Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 243–255

A Portable Method for Parallel and Comparable Document Alignment

Thierry ETCHEGOYHEN, Andoni AZPEITIA

Vicomtech-IK4, Donostia / San Sebastián, Gipuzkoa, Spain

{tetchegoyhen, aazpeitia}@vicomtech.org

Abstract. We present a document alignment method based on expanded lexical translation sets and document-level Jaccard similarity. We compare our approach to state-of-the-art methods on a variety of alignment tasks, showing that it outperforms alternative methods in most scenarios for both parallel and comparable corpora. The proposed method is highly portable, requiring only minimal seed information and no task-specific training, thus providing the means for an efficient exploitation of multilingual documents.

Keywords: Document alignment, Comparable corpora, Parallel corpora

1 Introduction

Multilingual document alignment is an important step in the creation of the parallel resources that are necessary for data-driven approaches to translation such as statistical machine translation (Brown et al. 1990). This part of the overall bitext creation process faces a variety of alignment scenarios. The input might for instance take the form of unordered collections of parallel documents in several languages, from which an alignment needs to be computed to pair those documents that are translations of each other. A second major scenario relates to the exploitation of comparable corpora, where large collections of multilingual documents need to be paired prior to mining parallel or similar sentences. With comparable corpora seen as a reservoir of training material for machine translation (Muntaneu and Marcu 2005), a particular effort needs to be placed on the development of efficient methods for comparable document alignment.

A single method that works efficiently for both parallel and comparable corpora would offer a flexible solution to the general issue of document alignment and reduce adaptation efforts from one alignment scenario to another. In this work, we present such a method, termed DOCAL, which explores the use of minimal information to enhance portability by relying only on automatically extracted lexical translations, expanded token sets and the Jaccard coefficient for the computation of document-level similarity. Components of the approach have been previously explored and used to evaluate document similarity, and we demonstrate the potential of their conjoined use for document

alignment on a variety of alignment tasks and corpora. We compare our approach to state-of-the-art methods for each alignment scenario, showing that DOCAL outperforms alternative methods in the majority of cases.

The paper is organised as follows. Section 2 presents related work in mulitlingual document alignment; Section 3 describes the DOCAL approach; in Section 4 we present controlled experiments with both parallel and comparable documents for various language pairs and domains; finally, Section 5 summarises the results and offers concluding remarks.

2 Multilingual document alignment

The alignment of multilingual documents has been performed with a variety of techniques, depending on their degree of parallelism and comparability.

For strictly parallel document alignment, simple approaches based on file name matching can be the most efficient methods, as they do not rely on any analysis of the content of documents. Unfortunately, this approach relies on uniform and consistent file naming conventions across languages, an assumption which is often defeated in practice, even in professional repositories (Tiedemann 2011). Thus, when filename-based alignment is part of a document alignment pipeline, it is often combined with content-based alignment methods (Chen et al. 2004). The usefulness of document metadata was 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. This approach too has the advantage of not requiring the examination of textual content to retrieve parallel documents, although it is tied to the assumed structural properties of the documents. Another approach based on document properties instead of content is Chen and Nie (2000), who developed the PTMINER system, a cross-language information retrieval system that exploits URL properties as well, along with document size and language identifiers.

Enright and Konrad (2007) describe a simple and fast method for parallel document alignment based on hapax legomena, i.e. aligning documents by counting the number of unique words that appear in both documents. Patry and Langlais (2005) train an Ada-Boost classifier that includes several features such as length, entities, and punctuation, achieving high results on their controlled experiments. Patry and Langlais (2011) describe their PARADOCS system in detail, which includes the following components: an information retrieval module based on hapaxes and numerical entities; a classifier that includes three features based on edit-distance between document representations, number of entities, and optimal edit-distance over the document collection; and a third filtering component with the ability to remove alignment duplicates, i.e. to remove all document pairs where alignments have been established between a given target document and more than one source document.

Chen et al. (2004) developed the Parallel Text Identification system, which includes a filename-based module and a semantic similarity component based on a vector space model with frequency-weighted term vectors. 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.

Several approaches have targeted comparable documents specifically. Muntaneu and Marcu (2005), for instance, proposed a binary classification approach to comparable sentence alignment, using date-aligned documents as input. Fung and Cheung (2004) 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. Uszkoreit et al. (2010) describe a large-scale parallel document mining method that involves translating all source documents into English then using n-gram matching through multiple scoring steps. Ion et al. (2011) describe the EMACC system, which uses an expectation-maximization algorithm to align textual units. Their approach is similar in spirit to the modelling computed at word level by IBM models and they use automatically created bilingual lexicons to apply the EM algorithm on document units. They report state-of-the-art results on a range of alignment scenarios that cover 6 language pairs and varied degrees of comparability between documents.

Li and Gaussier (2013) describe a comparability assessment method that measures the overall proportion of words for which a translation can be found in a comparable corpus using bilingual dictionaries. Their core methodology underlies the CCNUNLP system, which is amongst the approaches evaluated in Section 4. Besides EMACC and CCNUNLP, two other systems are part of the further described controlled experiments. LINA, described in Morin et al. (2015), is based on the hapax method of Enright and Konrad, extended with two strategies: they first use pigeon hole reasoning, where alignments mapping multiple sources to the same target are removed and only the pairs with the highest number of shared words are kept; cross-lingual information specific to Wikipedia is then exploited, breaking remaining alignment ties using the ordering provided by document pairs in a third language. The AUT system, described in Zafarian et al. (2015), employs four main components: a vectorisation module that maps documents to a common feature space, a topic model, a module for named entity detection, and a word feature mapping module based on machine translation.

The Jaccard coefficient (Jaccard 1901), which is a core component of the approach we present, is one of the standard similarity metrics used for text comparison and information retrieval (Manning and Schütze 1999), although cosine-based methods with weighted term vectors are often preferred to determine text similarity. Prochasson and Fung (2011), for instance, use the Jaccard coefficient to evaluate word association for rare word extraction from comparable corpora. Paramita et al. (2013) describe a comparable document similarity measure based on the Jaccard index computed over sentence pairs in the documents, filtering first sentences with large proportions of entities and numbers, and compute document similarity scores as the average of the sentence-based Jaccard similarity scores. In the next section, we describe an approach centred on this coefficient over expanded lexical sets.

3 DOCAL

DOCAL is a simple method to measure multilingual document similarity which aims for portability and ease of deployment. The core of the approach relies on expanded lexical

translation sets, defined at the document level, and the Jaccard coefficient computed on those sets. In other words, we extract token sets from each pair of documents, create two corresponding sets with the lexical translations of the tokens, augment the original sets through two operations of set expansion described below, and compute the ratio of intersection over union on the original token sets and their corresponding translation sets.

More specifically, let d_i and d_j be two tokenised documents in languages l_1 and l_2, respectively, S_i the set of tokens in d_i, S_j the set of tokens in d_j, T_{ij} the set of expanded lexical translations into l_2 for all tokens in S_i, and T_{ji} the 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:

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

That is, the score is defined as the average of the Jaccard similarity coefficients computed in both translation directions.

Lexical translations are extracted from seed parallel corpora, with translation probabilities computed according to the 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 provided ranking, i.e. all selected translations are equally considered in the computation of similarity. The main reason for this is the fact that, in most cases, the lexical translations are extracted from a different domain than the one at hand, and lexical distributions are likely to be different. We thus opted for simple set membership for all selected translation variants and used a default value for k.[2]

The Jaccard coefficient presents properties that are of interest for document alignment. As it results in a real value between 0 and 1, it allows for a bounded comparison of similarities within the document space. More importantly, when compared to a related measure such as the Dice index, the Jaccard coefficient penalizes more those sets with a small number of shared entries; this property is useful to penalize documents where the common terms are mostly functional words, for instance. As compared to cosine similarity, it provides for lesser tolerance over sets with large member disparity.

We now describe in turn the aforementioned set expansion operations and available optimisations of the core method. It is worth noting that no particular filtering is per-

[1] We used GIZA++ (Och and Ney 2003) to extract lexical translation tables. Although lexical translation modelling is sometimes based only on IBM model 1 in related work on comparable corpora, it is standard practice in statistical machine translation to use more sophisticated IBM models, usually up to model 4. We followed the latter approach, as the same tables can thus be used as components for both comparable corpora exploitation and SMT system development. We measured the impact of using one approach or the other on identical test sets and did not find any significant difference on document alignment results.

[2] We used 5 as a default for all language pairs, as a compromise between larger sets with less reliable translation candidates and smaller sets which may miss translation alternatives in comparable corpora. Note that optimal values for k could be empirically determined on domain-specific development sets for each language pair; such document-level tuning sets are however not usually available.

formed on the token sets, leaving punctuation marks alongside functional and content words, thus reducing document pre-processing to the minimal operation of tokenisation.

Out-of-vocabulary expansion. As we cannot guarantee that seed translations will cover the domain at hand satisfactorily, it is necessary to expand the translation sets with tokens that may be indicators of similarity although absent from translation tables. Since we do not lowercase the tokenised text,[3] case information is available throughout the text and all capitalised tokens are added to the sets if they are not found in the translation tables.[4] 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.

The effectiveness of this procedure varies between corpora, as it depends on both the amounts of entities in a given domain and the ability of the core method to discriminate between different documents without the inclusion of said entities. Thus, on the EITB test sets described in Section 4.3, its impact was not measurable, whereas on the French-English test sets of the Wikipedia task described in Section 4.2, it provided gains of over 30 points on all three metrics. We include this procedure in all experiments described in the remainder of the paper, as it would be unlikely to cause underperformance in any case.

Common prefix expansion. A common issue in statistical translation is morphological variation, with surface variants of a given lemma usually considered as independent unrelated units. This generates well-known data sparseness issues, which can be minimised through morphosyntactic analysis and lemmatisation. For under-resourced languages however, the resources for these processes might not be readily available or accurate enough, and a common approach relies on simple stemming through the use of manually created lists of endings. Surface forms are thus matched against possible endings and stemmed forms derived accordingly. For languages with rich inflectional morphology, however, these lists can contain hundreds of forms and their use is error prone, as several morphological phenomena need to be taken into account for a proper decomposition of endings and roots. Additionally, matching each surface form against large lists of endings is computationally costly.

To avoid both issues, we include a set expansion strategy that relies on longest common prefixes (LCP), which we compute over the minimal sets of elements that may have a common stem, defined to be the following two set differences: $T'_{ij} = T_{ij} - S_j$ and $T'_{ji} = T_{ji} - S_i$. Then 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 a minimal length of n characters,

[3] In all the results we present, the texts are not truecased either, to maintain the number of operations and required models to a minimum. Truecasing would provide a better treatment of sentence initial words but it remains to be tested whether this process would have a significant impact on document-level sets.

[4] 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*.

the prefix is added to both translation sets.[5] This approach reduces the problem to prefix comparison over the minimal necessary set of elements, as it exploits the nature of the alignment problem instead of generating stemmed candidates against large lists of endings.[6]

As is the case for the inclusion of surface-defined entities, the impact of LCP is expected to vary between corpora, for similar reasons. For all experiments reported in Section 4, the contribution of this operation was marginal at best and the reported results were obtained with a variant of DOCAL that does not include LCP. We nonetheless describe it here as an additional option with the potential to improve document alignment in other use cases.[7]

Document candidates. In some document alignment scenarios, a target-to-source 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. Experiments with DOCAL indicate limits of practicality being reached with over 260 million possible pairings on a single server with 64 Gigas of RAM and 16 cores.

For scenarios where the volume of documents renders an exhaustive comparison unsustainable, a standard cross-linguistic information retrieval (CLIR) approach is adopted. Target documents are first indexed using the Lucene search engine[8] and retrieved by building a query over the expanded translation sets created from each source document.

On the Spanish-English pair of the Europarl Version 7 corpus,[9] which contains 9.433 Spanish documents and 9.673 English documents, DOCAL performs alignment over the 91.235.976 possible pairings of the Cartesian product in 223 minutes and 13 seconds; the CLIR strategy over the same corpus executes the alignment process over 943.300 pairs in 33 minutes and 45 seconds, with an additional 1 minute and 5 seconds for indexing.

The two approaches are used in the experiments described in Section 4, with CLIR being used for the large volumes of documents in the Wikipedia task, and the Cartesian product being employed with the other datasets, where documents number in the thousands of documents at most.

Alignment filtering. As the alignment process is executed from source to target documents, a given target document can be taken as the best alignment for more than one source document. This results in correct alignments that end up hidden, often with

[5] Throughout the experiments we describe, n was set to 3, arbitrarily assumed to be the minimal length of a stem.

[6] To further improve the efficiency of the system, we use an implementation based on hash maps with minimal-length prefixes as keys and two sets as values for the original and translated tokens that have a given prefix in common. LCP is then computed on these reduced sets of elements.

[7] Experiments on internally available sets of parallel technical manuals showed improvements including LCP over the base version of DOCAL.

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

[9] Available at http://www.statmt.org/europarl/.

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 d_i and a target d_j if the latter is aligned to a different source document with a better similarity score.[10]

This process often produces large improvements, as it allows previously hidden good alignments to surface. On most of the experiments described in Section 4, the strategy led to significant improvement, with over 10 points gains in some scenarios. We include this alignment filtering as a default, although we will present some of the results with and without it to illustrate the gains obtained with this strategy.

4 Controlled experiments

To compare our approach with competing methods for document alignment, we evaluated system performance in three different scenarios that cover different language pairs, domains and degrees of comparability. We first performed document alignment on the EUROPARL corpus as a testbench for accuracy on parallel document alignment. We then applied DOCAL alignment to the WIKIPEDIA test sets selected for the 2015 BUCC shared task on similar document alignment, to measure the approach against recent results obtained by competing systems. Finally, we performed document alignment for a difficult language pair that includes one under-resourced language, namely Basque, using the EITB corpus of strongly comparable documents in the news domain. We describe the experimental setup and results for each scenario in turn.

In all experiments, DOCAL was tested with identical settings; as no training phase is necessary in the approach, those settings reduced to fixing the number of k-best lexical translations to 5. Unless otherwise specified, the lexical translation tables were created with GIZA++ on the JRC-Acquis Communautaire corpus.[11]

For the experiments on comparable corpora in Sections 4.2 and 4.3, we report results obtained with two different usages of the method: DOCAL.A refers to the core of the system, i.e. the basic translation sets augmented with capitalised tokens and numbers; DOCAL.B refers to the same system as DOCAL.A but augmented with the best alignment optimisation strategy described in Section 3.

4.1 EUROPARL

The Europarl corpus (Koehn 2005) is one of the main bitexts available, created from professional translations of parliamentary proceedings and covering the official EU languages. It is thus an appropriate resource to test parallel document alignment methods.

[10] Morin et al. (2015) refer to their similar removal of multiple source alignments as the *pigeonhole* method, following common terminology. To distinguish our version of the process from theirs, for presentation reasons we use the phrase *best alignment optimisation*.

[11] 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.

As various results have been reported on different versions of the corpus over the years, we applied DOCAL on two versions of the corpus, nameley versions 2 and 5.[12]

As a baseline, we implemented the previously described approach in Enright and Kondrak (2007), where hapaxes are defined over words with a minimal length of 4 characters. For the Spanish-English pair, the results match the accuracy reported in their paper, and the two methods gave highly accurate alignments: both methods fail on the one empty document in the corpus, while they differ on the one document with mixed language content, which only DOCAL aligns correctly. For later releases of the corpus however, results with the hapax method dropped sharply, either because of the larger sets of documents considered or because of the larger amount of documents with minimal textual content that can be found in the later releases (amounting to just one sentence in some cases). In contrast, results obtained with DOCAL were markedly better, although with a significant drop in F1 measure as well. The drops were similar across the three language pairs that we tested, as shown in Table 1. To investigate the specific impact of documents with minimal content, we prepared a variant of version 5 where all documents containing only one line of text were filtered.[13]. The results on this variant improved to higher scores, closer to those observed for version 2, with DOCAL providing markedly better results. The comparatively lower scores obtained on version 5 across the board do however show that the EUROPARL corpus can be useful to measure approaches to parallel document alignment, contrary to the conclusion reached by Enright and Kondrak (2007), which was based on version 2 of the corpus.

Table 1. Best F1 measures on Europarl

SYSTEM	CORPUS	ES-EN	FR-EN	NL-EN
HAPAX	EUPV2	99.6	99.6	99.6
DOCAL	EUPV2	**99.9**	**99.9**	**99.9**
HAPAX	EUPV5	54.2	54.5	50.3
DOCAL	EUPV5	**83.7**	**82.6**	**83.7**
HAPAX	EUPV5.2	72.9	72.5	67.2
DOCAL	EUPV5.2	**95.8**	**94.9**	**95.7**

Patry and Langlais (2011) applied their PARADOCS system on version 5 of the corpus, although the results they report render a direct comparison difficult: they use various corpus slices based on document length, a setup which we did not reproduce here,[14] and report percentage gains over the hapax-based method, and not absolute measures.

[12] We used the versions available as of February 2016 at the address: http://www.statmt.org/europarl/. We refer to version 2 as EUPV2 and to version 5 as EUPV5.

[13] We refer to this variant as EUPV5.2 in the table.

[14] They also indicate that the first sentences of each document were removed, which would directly eliminate the previously mentioned one-liner documents that are part of the version 5 we used.

They indicate F1 results of 95 and 93 for French-English and Dutch-English, respectively, for documents of at most 100 sentences. With all due caveats given experimental protocol differences, the DOCAL approach gave similar or better results on these pairs, as we reached 94.9 and 95.7 for these two language pairs on the full-length documents.

4.2 WIKIPEDIA

The 2015 Shared Task on Document Similarity, organised within the 8th workshop on Building and Using Comparable Corpora,[15] is the first task to provide a common testbench for the computation of similarity over a large collection of multilingual documents (Sharoff et al. 2015). The dataset is composed of static Wikipedia articles in three language pairs: English-French, English-Chinese and English-German. The articles were selected based on Wikipedia interlanguage links, provided the links were bidirectional and the documents matched size similarity criteria; the test sets were composed of articles with the interlanguage links removed. The task was to provide up to 5 target documents for each document in the test set, ranked according to the similarity scores assigned by the aligners to the document pairs.

Three measures were computed to assert the quality of the alignments, following standard TREC evaluation procedures:[16] SUCCESS@1 indicates the proportion of source articles for which the correct target article has been ranked in the top position; SUCCESS@5 measures the proportion of source articles for which the correct target has been ranked within the top 5 positions; finally, the MRR metric indicates the mean reciprocal range, i.e. the average over the 1/N scores assigned to correct target articles ranked at position N. For practical document alignment, the SUCCESS@1 results can be seen as the most significant, as these are the alignments that would be retained in actual usage of the systems.

Three systems participated in the task, all previously described in Section 2: LINA, whose results are reported with pigeon hole reasoning (referred to as LINA.P) and with alignment ties broken using the links from a third language (referred to as LINA.CL); the CCNUNLP system based on the approach in Li and Gaussier (2013); and the AUT system, described in Zafarian et al. (2015). The results for the AUT system having been reported as affected by a data processing bug, with all metric results close to zero, we do not include them in the tables below.

Results for the French-English and German-English pair are reported in Table 2. For the first pair, DOCAL performed markedly better than the alternative approaches, with an increase of almost 20 points on the SUCCESS@1 measure over the best reported system. It also obtained the best results on all other metrics in its variant where only the best alignments are retained. The base version of DOCAL also performs better on the SUCCESS@1 metric, although by a smaller margin.

As shown in Table 3, DOCAL performed markedly better for German-English as well, with gains of around 20 points on all three metrics over the best scoring version of LINA.[17]

[15] See: https://comparable.limsi.fr/bucc2015/.

[16] Text Retrieval Conference, see http://trec.nist.gov/.

[17] Results from the LINA system were the only ones available for this language pair.

Table 2. Results on the French-English Wikipedia task

SYSTEM	SUCCESS@1	SUCCESS@5	MRR
LINA.P	0.300	0.374	0.329
LINA.CL	0.577	0.606	0.590
CCNUNLP	0.607	0.764	0.669
DOCAL.A	0.636	0.693	0.659
DOCAL.B	**0.795**	**0.795**	**0.795**

Table 3. Results on the German-English Wikipedia task

SYSTEM	SUCCESS@1	SUCCESS@5	MRR
LINA.P	0.249	0.355	0.290
LINA.CL	0.607	0.639	0.622
DOCAL.A	0.649	0.621	0.688
DOCAL.B	**0.819**	**0.819**	**0.819**

Finally, results for the Chinese-English pair are shown in Table 4. For this language pair, the translation tables used by DOCAL were trained on 2 million parallel sentences extracted from the MULTIUN corpus, a collection of translated United Nations documents;[18] Chinese word segmentation was done with the Stanford segmenter (Tseng et al. 2005).[19] This is the only alignment scenario where one of the evaluated systems performs better than DOCAL among the selected tasks, with a marked difference for the SUCCESS@5 metric, a lower though significant difference for MRR, but a negligible difference of 0.014 points for the most important SUCCESS@1 metric. As the CCNUNLP system also relies on bilingual lexical information, it would be interesting to assess the difference in coverage and precision between the tables employed by each system; we did not however have the relevant information to perform this comparison. In further work, we will explore the impact of larger lexical translation tables for this language pair, as we only used a portion of the available training data for the experiment reported here.

Table 4. Results on the Chinese-English Wikipedia task

SYSTEM	SUCCESS@1	SUCCESS@5	MRR
CCNUNLP	**0.710**	**0.861**	**0.769**
DOCAL.A	0.576	0.601	0.576
DOCAL.B	0.696	0.696	0.696

[18] Available at: http://opus.lingfil.uu.se/MultiUN.php.
[19] http://nlp.stanford.edu/softhetware/segmenter.shtml

4.3 EITB

The final experiments were performed on the EITB corpus, a collection of strongly comparable documents in the news domain produced by the Basque public broadcaster Euskal Irrati Telebista.[20]

We manually aligned 299 documents in both languages as test set and applied DOCAL along with the EMACC expectation maximisation tool directly, i.e. on the cartesian product of documents; EMACC was set with default values. The lexical translation tables were created with GIZA++ using a parallel corpus of 645,223 aligned sentences extracted from the IVAP corpus, a collection of professional translations of public administration texts released by the Instituto Vasco de Administración Pública. Table 5 presents the results.

Table 5. Results on the Basque-Spanish EITB task

SYSTEM	PRECISION	RECALL	F1
EMACC	90.7	84.6	87.5
DOCAL.A	90.0	**90.0**	90.0
DOCAL.B	**91.1**	89.3	**90.2**

Although EMACC is a state-of-the-art aligner for comparable documents which obtained competitive results on this test set, the simpler DOCAL method reached better marks in terms of precision, recall and F1 measure. It also performed better in terms of execution time, as the complete alignments were computed in 1.568 seconds, as opposed to the 7 minutes and 9.693 seconds needed by EMACC, on the same environment with 48G of RAM and 16 cores. For a rather difficult language pair, we consider these results to be quite satisfactory.

5 Conclusions

We presented a simple approach to document alignment based on lexical translation sets derived from automatically created bilingual tables, simple set expansion procedures, and the Jaccard similarity coefficient.

To test the merits of this approach, we performed a series of controlled experiments on a varied set of corpora, with results that matched or outperformed those obtained with currently available methods. The comparison was performed over corpora exhibiting varying degrees of comparability, from fully parallel documents of the Europarl corpus to large document sets of comparable Wikipedia data. The experiments covered six different language pairs that included Germanic, Romance, Asian and under-resourced

[20] See Etchegoyhen et al. (2016) for a detailed description of this corpus, which will be made available in the META-SHARE repository (http://www.meta-share.eu/) as the $Basque\_Spanish\_EITB\_comparable\_corpus$, under the CC - BY - NC - SA licence for academic users and the MS - C - NO RED - FF licence for commercial users.

languages, thus indicating the strong potential of the proposed method to handle language and domain variation.

The DOCAL approach requires no specific adaptation process, nor rich feature sets, to improve over state-of-the-art results. Additionally, it can be successfully applied as is to parallel or comparable corpora without task-specific training phases. It is thus a highly portable and easy to deploy method, which proved effective for the alignment of parallel and comparable multilingual documents.

In future work, we will explore potential improvements for the proposed approach, with further evaluations of the impact of lexical translation coverage and precision.

Acknowledgements

This work was partially supported by the Basque Government under project TRADIN and the Spanish Ministerio de Economa y Competividad under project ADAPTA. We wish to thank the anonymous reviewers for their insightful comments.

References

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chen, J., Chau, R., and Yeh, C.-H. (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.

Chen, J. and Nie, J.-Y. (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 études internationales d'informatique documentaire.

Enright, J. and Kondrak, G. (2007). A fast method for parallel document identification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 29–32. Association for Computational Linguistics.

Etchegoyhen, T., Azpeitia, A., and Perez, N. (2016). Exploiting a large strongly comparable corpus. To appear in *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*.

Fung, P. and Cheung, P. (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.

Ion, R., Ceauşu, A., and Irimia, E. (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.

Jaccard, P. (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.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

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

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

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

Morin, E., Hazem, A., Boudin, F., and Clouet, E. L. (2015). Lina: Identifying comparable documents from Wikipedia. In *Eighth Workshop on Building and Using Comparable Corpora*.

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

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

Paramita, M. L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., and Sanderson, M. (2013). Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Patry, A. and Langlais, P. (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.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95. Association for Computational Linguistics.

Prochasson, E. and Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1327–1335. Association for Computational Linguistics.

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

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, pages 74–78.

Tiedemann, J. (2011). *Bitext alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Tseng, H., Chang, P., Andrew, G., Jurafsky, D. and Manning, C. (2005). A conditional random field word segmenter. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Uszkoreit, J., Ponte, J. M., Popat, A. C. and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Zafarian, A., Aghasadeghi, A., Azadi, F., Ghiasifard, S., Alipanahloo, Z., Bakhshaei, S., and Ziabary, S. M. M. (2015). AUT Document alignment framework for BUCC workshop shared task. *ACL-IJCNLP 2015*, page 79.

Received May 2, 2016 , accepted May 13, 2016