Identifying Word Translations from Comparable Corpora Using Latent Topic Models

Ivan Vulić, Wim De Smet and **Marie-Francine Moens**
Department of Computer Science
K.U. Leuven
Celestijnenlaan 200A
Leuven, Belgium
{ivan.vulic,wim.desmet,sien.moens}@cs.kuleuven.be

Abstract

A topic model outputs a set of multinomial distributions over words for each topic. In this paper, we investigate the value of bilingual topic models, i.e., a bilingual Latent Dirichlet Allocation model for finding translations of terms in comparable corpora without using any linguistic resources. Experiments on a document-aligned English-Italian Wikipedia corpus confirm that the developed methods which only use knowledge from word-topic distributions outperform methods based on similarity measures in the original word-document space. The best results, obtained by combining knowledge from word-topic distributions with similarity measures in the original space, are also reported.

1 Introduction

Generative models for documents such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are based upon the idea that latent variables exist which determine how words in documents might be generated. Fitting a generative model means finding the best set of those latent variables in order to explain the observed data. Within that setting, documents are observed as mixtures of latent topics, where topics are probability distributions over words.

Our goal is to model and test the capability of probabilistic topic models to identify potential translations from document-aligned text collections. A representative example of such a comparable text collection is Wikipedia, where one may observe articles discussing the same topic, but strongly varying in style, length and even vocabulary, while still sharing a certain amount of main concepts (or topics). We try to establish a connection between such latent topics and an idea known as the *distributional hypothesis* (Harris, 1954) - words with a similar meaning are often used in similar contexts.

Besides the obvious context of direct co-occurrence, we believe that topic models are an additional source of knowledge which might be used to improve results in the quest for translation candidates extracted without the availability of a translation dictionary and linguistic knowledge. We designed several methods, all derived from the core idea of using word distributions over topics as an extra source of contextual knowledge. Two words are potential translation candidates if they are often present in the same cross-lingual topics and not observed in other cross-lingual topics. In other words, a word w_2 from a target language is a potential translation candidate for a word w_1 from a source language, if the distribution of w_2 over the target language topics is similar to the distribution of w_1 over the source language topics.

The remainder of this paper is structured as follows. Section 2 describes related work, focusing on previous attempts to use topic models to recognize potential translations. Section 3 provides a short summary of the BiLDA model used in the experiments, presents all main ideas behind our work and gives an overview and a theoretical background of the methods. Section 4 evaluates and discusses initial results. Finally, section 5 proposes several extensions and gives a summary of the current work.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 479–484,
Portland, Oregon, June 19-24, 2011.

2 Related Work

The idea to acquire translation candidates based on comparable and unrelated corpora comes from (Rapp, 1995). Similar approaches are described in (Diab and Finch, 2000), (Koehn and Knight, 2002) and (Gaussier et al., 2004). These methods need an initial lexicon of translations, cognates or similar words which are then used to acquire additional translations of the context words. In contrast, our method does not bootstrap on language pairs that share morphology, cognates or similar words.

Some attempts of obtaining translations using cross-lingual topic models have been made in the last few years, but they are model-dependent and do not provide a general environment to adapt and apply other topic models for the task of finding translation correspondences. (Ni et al., 2009) have designed a probabilistic topic model that fits Wikipedia data, but they did not use their models to obtain potential translations. (Mimno et al., 2009) retrieve a list of potential translations simply by selecting a small number N of the most probable words in both languages and then add the Cartesian product of these sets for every topic to a set of candidate translations. This approach is straightforward, but it does not catch the structure of the latent topic space completely.

Another model proposed in (Boyd-Graber and Blei, 2009) builds topics as distributions over bilingual matchings where matching priors may come from different initial evidences such as a machine readable dictionary, edit distance, or the Pointwise Mutual Information (PMI) statistic scores from available parallel corpora. The main shortcoming is that it introduces external knowledge for matching priors, suffers from overfitting and uses a restricted vocabulary.

3 Methodology

In this section we present the topic model we used in our experiments and outline the formal framework within which three different approaches for acquiring potential word translations were built.

3.1 Bilingual LDA

The topic model we use is a bilingual extension of a standard LDA model, called bilingual LDA (BiLDA), which has been presented in (Ni et al., 2009; Mimno et al., 2009; De Smet and Moens, 2009). As the name suggests, it is an extension of the basic LDA model, taking into account bilinguality and designed for parallel document pairs. We test its performance on a collection of comparable texts which are document-aligned and therefore share their topics. BiLDA takes advantage of the document alignment by using a single variable that contains the topic distribution θ, that is language-independent by assumption and shared by the paired bilingual comparable documents. Topics for each document are sampled from θ, from which the words are sampled in conjugation with the vocabulary distribution ϕ (for language S) and ψ (for language T). Algorithm 3.1 summarizes the generative story, while figure 1 shows the plate model.

Algorithm
3.1: GENERATIVE STORY FOR BILDA()

for each document pair d_j
do $\begin{cases} \textbf{for each} \text{ word position } i \in d_{jS} \\ \quad \textbf{do} \begin{cases} \text{sample } z_{ji}^S \sim Mult(\theta) \\ \text{sample } w_{ji}^S \sim Mult(\phi, z_{ji}^S) \end{cases} \\ \textbf{for each} \text{ word position } i \in d_{jT} \\ \quad \textbf{do} \begin{cases} \text{sample } z_{ji}^T \sim Mult(\theta) \\ \text{sample } w_{ji}^T \sim Mult(\psi, z_{ji}^T) \end{cases} \end{cases}$

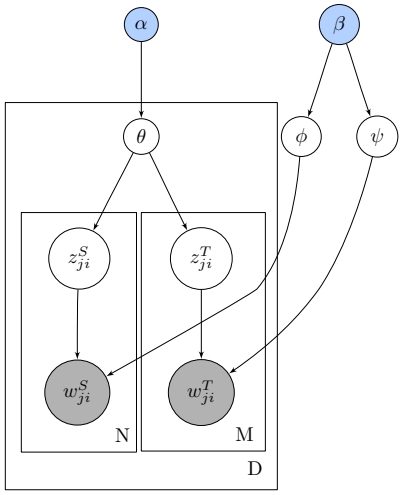

Figure 1: The standard bilingual LDA model

Having one common θ for both of the related documents implies parallelism between the texts. This observation does not completely hold for comparable corpora with topically aligned texts. To train the

model we use Gibbs sampling, similar to the sampling method for monolingual LDA, with parameters α and β set to $50/K$ and 0.01 respectively, where K denotes the number of topics. After the training we end up with a set of ϕ and ψ word-topic probability distributions that are used for the calculations of the word associations.

If we are given a source vocabulary W^S, then the distribution ϕ of sampling a new token as word $w_i \in W^S$ from a topic z_k can be obtained as follows:

$$P(w_i|z_k) = \phi_{k,i} = \frac{n_k^{(w_i)} + \beta}{\sum_{j=1}^{|W^S|} n_k^{(w_j)} + W^S\beta} \quad (1)$$

where, for a word w_i and a topic z_k, $n_k^{(w_i)}$ denotes the total number of times that the topic z_k is assigned to the word w_i from the vocabulary W^S, β is a symmetric Dirichlet prior, $\sum_{j=1}^{|W_S|} n_k^{(w_j)}$ is the total number of words assigned to the topic z_k, and $|W^S|$ is the total number of distinct words in the vocabulary. The formula for a set of ψ word-topic probability distributions for the target side of a corpus is computed in an analogical manner.

3.2 Main Framework

Once we derive a shared set of topics along with language-specific distributions of words over topics, it is possible to use them for the computation of the similarity between words in different languages.

3.2.1 KL Method

The similarity between a source word w_1 and a target word w_2 is measured by the extent to which they share the same topics, *i.e.*, by the extent that their conditional topic distributions are similar. One way of expressing similarity is the Kullback-Leibler (**KL**) divergence, already used in a monolingual setting in (Steyvers and Griffiths, 2007). The similarity between two words is based on the similarity between $\chi^{(1)}$ and $\chi^{(2)}$, the similarity of conditional topic distributions for words w_1 and w_2, where $\chi^{(1)} = P(Z|w_1)$[1] and $\chi^{(2)} = P(Z|w_2)$. We have to calculate the probabilities $P(z_j|w_i)$, which describe a probability that a given word is assigned to a particular topic. If we apply Bayes' rule, we get $P(Z|w) = \frac{P(w|Z)P(Z)}{P(w)}$, where $P(Z)$ and $P(w)$ are prior distributions for topics and words respectively. $P(Z)$ is a uniform distribution for the BiLDA model, whereas this assumption clearly does not hold for topic models with a non-uniform topic prior. $P(w)$ is given by $P(w) = P(w|Z)P(Z)$. If the assumption of uniformity for $P(Z)$ holds, we can write:

$$P(z_j|w_i) \propto \frac{P(w_i|z_j)}{Norm_\phi} = \frac{\phi_{j,i}}{Norm_\phi} \quad (2)$$

for an English word w_i, and:

$$P(z_j|w_i) \propto \frac{P(w_i|z_j)}{Norm_\psi} = \frac{\psi_{j,i}}{Norm_\psi} \quad (3)$$

for a French word w_i, where $Norm_\phi$ denotes the normalization factor $\sum_{j=1}^{K} P(w_i|z_j)$, *i.e.*, the sum of all probabilities ϕ (or probabilities ψ for $Norm_\psi$) for the currently observed word w_i.

We can then calculate the KL divergence as follows:

$$KL(\chi^{(1)}, \chi^{(2)}) \propto \sum_{j=1}^{K} \frac{\phi_{j,1}}{Norm_\phi} \log \frac{\phi_{j,1}/Norm_\phi}{\psi_{j,2}/Norm_\psi} \quad (4)$$

3.2.2 Cue Method

An alternative, more straightforward approach (called the **Cue** method) tries to express similarity between two words emphasizing the associative relation between two words in a more natural way. It models the probability $P(w_2|w_1)$, *i.e.*, the probability that a target word w_2 will be generated as a response to a cue source word w_1. For the BiLDA model we can write:

$$\begin{aligned} P(w_2|w_1) &= \sum_{j=1}^{K} P(w_2|z_j)P(z_j|w_1) \\ &= \sum_{j=1}^{K} \psi_{j,2} \frac{\phi_{j,1}}{Norm_\phi} \end{aligned} \quad (5)$$

This conditioning automatically compromises between word frequency and semantic relatedness (Griffiths et al., 2007), since higher frequency words tend to have higher probabilities across all topics, but the distribution over topics $P(z_j|w_1)$ ensures that semantically related topics dominate the sum.

[1] $P(Z|w_1)$ refers to a set of all conditional topic distributions $P(z_j|w_1)$

3.2.3 TI Method

The last approach borrows an idea from information retrieval and constructs word vectors over a shared latent topic space. Values within vectors are the *TF-ITF* (term frequency - inverse topic frequency) scores which are calculated in a completely analogical manner as the *TF-IDF* scores for the original word-document space (Manning and Schütze, 1999). If we are given a source word w_i, $n_{k,S}^{(w_i)}$ denotes the number of times the word w_i is associated with a source topic z_k. *Term frequency (TF)* of the source word w_i for the source topic z_k is given as:

$$TF_{i,k} = \frac{n_{k,S}^{(w_i)}}{\sum_{w_j \in W^S} n_{k,S}^{(w_j)}} \quad (6)$$

Inverse topical frequency (ITF) measures the general importance of the source word w_i across all source topics. Rare words are given a higher importance and thus they tend to be more descriptive for a specific topic. The inverse topical frequency for the source word w_i is calculated as[2]:

$$ITF_i = \log \frac{K}{1 + |k : n_{k,S}^{(w_i)} > 0|} \quad (7)$$

The final *TF-ITF* score for the source word w_i and the topic z_k is given by $TF{-}ITF_{i,k} = TF_{i,k} \cdot ITF_i$. We calculate the *TF-ITF* scores for target words associated with target topics in an analogical manner. Source and target words share the same K-dimensional topical space, where K-dimensional vectors consisting of the *TF-ITF* scores are built for all words. The standard cosine similarity metric is then used to find the most similar word vectors from the target vocabulary for a source word vector. We name this method the **TI** method. For instance, given a source word w_1 represented by a K-dimensional vector S^1 and a target word w_2 represented by a K-dimensional vector T^2, the similarity between the two words is calculated as follows:

$$cos(w_1, w_2) = \frac{\sum_{k=1}^{K} S_k^1 \cdot T_k^2}{\sqrt{\sum_{k=1}^{K} \left(S_k^1\right)^2} \cdot \sqrt{\sum_{k=1}^{K} \left(T_k^2\right)^2}} \quad (8)$$

4 Results and Discussion

As our training corpus, we use the English-Italian Wikipedia corpus of 18,898 document pairs, where each aligned pair discusses the same subject. In order to reduce data sparsity, we keep only lemmatized noun forms for further analysis. Our Italian vocabulary consists of 7,160 nouns, while our English vocabulary contains 9,166 nouns. The subset of the 650 most frequent terms was used for testing. We have used the *Google Translate* tool for evaluations. As our baseline system, we use the cosine similarity between Italian word vectors and English word vectors with *TF-IDF* scores in the original word-document space (**Cos**), with aligned documents.

Table 1 shows the Precision@1 scores (the percentage of words where the first word from the list of translations is the correct one) for all three approaches (**KL, Cue** and **TI**), for different number of topics K. Although **KL** is designed specifically to measure the similarity of two distributions, its results are significantly below those of the **Cue** and **TI**, whose performances are comparable. Whereas the latter two methods yield the highest results around the 2,000 topics mark, the performance of **KL** increases linearly with the number of topics. This is an undesirable result as good results are computationally hard to get.

We have also detected that we are able to boost overall scores if we combine two methods. We have opted for the two best methods (**TI+Cue**), where overall score is calculated by $Score = \lambda \cdot Score_{Cue} + Score_{TI}$.[3] We also provide the results obtained by linearly combining (with equal weights) the cosine similarity between *TF-ITF* vectors with that between *TF-IDF* vector (**TI+Cos**).

In a more lenient evaluation setting we employ the *mean reciprocal rank* (MRR) (Voorhees, 1999). For a source word w, $rank_w$ denotes the rank of its correct translation within the retrieved list of potential translations. MRR is then defined as follows:

[2] Stronger association with a topic is modeled by setting a higher *threshold* value in $n_{k,S}^{(w_i)} > threshold$, where we have chosen 0.

[3] The value of λ is empirically set to 10

K	KL	Cue	TI	TI+Cue	TI+Cos
200	0.3015	0.1800	0.3169	0.2862	0.5369
500	0.2846	0.3338	0.3754	0.4000	0.5308
800	0.2969	0.4215	0.4523	0.4877	0.5631
1200	0.3246	0.5138	0.4969	0.5708	0.5985
1500	0.3323	0.5123	0.4938	0.5723	0.5908
1800	0.3569	0.5246	0.5154	0.5985	**0.6123**
2000	0.3954	0.5246	**0.5385**	**0.6077**	0.6046
2200	0.4185	**0.5323**	0.5169	0.5908	0.6015
2600	0.4292	0.4938	0.5185	0.5662	0.5907
3000	0.4354	0.4554	0.4923	0.5631	0.5953
3500	**0.4585**	0.4492	0.4785	0.5738	0.5785

Table 1: Precision@1 scores for the test subset of the IT-EN Wikipedia corpus (baseline precision score: 0.5031)

$$MRR = \frac{1}{|V|} \sum_{w \in V} \frac{1}{rank_w} \tag{9}$$

where V denotes the set of words used for evaluation. We kept only the top 20 candidates from the ranked list. Table 2 shows the MRR scores for the same set of experiments.

K	KL	Cue	TI	TI+Cue	TI+Cos
200	0.3569	0.2990	0.3868	0.4189	0.5899
500	0.3349	0.4331	0.4431	0.4965	0.5808
800	0.3490	0.5093	0.5215	0.5733	0.6173
1200	0.3773	0.5751	0.5618	0.6372	0.6514
1500	0.3865	0.5756	0.5562	0.6320	0.6435
1800	0.4169	0.5858	0.5802	0.6581	**0.6583**
2000	0.4561	0.5841	**0.5914**	**0.6616**	0.6548
2200	0.4686	**0.5898**	0.5753	0.6471	0.6523
2600	0.4763	0.5550	0.5710	0.6268	0.6416
3000	0.4848	0.5272	0.5572	0.6257	0.6465
3500	**0.5022**	0.5199	0.5450	0.6238	0.6310

Table 2: MRR scores for the test subset of the IT-EN Wikipedia corpus (baseline MRR score: 0.5890)

Topic models have the ability to build clusters of words which might not always co-occur together in the same textual units and therefore add extra information of potential relatedness. Although we have presented results for a document-aligned corpus, the framework is completely generic and applicable to other topically related corpora.

Again, the **KL** method has the weakest performance among the three methods based on the word-topic distributions, while the other two methods seem very useful when combined together or when combined with the similarity measure used in the original word-document space. We believe that the results are in reality even higher than presented in the paper, due to errors in the evaluation tool (*e.g.*, the Italian word *raggio* is correctly translated as *ray*, but Google Translate returns *radius* as the first translation candidate).

All proposed methods retrieve lists of semantically related words, where synonymy is not the only semantic relation observed. Such lists provide comprehensible and useful contextual information in the target language for the source word, even when the correct translation candidate is missing, as might be seen in table 3.

(1) **romanzo** (novel)	(2) **paesaggio** (landscape)	(3) **cavallo** (horse)
writer	tourist	horse
novella	painting	stud
novellette	landscape	horseback
humorist	local	hoof
novelist	visitor	breed
essayist	hut	stamina
penchant	draftsman	luggage
formative	tourism	mare
foreword	attraction	riding
author	vegetation	pony

Table 3: Lists of the top 10 translation candidates, where the correct translation is not found (column 1), lies hidden lower in the list (2), and is retrieved as the first candidate (3); K=2000; TI+Cue.

5 Conclusion

We have presented a generic, language-independent framework for mining translations of words from latent topic models. We have proven that topical knowledge is useful and improves the quality of word translations. The quality of translations depends only on the quality of a topic model and its ability to find latent relations between words. Our next steps involve experiments with other topic models and other corpora, and combining this unsupervised approach with other tools for lexicon extraction and synonymy detection from unrelated and comparable corpora.

Acknowledgements

The research has been carried out in the framework of the TermWise Knowledge Platform (IOF-KP/09/001) funded by the Industrial Research Fund K.U. Leuven, Belgium, and the Flemish SBO-IWT project *AMASS++* (SBO-IWT 0060051).

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining*, pages 57–64.

Mona T. Diab and Steve Finch. 2000. A statistical translation model using comparable corpora. In *Proceedings of the 2000 Conference on Content-Based Multimedia Information Access (RIAO)*, pages 1500–1508.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 526–533.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review*, 114(2):211–244.

Zellig S. Harris. 1954. Distributional structure. In *Word 10 (23)*, pages 146–162.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition - Volume 9*, ULA '02, pages 9–16.

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

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference*, pages 1155–1156.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, ACL '95, pages 320–322.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Ellen M. Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of the Eighth TExt Retrieval Conference (TREC-8)*.