Using Term Position Similarity and Language Modeling for Bilingual Document Alignment

Thanh C. Le, Hoa Trong Vu, Jonathan Oberländer, Ondřej Bojar
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{thanhlct, hoavutrongvn, jonathan.oberlaender}@gmail.com
bojar@ufal.mff.cuni.cz

Abstract

The WMT Bilingual Document Alignment Task requires systems to assign source pages to their "translations", in a big space of possible pairs. We present four methods: The first one uses the term position similarity between candidate document pairs. The second method requires automatically translated versions of the target text, and matches them with the candidates. The third and fourth methods try to overcome some of the challenges presented by the nature of the corpus, by considering the string similarity of source URL and candidate URL, and combining the first two approaches.

1 Introduction

Parallel data play an essential role in training of statistical machine translation (MT) systems. While big collections have been already created, e.g. the corpus OPUS (Tiedemann, 2012), the World Wide Web remains a largely underexploited source. That is the motivation for the shared task "Bilingual Document Alignment" of the ACL 2016 workshop First Conference on Machine Translation (WMT16) which requires participants to align web page in one language to their translation counterparts in another language.

Given a large collection of documents, the first step in extracting parallel data is to organize the documents into heaps by the language they are written in. For two languages of interest, a brute-force approach would consider all pairs of documents from the two heaps. Since the number of possible pairings is too high, it is necessary to employ some broad and fast heuristics to filter out the obviously wrong pairs.

Some approaches to the task rely on document metadata (e.g. the similarity of document URLs or language tags within URLs), some emphasize more the actual content of the documents. Previous work (Rapp, 1999; Ma and Liberman, 1999) focused on document alignment by counting word co-occurrences between source and target documents in a fixed-size window. More recently, methods from cross-lingual information retrieval (CLIR) have been used (Snover et al., 2008; Abdul Rauf and Schwenk, 2011), ranking lists of target documents given a source document by a probabilistic model. Locality sensitive hashing has also been applied (Krstovski and Smith, 2011).

In this paper, we describe our attempt. The rest of the paper is organized as follows: In Section 2, we describe the methods we used in our four submitted systems. Section 3 describes our experimental setup and compares the results of the proposed methods. We conclude the paper and discuss possible future improvements in Section 4.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 710–716,
Berlin, Germany, August 11-12, 2016.

2 Methods

We submitted four different systems: UFAL-1 uses term position similarity (especially rare terms) between documents. UFAL-2 uses language modeling on automatically translated documents to perform the matching. UFAL-3 reorders the results of UFAL-2 to take into account the similarity in the URL structure, and UFAL-4 combines UFAL-3 and UFAL-1 to further improve the results.

2.1 Term position similarity (UFAL-1)

Two similar languages such as English and French can easily share a portion of their lexicons, especially proper names, some acronyms and numbers are likely to keep their forms after translation. If two documents are mutual translations, the sequence of positions of those terms should be correlated. Much past research (Ma and Liberman, 1999; Rapp, 1999) has exploited these features, using a fixed-size window and counting the co-occurrences in this range. This method, however, requires considerable tuning of parameters, and if two shared terms are located outside of the window, no credit will be added. In this work, we consider similarity which not only takes into account co-occurrences of terms but also their positions. This metric also assumes that co-occurrences of rare terms are more important than those of common terms. Experiments below show that our method performs much better than the fixed-window method.

Our term position similarity is defined as follows:

$$\rho(S,T) = \sum_{t \in S \cap T} \log(1 + \frac{\max(c)}{c_t}) \cdot \cdot \sum_{i}^{N_t} \frac{l_S - |p^i_{S_t} - p^i_{T_t}|}{l_S} \quad (1)$$

Here S, T are the source and target documents, respectively, $S \cap T$ is the set containing all terms which occurs in both documents, $N_t = \min(|S_t|, |T_t|)$ where S_t, T_t is the number of occurrences of term t in the respective document. The length of the source document is denoted l_S. $p^i_{S_t}$ is the position of i-th occurrence of the term t in the source document and similarly for the target document ($p^i_{T_t}$). Finally, c_t is the total number occurrences of t in the data set and $\max(c)$ is the total number of occurrences of the most frequent term in all the source documents. In sum, $\log(1 + \frac{\max(c)}{c_t})$ is a weight to promote the importance of rare terms and the inner sum $\sum_i^{N_t}$ measures the relative displacement of the term w in S compared to T.

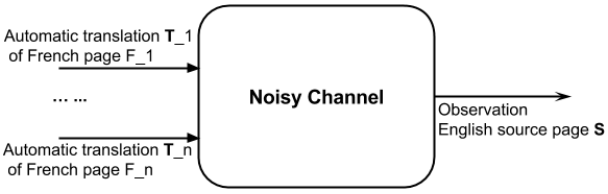

Figure 1: The noisy channel model for Bilingual Document Alignment

To increase the number of terms contributing to the metric result, we employ a bilingual dictionary and translate all words from target document that do not appear in the source into their most frequent translation.

The submission using this method is called **UFAL-1**.

2.2 Language model-based approach (UFAL-2)

In contrast to the method in Section 2.1, the approach labeled UFAL-2 relies on automatic translation from one side to the other (either source-to-target or vice versa). With documents on both sides converted to one language, we then treat the task as a noisy channel problem, similarly to many works of information retrieval based on language modelling techniques (Ponte and Croft, 1998; Zhai and Lafferty, 2001; Xu et al., 2001).

Specifically, we assume that the observed output is the source page S , damaged by noisy transfer of some target page T. Through decoding, we want to find the target page T that most likely lead to the observed output S. The process is visualized in Figure 1. Therefore, like in the noisy channel model (Brill and Moore, 2000), to decode the input T, we estimate the probability of T given the output observation S, $P(T|S)$. Following Bayes' rule, the problem is characterized by Equation 2:

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)} \quad (2)$$

(At this stage, it is no longer important, that T was the automatic translation of a French page into English and S was the original English source page.)

As our final aim is to find the best T that causes the output S, we can ignore the denominator $P(S)$ in Equation 2, since it is the same for every value of T. So we have the problem equation as follows:

$$T_{best} = \underset{T}{\operatorname{argmax}} \underbrace{P(S|T)}_{generative\ model} \overbrace{P(T)}^{prior} \quad (3)$$

Since estimating the generative model $P(S|T)$ in Equation 3 is intractable, we assume conditional independence of terms $t_i, t_j \in S$ given T:

$$P(S|T) = P(t_1, ..., t_{|S|}|T) \approx \prod_{i=1}^{|S|} P(t_i|T) \quad (4)$$

To slightly speed up the computation in Equation 4, we can group all occurrences of the same term together as in Equation 5. To avoid an underflow problem, we move the computation to log space, see Equation 6:

$$P(S|T) \approx \prod_{distinct\ t \in S} P(t|T)^{tf_S} \quad (5)$$

$$\log(P(S|T)) \approx \sum_{distinct\ t \in S} tf_S \log(P(t|T)) \quad (6)$$

where tf_S is the number of occurrences of the term t in S. The remaining problem is to estimate $P(t|T)$. Fortunately, this can be achieved simply using maximum likelihood estimation (Scholz, 1985) and it turns out to be the unigram language model (LM) as follows:

$$P(t|T) = \frac{tf_T}{|T|} \quad (7)$$

where tf_T is the number of occurrences of the term t in T. In order to avoid zero probabilities, a smoothing technique is necessary. We used Jelinek-Mercer smoothing (Jelinek, 1980). The estimation at document level in Equation 7 is smoothed with the estimation over the domain level, $P(t|D)$, where D is the set of all page translations available for webdomain D of page T. We additionally use add-one smoothing for $P(t|D)$ to make sure the model handles well also terms never seen in the webdomain data.

Back to prior in the problem equation (Equation 3), it may be used to integrate very useful information for each target French page. For example, a French page that has been selected to be a pair with another page should have a lower prior for the next prediction. The prior may also reflect the difference in length of T and S, avoiding the alignment of pages differing too much. Here, for simplicity, we use uniform distribution as the prior. The final equation ranking target French pages T with respect to a given English source document S is thus:

$$T_{best} = \underset{T}{\operatorname{argmax}} \sum_{distinct\ t \in S} tf_S \log(\lambda P(t|T) + (1-\lambda)P(t|D)) \quad (8)$$

where $P(t|D)$, as mentioned, is the probability of the term t occurring in the webdomain D and the parameter λ of Jelinek-Mercer smoothing is set to 0.5. We submit this method for evaluation under the label **UFAL-2**.

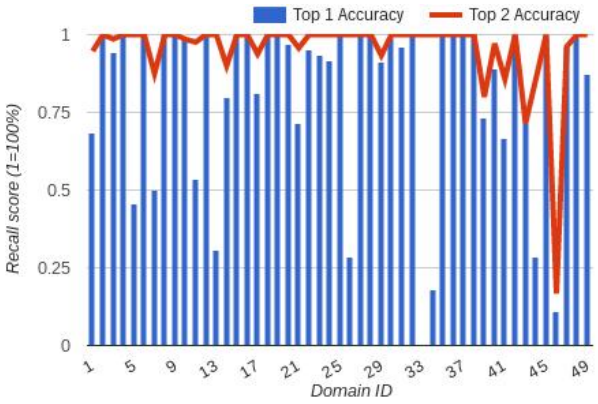

Figure 2: Performance of UFAL-2 on individual webdomains in the training set

2.3 Optimizing for top-1 evaluation (UFAL-3)

We noticed that there were many cases where several documents contained the same (or almost the same) text, which therefore get scored (roughly) the same by each of UFAL-1 and UFAL-2. This issue will create noise that can harm us in the evaluation of the shared task, as can be seen in Figure 2: There is a significant difference between the top 1 and top 2 accuracy of our UFAL-2 system from Section 2.2, see e.g. the webdomains 5, 7, 13 (`kusu.com`), or 34 (`www.eu2007.de`). While both the 1st best and the 2nd best top predictions could be assumed correct since the two predicted pages are not distinguishable or only differ in unimportant details (e.g. Google Ads), the offi-

cial scoring will be based on a single-best answer.[1]

A closer investigation reveals that the URLs that are marked correct in the training data are usually the ones most similar to the source URL. We therefore look at the top 10 candidates from the UFAL-2, and choose the candidate that is within some threshold of the top result and closest in Levenshtein distance from the source URL. The threshold value of **85** was obtained experimentally on the training data. The result after this refinement is submitted for the evaluation under the name **UFAL-3**.

2.4 Combining UFAL-3 and UFAL-1 into UFAL-4

We now have the outputs of two methods, UFAL-1 and UFAL-3 (as a replacement of UFAL-2), and we would like to combine them to one method. Since the result of UFAL-3 is very good (see Section 3.2), we decided to report UFAL-3 in most cases and resort to UFAL-1 only if we do not trust the proposal of UFAL-3.

To estimate the certainty of UFAL-3 prediction, we use Kullback-Leibler divergence (Kullback and Leibler, 1951) and measure how mismatching the predicted pair of documents is. To do so, we model the English source text and translation of the predicted candidate as multinomial distributions, and then compute the KL-divergence to see what their distance is. In particular, a higher KL-score presents a bigger distance between the pairs, in other words, they are less likely to be a correct pair.

Given the overall good performance of UFAL-3, there are not many negative examples to optimize the threshold for rejecting the predicted pair. We solve the issue by artificially creating new negative cases: we remove automatic translations of correct target French pages for two webdomains, rerun the predictions and then compute the KL-divergence for all predicted pairs. The result of 1624 pairs predicted is reported in Figure 3, in which the artificial negative examples are highlighted with a blue line.

Based on observations for the modified training data, we set the threshold to **0.35**. If the KL divergence for a pair of documents predicted by UFAL-3 exceeds this value, the pair is considered a wrong prediction. In that case, we use the method from Section 2.1 (UFAL-1) with the bilingual dictionary size of 5000 entries. Similar to the method from Section 2.3 (UFAL-3), we consider the top 2 candidates and choose the one with a lower Levenshtein distance. We call this combined method **UFAL-4** in the evaluation.

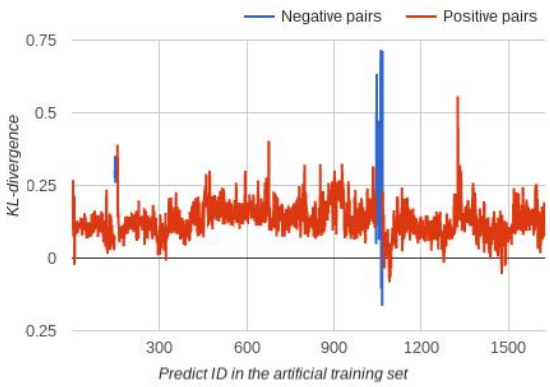

Figure 3: KL-divergence for all 1624 predicted pairs in the modified training set where two correct translations are removed.

3 Experiments

3.1 Experimental setup

We used the data published with the Shared Task on Bilingual Document Alignment (WMT 2016), containing roughly 4200 million pairs, in which 1624 pairs have been labeled as mutual translations to serve as a development set.

Work on information extraction typically uses precision and recall of the extracted information as an evaluation measure. However, in this task, manually classifying all possible pairs is impossible, so the true recall cannot be established. The organizers thus decided to evaluate the methods on the recall *within the fixed set of document pairs*, the development set released prior submission deadline and the official test set disclosed only with the final results.

While the official scores are top-1 recall (i.e. the recall taking the single-best prediction for each input sentence), we also evaluate our systems at top 2 and top 5 outputs because, as discussed in Section 2.3, the there are many documents with the same content, but the development set of pairs mentions only one of them.

All documents are tokenized by splitting on white-space and passed to a filter which prunes all pairs having a ratio of the lengths in tokens of two

[1] We were told by the organizers later that the test set does not suffer from this problem of many very similar pages.

Dictionary size	Systems		
	Baseline	Fixed window	Term position
0	67.92	78.94	88.30
1000	67.92	80.6	88.36
5000	67.92	81.9	89.53 (UFAL-1)
10000	67.92	85.71	91.63
25000	67.92	88.73	94.27
50000	67.92	90.76	**96.06**

Table 1: Recall measures by baseline system, system using fixed-size window method and system using term position similarity

Method	Recall		
	Top 1	Top 2	Top 5
Baseline	67.92		
UFAL-1	89.53		
UFAL-2	88.40	97.40	98.30
UFAL-3	93.70		
UFAL-4	94.70		

Table 2: Result on the development set

documents bigger than 2. Afterwards, all documents are ranked by the discussed methods. The first 1, 2 or 5 ranked documents with score higher than a threshold are reported.

In the first experiment, we prepare three systems for comparison. We use the provided baseline system in the mentioned shared task which simply finds matching URLs by discarding language identifiers, such as en, fr. We also implement a fixed-size window method as described in Ma and Liberman (1999). We compare the fixed-size window method with our term position similarity in 6 tests with increasing size of the underlying bilingual dictionary. This dictionary is obtained by running IBM Model 2 implemented by Dyer et al. (2013) on the translations of the data set provided by the organizers. We extract the 50000 most frequent word alignments $fr - en$ having $P(en \mid fr) > 0.7$ and then randomly draw a subset of this dictionary for each test. The variant with 5000 entries is our submission called UFAL-1. If two documents have an identical score, the one having a shorter URL is preferred.

In the second experiment, we compare the term position similarity method (UFAL-1) with the language model-based approach (UFAL-2 and UFAL-3) and the combination method (UFAL-4). The term position similarity method uses a bilingual dictionary containing 5000 entries. Automatic translations for all target documents were provided by the organizers who used a baseline Moses setup trained on Europarl and the News Commentary corpus.

3.2 Experiment result

The results for first experiment are in Table 1. From these results, we can clearly see that term position similarity outperforms the fixed-size window method and surpasses the baseline system with around 20% even without a bilingual dictionary. By increasing size of the bilingual dictionary up to 50000 entries, we can boost up the term position similarity method by 8% to **96.06%**. However, there are still a number of avenues for improvement. First, as we found that our method encountered many errors on the webdomain `www.luontoportti.com` that contains extremely specialized words not covered by our dictionary, this makes a domain-based bilingual dictionary one of the most desirable potential improvements. Secondly, the term position similarity method is very sensitive to the case when a target document contains source language text, because it increases the co-occurrence rate between two documents. Any errors in language identification can thus adversely affect the final extracted parallel corpus.

We present the results of the second experiment in Table 2. The improved methods UFAL-3 and UFAL-4 show significant gains, achieving 93.7% and 94.7% in recall. We also clearly see the remarkable changes in recall for the top match vs. top two matches caused by the similar documents in the corpus, as discussed in Section 2.3.

Finally, we report the official scores in Table 3. The official test set consists of 2402 document pairs and methods are evaluated in terms of the percentage of these pairs that they reported ("Recall"). The shared task winner NovaLincs-url-coverage (denoted "NovaLics" in the table for short) reached 94.96%, our best method UFAL-4 ranked about in the middle of the methods with the recall of 84.22%. As we see in the remaining columns, UFAL-4 produces by far the highest number of document pairs (more that 1M). The official scoring script filters this list and keeps only the pairs where neither the source URL nor the target URL was previously reported ("After 1-1").

Method	Official Recall [%]	Pairs Reported	Pairs After 1-1	Lenient Recall
NovaLincs	94.96	235812	235812	?
UFAL-4	84.22	1080962	268105	92.67
UFAL-1	81.31	592337	248344	87.89
UFAL-3	80.68	574434	207358	89.97
UFAL-2	79.14	574433	178038	88.43

Table 3: The winner and our methods on the official test set.

After this style of deduplication, the number of pairs reduces to about 268k, slightly higher than the number of pairs reported by the winner.

The official test set results are in line with our observation on the development set: term position similarity (UFAL-1) performs well (although not as well as on the development set) and the two variations of the noisy-channel approach are slightly worse, with UFAL-3 (URL similarity) better than UFAL-2. The combination (UFAL-4) is the best of our methods.

We note that for systems like ours that produce all URL pairs they deem good enough, the 1-1 deduplication may be too strict. We thus also report a lenient form of the recall: whenever a pair of URLs from the test set appears (as an unordered pair) among the pairs produced by our method, we give a credit for it. As seen in the last column of Table 3, the noisy-channel methods seem better than term position similarity in this measure. Considering that UFAL-2 and UFAL-3 produced slightly fewer pairs than UFAL-1, it may seem that they are more precise. This however need not be the case; the set of pairs produced by the systems is again too large for manual validation so the true precision cannot be evaluated.

4 Conclusion and future work

In this paper, we presented four systems for the Bilingual Document Alignment shared task. These system all perform well on the provided development set (roughly 90% accuracy for top 1 recall) as well as on the official test set (above 80%; about in the middle of all the participating methods). One system, UFAL-1, uses term position similarity. The second system, UFAL-2, uses a probabilistic model inspired by language modelling and the noisy channel model. Two others systems, UFAL-3 and 4, are improvements of the two former ones, where UFAL-3 tries to overcome the fact that content is repeated in a web-based corpus and UFAL-4 is a more advanced combination of UFAL-3 and 1.

Several refinements of the proposed approaches are worth further investigation. In particular, a systematic method of creating a bilingual dictionary dedicated for each specific webdomain should increase the accuracy of the term position similarity method. For the language model approach, it might be valuable to use a more comprehensive generative model (e.g. bi/tri-gram language model). Adding a prior might also enhance model accuracy. Another potential for the LM-based approach is, instead of depending on translations of target pages, to apply a bilingual dictionary or a translation model directly for the generative process.

The method of UFAL-3 still misses some of the straightforward cases of URL mapping. For instance, it might be advisable to use a more specific variant of edit distance variant, e.g. to penalize changes in special characters like "/" or "?" compared to normal word characters.

Beyond our submissions to the shared task, we suggest that more attention should be paid to the evaluation method. The problem of repeated or very similar content on the web is omnipresent, so any attempt to handle it is likely to improve the reliability of top-1 recall measurements, improving the bilingual alignment task itself.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 644402 (HimL).

Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

Sadaf Abdul Rauf and Holger Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation* 25(4):341–375.

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational*

Linguistics. Association for Computational Linguistics, pages 286–293.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. *Pattern recognition in practice* .

Kriste Krstovski and David A. Smith. 2011. A minimally supervised approach for detecting and ranking document translation pairs. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Stroudsburg, PA, USA, WMT '11, pages 207–216.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1):79–86.

Xiaoyi Ma and Mark Liberman. 1999. Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*. Citeseer, pages 538–542.

Jay M Ponte and W Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 275–281.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, pages 519–526.

FW Scholz. 1985. Maximum likelihood estimation. *Encyclopedia of Statistical Sciences* .

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 857–866.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 105–110.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 334–342.