Belgisch Staatsblad Corpus: Retrieving French-Dutch Sentences from Official Documents

Tom Vanallemeersch

Lessius University College, Antwerpen (Belgium)
Centre for Computational Linguistics, K.U.Leuven (Belgium)
tallem@ccl.kuleuven.be

Abstract

We describe the compilation of a large corpus of French-Dutch sentence pairs from official Belgian documents which are available in the online version of the publication *Belgisch Staatsblad/Moniteur belge*, and which have been published between 1997 and 2006. After downloading files in batch, we filtered out documents which have no translation in the other language, documents which contain several languages (by checking on discriminating words), and pairs of documents with a substantial difference in length. We segmented the documents into sentences and aligned the latter, which resulted in 5 million sentence pairs (only one-to-one links were included in the parallel corpus); there are 2.4 million unique pairs. Sample-based evaluation of the sentence alignment results indicates a near 100% accuracy, which can be explained by the text genre, the procedure filtering out weakly parallel articles and the restriction to one-to-one links. The corpus is larger than a number of well-known French-Dutch resources. It is made available to the community. Further investigation is needed in order to determine the original language in which documents were written.

1. Introduction

The Belgian authorities daily disclose a number of articles with official texts, such as laws, decrees etc., through a publication called the *Belgisch Staatsblad* in Dutch and *Moniteur belge* in French. It appears on paper and, since a number of years, online also[1]. The official languages of Belgium are Dutch, French and German. As the latter is the native language of less than 1 percent of the population, the publication contains mainly articles in French and Dutch, and relatively few in German[2]. Some articles are a translation of another article.

The online version of the Belgisch Staatsblad is targeted towards legal and other specialists looking for specific articles. It provides a search interface, allowing them to enter keywords, a range of dates, the language of the articles, etc. The online version is also interesting for translators, but for their purposes (e.g. finding out the possible Dutch equivalents of a French term), the search interface is inefficient, as articles need to be consulted one by one and no button is provided for switching to the equivalent article in another language. The online data are also potentially interesting for building a statistical machine translation system, creating a bilingual lexicon, performing translation studies etc. Therefore, we have built a French-Dutch parallel corpus from these data. We focused on these two languages because of their strong representation within the whole set of articles.

In the following sections, we present the procedure for obtaining and filtering online articles, describe the sentence alignment procedure, compare the corpus with other resources, and discuss the format in which it is made available. Finally, we present conclusions and future research. We have rounded some of the article, sentence and word counts for the sake of readability, using *K* as an abbreviation for thousands and *m* for millions.

2. Obtaining and Filtering Documents

We downloaded a large number of articles, in a similar fashion as the web crawling procedure which lead to the Europarl corpus (Koehn, 2005). As far as the intellectual property of the online version of the Belgisch Staatsblad is concerned, it is legally stated that the electronic files can be used freely, for personal or commercial use[3]. We created a list of URLs to be downloaded in batch by a web crawler[4]. By consulting some websites specialized in legal matter, we found out the form of a URL that directly leads to the summary of all articles which appeared during one day. Such a URL contains keywords whose values indicate language and date (year, month and day). We generated automatically a list of all possible URLs for a period of 10 years (1997 until 2006), for both languages, 1997 being the first year for which a substantial amount of summaries were digitally available.

In the summaries which we downloaded in batch using the automatically generated list, each article is tagged with a so-called *numac*, a unique code starting with a year. Web sites on legal matter provided information on the form of a URL that leads directly to a specific article. Based on the numacs in the daily summaries, we created a list of URLs containing keywords whose values indicate the numac, the date of the summary and the language[5]. By downloading those URLs in batch, we obtained a total of 199K articles. The whole download process took us several days. We converted the articles, downloaded as HTML files, into pure text using a utility[6], configuring it in such a way that para-

[1] `http://www.ejustice.just.fgov.be/cgi/welcome.pl` (last consultation: 16 March 2010)

[2] Figures of November 2009: 38% of the available articles written in French, 60% in Dutch and 2% in German.

[3] Article 477 of the law (*programmawet*) of 24 December 2002.

[4] `http://www.gnu.org/software/wget`

[5] Example of keywords and values: numac = 2006011348, article_lang = N, pub_date = 2006-09-04.

[6] `http://www.nirsoft.net`

graphs were stored as a singe line rather than a set of lines. We filtered out a number of downloaded articles. We applied the following cascade of filters (illustrated by Figure 1):

- We filtered out articles available in only one language (7K in French, 11K in Dutch), based on the fact that corresponding documents in French and Dutch have their numac in common.

- We filtered out pairs of articles with a substantial difference in length. Such difference is caused, for instance, by the fact that an article focuses on a language-specific political entity such as the *Communauté française* and provides the other language group with a less detailed translation. These article pairs could present difficulties during sentence alignment. To this purpose, we randomly selected 50 parallel articles, and verified whether the articles were completely monolingual (see next filter) and completely parallel to each other. On average, French articles appeared to be 5% shorter than their Dutch counterpart (in terms of characters, after removing redundant spaces); the biggest differences involved a French article which was 13% shorter and one which was 5% longer than its Dutch counterpart. We decided to filter out parallel pairs in which the French article is more than 20% shorter or longer than its counterpart. This resulted in a reduction by 599 parallel articles.

- We filtered out parallel pairs in which less than 90% of the French article consists of French text (e.g. a mix of French and German in texts concerning the German-speaking part of Belgium), or less than 90% of the Dutch article consists of Dutch text. For each of three languages (French, Dutch and German), we created a list of discriminating words, i.e. words that are unique to a language compared to the other two (such as certain function words). For each article pair, we estimated the portion written in French by comparing the number of occurrences of French discriminating words with the number of occurrences of any discriminating word, be it a French, Dutch or German one. Similarly, we estimated the Dutch and the German portion. We preferred this approach over standard language identification techniques (Padró and Padró, 2004), as the latter primarily deal with fully monolingual files. By setting a threshold of 90%, we didn't filter out articles with a sporadic text fragment in another language (e.g. a reference to a book). This resulted in a reduction by 6K parallel pairs, leaving us with 85K pairs.

3. Sentence Alignment

We wrote a script that converts the running text in the articles into a list of sentences (although, more accurately, we should talk about segments, as not all independent text units are sentences). The script disambiguates periods, for instance by recognizing abbreviations. As the average article size in terms of words is rather low for both languages (939 for French, 919 for Dutch), we didn't undertake paragraph alignment as a preparatory step. We performed sentence alignment using the GMA (Geometric Mapping and Alignment) system of Melamed (2000). This system applies two steps, SIMR and GSA. The SIMR (Smooth Injective Map Recognizer) algorithm creates a list of *anchors*. These are potential points of correspondence, which link identical words, *cognates* (similar orthography) or words that are equivalent according to a bilingual lexicon. The algorithm also uses stop word lists (e.g. function words) in order to avoid linking such words and causing a proliferation of anchors. The GSA (Geometric Segment Alignment) postprocessor links one or more source sentences to one or more target sentences by grouping anchors. We restricted language-specific knowledge to stop word lists, as an extensive bilingual lexicon was not at hand.

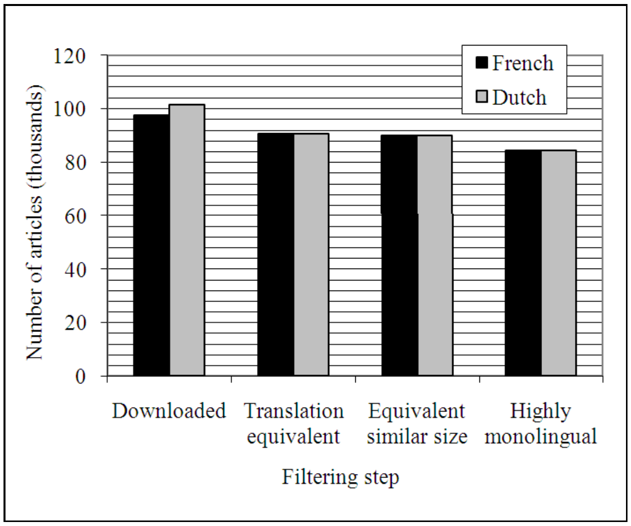

Figure 1: Reduction of number of articles through filtering

The alignment resulted in a parallel corpus with a total of 5m one-to-one links. We ignored links that involve more than one sentence in at least one language (one-to-many or many-to-many links) and *null* links (sentences without equivalent), assuming that they may be the product of a lack of alignment evidence.

We estimated the quality of the sentence alignment results by evaluating a small sample of aligned articles of different sizes, a sample of aligned portions of the two largest article pairs in the corpus, and a random sample of 500 sentence pairs taken from the whole corpus. It turned out that the alignment quality was almost perfect. Apart from a serious alignment problem caused by a glossary at the end of the largest article pair (the alphabetic order of the items in each language disturbs the positional correspondence of translation equivalents), we found only one completely incorrect link between two sentences, as well as a sporadic link that was partially incorrect due to segmentation errors caused by a colon inside brackets or an unrecognized abbreviation. The high alignment quality can be explained by the following factors: we had previously filtered out article pairs that are potentially hard to align, we restricted ourselves to one-to-one links, legally oriented translations are accurate rather than creative, and corresponding articles contain many identical words (proper names, dates, section numbers etc.).

4. Comparison with Other Resources

As a basis for comparing the size of the parallel corpus with that of other resources, we looked for the degree of repetition among sentences. Both the French and Dutch sentences contain on average 14 words. Among the 5m sentence pairs, there are 2.4m unique pairs, which contain a total of 52.3m French words and 52.6m Dutch words, and 22 words per sentence on average. This difference in average number of words indicates that especially shorter sentences are often repeated in the corpus. We also simplified the list of unique sentence pairs by removing the pairs in which one or both of the sentences consist of non-letters only (e.g. a date), by replacing non-letter sequences in the other sentences with a space, and by lowercasing the letters in those sentences. This lead to a total of 2.0m unique simplified sentence pairs, indicating a substantial amount of repetition caused by differences in punctuation and case (e.g. repetition among “Vroedvrouw.”, “vroedvrouw”, “Vroedvrouw :” etc.). Figure 2 shows the relation between number of words and number of sentences (averaged over both languages) before alignment, after alignment, after removing non-unique sentence pairs and after simplifying the list of unique sentence pairs.

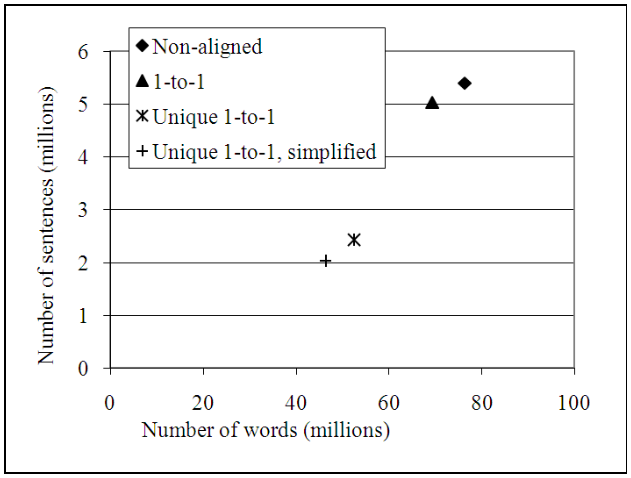

Figure 2: Relation words/sentences according to degree of corpus reduction

Even when we merely count unique sentence pairs in the corpus, its size is larger than a number of existing resources for the French-Dutch language pair. For instance, both the French-English subset and Dutch-English subset of the Europarl corpus contain 1.3m sentence pairs and around 40m words per language. The JRC-Acquis corpus (Steinberger et al., 2006) is based on the Acquis Communautaire (body of common rights and obligations binding all the Member States together within the EU) and was produced using two alignment tools which linked paragraphs that “can contain a small number of sentences, but they sometimes contain sentence parts (ending with a semicolon or a comma)” (p. 2144). The French-Dutch subset of the JRC-Acquis corpus contains 1.3m paragraph links, 35m French words and 33m Dutch words. The recently released Dutch Parallel Corpus (Rura et al., 2008) contains a total of 10m words. Its purpose is different from the above corpora, as it is a balanced corpus for two language pairs (Dutch-English and Dutch-French) and different text types, and contains linguistic annotations. The sentence alignment was performed by three tools, among which GMA; the results of the tools were merged.

As for the determination of the original language of a sentence pair, which is important for instance when we want to study translation effects (Johansson, 2007), the relevant information is coded in the Dutch Parallel Corpus but not or insufficiently in the other corpora mentioned. In Europarl, the tag indicating the language used by the speaker is not consistently coded on all speeches (van Halteren, 2008). In JRC-Acquis, the original language is not indicated at all. In case of our corpus, three alternatives apply, i.e. the original text was written in French, the original text was written in Dutch, or some parts of the original text were written in French and some in Dutch (source: personal communication with a translator working for the Belgian authorities). However, the Belgisch Staatsblad doesn’t indicate which of the three alternatives applied for a specific article pair. It may be worth investigating the approach by van Halteren (2008), who trained a classifier on Europarl speeches known to contain original sentences, in order to predict the source language of other speeches.

5. Availability

The corpus is made available to the community in the following formats[7]:

- Downloaded articles in HTML; their file names contain the date of publication, numac and language.
- One-to-one-links in TMX format, an open standard for exchanging translation memories[8]; metadata: each sentence pair is associated with a date of publication and a numac, and each sentence is associated with a language.
- Pairs of files containing the French and Dutch sentences that were aligned; the file names contain the date of publication, the numac and the language; a sentence in a French file has the same line number as its translation equivalent in the Dutch file.

6. Conclusions and Future Research

We have created a French-Dutch bilingual corpus containing legislative information, whose size (5m one-to-one sentence links, 2.4m unique sentence pairs) is larger than that of well-known existing resources for the language pair in question. Articles containing text in multiple languages were excluded from alignment by checking on words that are unique to a language compared to two other languages. Sample-based evaluation of the sentence alignment results indicates a near 100% accuracy, which can be explained by

[7] http://www.ccl.kuleuven.be/~tallem
[8] http://www.lisa.org/tmx

the text genre, the procedure for filtering out weakly parallel article pairs, and the restriction to one-to-one links. The corpus is made available for the community.

The fact that the original language of the articles is currently not known requires further investigation of the data in order to make the corpus apt for studying translation effects. Other future research on our Belgisch Staatsblad corpus will involve the construction of a statistical machine translation system, the extraction of a bilingual lexicon and term candidates, and word alignment based on a bilingual lexicon and word fragments (Vanallemeersch and Wermuth, 2008).

7. References

S. Johansson. 2007. *Seeing through Multilingual Corpora*. Studies in Corpus Linguistics 26. John Benjamins, Philadelphia.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit*, pages 79–86.

D. Melamed. 2000. Pattern recognition for mapping bitext correspondence. In J. Véronis, editor, *Parallel text processing: Alignment and use of translation corpora*, pages 25–47.

M. Padró and L. Padró. 2004. Comparing methods for language identification. *Procesamiento del Lenguaje Natural*, (33):155–162.

L. Rura, W. Vandeweghe, and M. Montero Perez. 2008. Designing a parallel corpus as a multifunctional translator's aid. In *Proceedings of XVIII FIT World Congress*. FIT, Translators Association of China.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 2142–2147.

H. van Halteren. 2008. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ, USA. Association for Computational Linguistics.

T. Vanallemeersch and C. Wermuth. 2008. Linguistics-based word alignment for medical translators. *Journal of Specialized Translation (Jostrans)*, (9).