The ADAPT Bilingual Document Alignment system at WMT16

Pintu Lohar, Haithem Afli, Chao-Hong Liu and Andy Way
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{FirstName.LastName}@adaptcentre.ie

Abstract

Comparable corpora have been shown to be useful in several multilingual natural language processing (NLP) tasks. Many previous papers have focused on how to improve the extraction of parallel data from this kind of corpus on different levels. In this paper, we are interested in improving the quality of bilingual comparable corpora according to increased document alignment score. We describe our participation in the bilingual document alignment shared task of the First Conference on Machine Translation (WMT16). We propose a technique based on source-to-target sentence- and word-based scores and the fraction of matched source named entities. We performed our experiments on English-to-French document alignments for this bilingual task.

1 Introduction

Parallel corpora (or "bitexts"), comprising bilingual/multilingual texts extracted from parallel documents, are crucial resources for building SMT systems. Unfortunately, parallel documents are a scarce resource for many language pairs with the exception of English, French, Spanish, Arabic, Chinese and some European languages included in *Europarl*[1] (Koehn, 2005) and OPUS (Tiedemann, 2012).[2] Furthermore, these existing available corpora do not cover some special domains or subdomains.

For the field of SMT, this can be problematic, because MT systems trained on data from a specific domain (e.g. parliamentary proceedings) perform poorly when applied to other domains, e.g. sports news articles. As a result, the area of domain adaptation has been a hot topic in MT over the past few years.

One way to overcome this lack of data is to exploit comparable corpora which are much more easily available (Munteanu and Marcu, 2005). A comparable corpus is a collection of texts composed independently in their respective languages and combined on the basis of similarity of content. These are bilingual/multilingual documents that are comparable in content and form to various degrees and dimensions. Potential sources of textual comparable corpora are the output from multilingual news organizations such as Agence France Presse (AFP), Xinhua, Reuters, CNN, BBC, etc. These texts are widely available on the Web for many language pairs (Resnik and Smith, 2003). Another example is *Euronews*, which proposes news text in several languages clustered by domain (e.g. sports, finance, etc.). The degree of parallelism can vary considerably, from noisy parallel texts, to 'quasi parallel' texts (Fung and Cheung, 2004).

No matter what data we are dealing with, if we want to automatically create large amounts of parallel documents for SMT training, the ability to detect parallel sentences or sub-sentences contained in these kinds of comparable corpus is crucial. However, for some specific domains, such as news, the problem of document alignment can drastically reduce the quantity of the final parallel data extracted. For example, Afli et al. (2012) showed that they were able to extract only 20% of an expected 1.9M-token parallel sentence collection using their automatic parallel data extraction method. For this reason, they tried to improve this method by exploiting parallel *phrases* (i.e. not just parallel *sentences*) which increased the quantity of extracted data (Afli et al., 2013, 2016).

However, the precision of such automatic meth-

[1] http://www.statmt.org/europarl/
[2] http://opus.lingfil.uu.se/

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 717–723,
Berlin, Germany, August 11-12, 2016.

ods is still much less than expected. We contend that the main problem comes from the document alignment of such comparable corpora. One of the challenges of our research is to build data and techniques for some under-resourced domains. We propose to investigate the improvement of alignment of bilingual comparable documents in order to solve this problem.

Accordingly, in this paper we describe an experimental framework designed to address a situation when we have large quantities of non-aligned parallel or comparable documents in different languages that we need to exploit. Our document alignment methods are based on a new scoring technique for parallel document detection based on the word-length and sentence-length ratio and named entity recognition (NER).

Apart from this, we also compared the total number of source and target named entities (NEs) so that they should not differ significantly which can play a major role in determining the comparability of two texts.

The remainder of the paper is structured as follows. The related work on parallel data extraction and comparability measures is briefly discussed in Section 2. In Section 3, we detail our proposed method and provide the results of our experiments on *WMT-2016* data in Section 4. In Section 5, we present the conclusion and directions for future work.

2 Related work

In the "Big Data" world that we now live in, it is widely believed that *there is no better data than more data* (e.g. Mayer-Schönberger and Cukier (2013)). In line with this idea, a considerable amount of work has taken place in the NLP community on discovering parallel sentences/fragments in a comparable corpus in order to augment existing parallel data collections. However, the extensive literature related to the problem of exploiting comparable corpora takes a somewhat different perspective than we do in this paper.

Typically, comparable corpora do not have any information regarding document-pair similarity. They are made of many documents in one language which do not have any corresponding translated document in the other language. Furthermore, when the documents are paired, they are not literal translations of each other. Thus, extracting parallel data from such corpora requires special algorithms. Many papers use the Web as a comparable corpus. An adaptive approach, proposed by Zhao and Vogel (2002), aims at mining parallel sentences from a bilingual comparable news collection collected from the Web. A maximum likelihood criterion was used by combining sentence-length models with lexicon-based models. The translation lexicon is iteratively updated using the mined parallel data to obtain better vocabulary coverage and translation probability estimation. Resnik and Smith (2003) propose a web-mining-based system called STRAND and show that their approach is able to find large numbers of similar document pairs. Yang and Li (2003) present an alignment method at different levels (title, word and character) based on dynamic programming (DP). The goal is to identify one-to-one title pairs in an English–Chinese corpus collected from the Web. They apply the longest common sub-sequence to find the most reliable Chinese translation of an English word. One of the main methods relies on cross-lingual information retrieval (CLIR), with different techniques for transferring the request into the target language (using a bilingual dictionary or a full SMT system). Utiyama and Isahara (2003) use CLIR techniques and DP to extract sentences from an English–Japanese comparable corpus. They identify similar article pairs, and having considered them as parallel texts, then align sentences using a sentence-pair similarity score and use DP to find the least-cost alignment over the document pair. Munteanu and Marcu (2005) use a bilingual lexicon to translate some of the words of the source sentence. These translations are then used to query the database to find matching translations using IR techniques. There have been only a few studies trying to investigate the formal quantification of how similar two comparable documents are. Li and Gaussier (2010) presented one of the first works on developing a comparability measure based on the expectation of finding translation word pairs in the corpus. Our approach follows this line of work based on a method developed by Sennrich and Volk (2010).

3 Aligning comparable documents

3.1 Processing the comparable documents

In this work, experiments are conducted on the test data[3] provided by the *WMT-2016* organizers, which comprised 203 web domains with more than 1 million documents in total. The data is provided in *.lett* format with following fields, 1) Language ID, 2) MIME type, 3) Encoding, 4) URL, 5) Complete content in Base64 encoding and 6) Main textual content in Base64 encoding. We extracted URLs and texts from this collection of data and converted them into *UTF-8* format.

3.2 Basic Idea

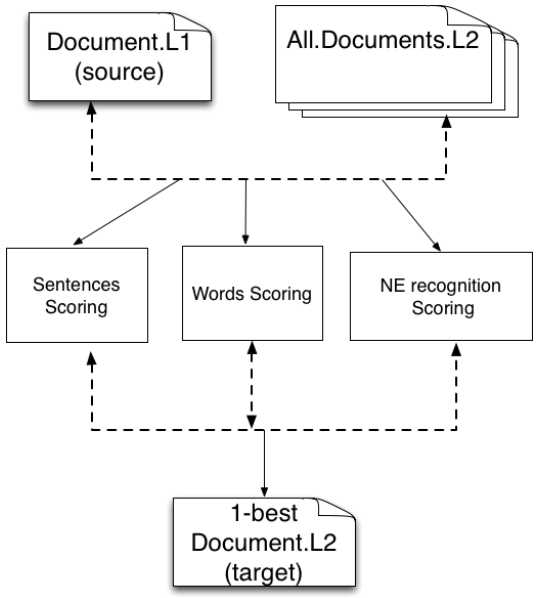

Figure 1: Architecture of comparable alignment system

In this work we propose an extension of the method described in Sennrich and Volk (2010). The basic system architecture is described in Figure 1. We begin by removing those documents that have very little contents in order to avoid all possible comparisons. Subsequently, we introduce three steps: sentence-based scoring, word-based scoring and NE-based scoring. Finally we used a combined weighted score of the three scores to select the target document with highest value.

3.3 Sentence-based scoring

Since there are a large number of source and target documents, there are billions of possible comparisons required to complete the calculations of finding possible document alignments. Therefore, we have to restrict the comparison calculations only to those source-target text pairs that have a close sentence-length ratio, otherwise they are less likely to be comparable texts. This is necessary since comparing each source with each target text would result in an undesirably large number of comparisons and thus a very long time to process all steps even for a single domain. Let us assume that S_s and S_t are the number of sentences in the source and target texts, respectively. We then follow a very simple formula to calculate source-target sentence-length ratio(R_{SL}), as in (1) :

$$R_{SL} = \frac{\min(S_s, S_t)}{\max(S_s, S_t)} \quad (1)$$

We construct this equation in order to confine the value between 0 and 1 which implies that if either of the source or target text contains no sentences, R_{SL} will be 0, and 1 if they have the same number of sentences. Therefore, a value of 1 or even very close to it has a positive indication towards being comparable but this is not the only requirement, as there are many documents with the same (or very similar) number of sentences. For this reason, we consider word and NE-based scoring in Sections 3.4 and 3.5, respectively.

3.4 Word-based scoring

The reason behind this step is very similar to the step discussed in Section 3.3, but here it is based on word-length comparison. Let us assume that W_s and W_t are the number of words in the source and target texts, respectively. Hence our equation for calculating source-target word-length ratio (R_{WL}) is (2):

$$R_{WL} = \frac{\min(W_s, W_t)}{\max(W_s, W_t)} \quad (2)$$

3.5 NE-based scoring

Having studies the comparable documents from a linguistic point of view, it appeared that looking for NEs present in both source and target texts might be a good way to select the 1-best target document. We extracted NEs from all the documents to be compared. Let us assume that the number of NEs in a source text and in a target text are NE_S and NE_T, respectively. Initially we calculate source-target NE-length ratio (R_{NL}) as in (3):

[3] http://www.statmt.org/wmt16/bilingual-task.html

$$R_{NL} = \frac{\min(NE_S, NE_T)}{\max(NE_S, NE_T)} \quad (3)$$

Then we calculated the ratio of the total number of source-target NE matches to the total number of source NEs, which we call R_{SNM}. Let us assume that the total number of NEs matched is M_{NE}. Considering this, R_{SNM} can be calculated as shown in (4):

$$R_{SNM} = \frac{M_{NE}}{NE_S} \quad (4)$$

In many cases a text-pair in a comparison can have a huge difference between the number of NEs present in both documents. For example, if NE_S and NE_T are 5 and 50, respectively, and all of the source NEs match the target NEs, we might not necessarily want to link them. Accordingly, therefore, (3) is also taken into account, and we multiply R_{SNM} by $R_{NL}$to give our overall NE-based score(SC_{NE}) in (5) :

$$SC_{NE} = R_{SNM} * R_{NL} \quad (5)$$

3.6 Combining all scores

We propose to re-rank our possible alignments based on adding sentence-, word- and NE-based scores and call this our alignment-score (SC_A), as in (6) :

$$SC_A = R_{SL} + R_{WL} + SC_{NE} \quad (6)$$

Using equation (6), we calculate scores for each possible document pair and retain the 1-best pair with the maximum value.

4 Experimental results

4.1 Data and systems

In order to test our proposed techniques we conducted experiments on the provided development data and corresponding references. As discussed in Sections 3.4 and 3.5, we selected only those document pairs for comparison that have a sentence-length and word-length ratio of 1 (or very close to it).

It is usually seen that on average a French translation of an English document has 1.2 words for every English word in the original. In this work, since we are dealing with the comparable texts that are usually not proper translations of each other but contain similar information, we choose to set this ratio closer to 1.

In addition to this, we applied different weighted scores for the three features (i.e. sentence-based, word-based and NE-based scoring). The weights applied on the test data were extracted from our experiments on the development data. We held out the documents randomly selected from 10 web-domains in the training data. We assigned different sets of weights to the three features and conducted experiments on the development set using these weighted scores.

The Stanford Named Entity Recognizer[4] was used to detect NEs in our system.

4.2 Results

We assigned weights to the three features in five different combinations (termed as C_n, where n=1, 2 ...5) as shown in Table 1. The summation of these weights is always 1.

Feature	Weight assigned				
	C_1	C_2	C_3	C_4	C_5
R_{SL}	0.33	0.25	0.15	0.1	0
R_{WL}	0.33	0.25	0.15	0.1	0
SC_{NE}	0.33	0.5	0.7	0.8	1

Table 1: Weights assigned to different features with different combinations.

As can be seen in Table 1, C_1 represents the combination where all features are assigned an equal score. Subsequently, the weights of R_{SL} and R_{WL} are decreased but for SC_{NE} it is increased. C_5 indicates that the whole weight is assigned to SC_{NE} whereas R_{SL} and R_{WL} are not taken into account. Let us assume that the weights assigned to the sentence-based, word-based and NE-based features are λ_1, λ_2 and λ_3, respectively. Taking these weights into account, the overall alignment score of a document-pair is calculated as shown in equation (7):

$$SC_A = \lambda_1 R_{SL} + \lambda_2 R_{WL} + \lambda_3 SC_{NE} \quad (7)$$

where, $\lambda_1 + \lambda_2 + \lambda_3 = 1$

The experimental results on the development data with different scoring combinations are given in Table 2.

Table 3 shows the detailed results using C_3 combinations. Prior to tuning the feature weights in the development phase, our published result on

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

Combination of Weights	References	System output	Recall
C_1	247	147	59.51
C_2	247	152	61.53
C_3	247	153	61.94
C_4	247	153	61.94
C_5	247	147	59.51

Table 2: Results of document alignment method used in our experiments.

Web-domain name	Ref.	Sys. o/p	Recall
bugadacargnel.com	19	9	47.36
cbsc.ca	20	12	60.0
cineuropa.mobi	73	58	79.45
creationwiki.org	22	4	18.18
eu2007.de	11	4	36.36
eu.blizzard.com	10	8	80.0
forcesavenir.qc.ca	8	3	37.5
galacticchannelings.com	9	1	11.11
golftrotter.com	8	8	100.0
iiz-dvv.de	67	45	67.16

Table 3: Detailed results of 10 web-domains of the development data.

the test data was based on simple addition of the three features we used. The result is published on the basis of recall and contains 2,402 alignment pairs. We extended the published results with the precision values which is shown in Table 4.

Subsequently, we tuned the feature weights in the development phase and selected the weight combination C_3 to apply on the test data. Table 5 shows the results.

It can be observed from Table 5 that applying the tuned feature weights helps in increasing the recall value by up to 2% compared to our initial results ('ADAPT' in Table 4). The precision value is also slightly increased from 1.05% (in ADAPT-2) to 1.1%. However, in both Table 4 and Table 5, it is obvious that both of our systems produced much lower recall value than the top-ranked systems (e.g. NovaLincs, UEdin1_cosine etc.). In contrast, our precision is quite competitive to these systems and higher than most of the submitted systems.

Another very important observation is that our results on the development data are much better than on the test data. The main reason for this is

System submitted	Rec.	Prec.	Num. found	1-1 pairs
ADAPT	27.10	0.93	651	69,518
ADAPT-2	26.81	1.05	644	61,094
arcpv42	84.92	0.7	2040	287,860
ITRI-DCU	0.49	0.008	12	146,566
DOCAL	88.59	1.1	2128	191,993
Jakubina-Langlais	79.30	0.72	1905	263,133
JIS	1.99	0.16	48	28,903
Meved	79.39	1.22	1907	155,891
NovaLincs	85.76	0.99	2060	207,022
NovaLincs-2	88.63	0.9	2129	235,763
NovaLincs-3	**94.96**	0.96	2281	235,812
UA_bitextor_4.1	31.14	0.78	748	95,760
UA_bitextor_5.0	83.30	**1.26**	2001	157,682
UEdin1_cosine	89.09	0.58	2140	368,260
UEdin2_LSI-v2	87.63	0.57	2105	367,948
UEdin2_LSI	85.84	0.75	2062	271,626
UFAL-1	81.30	0.78	1953	248,344
UFAL-2	79.14	1.06	1901	178,038
UFAL-3	80.68	0.93	1938	207,358
UFAL-4	84.22	0.75	2023	268,105
Yandex	84.13	0.72	2021	277,896
YODA	93.92	0.7	2256	318,568

Table 4: Published results with an extension of precision values.

Combination of weights	Rec.	Prec.	Num. found	1-1 pairs
C_3	29.1	1.1	699	63,255

Table 5: Results obtained after applying tuned feature weights.

that we strictly pruned out many of the possible comparisons for the web-domains in the test set having a large number of texts in order to reduce the runtime of the whole process. It would have consumed a lot of time if we had considered all the documents (i.e. more than one million document pairs). Therefore, we removed those documents that contain only a few lines of text which resulted in discarding many possible alignments. In contrast, we applied a much softer pruning technique on the development data and produced much better recall values than that on the test data.

Finally, analysing the source of the problem of misalignments, we found that in our data we have many articles that deal with similar topics in dif-

ferent documents. Hence it may not always be helpful to rely mostly on NE-matching.

5 Conclusion

Despite the fact that phrase-based models of translation obtain state-of-the-art performance, sufficient amounts of good quality training data do not exist for many language pairs. Even for those language pairs where large amounts of data are available, these do not always occur in the required domain of application. Accordingly, many researchers have investigated the use of comparable corpora either to generate initial training data for SMT engines, or to supplement what data is already available.

In this paper, we seek to improve the quality of the multilingual comparable documents retrieved. In our approach, we actually quantify the amount of correct target-language documents retrieved. Here we propose a technique combining three features. The first one is based on matched source-to-target sentence scoring, the second on matched source-to-target sentence scoring and the third on NE-based scoring.

Analysing this result, in future work we would like to add more semantic features to our system and apply these techniques to other language pairs and data types. In addition to this, we would also like to automatically determine the weighted scores, for instance by using n-fold cross-validation. Our proposed method does not consider the difference between translation ratio of languages as we are dealing with different qualities of comparable corpora in this task, but we plan to investigate this problem with a specific corpus in different languages for our future work.

Acknowledgments

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

References

Afli, H., Barrault, L., and Schwenk, H. (2012). Parallel texts extraction from multimodal comparable corpora. In *JapTAL*, volume 7614 of *Lecture Notes in Computer Science*, pages 40–51, Kanazawa, Japan.

Afli, H., Barrault, L., and Schwenk, H. (2013). Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. In *International Joint Conference on Natural Language Processing*, pages 286–292, Nagoya, Japan.

Afli, H., Barrault, L., and Schwenk, H. (2016). Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22(04):603 – 625.

Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, pages 1051–1057, Geneva, Switzerland.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, volume 5, pages 79–86, Phuket, Thailand.

Li, B. and Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 644–652, Beijing, China.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Houghton Mifflin Harcourt.

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

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

Sennrich, R. and Volk, M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado, USA.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218, Istanbul, Turkey.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the*

41st Annual Meeting on Association for Computational Linguistics - Volume 1, ACL '03, pages 72–79, Sapporo, Japan.

Yang, C. C. and Li, K. W. (2003). Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742.

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.