English-French Document Alignment Based on Keywords and Statistical Translation

Marek Medveď **Miloš Jakubíček** **Vojtech Kovář**
Lexical Computing CZ s.r.o.
&
Centre of Natural Language Processing, Faculty of Informatics, Masaryk University,
Botanická 68a 602 00 Brno
firstname.lastname@sketchengine.co.uk

Abstract

In this paper we present our approach to the Bilingual Document Alignment Task (WMT16), where the main goal was to reach the best recall on extracting aligned pages within the provided data.

Our approach consists of tree main parts: data preprocessing, keyword extraction and text pairs scoring based on keyword matching.

For text preprocessing we use the TreeTagger pipeline that contains the Unitok tool (Michelfeit et al., 2014) for tokenization and the TreeTagger morphological analyzer (Schmid, 1994).

After keywords extraction from the texts according TF-IDF scoring our system searches for comparable English-French pairs. Using a statistical dictionary created from a large English-French parallel corpus, the system is able to find comaparable documents.

At the end this procedure is combined with the baseline algorithm and best one-to-one pairing is selected. The result reaches 91.6% recall on provided training data.

After a deep error analysis (see section 5) the recall reached 97.4%.

1 Introduction

In this paper we describe our approach to solve the Bilingual Document Alignment Task (WMT16). It consists of tree main parts: data preprocessing, keyword extraction and text pairs scoring based on keyword matching.

According to these steps, the text is divided into three main sections. Section 2 describes the data preprocessing that was crucial for key-word extraction. In the next section we describe the keyword extraction process, and Section 4 describes scoring of comparable English-French pairs.

The final results on the training data are summarized in Section 5 where we also discuss errors of our system and problematic features of the provided data.

2 Preprocessing

The training and testing data were provided in the `.lett` format. Each `.lett` file consists of lines where each line contains these six parts:

- Language ID (e.g. “en”)
- Mime type (always “text/html”)
- Encoding (always “charset=utf-8”)
- URL
- HTML in Base64 encoding
- Text in Base64 encoding

We pick up language id, URL and text as a input for our system. To obtain keywords for each text, our system converts plain text into a so-called vertical text, or word-per-line format. This format contains each word on a separate line together with morphological information, namely lemma (base form of the word) and morphological tag. For text tokenization we use the Unitok tool (Michelfeit et al., 2014) that splits sentences into tokens according to a predefined grammar. Unitok has a special grammar model for each language that was created using information extracted from large corpora. An example of Unitok output is the first column of Figure 1. The Unitok output is enhanced by a sentence boundaries recognizer (we use `<s>` and `</s>` for marking sentence boundaries).

After tokenization and sentence boundary detection, lemmatization and morphological analysis follows. For both we use TreeTagger

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 728–732,
Berlin, Germany, August 11-12, 2016.

(Schmid, 1994) with language dependent models (i.e. French model for French texts, English for English texts). Figure 1 contains an example of a morphologically analyzed sentence in the vertical format.

Unitok and TreeTagger, together with sentence boundary detection and few other small pre- and post-processing scripts, form the TreeTagger pipeline that is used in the Sketch Engine (Kilgarriff, 2014) corpus query and management system.

```
word        tag     lemma
<s>
A           DT      a
web         NN      web
page        NN      page
is          VBZ     be
a           DT      a
web         NN      web
document    NN      document
<g/>
.           SENT    .
</s>
```

Figure 1: TreeTagger morphological analysis

3 Keyword Extraction

In the previous section, we described the text preprocessing needed for the next part of our system, the keyword extraction.

The lemma (base form) information from the morphological analysis was used for computing "keyness", or specificity scores for each word in the text. For this, we used three different variants of the standard TF-IDF score (Equation 1, 2, 3)[1] and a Simple math score[2] (Kilgarriff, 2009) used in keywords extraction in Sketch Engine (Equation 4):

$$key_t = 1 * log\left(\frac{N}{n_t}\right) \quad (1)$$

$$key_t = (1 + log(f_{t,d})) * log\left(\frac{N}{n_t}\right) \quad (2)$$

$$key_t = \left(\frac{f_{t,d}}{f_d}\right) * log\left(\frac{N}{n_t}\right) \quad (3)$$

[1]The difference between Equations 1,2 and 3 is in TF weight score.

[2]Variant of statistic that choose keywords according rule: 'word W is N times as frequent in document/corpus X vs document/corpus Y'.

$$key_t = \left(\frac{fpm_{t,d} + 1}{fpm_{t,ref} + 1}\right) \quad (4)$$

Legend:

- N: number of documents in corpus
- n_t: number of documents containing a particular word (token) *t*
- $f_{t,d}$: frequency of token *t* in document *d*
- f_d: size (length) of document *d*
- $fpm_{t,d}$: frequency per million of token *t* in document *d*
- $fpm_{t,ref}$: frequency per million of token *t* in a reference corpus (large, representative sample of general language)

As reference corpora, the TenTen web corpora in Sketch Engine for English and French were used (Jakubíček et al., 2013), in particular enTenTen 2013 and frTenTen 2012.

Sometimes the TF-IDF scoring can score some of the most common words (like "the", "a", ...) very high. These so-called stop-words do not have any value when finding match between two texts, as practically all of the texts will contain them. Therefore, we created stop-word lists for English and French (from enTenTen and frTenTen corpus) that filter out these most frequent words so they are never considered keywords.

As we will see, the Equation 3 gives the best results on the training data, therefore we chose it for the final evaluation.

4 Scoring

After obtaining the keyword list from each text, the final step was to find matches between English and French texts.

We used top 100 keywords from each text (this number was estimated during the experiments). Then we consulted a statistical dictionary which contains 10 most probable French translations for each English lemma (see below for more information about this dictionary).

We translated the English keywords into all of their French variants, and intersected this list of translations with the keyword lists etracted from all of the French documents. The French document with the biggest intersection was selected as the best candidate.

This procedure was combined with the baseline algorithm[3] based on finding language identification in the URLs of the documents – firstly, the baseline was applied, then (if no matching document was found) the matching by keywords was performed.

The data processing flow is on Figure 2.

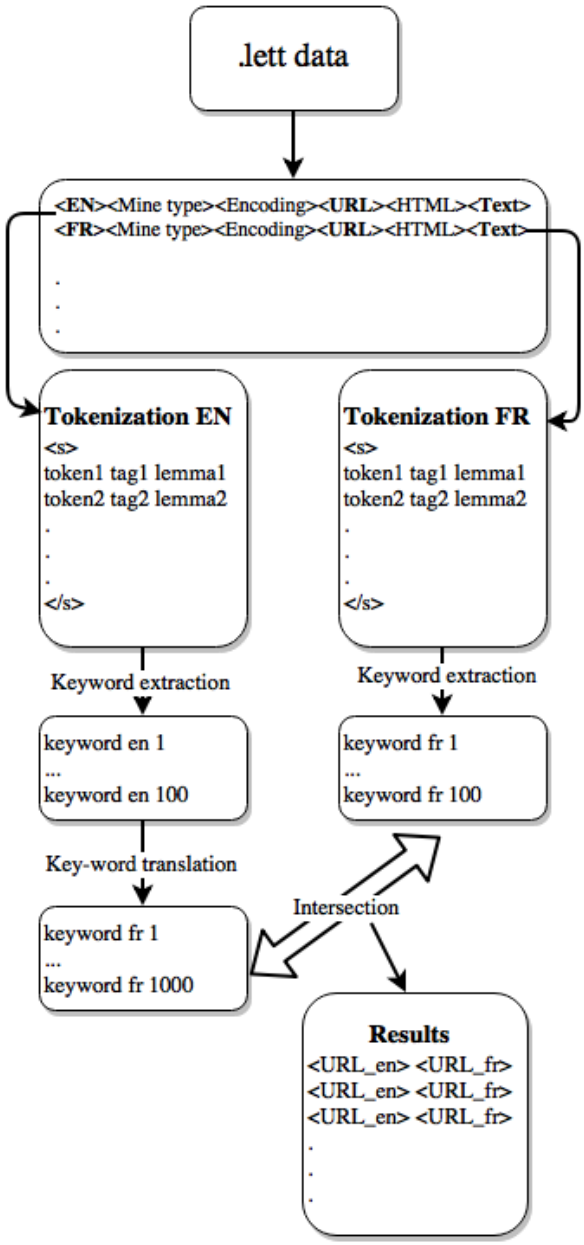

Figure 2: System data flow

4.1 Statistical translation dictionary

Sentence alignment in some of the available parallel corpora enables us to compute various statistics over the number of aligned pairs, and to quantify the probability (or other metric) that word X translates to word Y, for each pair of words in the corpus. The procedure is similar to training a translation model in statistical machine translation (Och and Ney, 2003). Our implementation uses the logDice association score (Rychlý, 2008) which is the same measure that is used in scoring collocational strength in word sketches, the key feature of the Sketch Engine system. It depends on

- frequency of co-occurrence of the two words (e.g. "chat" and "cat") – the higher this frequency, the higher the resulting score; co-occurrence here means that the words occured in a pair of aligned sentences
- standalone frequencies of the two words – the higher these frequencies, the lower the resulting score

By computing these scores for all word pairs across the corpus, we are able to list the strongest "translation candidates" for each word, according to the score; for our purposes, we store 10 best candidates.

The procedure is computationally demanding – quadratic to the number of types (different words) in the corpus – and we exploit an algorithm for computing bi-grams to make it feasible even for very large corpora.

The statistical dictionary for this task was extracted from the English-French Europarl 7 corpus (Koehn, 2005).

5 Evaluation

The goal of this task was to find English-French URL pairs. Some training pairs were provided by authors of this task. Our procedure does not include any learning from the training data, therefore we can use them for quite a reliable evaluation. With regard to that data, our solution reached 91.6% recall, using the most successful TF-IDF equation 3; the results for the other equations are comparable and are summarized in Table 1.

If we did not include the baseline algorithm into the procedure, the recall was 82%.

After a detailed error analysis we found out that the provided data **contain duplicate web pages with different URLs**. This is an important problem – our error analysis shows that we have found

[3]The baseline algorithm iterates through all URLs and search for language identifiers inside URLs and then produces pairs of URLs that have the same language identifiers.

Expected	http://cineuropa.mobi/interview.aspx?lang=en&documentID=65143
	http://cineuropa.mobi/interview.aspx?lang=fr&documentID=65143
Found	http://cineuropa.mobi/interview.aspx?documentID=65143
	http://cineuropa.mobi/interview.aspx?lang=fr&documentID=65143
Expected	http://creationwiki.org/Noah%27s_ark
	http://creationwiki.org/fr/Arche_de_No%C3%A9
Found	http://creationwiki.org/Noah%27s_Ark
	http://creationwiki.org/fr/Arche_de_No%C3y%A9
Expected	http://pawpeds.com/pawacademy/health/pkd/
	http://pawpeds.com/pawacademy/health/pkd/index_fr.html
Found	http://pawpeds.com/pawacademy/health/pkd/index.html
	http://pawpeds.com/pawacademy/health/pkd/index_fr.html

Figure 3: Examples of false errors

Equation	Recall in %
1	89.2
2	89.5
3	**91.6**
4	88.7
Baseline	67.92

Table 1: Overall results according to "keyness" Equations

a correct document pair in many cases, but a document with a different URL (and identical text) was marked as correct in the data.

We went through the document pairs marked as errors of our algorithm and manually evaluated them for correctness. If we exclude the false errors (correct document pairs evaluated as incorrect), the recall is 97.4%. Some examples of these URL pairs are given in Figure 3 – as we can see, in many cases the duplicity is clear directly from the URL.

Unfortunately, we were unable to assess the number of duplicates in the data by the submission deadline. However, we believe it will be done, as the mentioned duplicates significantly reduce the soundness of such evaluation.

6 Conclusion

We have described a method for finding English-French web pages that are translations of each other. The method is based on statistical extraction of keywords and comparing them, using a translation dictionary. The results are promising, but detailed error analysis shows there are significant problems in the testing data, namely unmarked duplicate texts with different URLs.

Acknowledgments

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

References

Jan Michelfeit, Jan Pomikálek, Vít Suchomel. Text tokenisation using unitok. In: 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU, pp. 71-75, 2014

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the international conference on new methods in language processing, pp. 44-49, 1994.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: ten years on. Lexicography, pp. 7-36, 2014

Adam Kilgarriff. Simple maths for keywords. In Proceedings of Corpus Linguistics Conference CL2009, Mahlberg, M., González-Díaz, V. & Smith, C. (eds.), University of Liverpool, UK, 2009.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, Vít Suchomel. The TenTen corpus family. The 7th International Corpus Linguistics Conference, Lancaster, 2013.

Franz Josef Och, Hermann Ney. A systematic comparison of various statistical alignment models, Computational Linguistics, volume 29, number 1, pp. 19-51, 2003.

Pavel Rychlý. A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN, pp. 6-9, 2008.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation, MT Summit, 2005.