Bilingual Document Alignment with Latent Semantic Indexing

Ulrich Germann
School of Informatics
University of Edinburgh
ugermann@inf.ed.ac.uk

Abstract

We apply cross-lingual Latent Semantic Indexing to the Bilingual Document Alignment Task at WMT16. Reduced-rank singular value decomposition of a bilingual term-document matrix derived from known English/French page pairs in the training data allows us to map monolingual documents into a joint semantic space. Two variants of cosine similarity between the vectors that place each document into the joint semantic space are combined with a measure of string similarity between corresponding URLs to produce 1:1 alignments of English/French web pages in a variety of domains. The system achieves a recall of ca. 88% if no in-domain data is used for building the latent semantic model, and 93% if such data is included.

Analysing the system's errors on the training data, we argue that evaluating aligner performance based on exact URL matches under-estimates their true performance and propose an alternative that is able to account for duplicates and near-duplicates in the underlying data.

1 Introduction

Identifying document pairs that are mutual translations of one another in large multilingual document collections is an important processing step in harvesting parallel bilingual data from web crawls. The *Shared Task on Bilingual Document Alignment* at the *First Conference on Machine Translation (WMT16)* provides a common framework to investigate and compare approaches to solving this problem: given a collection of web site crawls, and a list of known matches, identify additional document pairs in the collection.

This paper explores the use of cross-lingual Latent Semantic Indexing (Berry and Young, 1995) in combination with a URL matching scheme for this task.

2 Latent Semantic Indexing

2.1 Singular Value Decomposition

Latent Semantic Indexing (LSI; Dumais et al., 1988; Deerwester et al., 1990) is a well-known indexing technique in information retrieval. It relies on reduced-rank singular value decomposition to map a high-dimensional *term-document matrix* into a "semantic" space of much lower dimensionality.

The term-document matrix is set up by counting word occurrence in documents. Each row in the matrix corresponds to a term in the vocabulary, each column to a document. The individual values in the matrix are weighted term counts of the respective term in the respective document. For this work, we use log-normalised term counts (tf – term frequency) weighted by term specificity as measured by the inverse document frequency (idf; Spärck-Jones, 1972[1]):

$$w_{t;d} = \mathsf{tf} \cdot \mathsf{idf} \quad (1)$$

$$\text{with} \quad \mathsf{tf} = 1 + \log \text{count}_d(t) \quad (2)$$

$$\mathsf{idf} = \log \frac{|\mathcal{C}|}{\text{count}_{\mathcal{C}}(d : t \in d)} \quad (3)$$

where t is a term from the vocabulary $\mathcal{V}$ and d a document from the document collection $\mathcal{C}$.

Singular value decomposition (SVD; cf., for example, Manning and Schütze, 1999) is then used to factorise this term-document matrix M with $m = |\mathcal{V}|$ rows and $n = |\mathcal{C}|$ columns into three matrices $\mathsf{T}_{m\times k}$,$\mathsf{S}_{k\times k}$, and $\mathsf{D}_{n\times k}$ (with with $k = \min(m, n)$), such that $\mathsf{T}\ \mathsf{S}\ \mathsf{D}\ ^{\mathsf{T}}$=$\mathsf{M}$.

The column vectors of T and D are orthonormal bases of a k-dimensional vector space; S is a diagonal matrix with the Eigenvalues of M in descending order. In other words, dimensions in which the data differs the most come first, dimensions in which the data differs little come last. By truncating each of the SVD output matrices to the

[1] The technique was proposed by Spärck-Jones; the term idf was coined later.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 692–696,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

respective first $r \ll k$ columns, we obtain a low-rank representation that approximates the original term-document Matrix: $\mathsf{T}'_{m\times r}\mathsf{S}'_{r\times r}(\mathsf{D}'_{n\times r})^{\mathsf{T}} \approx \mathsf{M}$. (Note, by the way, that $D'\mathsf{S}'\mathsf{D}'^{\mathsf{T}}$ is the cosine similarity matrix in the new low-dimensional vector space.)

2.2 Document fold-in

To map a new document into this vector space, we compute the corresponding new row to be added to D' as $d_q = \mathsf{M}_q^{\mathsf{T}}\mathsf{T}'\mathsf{S}'^{-1}$, where M_q is an additional column in M that contains the weighted counts of terms in the respective document.

3 Alignment of multilingual web pages via cross-lingual LSI

3.1 Introduction

Web pages as delivered by web servers are a mix of data: HTML markup, which structures the document and pulls in additional resources such as cascading style sheets, JavaScript libraries, images, and video; scripts executable in the web browser that influence and extend its interactive behaviour and functionality; embedded images and videos, and, finally, text visible to the human user. Visible text comprises *boilerplate* and *payload*. Boilerplate text is text that appears repeatedly across a web site in the form of menus, page headers and footers, etc. While usually highly distinctive of a specific web site, boilerplate contributes little to being able to distinguish individual web pages on a specific site. Web site readers will usually not pay much attention to boilerplate text except when navigating the web site; it is nothing that they will actively read in order to satisfy information needs other than how to navigate the web site. Payload text, on the other hand, is text that users visit the specific page for.

While document structure, embedded links, etc. can provide valuable clues for the alignment of web pages, this work focuses on the text extracted from the original HTML, as provided by the workshop organisers as part of the data set.

3.2 Approach

The central idea in our approach is to use cross-lingual LSI to map monolingual documents into a joint vector space and use similarity between the corresponding embedding vectors to perform bipartite alignment of pairs of documents in different languages.

To obtain a cross-lingual model of latent semantics, we first set up a bilingual term-document matrix M using parallel documents, keeping the vocabularies of the two languages separate, so that identical word forms in the two languages correspond to different rows in M. Rank-reduced SVD is then performed on this bilingual matrix to map the terms of the two languages into a common semantic space with 1,000 dimensions.[2] Via fold-in, all monolingual documents from the collection that have been labelled by the language recogniser as being in one or the other of the language in question are also be mapped into this common space.

We then use *Competitive Linking* (Melamed, 1997) to obtain a bipartite alignment of documents: first, we rank all possible bipartite alignment hypotheses by score. Processing the list of hypotheses in descending order, we keep all hypotheses that do not overlap or conflict with higher-ranking hypotheses and discard the others. (In fact, competitive linking is what the official evaluation procedure for this shared task does; for the purpose of participation in the Shared Task, it is sufficient to produce a ranked list).

3.3 Term Weighting

As mentioned above, text extracted from a web page consists of boilerplate and payload text. To reduce the influence of the former and boost the impact of the latter on the document vectors, we compute idf separately for each domain in the set (rather than globally across all domains). Thus, terms that occur frequently across a particular web site will receive a low specificity score (i.e., idf) on pages from that web site, yet may receive a high score if they appear elsewhere.

3.4 Scoring functions

In our experiments, we explored and combined the following scoring functions:

3.4.1 Cosine Similarity (cos)

This is the classical measure of similarity in LSI-based Information Retrieval. It computes the cosine of the angle between the two vectors that embed two candidate documents in the joint semantic vector space.

3.4.2 “Local” cosine similarity (lcos)

The intuition behind the local cosine similarity measure is this: since we perform SVD on a bilingual term-document matrix that consists of document column vectors for documents from a large collection of web sites, web pages from each specific web site will still appear quite similar if the web site is dedicated to a particular topic area (which the vast majority of web sites are). Similarity scores will thus be dominated by the general domain of the web site rather than the differences between individual pages within a given web site. The local cosine similarity measure tries to mediate this phenomenon by shifting the origin of the vector space to the centre of the sub-space in which the pages of

[2] We used the open-source software package *redsvd* (randomised SVD; Okanohara, 2010) to perform the singular value decomposition.

a particular web site reside before computing cosine similarity. In practice this is accomplished by subtracting the mean embedding vector for the domain in question from each individual embedding vector for pages in that domain. Note that we are only comparing pages that belong to the same web site within the context of the shared task.

3.4.3 URL similarity (url)

The data provided for the Shared Task contains many duplicates and near-duplicates of web pages. Duplicates occur when multiple URLs lead to exactly the same content (e.g. `www.domain.com` and `www.domain.com/index.html`); near-duplicates are often the result of dynamically created content, such as results of database look-up (e.g., calendars, stock price trackers), embedded page counts, or different boilerplate due to different language settings delivering the same payload (e.g., an English article delivered under two different country-specific user interfaces using different boilerplate text). Not knowing how the reference set for evaluation within the Share Task was constructed, we conjectured that the gold standard used for evaluation might be biased towards URL matches.

Hence, we devised the following match score for pairs of URLs.

1. All URLs within a domain are tokenised into blocks of either all letters or all numbers relying on POSIX UTF-8 character classes; punctuation is discarded.

2. For a given pair of candidate URLs, we determine via the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) the cumulative score of the longest match sequence between the token sequences corresponding to the two URLs. The match score for each individual token pair $\langle t_1, t_2\rangle$ in the alignment is computed as follows.
   - $\text{score}(t_1, t_2) = 0$ if $t_1 \neq t_2$ and at least one of them is a number
   - $\text{score}(t_1, t_2) = \frac{1}{\text{cnt}(t_1)^2}$ if $t_1 = t_2$, where $\text{cnt}(t)$ is the position-independent count of token t in all the URLs in the collection. The match weighting based on relative frequency in the domain serves to discount very frequently occurring URL components, (such as *http* or *www*) and boost components that are rare in the URLs for this domain, such as, for example, article IDs.
   - $\text{score}(t_1, t_2) = \frac{2 * \text{lcss}(t_1,t_2)}{\text{len}(t_1)+\text{len}(t_2)} \cdot \frac{1}{\text{cnt}(t_1)\cdot\text{cnt}(t_2)}$ if t_1 and t_2 both are sequences of letters, where $\text{lcss}(t_1, t_2)$ is the length of the longest common letter sub-sequence between t_1 and t_2. The idea behind this soft match score is to reward cognates over candidate pairs that have no semblance of one another whatsoever. For example, the lcss score component for the pair $\langle \textit{London}, \textit{Londres}\rangle$ would be ca. 0.62 ($\frac{2*4}{\text{len}(\textit{“London”})+\text{len}(\textit{“Londres”})}$), whereas the pair $\langle \textit{London}, \textit{Paris}\rangle$ would receive a match score of 0, each of the scores yet to be weighted by $\frac{1}{\text{cnt}(t_1)\cdot\text{cnt}(t_2)}$. This soft matching score serves to accommodate web sites that base their URLs on, for example, the headlines of articles or posts.

4 Evaluation

4.1 Recall on training and test data

To rank alignment hypotheses, we investigated all uniform linear combinations of the three individual scoring functions. Table 1 shows the results for the training set, and, in the last row, the performance of the best feature combination on the test set. In the first set of experiments on the training set, whose results are shown in the left half of the table, we used the list of known matches in the training data both for seeding cross-lingual LSI and evaluation. These numbers give us a sense how well monolingual documents are mapped into the joint semantic space by LSI and document fold-in. The first column of the recall numbers (“strict”) follows the official evaluation procedure, counting only exact URL matches as correct. The following columns show the performance if a more lenient notion of “matching documents” is applied. This more lenient measures computes the similarity between the expected and a proposed target document for a given source document (and vice versa) as follows:

$$\text{score}(\text{text}_1, \text{text}_2) = \frac{2 \cdot \text{lcss}(\text{text}_1, \text{text}_2)}{|\text{text}_1| + |\text{text}_2|} \tag{4}$$

The length of the longest common sub-sequence (lcss) is here measured in terms of space-separated tokens as they occur in the text. No more sophisticated tokenisation is performed. The content-based evaluation measure counts a proposed match as correct if the similarity between a proposed target (or source) document and the expected document is greater or equal to the threshold indicated in the column header.

The right half of the table shows the results for the same evaluation performed on the basis of original bilingual term-document matrices that *exclude* all known matches from the domain in question, relying only on known matches from other web domains. This leads to fewer vocabulary matches, as terms specific to the web site in question may not be included in the model. As expected, we see a drop in performance, but we are still able to recover

Table 1: Recall on the training and test data with known in-domain document pairs included in / excluded from the initial term-document matrix.

performance on the training data										
	included					**excluded**				
features used	strict[a]	1.00[b]	0.99[b]	0.95[b]	0.90[b]	strict[a]	1.00[b]	0.99[b]	0.95[b]	0.90[b]
cosine (cos)	86.7	93.4	95.4	96.7	97.6	82.5	88.9	91.3	92.9	93.7
"local" cos. (lcos)	86.7	92.8	94.7	95.8	96.9	83.3	88.9	91.4	92.8	93.6
URL similarity (url)	83.6	87.8	88.1	88.2	88.2	83.6	87.8	88.1	88.2	88.2
cos lcos	87.2	93.7	95.6	96.6	97.5	83.3	89.7	92.1	93.6	94.4
cos url	90.6	94.7	95.6	96.4	97.1	86.3	90.6	91.4	92.7	93.5
lcos url	91.3	95.4	96.3	97.2	97.8	86.8	91.3	92.2	93.4	94.2
cos lcos url	92.8	96.7	97.6	98.5	99.1	88.0	92.5	93.4	94.7	95.5
performance on the test data										
cos lcos url						87.6	87.6	94.1	95.5	96.0

[a] exact string match with the reference ULR pairs
[b] soft match based on document similarity with different similarity thresholds.

Table 2: Distribution of missed pairs over domains with a soft similarity threshold of .95. Domains with a single miss are aggregated under "other".

domain	**missed pairs**
www.lagardere.com	20
meatballwiki.org	12
www.toucherdubois.ca	8
www.rfimusique.com	8
www.taize.fr	6
www.lalettrediplomatique.fr	4
www.publictendering.com	3
www.iisd.ca	3
hrcouncil.ca	3
arabpressnetwork.org	3
www.technip.com	2
www.kinnarps.com	2
www.gameonly.com	2
www.eufic.org	2
other	17

about 92.5% (down from 96.7%) of the known matches, even when counting only full matches and matches with exact duplicates.

4.2 Error analysis

Table 2 shows the distribution of missed page pairs over the respective domains in the test data. As we can see, errors are concentrated in only a few of the 203 domains in the test set. We will briefly discuss the top five here. The errors in `www.lagardere.com` originate from mixed-language pages, typically pages with the boilerplate text for the user interface in one language and the actual content in the other. The missed pairs in `meatballwiki.org` can be attributed to *red herrings*: URL pairs that erroneously suggest a correspondence between the two pairs in question. The web site `www.toucherdubois.ca` provides teaching resources (including images and lesson plans) for teaching students about "the sociocultural heritage of the people of Madawaska" in Canada and the US. Some of the pages consist of little text wrapped around image resources; lesson plans are often very similar in terms of the vocabulary used, thus confusing the LSA model. The missing pairs from `www.rfimusique.com` and `www.taize.fr` are pairs of pages with a low payload-to-boilerplate (or near-boilerplate) ratio, i.e., they are dominated by text that can be found on multiple pages, thus leading to document alignment errors.

5 Related work

One of the first systematic approaches to identifying parallel data on the web is the STRAND algorithm (Resnik, 1999). It is a pipeline process that first generates candidate pairs via a web search (or by link analysis if a complete download of a web site is available). It then performs language identification on the retrieved pages and analyses the HTML structure of candidate documents in order to filter out document pairs that are too dissimilar in their document structure. Resnik and Smith (2003) extend this approach by adding content-based analysis. They use probabilistic word translation lexicons to assess the probability that two pages are translations of each other.

Very similarly to the work presented in this paper, Saad et al. (2014) use LSI for identification of parallel and comparable corpora. In addition to the cross-lingual LSI approach taken here, they also investigate monolingual LSI after document translation. They conclude that cross-lingual LSI is competitive with monolingual LSI of automatically translated texts.

6 Conclusion

We have investigated the feasibility of using cross-lingual LSI for identifying parallel documents in large collections of text. Our results suggest that this is a viable approach to harvesting parallel data from web crawls. We achieve the best performance with a combination of classical cosine measure, "local" cosine measure, and URL matching.

The existence of duplicate and near-duplicate documents in the data raises the question whether it is reasonable to measure performance in terms of URL matches, or whether evaluation should be based on the distance between retrieved and expected documents.

Acknowledgements

This work was conducted within the scopes of the Innovation Action *MMT* and the Research and Innovation Action *SUMMA*, which have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 645487 and 688139, respectively.

References

Berry, Michael W. and Paul G. Young. 1995. "Using latent semantic indexing for multilanguage information retrieval." *Computers and the Humanities*, 29(6):413–429.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*, 41(6):391–407.

Dumais, Susan T., George W. Furnas, Thomas K. Landauer, Scot Deerwester, and Richard Harshman. 1988. "Using latent semantic analysis to improve access to textual information." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, 281–285. New York, NY, USA.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press.

Melamed, I. Dan. 1997. "A word-to-word model of translational equivalence." *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 490–497. Madrid, Spain.

Needleman, Saul B. and Christian D. Wunsch. 1970. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology*, 48(3):443–453.

Okanohara, Daisuke. 2010. "redsvd." `http://mloss.org/software/view/274/`.

Resnik, Philip. 1999. "Mining the web for bilingual text." *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 527–534. College Park, Maryland, USA.

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

Saad, Motaz, David Langlois, and Kamel Smaïli. 2014. "Cross-lingual semantic similarity measure for comparable articles." *Proceedings of the 9th International Conference on NLP, PolTAL 2014*, 105–115. Warsaw, Poland.

Spärck-Jones, Karen. 1972. "A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation*, 28(1):11–21.