Identifying Comparable Corpora Using LDA

Judita Preiss
j.preiss@sheffield.ac.uk
Department of Computer Science, University of Sheffield, Regent Court
211 Portobello, Sheffield, S1 4DP, United Kingdom

Abstract

Parallel corpora have applications in many areas of Natural Language Processing, but are very expensive to produce. Much information can be gained from comparable texts, and we present an algorithm which, given any bodies of text in multiple languages, uses existing named entity recognition software and topic detection algorithm to generate pairs of comparable texts without requiring a parallel corpus training phase. We evaluate the system's performance firstly on data from the online newspaper domain, and secondly on Wikipedia cross-language links.

1 Introduction

Manual alignment or creation of parallel corpora is exceedingly expensive, requiring highly skilled annotators or professional translators. Methods exist for aligning parallel corpora, and extracted parallel segments can be used to, for example, augment machine translation phrase tables, but the amount of genuinely parallel data is limited. However, parallel segments can also be extracted from comparable corpora (a comparable corpus is one which contains similar texts in more than one language). Comparable documents, if produced with a confidence value, could also be used to prioritize translation (manual or automatic) when one is searching for further information (which may only be available in a foreign language) to augment information given in an article in the source language. We present a technique to automatically detect comparable corpora in existing data, and we demonstrate the applicability of our method to any genre by evaluating on crawled online newspaper text, as well as Wikipedia articles.

Clearly, texts need to contain some of the same data in order to be comparable (Harris, 1954), and we assume:

- To be similar, texts need to share some **named entities**, e.g., Tóth et al., (2008).
- Comparable texts need to be on the same **topic**.

Construction of multilingual topic models usually requires either parallel data or some number of aligned documents across multiple languages. Zhao and Xing (2007) create bilingual topic models from (at least 25%) of parallel data. Mimno et al., (2009) start from tuples of equivalent documents to build models, and then the same distribution over topics holds in both source and target languages.

While Zhao and Xing (2007) used their topic models for word alignment from comparable corpora (combined with underlying parallel data), multilingual topic models are usually applied to data to automatically detect word translations based on parallel data, e.g., Vulić et al., (2011) exploit a shared language independent topic distribution to measure the similarity between topics pertaining to words.

The novelty of our work is the transformation of a source language topic model rather than the creation of a language independent model from parallel data. Transforming the source language model to the target language allows the classification of the target language documents to source language topics. The translated model is applied to two document collections to demonstrate its ability to detect comparable

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 558–562,
Montréal, Canada, June 3-8, 2012.

corpora. Our system can be applied to any pair of languages for which there is a dictionary.

Section 2 describes the tools we employ. Section 3 contains a description of our system: the method for employing NE recognition across languages is presented in Section 3.1, while Section 3.2 outlines our technique for employing LDA across languages. Our experiments and their results are described in Section 4. Section 5 draws our conclusions and indicates avenues for future work.

2 Tools

2.1 Named entity recognition

The Stanford named entity recognition (NER) software[1] (Finkel et al., 2005) is an implementation of linear chain Conditional Random Field (CRF) sequence models, which includes a three class (person, organization, location and other) named entity recognizer for English.

2.2 Topic detection

LDA (Blei et al., 2003) is a generative probabilistic model where documents are viewed as mixtures over underlying topics, and each topic is a distribution over words. Both the document-topic and the topic-word distributions are assumed to have a Dirichlet prior. Given a set of documents and a number of topics, the model returns θ_i, the topic distribution for each document i, and ϕ_{ik}, the word distribution for topic k. We employ the publicly available implementation of LDA, JGibbLDA[2] (Phan et al., 2008), which has two main execution methods: parameter estimation (model building) and inference for new data (classification of a new document). Both invocations produce the following:

ϕ_{ij}: $p(\text{word}_i|\text{topic}_j)$

θ_{jk}: $p(\text{topic}_j|\text{document}_k)$

tassign: a deterministic topic-word assignment for each word in every document

The LDA topic models are created from a randomly selected tenth of the Reuters corpus (Rose et al., 2002).[3]

[1] http://nlp.stanford.edu/ner/index.shtml

[2] http://jgibblda.sourceforge.net/

[3] LDA modeling can abstract a model from a relatively small corpus and a tenth of the original Reuters corpus is much more manageable in terms of memory and time requirements.

2.3 Indexing

To provide quick searching access to the large text collections, we utilize the high-performance search engine library Lucene.[4] The stemmed and stoplisted documents are stored along with the frequency of occurrence of each word within a document.

2.4 Lemmatization / stemming

English text is lemmatized using the lemmatizer available within RASP[5] (Briscoe et al., 2006). Stemming is provided for all the non-English languages included in our work within Lucene.

3 Identifying comparable corpora

3.1 Cross language NER

NEs extracted from the English text collections are automatically translated into the target languages using the BING Translation API[6] yielding a single translation, which is retained. The stemmed, translated version of each NE in the source text is sought in the indexed form of the target language document collection, and the frequency of occurrence of the NE is returned.

Filtering is applied based on the proportion of source language document's NEs found in the target document (we do not expect all the NEs to be present in the target language: NEs could be mis-translated, and not all NEs would necessarily be mentioned even in a comparable document). The proportions of all types of NEs required were optimized over a small manually created set. While we could assign a weight and not filter documents, this is not believed to be adequate: e.g., a newspaper article containing all the source location mentions (and thus having a high weight), but none of the same people, is likely to be a news story about the same area but a different event.

[4] http://lucene.apache.org

[5] http://ilexir.co.uk/applications/rasp/download

[6] The translations could also be retrieved from NE mapping lists, dictionaries (if these are available) or manually translated – we therefore do not see this step as violating the lack of need for a parallel corpus.

3.2 Cross language topic identification

Being non-deterministic, multiple executions of the LDA algorithm are not guaranteed to (and do not) give rise to identical topics (even within one language). It is therefore not possible to build a topic model in the source language and the target language separately, as there is no clear alignment between their respective topics. Traditionally, parallel corpora are used to generate a language independent topic-document distribution, from which polylingual topic models can be created so the underlying topics are shared.

We propose to translate each word from the source language topic model using the BING API and substitute the new wordmap thus creating a target language topic model. While word distributions are clearly different across languages, and building a shared topic-document distribution to sample words from allows words to retain their language specific distributions, our technique completely avoids the need for parallel corpora, and merely requires the translation of the words in the LDA model (which can be performed using dictionary lookup, or NE lists instead of the BING API).

3.3 Selecting comparable corpora

Target language candidate documents found to share sufficient proportions of NEs are classified using the translated target language LDA model. This yields θ_{jk} (the probability distribution of topic given document) and classifying the original document using the source language LDA model gives θ'_{jk}. The candidate documents are ranked according to the cosine similarity between the two vectors:

$$similarity = \frac{\theta_{jk} \cdot \theta'_{jk}}{\|\theta_{jk}\|\|\theta'_{jk}\|}$$

By definition, cosine similarity ranges between -1 and 1. Similarity of 1 indicates two documents with $\theta = \theta'$, and thus the higher the similarity, the higher we rank the document.

4 Experiments

We present two evaluations: firstly, we manually evaluate the comparable documents generated from online newspaper text in two languages, while the second evaluation finds comparable articles in source and target versions of Wikipedia with results evaluated against the cross-language links present in Wikipedia.

4.1 Online newspaper documents

Simple Google search yields a number of links to online newspapers in any language, these lists (automatically retrieved) are used to seed a crawler. Documents from newspaper sites which allow crawling are retrieved and only well formed HTML documents are retained,[7] and the language of the documents is verified using a Perl implementation of Lingua::Ident (Dunning, 1994), an n-gram based model for language identification.[8]

A single annotator evaluated 10 randomly selected English documents and the comparable documents returned for them from 40,528 Czech newspaper articles (total retrieved within a 24 hour period). Since there is no current scheme available for judging comparability, we employed a four category scale:

Strong: The documents are about the same news event, in a similar style. (Articles about the same news event, but elaborating, would be included here.)

Medium: The documents are about related news events.

Weak: The documents refers to similar events.

None: No overlap in topic in the two documents.

Results of the evaluation are presented in Table 1; the top document is scored for each pair, showing the high precision of the technique. The 10 English documents were selected subject to the constraint that a comparable corpus was retrieved for them: the imposed constraints on NEs make this a high precision / low recall technique. Many articles found using the crawling approach on news sites (rather than an RSS feed gathering approach) were discussions,

[7] Note that the crawler is not permitted to leave the domain of the newspaper.

[8] The Lingua::Ident Perl module is available from `http://search.cpan.org/~mpiotr/Lingua-Ident-1.7/Ident.pm`. We build the models for the language identification system from downloaded Wikipedia content for each language.

Strong	Medium	Weak	None
4	4	1	1

Table 1: Results for English-Czech documents

for example discussions of strategies in sports, interviews with actors, rather than topical news stories. From a manual inspection of the target language articles, many of these articles do not appear to have comparable equivalents. Also, enforcing a high proportion of NEs shared between the source and target languages frequently rules out documents which are subsets of each other (this was also apparent in the second evaluation).

4.2 Wikipedia

Information within Wikipedia is connected across languages using cross-language links. While the lists of links are not necessarily complete, and the articles they link may not contain large parallel segments, the linked documents should be comparable (under the definition), and thus provide an empirical measure of the utility of our method.

The top comparable articles in Czech were generated for 100 randomly selected English Wikipedia articles (subject to the constraint that they have cross-language links). As in our first evaluation, the system had a low recall (35%), however precision was 83%. By the design of the experiment, an article about the same subject has to exist in both languages, and therefore the low recall value is surprising. Rather than a low cosine value, the low recall is mainly due to the NE filtering step removing the 'correct' article from consideration. A brief inspection of a small number of articles which had been filtered out was performed and substantial differences between the pages were found – for example, a significant portion of the Wikipedia page for *Equinox* in English contains descriptions of Equinox commemorations all over the world, which are missing in the Czech version of the Wikipedia article (leading to a large number of missing NEs). Similar length of articles appeared to be a good indicator of both articles containing similar data, and our system detecting the two texts to be comparable.

Please note that while the NE filtering step is removing texts from consideration, it is not possible to compute cosines of the topic vectors of all documents and thus some candidate selection step is necessary.

4.3 Baseline

There are no standard baselines for the task of creating comparable corpora. It is possible to translate the source language text into the target language using BING, however, a cosine comparison of the stemmed, automatically translated document with all documents in the target language collection is extremely time consuming. Applying NE filtering, automatically translating the remaining target language candidate texts into the source language using BING, and ranking according to cosine similarity gives a precision of 69% for the collection discussed in Section 4.2.

5 Conclusion

We have presented an LDA based algorithm applicable to large document collections to find comparable documents across multi-lingual corpora without the needing to train with parallel data. We show, using a human judge as well as Wikipedia cross-language links, that the system achieves high precision in finding comparable documents.

The technique strongly relies on the named entity method selected, and another technique may be more suitable. A comparison with a bilingual topic model created from parallel data would also prove interesting.

Acknowledgments

This research was supported by EU grant 248347 on Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation (ACCURAT). My thanks also go to the three reviewers whose comments strengthened the findings of this work.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Briscoe, T., Carroll, J., and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

Dunning, T. (1994). Statistical identification of language. Technical Report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146162.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 880889.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of The 17th International World Wide Web Conference (WWW 2008)*, pages 91–100.

Rose, T. G., Stevenson, M., and Whitehead, M. (2002). The Reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 827–832.

Tóth, K., Farkas, R., and Kocsor, A. (2008). Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm. *Acta Cybernetica*, pages 463–478.

Vulic, I., Smet, W. D., and Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of ACL*, pages 479–484.

Zhao, B. and Xing, E. P. (2007). HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *NIPS*.