DOCAL - Vicomtech's Participation in the WMT16 Shared Task on Bilingual Document Alignment

Andoni Azpeitia and **Thierry Etchegoyhen**
Vicomtech-IK4
Mikeletegi Pasalekua, 57
Donostia / San Sebastian, Gipuzkoa, Spain
{aazpeitia, tetchegoyhen}@vicomtech.org

Abstract

This article presents the DOCAL system for document alignment, which took part in the WMT 2016 shared task on bilingual document alignment. The system is meant to offer a portable solution for varied document alignment scenarios, from parallel to comparable corpora, with minimal deployment effort. Its main goal is to provide an optimal balance between alignment precision and recall using minimal resources and adaptation across alignment scenarios. We describe and discuss the performance of the system in the recall-oriented shared task.

1 Introduction

Parallel corpora are essential to the development of data-driven approaches to translation such as statistical machine translation (Brown et al., 1990). As it feeds further processes in the creation of bitexts, multilingual document alignment plays an important role in building accurate resources.

This article presents the DOCAL system for document alignment, which took part in the WMT 2016 shared task on bilingual document alignment. The system is meant to offer a portable solution for varied document alignment scenarios, from parallel to comparable corpora.

The alignment of multilingual documents has been performed with a variety of techniques over the years, with alternatives targeting various scenarios, from parallel to weakly comparable corpora.

Simple approaches based on file name matching can provide fast document pairing, as they do not rely on any analysis of the content of documents. Unfortunately, these approaches rely on consistent file naming conventions, an assumption which is often defeated in practice (Tiedemann, 2011). This approach is thus often complemented with content-based alignment methods, as in (Chen et al., 2004), whose system includes a filename-based module and a semantic similarity component based on a vector space model with frequency-weighted term vectors.

The usefulness of document metadata for document alignment has been explored in depth by (Resnik and Smith, 2003), who exploit URL properties and structural tags to gather bilingual corpora from HTML pages on the Web. (Chen and Nie, 2000) is another example of an approach that exploits URL properties, along with document size and language identifiers. (Munteanu and Marcu, 2005) use date-aligned documents as input for their binary classification approach to comparable sentence alignment.

To address comparable corpora specifically, different types of content-based approaches have been proposed. (Fung and Cheung, 2004), for instance, present the first exploration of very non-parallel corpora using a document similarity measure based on bilingual lexical matching defined over mutual information scores on word pairs. (Patry and Langlais, 2005) present a feature-based method based on an Ada-Boost classifier that includes features such as length, entities, and punctuation, along with a filtering component to remove alignment duplicates. The BITS system is another alternative proposed by (Ma and Liberman, 1999) for bilingual text mining on the Web, measuring content similarity by counting the ratio of token translation pairs over the total number of tokens in the source document, where translation pairs are determined within fixed windows of text.

Other general methods include (Ion et al., 2011), who propose an approach based on expectation-maximization using bilingual lexi-

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 666–671,
Berlin, Germany, August 11-12, 2016.

cons, and (Li and Gaussier, 2013), whose comparability metric measures the overall proportion of words for which a translation can be found in a comparable corpus using bilingual dictionaries.

The Jaccard coefficient (Jaccard, 1901), which is a core component of DOCAL, has been used for instance by (Paramita et al., 2013) whose comparable document similarity measure is partially based on this metric computed over a subset of sentence pairs in the documents.

DOCAL (Etchegoyhen and Azpeitia, 2016) is a simple method to measure multilingual document similarity, whose main goal is to provide an optimal balance between alignment precision and recall with minimal resources and adaptation across alignment scenarios. The next sections describe the system and its performance in the recall-oriented shared task.

2 DOCAL

The core of the DOCAL approach relies on expanded lexical translation sets, defined at the document level, and the Jaccard coefficient computed on those sets. Two token sets are thus extracted from each pair of documents, along with two corresponding sets containing lexical translations of the tokens. The translation sets are then augmented through set expansion operations, described below, and similarity is computed as the ratio of intersection over union on the original token sets and their corresponding translation sets.

Formally, the following components are generated for each document pair:

- d_i and d_j: tokenised documents in languages l_1 and l_2, respectively.
- S_i: set of tokens in d_i.
- S_j: set of tokens in d_j.
- T_{ij}: set of expanded lexical translations into l_2 for all tokens in S_i.
- T_{ji}: set of expanded lexical translations into l_1 for all tokens in S_j.

From these elements, the similarity score is computed as in Equation 1:

$$sim_{docal} = \frac{\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}}{2} \quad (1)$$

In other words, the score is defined as the average of the document-level Jaccard similarity coefficients computed in both translation directions.

Lexical translations are extracted from seed parallel corpora, with translation probabilities computed according to IBM models (Brown et al., 1993).[1] For each token, the k-best translation options are selected among the alternatives ranked according to their lexical translation probability. The actual probability values are not used beyond the ranking they enable, i.e. all selected translations are equally considered in the computation of similarity. This is meant to abstract away from differences in lexical distributions between the seed corpora used to create translation tables and the data in the domain at hand, which is often of a different nature.

No filtering is performed on the token sets, leaving punctuation marks alongside functional and content words, and the text is preserved with its original capitalisation. Pre-processing is thus reduced to the minimal operation of tokenisation.

We now describe in turn the aforementioned set expansion operations, the retrieval of alignment candidates, and the available optimisations of the core method.

2.1 Set Expansion

Since lexical translation tables cannot be expected to cover a given domain satisfactorily, the translation sets are expanded with tokens that may be indicators of similarity, although absent from translation tables. First, all capitalised tokens are added to the sets if they are not found in the translation tables.[2] This simple operation, which we perform at set creation time, provides coverage for named entities, which can be viewed as important indicators of content similarity given their low relative frequency. The same process applies to numbers as well, which can also be strong indicators of similarity, in particular when they denote dates.

DOCAL includes an additional set expansion operation based on longest common prefixes (LCP), which are computed over the minimal sets of elements that may have a common stem, defined to be the following two set differences: $T'_{ij} = T_{ij} - S_j$

[1] We use GIZA++ (Och and Ney, 2003) to extract lexical translation tables.

[2] Checking for their presence in lexical translation tables allows one to distinguish between out-of-vocabulary tokens and entities with an existing translation, e.g. *Germany* translated into Spanish *Alemania*.

and $T'_{ji} = T_{ji} - S_i$. For each element in T'_{ij} (respectively T'_{ji}) and each element in S_j (respectively S_i), if a common prefix is found with an empirically set minimal length of n characters, the prefix is added to both sets. This specific expansion operation is not included by default in the actual usage of the system, as it increases the overall computational cost and its benefits are largely dependent on the specifics of the corpora and language pairs at hand.

2.2 Alignment Candidates

Alignments are computed from source to target documents, with the additional filtering described in Section 2.3.

In some document alignment scenarios, an alignment process based on the Cartesian product of the document sets might be the optimal approach, as the alignment space is guaranteed to be searched exhaustively. Since this approach has quadratic complexity, it is however computationally prohibitive if the volumes of documents reach a certain amount.

For scenarios where the volume of documents renders an exhaustive comparison unsustainable, a standard cross-linguistic information retrieval (CLIR) approach is provided. Target documents are first indexed using the Lucene search engine[3] and retrieved by building a query over the expanded translation sets created from each source document. This strategy drastically reduces the overall processing time and resource consumption, at the cost of missing some correct alignment pairs.[4]

2.3 Alignment Filtering

As the alignment process is executed from source to target documents, a given target document can be selected as the best alignment for more than one source document. This results in hidden correct alignments, often with scores that are marginally lower than the top alignment scores assigned by the similarity metric.

A simple solution to this issue consists in removing all alignments between a source document and a target if the latter is aligned to a different source document with a better similarity score. That is, we remove alignment tuples (d_i, d_j, sim_{ij}) between any two documents d_i and d_j if there exists a different tuple (d_k, d_j, sim_{kj}) such that $sim_{kj} > sim_{ij}$.

This process often produces large improvements, as it allows previously hidden correct alignments to surface, and is included by default in DOCAL.

3 WMT 2016 Bilingual Document Alignment Task

The WMT 2016 shared task on multilingual document alignment[5] consists in identifying pairs of English and French documents from a given collection of documents such that one document is the translation of the other. Candidate pairs were defined as all pairs of documents from the same web domain for which the source side has been identified as mostly English and the target side as mostly French.

Participants were to submit a list of possible pairings, with each source URL matched with at most one target URL and vice-versa. The evaluation metric was selected to be recall on the test set, i.e. the percentage of the test-set pairs that a participating system could find after enforcing the 1-1 alignment rule.

Our participation in the shared task was meant to check the effectiveness of DOCAL in a new large-scale document alignment task with no task-specific adaptation, in accordance with our stated aim at portability and ease of deployment across document alignment scenarios. Thus, the system was applied in its default configuration and the provided training datasets were not used beyond testing the processing tools provided for the task. Document metadata or URL properties were not exploited either, to strictly measure our content-based approach to document alignment.

In the next section, we describe the setup for our system, with results presented in Section 3.2.

3.1 System Setup

As mentioned above, DOCAL was applied in its default configuration. Lexical translation tables were created with GIZA++ on the JRC-Acquis Communautaire corpus.[6] For the English-French pair, the training corpus consisted in 708.896 aligned sentences. No experiments were made with different translation tables, larger or more varied, although

[3] https://lucene.apache.org.

[4] In experiments on different datasets, the loss of correct alignment pairs was minimal, at around 1% per test set.

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

[6] We used the latest available version of the corpus, as of November 2015, in the OPUS repository: http://opus.lingfil.uu.se/JRC-Acquis.php.

we view this research path as worth exploring in future work.

We set $k = 5$ to define the range of k-best lexical translations, as a compromise between larger sets with less reliable translation candidates and smaller sets which may miss translation alternatives. Note that this value could have been tuned on the provided training data, thus optimising the setting to this specific task. However, as previously mentioned, our goal was to evaluate the approach with portability in mind, where no particular adaptation is performed; we therefore used this default value for the k parameter.

Document content was tokenised using the scripts provided in the Moses toolkit (Koehn et al., 2007). For all but four web domains in the test set, the set of possible alignment pairs was computed using the Cartesian product of source-target documents, as this guaranteed an exhaustive search in the alignment space and the computation was deemed practical for up to 260 million possible pairings.[7] The remaining four domains featured potential pairs above the 300 million mark and the CLIR approach using Lucene was used in those cases.[8]

Finally, DOCAL was used with alignment filtering, as described in Section 2.3, and without the set expansion operation based on longest common prefixes described in Section 2.1.

3.2 Results and Discussion

Overall, DOCAL ranked in 5th place on the official test set, with 2128 pairs retrieved out of 2402 for a recall score of 88.59%. It is interesting to note that several systems, and in particular all four systems with better scores, have submitted a significantly larger number of pairs than DOCAL, which is indicative of underlying differences in terms of precision and f-measure. However, without knowing the correctness of the alignments outside the test set pairs, it is obviously not possible to determine whether these differences show better precision on the part of DOCAL or not.

While performing an error analysis of the cases where our system had retrieved the incorrect pair according to the test set, we found 100 cases where the test set contained what we consider to be incorrect alignments. That is, in all 100 cases, shown in Table 1,[9] the target pair found by DOCAL seems to us to be the correct one. In most of these cases, the French documents in the test set and the one retrieved by DOCAL were nearly identical, with only minor differences where the test set document was missing a small portion of information from the source document.[10]

These cases account for 4.16% of the test, and impact the final results, as shown in Table 2.[11] On the corrected test set, DOCAL reaches a score of 92.76%, significantly better than its result on the original test set.

It is of course entirely possible that other participating systems had actually retrieved the correct target documents as well in those cases, and that the final ranking of systems would thus be unaffected. Whether this is actually the case or not is unknown to us at the time of this writing.

4 Conclusion

Overall, we found the results obtained by DOCAL on the shared task to be satisfactory, in particular as a test case for the portability of the default method in a new large-scale alignment scenario.

The system was developed to seek an optimal balance between precision and recall, and has shown promising results along these lines in different scenarios involving both parallel and comparable corpora (Etchegoyhen and Azpeitia, 2016). In future tasks, it would be interesting to compare our approach to alternatives in terms of f-measure as well, to fully assess the usefulness of available methods for multilingual document alignment.

Acknowledgments

This work was partially funded by the Spanish Ministry of Economy and Competitiveness and the

[7] The documents were processed on a single server with 64G of RAM and 16 cores.

[8] The domains were: www.domainepechlaurier.com; www.desmarais-robitaille.com; italiasullarete.it; and: egodesign.ca.

[9] As many of the erroneous cases came from a single domain, namely www.lalettrediplomatique.fr, we indicate the URL structure once where replacing the place-holder X with one of the values in the last line forms the actual URL. Note also that we indicate ranges with a dash, e.g., X = 15-17 indicates that all values from 15 to 17 (included) lead to a URL that is in the set of identified errors.

[10] For instance, 94 of the cases came from the domain www.lalettrediplomatique.fr, where the English source document content contains a date which is accurately translated in the document retrieved by DOCAL, and incorrect in the target document in the test set.

[11] *wmt2016_corr* denotes the corrected version of the test set.

Source:	http://artfactories.net/Espace-Linga-Tere.html
Test set:	http://artfactories.net/-Republique-centrafricaine-.html
Correct:	http://artfactories.net/Espace-Linga-Tere-Bangui.html
Source:	http://www.ipu.org/hr-e/169/Co121.htm
Test set:	http://www.ipu.org/hr-f/168/Co121.htm
Correct:	http://www.ipu.org/hr-f/169/Co121.htm
Source:	http://www.lifegrid.fr/en/projets/projects/biomedicale-search.html
Test set:	http://www.lifegrid.fr/fr/projets/appel-a-projets-e-nnovergne-lifegrid-2006/recherche-biomedicale.html
Correct:	http://www.lifegrid.fr/fr/projets/31-recherche-biomedicale.html
Source:	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profils_eng.asp?ID=1008
Test set:	http://www.nserc-crsng.gc.ca/Prizes-Prix/Herzberg-Herzberg/Profiles-Profils_fra.asp?ID=1003
Correct:	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profils_fra.asp?ID=1008
Source:	http://www.rfimusique.com/musiqueen/articles/060/article_6465.asp
Test set:	http://www.rfimusique.com/musiquefr/articles/060/article_14625.asp
Correct:	http://www.rfimusique.com/musiquefr/articles/060/article_13250.asp
Source:	http://www.rfimusique.com/musiqueen/articles/129/article_8397.asp
Test set:	http://www.rfimusique.com/musiquefr/articles/128/article_18057.asp
Correct:	http://www.rfimusique.com/musiquefr/articles/129/article_18094.asp
Source:	http://www.lalettrediplomatique.fr/contribution.php?choixlang=2&id=10&idrub=**X**
Test set:	http://www.lalettrediplomatique.fr/contribution.php?id=10&idrub=**X**
Correct:	http://www.lalettrediplomatique.fr/contribution.php?choixlang=1&id=10&idrub=**X**
X =	5, 7, 11-12, 15-17, 23, 28-31, 35, 37-39, 43, 45-46, 50-52, 56-58, 61-65, 69, 83-84, 86, 89, 91-94, 96-100, 103-107, 109-111, 114-115, 119-120, 123-125, 127-130, 133-135, 137-141, 144, 146, 149-152, 155-156, 158, 160-163, 165-167, 169, 173, 175, 177, 194, 197

Table 1: Identified likely errors in the test set

TEST SETS	FOUND PAIRS	SUBMITTED PAIRS	PAIRS AFTER 1-1 RULE	RECALL
wmt2016	2.128	191.993	191.993	88.592839
wmt2016_corr	2.228	191.993	191.993	92.756037

Table 2: DOCAL results

Department of Economic Development and Competitiveness of the Basque Government through the AdapTA (RTC-2015-3627-7) and TRADIN (IG-2015/0000347) projects. We would like to thank MondragonLingua Translation & Communication for their support as coordinator of these projects, and the organisers of the shared task for their work and support.

References

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Jiang Chen and Jian-Yun Nie. 2000. Parallel Web Text Mining for Cross-language IR. In *Content-Based Multimedia Information Access - Volume 1*, RIAO '00, pages 62–77, Paris, France, France. Centre des hautes tudes internationales d'informatique documentaire.

Jisong Chen, Rowena Chau, and Chung-Hsing Yeh. 2004. Discovering parallel text from the world wide web. In *Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation-Volume 32*, pages 157–161. Australian Computer Society, Inc.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. A Portable Method for Parallel and Comparable Document Alignment. *Baltic Journal of Modern Computing*, 4(2):243–255. *Special Issue: Proceedings of EAMT 2016*.

Pascale Fung and Percy Cheung. 2004. Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and E.M. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 57–63.

Radu Ion, Alexandru Ceauşu, and Elena Irimia. 2011. An expectation maximization algorithm for textual unit alignment. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 128–135. Association for Computational Linguistics.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Bo Li and Eric Gaussier. 2013. Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In *Building and Using Comparable Corpora*, pages 131–149. Springer.

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

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Monica Lestari Paramita, David Guthrie, Evangelos Kanoulas, Rob Gaizauskas, Paul Clough, and Mark Sanderson. 2013. Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Alexandre Patry and Philippe Langlais. 2005. Automatic identification of parallel documents with light or without linguistic resources. In *Proceedings of the 18th Canadian Society Conference on Advances in Artificial Intelligence*, AI'05, pages 354–365, Berlin, Heidelberg. Springer-Verlag.

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

Jörg Tiedemann. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.