data preparation workflow has been modified as follows. We work with lowercased texts instead of truecased ones and filter out sentences longer than 50 tokens (instead of 80). We build the individual language models (e.g. Europarl, Text+Berg, Wikipedia Shares) with KenLM (Heafield et al., 2013), which implements interpolated modified Kneser-Ney estimation, using n-gram length 5 (instead of 3).

[1] http://www.statmt.org/moses/?n=Moses.Baseline

The Text+Berg language model is trained on the monolingual French side of the corpus, which sums up to 14 million tokens. All other language models are trained on the parallel side of the respective corpora (e.g. Europarl, Web Share3 etc.). We train 5-gram phrase-based translation models, computing the word alignments with MGIZA++ (Gao and Vogel, 2008).

The combined language models, which represent linear interpolations of the component LMs, are built with SRILM (Stolcke, 2002). The combined phrase tables are optimized for minimal perplexity on the in-domain development set with the tools available in the Moses distribution (Sennrich, 2012). The parameters of the global models (phrase table, reordering model, language model) are optimized (after combination) through Minimum Error Rate Training (MERT) on the same in-domain development set (Och, 2003).

Unless stated otherwise, the translation direction is from German into French. The translation performance is measured using several evaluation metrics (BLEU, METEOR 1.4 and TER) on a single reference translation. We choose to report several evaluation scores because they consider different comparison criteria, e.g. BLEU - n-gram precision, METEOR - precision and recall for exact, stemmed and synonym matches, TER - number of edits required to correspond to the reference. Additionally, we analyze the out-of-vocabulary (OOV) rate of the filtered translation models on the test set as an indicator of data sparseness. We also apply statistical significance tests, in order to indicate the validity of the comparisons between the SMT systems (Riezler and Maxwell, 2005). We consider the score differences significant if the computed p-value is below 0.05.

6.2 SMT Results

We set up different SMT experiments aiming to assess the impact of the extracted data on SMT performance. The purpose of the first experiment is to compare the data extracted with the approaches described in Chapters 4 and 5 in terms of usefulness for SMT. Specifically, we use the extracted data as additional training data for the same baseline system and compare the respective performances. In the following experiments, we add data gradually to the baseline systems in order to identify the optimal trade-off between the amount and the quality of the additional data required to improve a given baseline system. We perform the experiments with two different baselines, an out-of-domain and an in-domain one. In both cases, the procedure is the same: we first add the segments with the highest similarity scores (as defined during the corresponding extraction process) and then progressively add less similar segments (in accordance with the decreasing similarity scores).

6.2.1 Comparison of the Extraction Approaches

In the first experiment, we add two different samples of parallel segments extracted from Wikipedia on top of an in-domain baseline. The data samples are comparable in size, but are the outcome of different extraction approaches and have very little overlap (approximately 1%). The results are summarized in Table 6.2.

System configuration (TM, LM)	**BLEU ↑**	**METEOR ↑**	**TER ↓**
Text+Berg (TB)	18.5	37.4	67.8
TB+Wikipedia_bleu	18.3	37.2	68.1
TB+Wikipedia_meteor	18.3	37.2	67.5

Table 6.2: SMT results (DE-FR) for system configurations using in-domain data and Wikipedia data extracted with different approaches.

In this case, no evaluation metric can pinpoint a significant difference between the systems performance. However, a mere drop of 0.3% TER can be observed between the baseline system and one of the systems with additional Wikipedia data, extracted by means of the clause-level approach (*TB+Wikipedia_meteor*). Additionally, we compute the OOV rate of the respective phrase tables against the test set and find only a 0.1% drop between the baseline system and the combined ones. These findings imply that the amount of additional data is too small to improve the existing in-domain SMT system. Moreover, it seems that the quality of the additional data does not play a significant role in this case, since the overall quality stays the same regardless of the method used to extract the additional data. Therefore we cannot make any assumption regarding the influence of the extraction method on the SMT performance.

6.2.2 Comparison to the Out-of-domain Baseline

The purpose of this experiment is to simulate the case where no in-domain data is previously available, therefore we add potential in-domain data from Wikipedia on top of an out-of-domain system trained on Europarl. The results are summarized in Table 6.3. We notice that the additional in-domain data significantly improves the baseline system (EP), as reflected by all evaluation metrics. The improvement is visible when using as little as 10,000 parallel segments on top of the existing 1.7 million segments. This is a strong indicator that the extracted data is similar to the topical domain of the test data (in-domain data).

Adding more extracted data slightly improves the SMT performance, as measured by the MT evaluation metrics, but the differences are moderate. This means that the plateau

System configuration	**BLEU** ↑	**METEOR** ↑	**TER** ↓	**TM OOV (types)** ↓	**TM OOV (tokens)** ↓
Europarl (EP)	11.0	27.9	77.3	32.3%	11.3%
EP+Wiki-Share1	12.1	28.9	75.1	31.2%	10.4%
EP+Wiki-Share2	11.9	29.0	75.0	29.1%	9.4%
EP+Wiki-Share3	12.1	29.3	74.9	27.8%	8.8%
EP+Wiki-Share4	12.2	29.4	75.5	27.1%	8.6%

Table 6.3: SMT results (DE-FR) for system configurations using Wikipedia-extracted data and out-of-domain data.

of the learning curve of the SMT systems is reached rather soon, compared to the trends illustrated in (Abdul Rauf and Schwenk, 2011). The overall SMT performance trends are similar to other approaches in the literature employing standard domain adaptation techniques (e.g. (Munteanu and Marcu, 2005, IMS, 2013, Pecina et al., 2015)), but on a lower scale. This effect is strongly related to the chosen language pair and the chosen domain. Although our BLEU scores are much lower than the ones reported for other language pairs, they represent the state of the art for German-French and are in agreement with the scores obtained in the TTC project for the same languages, but for a different sort of texts.

Since BLEU scores below 20 are not considered reliable, we also computed the OOV rate of the developed SMT systems, similar to Jehl et al. (2012). In our case, we noticed a steady drop of the OOV rate, which indicates that the SMT systems have to deal with less unknown words. This represents a clear indicator that the additional data contributes to improving the SMT performance.

Another way to analyze these results is by plotting the relative improvement of the automatic scores between the system combinations. We define the relative improvement as the relative change multiplied by plus one if an increase implies an improvement and by minus one if a decrease implies an improvement, where the relative change between the new value x and the reference value x_{ref} is

$$\text{Rel. change}(x, x_{\mathrm{ref}}) = \frac{x - x_{\mathrm{ref}}}{x_{\mathrm{ref}}} \,. \tag{6.1}$$

In our case the reference value x_{ref} is given by the score of the baseline and the new value x is the score of the combined systems, e.g. EP+Wiki-Share3. By score we mean the automatic evaluation scores such as BLEU, METEOR, TER or OOV rate. It is important to notice that a performance improvement triggers always a positive relative improvement, thus a decrease of the TER score (which implies a quality improvement) corresponds to a positive relative improvement. We compute the relative improvement

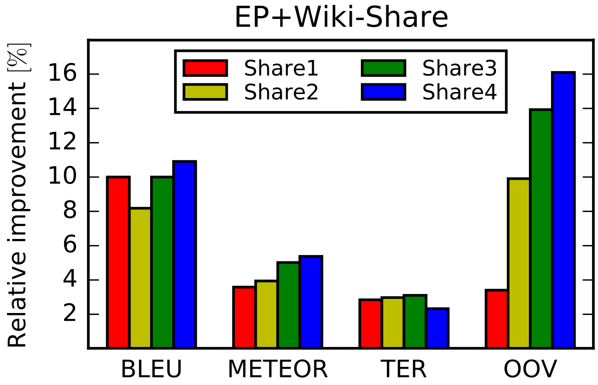


FIGURE 6.1: The relative improvement between the out-of-domain baseline and the system combinations built on top of it with Wikipedia-extracted data, expressed in terms of BLEU, METEOR, TER and OOV scores.

in percentages, hence

$$\text{Rel. imp.}(Sys, Base|Score) = \text{sgn}(Score)\frac{Score(Sys) - Score(Base)}{Score(Base)} \times 100\ , \qquad (6.2)$$

where in practice sgn(BLEU) = sgn(METEOR) = 1 and sgn(TER) = sgn(OOV) = −1 .

Figure 6.1 illustrates the relative improvement between the out-of-domain baseline and the system combinations built on top of it with Wikipedia-extracted data. The changes are measured between the BLEU, METEOR, TER and the type-level OOV (further referred as OOV) scores. Except for the BLEU scores (where we could not identify a clear trend), the automatic scores are in agreement and show a gradual growth when more data is used. These findings imply that the increasing amounts of additional data have a positive effect on the baseline.

Table 6.4 illustrates the influence of the data extracted from the Common Crawl on the same out-of-domain baseline. All system combinations significantly outperform the baseline (with up to 1.6 BLEU). Even the smallest share of extracted data (12,000 segments) triggers a considerable improvement of 1.1 BLEU on top of the existing system (trained on 1.7 million parallel segments). The performance is similar to the one obtained by the smallest share of Wikipedia-extracted data, which is of comparable size. Adding more data steadily improves the SMT performance (up to the the third additional share). Unlike for the systems trained with Wikipedia data (see Table 6.3), the BLEU score differences between the system combinations with data from the Common Crawl (Share1-4) are statistically significant.

The lexical coverage of the systems trained with additional data also improves, as we can infer from the decrease of the OOV rate. Moreover, the current system combinations

System configuration	**BLEU** ↑	**METEOR** ↑	**TER** ↓	**TM OOV (types)** ↓	**TM OOV (tokens)** ↓
EP	11.0	27.9	77.3	32.3%	11.3%
EP+Web-Share1	12.1	29.0	74.2	30%	9.9%
EP+Web-Share2	12.3	29.7	74.9	28.9%	9.3%
EP+Web-Share3	12.8	30.3	73.6	26.6%	8.4%
EP+Web-Share4	12.7	30.5	74.1	24.4%	7.6%

Table 6.4: SMT results (DE-FR) for system configurations using Web-extracted data and out-of-domain data.

show a steeper drop of the OOV rate than the ones using additional data from Wikipedia. This possibly implies that the texts from the Common Crawl are more variate than Wikipedia texts, hence the systems using them learn to translate more new words.

The following example illustrates the improvement of the lexical coverage as an effect of the additional training data. For this purpose we compared the output of the out-of-domain baseline with the output of the best performing system combination for each data set (Wikipedia and Common Crawl, respectively). The purely out-of-domain system fails to translate the past participle *bestiegen* (EN: climbed), leaving it untranslated, whereas the system *EP+Web3* omits it entirely. On the other hand, the system using Wikipedia data *EP+Wiki1* gets the correct translation into French (*gravi*). This comes with some downsides, such as the disagreement with the subject, the position in the sentence, or the insertion of additional tokens (*il est* - EN: it is). Nevertheless, we consider that the positive effect overpowers the negative aspects, since the latter ones can be overcome by using the same data with additional preprocessing (e.g. reordering, building syntactic models), but the lexical coverage can only be improved by means of additional data.

DE orig	Wir hatten sechzehn Gipfel **bestiegen.**
FR ref	Nous avions **gravis** 16 sommets.
Gloss	We climbed 16 peaks.
EP	Nous avons eu seize sommet **bestiegen.**
EP+Web3	Nous avons eu seize sommet.
EP+Wiki1	Nous avions 16 sommet **il est gravi.**

The relative improvements caused by the data extracted from the Common Crawl to the out-of-domain baseline are depicted in Figure 6.2. We notice that there is an agreement between the BLEU and the TER scores, and, METEOR and OOV, respectively, but not between all of them at the same time. This may be an effect of the way we extracted the data shares, namely by their similarity to the domain. The SMT performance peaks with the addition of the third share and then decreases with the fourth share, which implies that Share4 contains data that is not relevant for the domain, possibly interfering with the in-domain data.

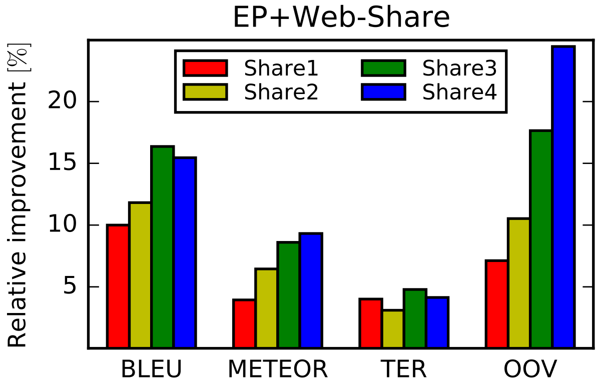


Figure 6.2: The relative improvement between the out-of-domain baseline and the system combinations built on top of it with Web-extracted data, expressed in terms of BLEU, METEOR, TER and OOV scores.

6.2.3 Comparison to the In-domain Baseline

In the case where we have a decent in-domain data set, the additional data either from Wikipedia or from the Common Crawl cannot bring sizable improvements. Since the automatic evaluation is performed after the optimization step and it is known that MERT optimization is not deterministic, we conclude that the variations of the BLEU score (of maximum 0.2 BLEU) are not statistically significant. This assumption is in agreement with the results of the statistical significance tests. The results of the SMT experiments with in-domain data can be found in Table 6.5 (for Wikipedia data) and Table 6.6 (for Web data).

Munteanu and Marcu (2005) encountered a similar situation when adding their extracted in-domain texts on top of a big collection of both in- and out-of-domain texts. Since the baseline systems were trained on considerable amounts of texts, the additional in-domain ones could not make any significant change. They also illustrated the evolution of the SMT performance with the growth of the baseline corpus and showed that the performance stops improving after a given corpus size. We think that in our case the plateau has been reached earlier, making it difficult to outperform the baseline. This might be the reason why many experiments in the literature choose out-of-domain baseline systems for comparison.

The automatic evaluation metrics do not reflect the small changes in the MT output, especially when both outputs differ from the reference. The OOV rate, instead, is a better indicator, since it quantifies the lexical coverage. The drop of the OOV rate indicates that the combined systems have a better lexical coverage. Although the drop of the OOV rate (0.1% − 0.3%) is smaller than the one between the out-of-domain baseline and the system combinations built on top of it (0.9% − 3.7%), we can nevertheless

System configuration	**BLEU** ↑	**METEOR** ↑	**TER** ↓	**TM OOV (types)** ↓	**TM OOV (tokens)** ↓
TB	18.5	37.4	67.8	14.0%	4.1%
TB+Wiki-Share1	18.3	37.3	68.1	13.8%	4.0%
TB+Wiki-Share2	18.5	37.4	68.2	13.6%	3.9%
TB+Wiki-Share3	18.3	37.3	68.1	13.8%	4.0%
TB+Wiki-Share4	18.4	37.2	68.2	13.6%	4.0%

Table 6.5: SMT results (DE-FR) for system configurations using Wikipedia-extracted data and in-domain data.

conclude that the additional data brings added value to the baseline system. Moreover, our relatively low OOV rates indicate that unknown words are not the main cause of the poor SMT performance, but rather the different grammatical structures between morphologically rich languages, such as German and French.

We illustrate the above discussed trends in Figure 6.3. We notice that the automatic evaluation scores BLEU, METEOR and TER (with one exception) are in agreement, showing similar rise and fall trends. Except for the first system combination *TB+WebShare1*, all system combinations achieve negative percentage changes. This implies that the SMT performance decreases when more data is added to the system, probably because the additional data modifies the translation probabilities of in-domain terms, thus influencing the lexical choices of the system.

Although it was not possible to identify a global improvement trend of the translation quality after supplementing the training data, we could still pinpoint local improvements. In the following example, the additional data contributes to the improvement of the lexical choices and at the same time reduces the number of unknown words.

The first thing that draws our attention is the noun phrase *unser schwächster skifahrer* (EN: our weakest skier). The closest translation to the reference is obtained by the system *TB+Wiki3*, the only system which is able to translate the adjective *schwächster*. On the other hand, this system replaces the possessive *unser* (EN: our) with the determiner *le* (EN: the). The other systems wrongly translate the noun in plural, as the German

System configuration	**BLEU** ↑	**METEOR** ↑	**TER** ↓	**TM OOV (types)** ↓	**TM OOV (tokens)** ↓
TB	18.5	37.4	67.8	14.0%	4.1%
TB+Web-Share1	18.6	37.6	67.6	14.0%	4.0%
TB+Web-Share2	18.3	37.1	67.2	13.9%	4.0%
TB+Web-Share3	18.0	37.0	68.5	13.3%	3.9%
TB+Web-Share4	18.4	37.4	68.4	12.7%	3.7%

Table 6.6: SMT results (DE-FR) for system configurations using Web-extracted data and in-domain data.

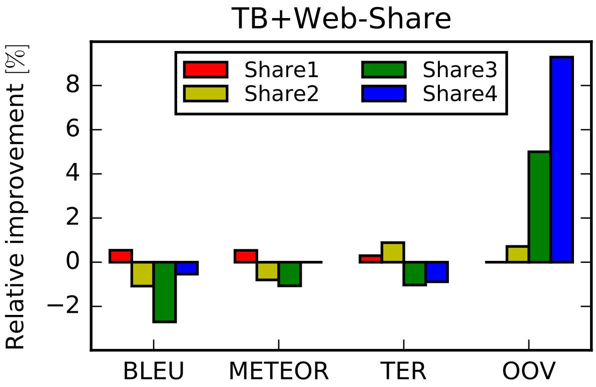


FIGURE 6.3: The relative improvement between the in-domain baseline and the system combinations built on top of it with Web-extracted data, expressed in terms of BLEU, METEOR, TER and OOV scores.

DE orig	**Unser schwächster Skifahrer**, der die Ski längst abgezogen hat, erreicht den Talboden mit grossem Vorsprung als erster.
FR ref	**Notre skieur le plus faible**, qui a enlevé ses skis depuis longtemps, atteint le fond de la vallée le premier et avec une grande avance.
TB	**Notre schwächster skieurs**, les skis depuis longtemps en débandade, a atteint le fond de la vallée où le premier grand ressaut.
TB+Web2	**Notre schwächster skieurs**, les skis depuis longtemps en débandade, a atteint le fond de la vallée, avec le premier grand ressaut.
TB+Wiki3	**Le skieur le plus faible**, les skis depuis longtemps en débandade, a atteint le fond de la vallée où le premier grand ressaut.

surface form is the same in singular and plural, and, even worse, leave the adjective *schwächster* untranslated.

The relative clause which follows poses difficulties to all considered systems. The verb is translated wrongly with the prepositional phrase *en débandade* due to the fact that the German verb *abziehen* can build perfect tenses with both auxiliary verbs “to be” or “to have”, in which cases it has different meanings. The selected translation of the participle *abgezogen* originated from a phrase with the inappropriate auxiliary (*das militär ist abgezogen* - EN: the military withdrew / stepped back), which was aligned to the French phrase *les militaires sont en débandade.* Since this is in fact a paraphrase of the first fragment, the system learned wrong phrase alignments, which lead to the erroneous output.

All in all, we consider the system which left no words untranslated better than both the baseline and its competitor. This result would not have been possible without the additional texts extracted from Wikipedia. The remaining errors could still be resolved if we provided the system with additional information, such as morphological information or input sentences reordered according to the word order in the target language.

6.3 Comparison with Commercial SMT Systems

Since the previous experiments date a few years back, we compared the performance of the in-domain SMT system from Section 6.2.3 with commercial systems using the latest MT technology, such as Google Translate and Systran. The results, as shown in Table 6.7, suggest that our system based on a moderate amount of in-domain texts is better than the commercial systems using considerably bigger amounts of general texts and possibly more recent MT technology, such as neural translation models. This finding is reflected by all the considered evaluation metrics, regardless if they are only based on exact matches (e.g. BLEU, TER) or also on word stems and paraphrases (e.g. METEOR). Moreover, this demonstrates that domain adaptation is still a powerful approach in Statistical MT and that the collected texts are valuable over time.

System	**BLEU ↑**	**METEOR ↑**	**TER ↓**
In-domain (TB)	18.5	37.4	67.8
Google	16.0	35.0	71.5
Systran	13.3	33.5	75.3

Table 6.7: SMT results (DE-FR) for system configurations using Web-extracted data and in-domain data.

In order to gain a better understanding of these figures, we analyzed a few translations from the considered test set. We chose sentences with various translation performances on the same system, in order to avoid biased results. The examples reveal two main tendencies. As expected, the commercial SMT systems generally have difficulties in translating words from the Alpine domain, as it is clear from the first three examples.

In the first one, the noun phrase *eine steile, eher griffarme Rampe* (EN: a steep ramp, with relatively few grips) poses difficulties: Google translates all words literally, but cannot convey the meaning of the adjective *griffarm*, whereas Systran even leaves it untranslated. In the second example, the commercial systems translate the noun *Angriff* with its most common equivalent in French *attaque* (EN: attack), whereas the desired translation would have been *tentative* (EN: attempt). In the third example, the word *flüssig* is interpreted by the commercial system as an adjective and thus translated as *liquide* (EN: liquid), but in the original texts it was used as an adverb, as to express a continuous movement. In all these cases, our in-house MT system provided the correct translation in place.

DE orig	Es ist dies eine steile, eher griffarme Rampe.
EN gloss	This is a steep ramp, with relatively few grips.
FR ref	C'est une rampe raide, plutôt pauvre en prises.
TB	C'est une rampe raide, plutôt pauvre en prises.
Google	Ceci est une plutôt faible poignée rampe raide.
Systran	C'est cela une piste raide, plutôt griffarme.
DE orig	20. Juni: unser dritter Angriff auf das Gross Grünhorn (4044 m).
EN gloss	June 20th: our third attempt to the Gross Grünhorn (4044 m).
FR ref	Le 20 juin eut lieu notre troisième tentative au Gross Grünhorn (4044 m).
TB	Le 20 juin, notre troisième tentative au Gross Grünhorn (4044 m).
Google	20 juin: notre troisième attaque du Grünhorn (4044 m).
Systran	20. juin: notre troisième attaque sur grandement la corne d' écologiste (4044 m)
DE orig	Wir klettern flüssig weiter, schräg aufwärts, der Wand entlang.
EN gloss	We climb up steadily along the mountain face.
FR ref	Nous continuons à grimper en nous élevant en diagonale le long de la paroi.
TB	Nous grimpons vivement, en diagonale, le long de la paroi.
Google	Nous montons encore liquide, obliquement vers le haut le long du mur.
Systran	Nous montons plus loin, en diagonale vers le haut, la paroi liquide.
DE orig	Die nächsten Montblanc-Anwärter sind bereits eingetroffen.
EN gloss	The next candidates to the Mont Blanc have already arrived.
FR ref	Les prochains candidats au Mont Blanc sont déjà arrivés.
TB	Les prochaines Montblanc-Anwärter sont déjà arrivés.
Google	Montblanc prétendants à venir sont déjà arrivés.
Systran	Les prochains candidats de Mont-Blanc sont déjà arrivés.
DE orig	Gegen Ende der Woche aber besserten sich die Verhältnisse zusehends.
EN gloss	By the end of the week the conditions improved significantly.
FR ref	Mais, vers la fin de la semaine, les conditions s'améliorèrent sensiblement.
TB	Vers la fin de la semaine, mais les conditions se besserten à vue d'œil.
Google	Plus tard dans la semaine, mais les conditions améliorées visiblement.
Systran	Vers la fin de la semaine, les relations se sont toutefois améliorées sensiblement.

Table 6.8: Example translations generated with our in-domain system and other commercial SMT systems.

The latter two examples illustrate the problem of untranslated words. In the fourth example, the German compound *Montblanc-Anwärter* was translated correctly only by the Systran system. Google successfully split it into parts (*Montblanc prétendants*), but translated them literaly and thus failed to generate a sound noun phrase, such as *prétendants au Montblanc*. Since our system trained on Alpine texts did not see the compound in the training phase, nor included a module for compound splitting, the compound was left untranslated. We think that our system would have been able to translate the individual words forming the compound if they would have been entered separately (*Montblanc Anwärter*).

In the last example, the verb *sich bessern* (EN: to improve) could not be translated by our domain-specific system. Although the verb occured two times in the training corpus in the exactly same form, the automatic word alignments did not assign an equivalent for it. Google translated the main verb correctly, but omitted the reflexive pronoun and the auxiliary verb. Systran generated the translation most similar to the reference, though it included an inappropriate translation of the noun *Verhältnisse* (EN: conditions, relations), which is otherwise translated correctly by the other systems. A solution in this case would be to combine the available translation hypotheses in order to correct the shortcomings of the in-domain system, as suggested by Sennrich (2013).

These examples provide further support of our idea that in-domain texts are extremely valuable when training SMT systems, as they represent the basis for learning the correct translations in the context. This is of utmost importance for words with several meanings, one of which is frequent in general texts and another one occurring mostly in domain-specific texts. When enough in-domain data is used to train the SMT system, the translation probabilities will be automatically adjusted in order to give preference to domain-specific translations.

In Section 1.6 we showed that the vocabularies of an in-domain and an out-of-domain corpus vary a lot, in our case having an overlap of merely 26-35% (see Figure 1.6). This means that in-domain texts also contribute to the high lexical coverage of the SMT systems using them. This is also reflected by the higher evaluation scores achieved by the in-domain system. Although in some of the cases, the domain-specific system fails to translate words with little or no evidence in the underlying training corpus (e.g. compounds), its performance can be improved by using translations from other MT sytems.

6.4 Discussion

The previous experiments shed light upon the effect of data extracted from comparable corpora on existing SMT systems. First, the data extracted from either Wikipedia or the Common Crawl can bring significant improvements to an out-of-domain SMT system.

We showed that small amounts of quality data can improve the performance of the baseline system, even for a ratio of 1:160 of in-domain vs. out-of-domain data. This means that it is possible to build SMT systems for specific topical domains based on comparable texts from the domain of interest. The described extraction methods are therefore expected to mitigate the bottleneck of lacking parallel in-domain data, which are crucial for building a SMT system for a new language pair.

Surprisingly, the extracted data does not have a clear positive influence on the chosen in-domain system. The automatic metrics are not sensitive to the small changes produced by the additional data. However, improvements can be seen in terms of lexical coverage, as shown by the out-of-vocabulary rates. By analyzing the influence of varying amounts of additional training data on SMT performance, we demonstrate that, in this particular case, the assumption that *More data is better data* does not hold when combining a strong in-domain data set with data sets extracted from comparable corpora.

The presented results are strongly correlated with the test set used for evaluating the systems. The current test set, consisting of sentences from the Text+Berg corpus, contains many long, elaborated phrases. The average sentence length in the test set is 19.40 in German and 21.8 in French, higher than the averages computed for Common Crawl texts, for example. It is thus possible to obtain more significant results when comparing the systems against a different test set, for example one containing shorter and simpler sentences.

We envision that the same evaluation setting can pinpoint clearer trends for topical domains where more data is available, such as automotive texts. The experiments conducted by Ştefănescu et al. (2012) showed that significant amounts of parallel segments can be extracted from Web pages related to this domain. Moreover, we think that the repetitive language used in these texts (similar to other technical documents) is an advantage for the translation system, because it mostly deals with text constructions which it has seen before. The performance improvement is then accordingly high even for a difficult translation direction such as English-German.

On the other hand, the performance of our systems is in agreement with the results reported for German-French even for other topical domains. This means that the performance is limited mainly by the performance of state of the art statistical translation methods for this particular language pair. Moreover, the domain adaptation techniques seem to perform similarly for different narrow topical domains, such as alpinism or wind energy. However, a significant performance improvement has been reached for German-French by employing the most recent approaches in MT, pivot-based neural machine translation (Cheng et al., 2017). This result supports our previous claim that the statistical approach reached its limits and that a new paradigm becomes established in MT.

However, the reported results concern only open domain translations. In terms of domain-specific translations and in particular for Alpine texts, our system outperforms state of the art commercial systems, such as Google Translate or Systran, as shown in Table 6.7. This finding suggests that domain-adaptation is a prerequisite to reliably translating in-domain texts, regardless of the translation paradigm used. Further research is needed to find the most effective way of applying domain adaptation to neural translation models.

Chapter 7

Conclusions

This work introduced three original methods for extracting parallel text segments from comparable text collections (in particular from Wikipedia and the Common Crawl). We particularly focused on texts close to the Alpine domain in German and French. The main part of the work is the extraction of domain-specific texts from Wikipedia described in Chapters 3 and 4. The methods have a common ground, but differ in the granularity of their output. The third method, described in Chapter 5, was designed for the generic extraction of parallel segments and has been subsequently tailored for the Alpine domain.

We evaluated the extracted texts both intrinsically, i.e. in terms of parallelism, and extrinsically, i.e. in the context of domain-specific Statistical Machine Translation (SMT). Our results demonstrated that the considered comparable corpora contain parallel segments and we were able to identify some of them. An estimation of the recall values was not possible due to the large size of the corpora, but the precision values computed for the test sets range from 57% to 93%. Chapter 6 is entirely dedicated to the SMT experiments with the texts extracted from the various corpora. Whilst some of the conducted experiments (e.g. adding the extracted texts on top of an out-of-domain baseline) led to success, in other cases the extracted texts were not good enough to improve significantly the SMT performance. The poor performance is most probably due to the fact that the extracted texts use an unvaried vocabulary, as shown in Table 5.2.

Moreover, it is conceivable that the extracted texts can be used for other practical applications as well, such as computer-assisted language learning and translation (Delpech, 2014), cross-linguistic translation studies (Bernardini and Ferraresi, 2013) or terminology extraction (Morin et al., 2013). The latter research direction has received particular attention in the past twenty years, as it offered an effort-saving alternative to the manual compilation of dictionaries.[1] Moreover, language professionals show increased interest for automatic technologies, which have the potential to minimize their workload.

In the following paragraphs we will discuss our main findings in detail, at the same time seeking to answer the research questions from the beginning of this work.

Our first research question was related to the amount of translated text segments available in comparable corpora. As mentioned in the introductory chapter, comparable corpora represent collections of texts on similar topics written independently from each other. Our hypothesis was that texts on the same topic are likely to contain overlapping information, although their source is different. We found that the examined comparable corpora indeed contain semantically equivalent pieces of text, including parallel texts (i.e. translations). The amount of parallel text varies with the source texts and with the accepted similarity degree between the text segments. For example, we found Wikipedia pages with no or only a handful of translated segments (e.g. the German and French versions of the article about the Appenzell Alps) and others which represent almost accurate translations of each other (e.g. the German and French versions of the article about Jedediah Smith, an American explorer). Our experiments were able to extract shares of parallel texts representing 3% to 10% of the considered corpora.

A more precise estimate of the parallel data available in Wikipedia is provided by the statistics over the OPUS corpus, a freely available collection of parallel text from various domains (Tiedemann, 2012). The corpus includes a collection of 36 general domain bitexts extracted from Wikipedia [2]. Unfortunately, the language pair German-French is not present in the corpus, but we could have a rough estimate of its size if we consider the size of the German and French bitexts (paired with English and Polish, respectively). We think that the size of the German-French bitext could lie in the range 0.2-0.8 million sentence pairs, which represent intermediate values between the size of the individual bitexts paired with Polish and with English, respectively. Our method can extract up to 0.25 million parallel segments similar to the Alpine domain (whereby we expect the texts to cover also bordering domains). Therefore we consider the above estimate of the available parallel texts to be legitimate.

Once demonstrated that comparable corpora also contain translated pieces of text, the next question was how to extract the parallel pieces accurately. The first step towards this goal was to identify the particularities of these corpora which signal the presence of parallel texts. The proposed extraction methods exploited the particularities of the considered corpora. In the case of Wikipedia, reliable anchor points were the interlanguage links or the outgoing internal links, whereas in the Common Crawl, the URLs and

[1] See Chapter 1 of (Sharoff et al., 2013) for a detailed overview of existing approaches in this field.

[2] http://opus.lingfil.uu.se/Wikipedia.php

the HTML structure of the documents represented useful cues. For example, document matching was trivial for Wikipedia texts, since the correspondence between documents was given by the interlanguage links. In the case of the Common Crawl, we had to use URL matching to find the corresponding documents, but this method would miss possible candidates in case the naming conventions are not consistent across languages. On the other hand, the alignment based on the document structure was not reliable in the case of Wikipedia, since there was no 1-1 correspondence between the article sections in different language variants and even if it would exist, the sections could occur at different positions in the text. In the case of the Common Crawl, the document structure (i.e. HTML structure) played a key role in the alignment process.

The extraction methods followed roughly the same steps, namely the identification of matching documents, the keyword-based extraction of in-domain documents, the extraction of parallel segments and the filtering of the extracted segments. The identification of matching documents followed according to the criteria mentioned above (interlanguage links and URL matching). For the selection of in-domain documents we ran IR queries containing frequent mountaineering keywords extracted from an in-domain corpus, Text+Berg. We then removed formatting information, figures, tables etc. and split the remaining plain text into sentences and clauses, respectively. The extraction of parallel segments was modeled as an alignment problem, whereby segments represented either sentences or clauses, depending on the extraction method. Finally the candidates were filtered by means of a similarity threshold, which in case of Wikipedia texts reflected the similarity between the text segments, whereas in case of Common Crawl the similarity to the topical domain.

The proposed extraction methods differed in the modeling of the extraction steps (e.g. alignment, similarity), as well as in the sequence in which they were applied. For example, the methods applied on Wikipedia articles used an intermediate translation of one of the texts to be aligned, thus the alignment was performed between texts in the same language, whereas the method applied on the Common Crawl aligned bilingual texts directly. As a consequence, the similarity metrics for comparing the texts also differed, even the ones used for monolingual comparisons. The bilingual similarity metric relied on the segment length, whereas the monolingual similarity metrics relied to a great extent on automatic evaluation metrics used in SMT. Another difference concerned the granularity of the extracted texts: two methods returned parallel sentences and the third one parallel sub-sentential fragments (clauses).

A delicate question when including automatic translations in a more complex workflow is the quality of the translations, particularly in the case of matching translations against natural language. To improve the precision of the matching process (and therefore of the whole extraction process), we chose the best translation direction for the considered language pair (in our case, from German into French). A further refinement of the method would be to consider not only the best translation generated by the system, but the n-best translation variants. This would likely improve the extraction recall, as the search space would increase by various n-grams.

A limitation of our approach is that we discover sentence pairs which contain, to a great extent, words and phrases that were already known to the MT system. This potentially affects the recall of the extraction approach, as we lose sight of sentence pairs which contain, for example, words unknown to the MT system. Moreover, if we retrain the MT system using additionally the extracted pairs, the chances to obtain improvements are very small due to the reduced amounts of new information. A possible solution to improve the translations would be the integration of several MT engines, such as Google Translate or Personal Translator, with the purpose of reducing the number of unknown words while still giving preference to in-domain translations. This technique, applied through instance weighting, yielded significant improvements of 2 BLEU points over an in-domain baseline trained on an earlier version of the Text+Berg corpus (Sennrich, 2013).

The third research question aimed to identify ways of measuring the similarity between two candidate sentences. As mentioned before, the chosen alignment methods are strongly correlated with the similarity metrics for comparing the candidate sentence pairs. For the monolingual alignment method we used similarity metrics inspired from the automatic evaluation in MT. In the first place, we used the string-based metric BLEU (Papineni et al., 2002) to compare the automatic translation of a source article into the target language with the original article in the target language. Since we were not happy with the results, we defined a customized similarity metric based on a more informed evaluation metric, METEOR (Denkowski and Lavie, 2011), and also on the word alignments. The third method, which employed bilingual alignment, measured the similarity in terms of sentence length, while also considering the a priori probability of the resulting type of alignment, as suggested by Gale and Church (1993). Other similarity criteria proposed in the literature include the percentage of aligned or translated words, the word fertility, the length of the longest contiguous span, the Jaccard similarity.

The fourth research question was concerned with the evaluation of the extracted pieces of text. We first performed a manual intrinsic evaluation to measure if the extracted texts represent translations. Since some of the extracted pairs contained only partial alignments, we made the distinction between strict matches (sentence pairs representing perfect translations) and lax matches (sentence pairs with partial alignments). The rate of strict matches ranged from 43% to 68%, whereas the rate of lax matches ranged from 54% to 93%, depending on the extraction approach. The poorest performance was obtained at sentence level, whereas the best one was achieved on clause level, both applied on Wikipedia-extracted texts. The false positives in the case of the texts extracted from the Common Crawl were mostly due to language identification problems. In the remaining cases, as well as in the case of Wikipedia-extracted texts, the false positives were due to pitfalls of the similarity metric used to compare the bilingual texts.

Given these accuracy figures, we decided to also evaluate the extracted sentence pairs extrinsically, namely in an SMT scenario. This was the object of the last research question. The purpose was to obtain a domain-adapted SMT system able to translate texts from the Alpine domain between German and French. We used the extracted texts (in different amounts) for training several translation and language models, which we combine with models trained on existing bitexts. We demonstrated that the extracted texts, either from Wikipedia or from the Common Crawl, had a significant influence on top of out-of-domain models. However, only small improvements were pinpointed on top of an in-domain system. These results are in direct correlation with the chosen textual domain (Alpinism) and with the texts used to test the system. On the other hand, the automatic evaluation scores are in the same range as the ones reported for the same language pair, but for a different domain (IMS, 2013), which indicates a certain limitation of existing SMT models for the considered language pair.

Since this work dates a few years back, a legitimate question is whether the proposed approaches are relevant for the current state of the art in Machine Translation. This research area underwent in the meantime a paradigm shift: whilst a few years back phrase-based SMT was the leading approach to Machine Translation, during the past two years more and more researchers employed neural networks to approach this goal, developing what is called Neural Machine Translation (NMT). The main reason for this shift is that NMT seems to generalize better the statistical evidence from the texts and also to handle better rich contexts. NMT-based approaches obtained the best performance for the language pairs and the text types tested so far (written texts and spoken dialogues) and have therefore become the state of the art approaches in MT. Moreover, commercial MT vendors such as Google or Systran also deployed their first NMT engines, confirming the improved performance of this new paradigm (Wu et al., 2016, Crego et al., 2016).

In order to train the translation models, NMT, as well as SMT, requires considerable amounts of parallel texts, which are still scarce for many language pairs. However, experiments show that NMT requires less data to learn from, empirically estimated by some MT practitioners to 1/5 of the amount of training data used for SMT[3]. On the other hand, for small collections of training texts, one runs the risk of overfitting. Therefore our approach to extract parallel texts from comparable corpora can also be relevant in the context of NMT. The extracted texts can either be concatenated directly with the existing training texts (in case a collection of in-domain parallel texts already exists) or be used for fine-tuning in the end of the training process (in case of lacking in-domain texts).

On the other hand, noisy texts can seriously influence the NMT models (Chen et al., 2016), in contrast to phrase-based models, where random noise seldom modifies the most probable translation of a phrase. Since the data extracted with our extraction approaches becomes noisy when the selection threshold drops, we should use a high threshold in order to ensure the quality of the extracted data. It is therefore preferable to extract less, but high quality data. Such data can be used for domain-adaptation in NMT by retraining the models with the additional in-domain data (Servan et al., 2016).

In addition to the extracted parallel data, we can also exploit the in-domain texts selected from our comparable corpora which are only available in the target language. For example, we can use them to generate synthetic parallel data or for fine-tuning whilst leaving the neural network architecture unchanged, as Sennrich et al. (2016) suggest. Another possibility is to train a separate neural language model on the monolingual data and then integrate it in the NMT architecture, similar to Gülçehre et al. (2015).

Returning to the approaches proposed in this thesis, we believe that they can have a significant impact when applied to other domains or genres and/or language pairs. The developed methods are domain and language independent, but expect a list of domain-specific keywords for the in-domain selection and a small bitexts collection for training a basic MT system or at least a bilingual dictionary. In case no bilingual resources exist, a convenient way to generate them is by crowdsourcing, i.e. asking native speakers to contribute translations. Other possibilities include the automatic dictionary induction from comparable corpora or corpora generation by means of a pivot language. Further approaches to deploy SMT systems for low-resource languages are described in (Irvine, 2014). These methods could be used conjointly with our adaptation methods described in Chapter 6.

Moreover, we believe that the extracted data is relevant in the context of SMT, especially when no or little parallel, domain-specific data is previously available. This has been demonstrated in the experiments with out-of-domain data. We think that the translation performance would furthermore improve for texts using a standardized language, yet including domain-specific terms. For example, user manuals which use a

[3] http://kv-emptypages.blogspot.ch/2016/09/comparing-neural-mt-smt-and-rbmt.html

small variety of phrases should be easier to translate due to their repetitive language. Domain-specific terms occurring in these texts would be then easily translated by means of bilingual terminologies, which can be extracted from comparable corpora with little effort (Erdmann et al., 2008). If we leave aside the domain restriction, we can obtain even more parallel texts, which can be explored in various scenarios. We envision that the SMT performance on general domain or speech translation tasks would also improve compared to the figures reported in this thesis, similar to the results in (Smith et al., 2010).

A possible extension of the current approaches is to drop out the alignment based on fixed base segmentation (Tiedemann, 2011), such as paragraphs, sentences or clauses. Instead, we could split the texts into sequences of n-grams of variable length and try to align them. In this way we could avoid the cases in which one alignment candidate contains words which have no equivalent in the segment it has been aligned with, yet the rest represents mutual translations. Since this segmentation would considerably increase the search space, we should also implement some search space restrictions in order to reduce the computational complexity.

It is conceivable that a supervised framework would alleviate the problem of finding the threshold which would minimize the number of false alignment pairs. The question that arises is which functions are relevant for identifying parallel fragments in comparable corpora. In the literature, various features have been proposed to solve this task, from length-based ones to lexico-syntactic ones. We think that the similarity criteria proposed in the current thesis also represent reliable features for a classifier, therefore we would try to combine as many features as possible in order to take the best out of each of them. In this way we might obtain a robust automatic distinction between aligned and not aligned pieces of text.

We think that comparable corpora will continue to play an important role in future linguistic research. Statistical machine translation, addressed in this thesis, is by far the biggest beneficiary of bilingual texts extracted from comparable corpora. Neural MT has so far only marginally made use of comparable corpora (Karakanta et al., 2017), but further experiments are to be expected. Lexicography is another field where the terminologies extracted from comparable corpora will play an important role, as they reduce the manual efforts. Other application scenarios include language teaching and practice, as well as cross-linguistic experiments. Recently, Grave et al. (2018) extracted a sizable collection of word vectors from Wikipedia and the Common Crawl, which can be exploited in various NLP applications.

Parallel sentence mining from comparable corpora will continue to be a popular research avenue, first because corpora like the Web are continuously growing and will facilitate the extraction of similar, if not parallel texts for a wide range of language pairs. And secondly because the Web content will become more structured (e.g. in terms of metadata), thus the automatic processing of such resources will become easier. It is very likely that the additional content will still cover predominantly rich resource languages, but we expect moderate grows for the other languages as well.

A suggestive example in this sense is the development of Wikipedia. As of February 2017, the most representative Wikipedia versions are written in rich resource languages such as English, German, French or Spanish, but also in barely known languages such as Cebuano or Waray (languages of the Philippines). The growth of the latter two Wikipedia versions is due to automatically generated texts written by Lsjbot, an Internet bot[4]. According to the bot author, the automatically generated articles come from monolingual texts in the respective language.

It is certainly conceivable that such articles could have been automatically translated from another language. Since most machine translation systems are far worse than the human translation performance, using such texts for parallel sentence mining will probably result in erroneous translations. Such results might cause more harm than good to any application in which they would be employed later on. Particularly in the case of SMT, it could happen that the existing errors propagate to the output. Therefore an important research topic for future extraction approaches should be the identification of automatically translated content.

[4] https://en.wikipedia.org/wiki/Lsjbot

Bibliography

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

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a Treebank for French. *Building and Using Parsed Corpora*, Text, Speech and Language Technology(20):165–187.

Adafre, S. F. and de Rijke, M. (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.

Afantenos, S., Denis, P., Muller, P., and Danlos, L. (2010). Learning Recursive Segments for Discourse Parsing.

Axelrod, A., Xiaodong, H., and Jianfeng, G. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Bernardini, S. and Ferraresi, A. (2013). *Old Needs, New Solutions: Comparable Corpora for Language Professionals*, pages 303–319. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Carreras, X. and Màrquez, L. (2001). Boosting trees for clause splitting. In *Proceedings of the 2001 Workshop on Computational Natural Language Learning - Volume 7*, pages 26:1–26:3, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chen, B., Kuhn, R., Foster, G., Cherry, C., and Huang, F. (2016). Bilingual methods for adaptive training data selection for machine translation. In *Proceedings of the The Twelfth Conference of The Association for Machine Translation in the Americas: Volume 1*, AMTA 2016, pages 93–106.

Cheng, Y., Liu, Y., Yang, Q., Sun, M., and Xu, W. (2017). Neural machine translation with pivot languages. Retrieved from `https://arxiv.org/abs/1611.04928v2`.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D.-C., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., and Zoldan, P. (2016). SYSTRAN's Pure Neural Machine Translation Systems.

de Groc, C. (2011). Babouk: Focused web crawling for corpus compilation and automatic terminology extraction. In Boissier, O., Benatallah, B., Papazoglou, M. P., Ras, Z. W., and Hacid, M.-S., editors, *Web Intelligence*, pages 497–498. IEEE Computer Society.

Delpech, E. M. (2014). *Leveraging Comparable Corpora for Computer-assisted Translation*, pages 1–39. John Wiley & Sons, Inc.

Denkowski, M. and Lavie, A. (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.*

Erdmann, M., Nakayama, K., Hara, T., and Nishio, S. (2008). *An Approach for Extracting Bilingual Terminology from Wikipedia*, pages 380–392. Springer Berlin Heidelberg, Berlin, Heidelberg.

Esplà-Gomis, M. and Forcada, M. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation.*

Fraser, A. M., Weller, M., Cahill, A., and Cap, F. (2012). Modeling inflection and word-formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674.

Fung, P. and Cheung, P. (2004). Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

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

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19:75–102.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages.

Gülçehre, C. et al. (2015). On using monolingual corpora in neural machine translation. Retrieved from https://arxiv.org/abs/1503.03535.

Hardmeier, C. and Volk, M. (2009). Using linguistic annotations in statistical machine translation of film subtitles. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NoDaLiDa).*

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

IMS (2013). Evaluation of the impact of TTC on statistical MT. Technical report, IMS.

Irvine, A. (2014). *Using Comparable Corpora to Augment Statistical Machine Translation Models in Low Resource Settings.* PhD thesis, Johns Hopkins University.

Jehl, L., Hieber, F., and Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 410–421, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karakanta, A., Dehdari, J., and van Genabith, J. (2017). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation.*

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Lüngen, H., Puskás, C., Bärenfänger, M., Hilbert, M., and Lobin, H. (2006). *Discourse Segmentation of German Written Texts*, pages 245–256. Springer Berlin Heidelberg, Berlin, Heidelberg.

Matsoukas, S., Rosti, A.-V. I., and Zhang, B. (2009). Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 708–717, Stroudsburg, PA, USA. Association for Computational Linguistics.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Morin, E., Daille, B., and Prochasson, E. (2013). *Bilingual Terminology Mining from Language for Special Purposes Comparable Corpora*, pages 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Pecina, P., Toral, A., Papavassiliou, V., Prokopidis, P., Tamchyna, A., Way, A., and van Genabith, J. (2015). Domain adaptation of statistical machine translation with domain-focused web crawling. *Language Resources and Evaluation*, 49(1):147–193.

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

Plamada, M. and Volk, M. (2013). Mining for domain-specific parallel text from Wikipedia. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 112–120, Sofia, Bulgaria. Association for Computational Linguistics.

Puşcaşu, G. (2004). A multilingual method for clause splitting. In *Proceedings of the 7th annual colloquium for the UK Special interest group for computational linguistics (CLUK 2004)*.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

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

Riezler, S. and Maxwell, J. T. (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. Association for Computational Linguistics.

Sennrich, R. (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.

Sennrich, R. (2013). *Domain adaptation for translation models in statistical machine translation*. PhD thesis, University of Zurich.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Sennrich, R. and Volk, M. (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)*.

Servan, C., Crego, J., and Senellart, J. (2016). Domain specialization: a post-training domain adaptation for neural machine translation. Retrieved from https://arxiv.org/abs/1612.06141.

Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P., editors (2013). *Building and Using Comparable Corpora*. Springer-Verlag Berlin Heidelberg.

Smith, J., Quirk, C., and Toutanova, K. (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.

Smith, J., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt Cheap Web-Scale Parallel Text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia.

Ştefănescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation EAMT 2012*, pages 137–144.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 257–286.

Tiedemann, J. (2011). *Bitext Alignment.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).*

Tillmann, C. and Xu, J. (2009). A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 93–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tufiş, D., Ion, R., Ştefan Dumitrescu, and Ştefănescu, D. (2014). Large SMT data-sets extracted from Wikipedia. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Uszkoreit, J., Ponte, J. M., Popat, A. C., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.

Volk, M. (2001). *The Automatic Resolution of Prepositional Phrase - Attachment Ambiguities in German*. Habilitation thesis, University of Zurich.

Wu, D. and Fung, P. (2005). Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*, pages 257–268. Springer Berlin Heidelberg.

Wu, Y. et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. Retrieved from https://arxiv.org/abs/1609.08144.

Zhao, B. and Vogel, S. (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.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora (BUCC)*.