Orthonormal Explicit Topic Analysis for Cross-lingual Document Matching

John Philip McCrae
University Bielefeld
Inspiration 1
Bielefeld, Germany

Philipp Cimiano
University Bielefeld
Inspiration 1
Bielefeld, Germany

Roman Klinger
University Bielefeld
Inspiration 1
Bielefeld, Germany

{jmccrae,cimiano,rklinger}@cit-ec.uni-bielefeld.de

Abstract

Cross-lingual topic modelling has applications in machine translation, word sense disambiguation and terminology alignment. Multilingual extensions of approaches based on latent (LSI), generative (LDA, PLSI) as well as explicit (ESA) topic modelling can induce an interlingual topic space allowing documents in different languages to be mapped into the same space and thus to be compared across languages. In this paper, we present a novel approach that combines latent and explicit topic modelling approaches in the sense that it builds on a set of explicitly defined topics, but then computes latent relations between these. Thus, the method combines the benefits of both explicit and latent topic modelling approaches. We show that on a cross-lingual mate retrieval task, our model significantly outperforms LDA, LSI, and ESA, as well as a baseline that translates every word in a document into the target language.

1 Introduction

Cross-lingual document matching is the task of, given a *query* document in some source language, estimating the similarity to a document in some target language. This task has important applications in machine translation (Palmer et al., 1998; Tam et al., 2007), word sense disambiguation (Li et al., 2010) and ontology alignment (Spiliopoulos et al., 2007). An approach that has become quite popular in recent years for cross-lingual document matching is Explicit Semantics Analysis (ESA, Gabrilovich and Markovitch (2007)) and its cross-lingual extension CL-ESA (Sorg and Cimiano, 2008). ESA indexes documents by mapping them into a topic space defined by their similarity to predefined explicit topics – generally articles from an encyclopaedia – in such a way that there is a one-to-one correspondence between topics and encyclopedic entries. CL-ESA extends this to the multilingual case by exploiting a background document collection that is aligned across languages, such as Wikipedia. A feature of ESA and its extension CL-ESA is that, in contrast to latent (e.g. LSI, Deerwester et al. (1990)) or generative topic models (such as LDA, Blei et al. (2003)), it requires no training and, nevertheless, has been demonstrated to outperform LSI and LDA on cross-lingual retrieval tasks (Cimiano et al., 2009).

A key choice in Explicit Semantic Analysis is the document space that will act as the topic space. The standard choice is to regard all articles from a background document collection – Wikipedia articles are a typical choice – as the topic space. However, it is crucial to ensure that these topics cover the semantic space evenly and completely. In this paper, we present an alternative approach where we remap the semantic space defined by the topics in such a manner that it is orthonormal. In this way, each document is mapped to a topic that is distinct from all other topics. Such a mapping can be considered as equivalent to a variant of Latent Semantic Indexing (LSI) with the main difference that our model exploits the matrix that maps topic vectors back into document space, which is normally discarded in LSI-based approaches. We dub our model *ONETA* (OrthoNormal Explicit Topic Analysis) and empirically show that on a cross-lingual retrieval

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1732–1740,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

task it outperforms ESA, LSI, and Latent Dirichlet Allocation (LDA) as well as a baseline consisting of translating each word into the target language, thus reducing the task to a standard monolingual matching task. In particular, we quantify the effect of different approximation techniques for computing the orthonormal basis and investigate the effect of various methods for the normalization of frequency vectors.

The structure of the paper is as follows: we situate our work in the general context of related work on topic models for cross-lingual document matching in Section 2. We present our model in Section 3 and present our experimental results and discuss these results in Section 4.

2 Related Work

The idea of applying topic models that map documents into an interlingual topic space seems a quite natural and principled approach to tackle several tasks including the cross-lingual document retrieval problem.

Topic modelling is the process of finding a representation of a document d in a lower dimensional space $\mathbb{R}^K$ where each dimension corresponds to one *topic* that abstracts from specific words and thus allows us to detect deeper semantic similarities between documents beyond the computation of the pure overlap in terms of words.

Three main variants of document models have been mainly considered for cross-lingual document matching:

Latent methods such as Latent Semantic Indexing (LSI, Deerwester et al. (1990)) induce a decomposition of the term-document matrix in a way that reduces the dimensionality of the documents, while minimizing the error in reconstructing the training data. For example, in Latent Semantic Indexing, a term-document matrix is approximated by a partial singular value decomposition, or in Non-Negative Matrix Factorization (NMF, Lee and Seung (1999)) by two smaller non-negative matrices. If we append comparable or equivalent documents in multiple languages together before computing the decomposition as proposed by Dumais et al. (1997) then the topic model is essentially cross-lingual allowing to compare documents in different languages once they have been mapped into the topic space.

Probabilistic or generative methods instead attempt to induce a (topic) model that has the highest likelihood of generating the documents actually observed during training. As with latent methods, these topics are thus interlingual and can generate words/terms in different languages. Prominent representatives of this type of method are Probabilistic Latent Semantic Indexing (PLSI, Hofmann (1999)) or Latent Dirichlet Allocation (LDA, Blei et al. (2003)), both of which can be straightforwardly extended to the cross-lingual case (Mimno et al., 2009).

Explicit topic models make the assumption that topics are explicitly given instead of being induced from training data. Typically, a background document collection is assumed to be given whereby each document in this corpus corresponds to one topic. A mapping from document to topic space is calculated by computing the similarity of the document to every document in the topic space. A prominent example for this kind of topic modelling approach is Explicit Semantic Analysis (ESA, Gabrilovich and Markovitch (2007)).

Both latent and generative topic models attempt to find topics from the data and it has been found that in some cases they are equivalent (Ding et al., 2006). However, this approach suffers from the problem that the topics might be artifacts of the training data rather than coherent semantic topics. In contrast, explicit topic methods can use a set of topics that are chosen to be well-suited to the domain. The principle drawback of this is that the method for choosing such explicit topics by selecting documents is comparatively crude. In general, these topics may be overlapping and poorly distributed over the semantic topic space. By comparison, our method takes the advantage of the pre-specified topics of explicit topic models, but incorporates a training step to learn latent relations between these topics.

3 Orthonormal explicit topic analysis

Our approach follows Explicit Semantic Analysis in the sense that it assumes the availability of a background document collection $B = \{b_1, b_2, ..., b_N\}$ consisting of textual representations. The mapping into the explicit topic space is defined by a language-specific function Φ that maps documents into $\mathbb{R}^N$ such that the j^{th} value in the vector is given by some *association measure* $\phi_j(d)$ for each background document b_j. Typical choices for this association measure ϕ are the sum of the TF-IDF scores or an information retrieval relevance scoring function such as BM-25 (Sorg and Cimiano, 2010).

For the case of TF-IDF, the value of the j-th element of the topic vector is given by:

$$\phi_j(d) = \overrightarrow{\text{tf-idf}}(b_j)^{\mathrm{T}}\ \overrightarrow{\text{tf-idf}}(d)$$

Thus, the mapping function can be represented as the product of a TF-IDF vector of document d multiplied by an $W \times N$ matrix, $\mathbf{X}$, each element of which contains the TF-IDF value of word i in document b_j:

$$\Phi(d) = \begin{pmatrix} \overrightarrow{\text{tf-idf}}(b_1)^{\mathrm{T}} \\ \vdots \\ \overrightarrow{\text{tf-idf}}(b_N)^{\mathrm{T}} \end{pmatrix} \overrightarrow{\text{tf-idf}}(d) = \mathbf{X}^{\mathrm{T}} \cdot \overrightarrow{\text{tf-idf}}(d)$$

For simplicity, we shall assume from this point on that all vectors are already converted to a TF-IDF or similar numeric vector form. In order to compute the similarity between two documents d_i and d_j, typically the cosine-function (or the normalized dot product) between the vectors $\Phi(d_i)$ and $\Phi(d_j)$ is computed as follows:

$$\text{sim}(d_i, d_j) = \cos(\Phi(d_i), \Phi(d_j)) = \frac{\Phi(d_i)^{\mathrm{T}}\Phi(d_j)}{||\Phi(d_i)||||\Phi(d_j)||}$$

If we represent the above using our above defined $W \times N$ matrix $\mathbf{X}$ then we get:

$$\text{sim}(d_i, d_j) = \cos(\mathbf{X}^{\mathrm{T}} d_i, \mathbf{X}^{\mathrm{T}} d_j) = \frac{d_i^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}} d_j}{||\mathbf{X}^{\mathrm{T}} d_i||||\mathbf{X}^{\mathrm{T}} d_j||}$$

The key challenge with ESA is choosing a good background document collection $B = \{b_1, ..., b_N\}$. A simple minimal criterion for a good background document collection is that each document in this collection should be maximally similar to itself and less similar to any other document:

$$\forall i \neq j \;\; 1 = \text{sim}(b_j, b_j) > \text{sim}(b_i, b_j) \geq 0$$

While this criterion is trivially satisfied if we have no duplicate documents in our collection, our intuition is that we should choose a background collection that maximizes the *slack margin* of this inequality, i.e. $|\text{sim}(b_j, b_j) - \text{sim}(b_i, b_j)|$. We can see that maximizing this margin for all i,j is the same as minimizing the *semantic overlap* of the background documents, which is given as follows:

$$\text{overlap}(B) = \sum_{\substack{i = 1, \dots, N \\ j = 1, \dots, N \\ i \neq j}} \text{sim}(b_i, b_j)$$

We first note that we can, without loss of generality, normalize our background documents such that $||Xb_j|| = 1$ for all j, and in this case we can redefine the semantic overlap as the following matrix expression[1]

$$\text{overlap}(\mathbf{X}) = ||\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{I}||_1$$

It is trivial to verify that this equation has a minimum when $\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$. This is the case when the topics are *orthonormal*:

$$(\mathbf{X}^{\mathrm{T}} b_i)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}} b_j) = 0 \quad \text{if } i \neq j$$

$$(\mathbf{X}^{\mathrm{T}} b_i)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}} b_i) = 1$$

Unfortunately, this is not typically the case as the documents have significant word overlap as well as semantic overlap. Our goal is thus to apply a suitable transformation to $\mathbf{X}$ with the goal of ensuring that the orthogonality property holds.

Assuming that this transformation of $\mathbf{X}$ is done by multiplication with some other matrix $\mathbf{A}$, we can define the learning problem as finding that matrix $\mathbf{A}$ such that:

$$(\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{X}) = I$$

[1] $||\mathbf{A}||_p = \sum_{i,j} |a_{ij}|^p$ is the p-norm. $||\mathbf{A}||_{\mathcal{F}} = \sqrt{||\mathbf{A}||_2}$ is the Frobenius norm.

If we have the case that $W \geq N$ and that the rank of $\mathbf{X}$ is N, then $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible and thus $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ is the solution to this problem.[2]

We define the projection function of a document d, represented as a normalized term frequency vector, as follows:

$$\Phi_{\mathrm{ONETA}}(d) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}d$$

For the cross-lingual case we assume that we have two sets of background documents of equal size, $B^1 = \{b_1^1, \ldots, b_N^1\}$, $B^2 = \{b_1^2, \ldots, b_N^2\}$ in languages l_1 and l_2, respectively and that these documents are aligned such that for every index i, b_i^1 and b_i^2 are documents on the same topic in each language. Using this we can construct a projection function for each language which maps into the same topic space. Thus, as in CL-ESA, we obtain the cross-lingual similarity between a document d_i in language l_1 and a document d_j in language l_2 as follows:

$$\mathrm{sim}(d_i, d_j) = \cos(\Phi^{l_1}_{\mathrm{ONETA}}(d_i), \Phi^{l_2}_{\mathrm{ONETA}}(d_j))$$

We note here that we assume that Φ could be represented as a symmetric inner product of two vectors. However, for many common choices of association measures, including BM25, this is not the case. In this case the expression $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ can be replaced with a kernel matrix specifying the association of each background document to each other background document.

3.1 Relationship to Latent Semantic Indexing

In this section we briefly clarify the relationship between our method ONETA and Latent Semantic Indexing. Latent Semantic Indexing defines a mapping from a document represented as a term frequency vector to a vector in $\mathbb{R}^K$. This transformation is defined by means of calculating the singular value decomposition (SVD) of the matrix $\mathbf{X}$ as above, namely

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

Where $\mathbf{\Sigma}$ is diagonal and $\mathbf{U}$ $\mathbf{V}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$., respectively. Let $\mathbf{\Sigma}_K$ denote the $K \times K$ submatrix containing the largest eigenvalues, and $\mathbf{U}_K$,$\mathbf{V}_K$ denote the corresponding eigenvectors. Thus LSI can be defined as:

$$\Phi_{\mathrm{LSI}}(d) = \mathbf{\Sigma}_K^{-1}\mathbf{U}_K d$$

With regards to orthonormalized topics, we see that using the SVD, we can simply derive the following:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} = \mathbf{V}\mathbf{\Sigma^{-1}}\mathbf{U}^{\mathrm{T}}$$

When we set $K = N$ and thus choose the maximum number of topics, ONETA is equivalent to LSI modulo the fact that it multiplies the resulting topic vector by $\mathbf{V}$, thus projecting back into document space, i.e. into explicit topics.

In practice, both methods differ significantly in that the approximations they make are quite different. Furthermore, in the case that $W \gg N$ and $\mathbf{X}$ has n non-zeroes, the calculation of the SVD is of complexity $\mathcal{O}(nN + WN^2)$ and requires $\mathcal{O}(WN)$ bytes of memory. In contrast, ONETA requires computation time of $\mathcal{O}(N^a)$ for $a > 2$, which is the complexity of the matrix inversion algorithm[3], and only $\mathcal{O}(n + N^2)$ bytes of memory.

3.2 Approximations

The computation of the inverse has a complexity that, using current practical algorithms, is approximately cubic and as such the time spent calculating the inverse can grow very quickly. There are several methods for obtaining an approximate inverse. The most commonly used are based on the SVD or eigendecomposition of the matrix. As $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is symmetric positive definite, it holds that:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^{\mathrm{T}}$$

Where $\mathbf{U}$ are the eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{\Sigma}$ is a diagonal matrix of the eigenvalues. With $\mathbf{U}_K$,$\mathbf{\Sigma}_K$

[2] In the case that the matrix is not invertible we can instead solve $||\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{A} - \mathbf{I}||_{\mathcal{F}}$, which has a minimum at $\mathbf{A} = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^{\mathrm{T}}$ where $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ is the singular value decomposition of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

As usual we do not in fact compute the inverse for our experiments, but instead the LU Decomposition and solve by Gaussian elimination at test time.

[3] Algorithms with $a = 2.3727$ are known but practical algorithms have $a = 2.807$ or $a = 3$ (Coppersmith and Winograd, 1990)

as the first K eigenvalues and eigenvectors, respectively, we have:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \simeq \mathbf{U}_K \mathbf{\Sigma}_K^{-1} \mathbf{U}_K^{\mathrm{T}} \tag{1}$$

We call this the *orthonormal eigenapproximation* or *ON-Eigen*. The complexity of calculating $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ from this is $\mathcal{O}(N^2 K + Nn)$, where n is the number of non-zeros in $\mathbf{X}$.

Similarly, using the formula derived in the previous section we can derive an approximation of the full model as follows:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \simeq \mathbf{U}_K \mathbf{\Sigma}_K^{-1} \mathbf{V}_K^{\mathrm{T}} \tag{2}$$

We call this approximation *Explicit LSI* as it first maps into the latent topic space and then into the explicit topic space.

We can consider another approximation by noticing that $\mathbf{X}$ is typically very sparse and moreover some rows of $\mathbf{X}$ have significantly fewer non-zeroes than others (these rows are for terms with low frequency). Thus, if we take the first N_1 columns (documents) in $\mathbf{X}$, it is possible to rearrange the rows of $\mathbf{X}$ with the result that there is some W_1 such that rows with index greater than W_1 have only zeroes in the columns up to N_1. In other words, we take a subset of N_1 documents and enumerate the words in such a way that the terms occurring in the first N_1 documents are enumerated $1, \ldots, W_1$. Let $N_2 = N - N_1$, $W_2 = W - W_1$. The result of this row permutation does not affect the value of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and we can write the matrix $\mathbf{X}$ as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

where $\mathbf{A}$ is a $W_1 \times N_1$ matrix representing term frequencies in the first N_1 documents, $\mathbf{B}$ is a $W_1 \times N_2$ matrix containing term frequencies in the remaining documents for terms that are also found in the first N_1 documents, and $\mathbf{C}$ is a $W_2 \times N_2$ containing the frequency of all terms not found in the first N_1 documents.

Application of the well-known divide-and-conquer formula (Bernstein, 2005, p. 159) for matrix inversion yields the following easily verifiable matrix identity, given that we can find $\mathbf{C}'$ such that $\mathbf{C}'\mathbf{C} = \mathbf{I}$.

$$\begin{pmatrix} (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}} & -(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{B}\mathbf{C}' \\ \mathbf{0} & \mathbf{C}' \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} = \mathbf{I} \tag{3}$$

We denote the above equation using a matrix $\mathbf{L}$ as $\mathbf{L}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$. We note that $\mathbf{L} \neq (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}$, but for any document vector d that is representable as a linear combination of the background document set (i.e., columns of $\mathbf{X}$) we have that $\mathbf{L}d = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}d$ and in this sense $\mathbf{L} \simeq (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$.

We further relax the assumption so that we only need to find a $\mathbf{C}'$ such that $\mathbf{C}'\mathbf{C} \simeq \mathbf{I}$. For this, we first observe that $\mathbf{C}$ is very sparse as it contains only terms not contained in the first N_1 documents and we notice that very sparse matrices tend to be approximately orthogonal, hence suggesting that it should be very easy to find a left-inverse of $\mathbf{C}$. The following lemma formalizes this intuition:

Lemma: If $\mathbf{C}$ is a $W \times N$ matrix with M non-zeros, distributed randomly and uniformly across the matrix, and all the non-zeros are 1, then $\mathbf{D}\mathbf{C}^{\mathrm{T}}\mathbf{C}$ has an expected value on each non-diagonal value of $\frac{M}{N^2}$ and a diagonal value of 1 if $\mathbf{D}$ is the diagonal matrix whose values are given by $||c_i||^{-2}$, the square of the norm of the corresponding column of $\mathbf{C}$.

Proof: We simply observe that if $\mathbf{D}' = \mathbf{D}\mathbf{C}^{\mathrm{T}}\mathbf{C}$, then the $(i, j)^{\mathrm{th}}$ element of $\mathbf{D}'$ is given by

$$d_{ij} = \frac{c_i^{\mathrm{T}} c_j}{||c_i||^2}$$

If $i \neq j$ then the $c_i^{\mathrm{T}} c_j$ is the number of non-zeroes overlapping in the i^{th} and j^{th} column of $\mathbf{C}$ and under a uniform distribution we expect this to be $\frac{M^2}{N^3}$. Similarly, we expect the column norm to be $\frac{M}{N}$ such that the overall expectation is $\frac{M}{N^2}$. The diagonal value is clearly equal to 1.■

As long as $\mathbf{C}$ is very sparse, we can use the following approximation, which can be calculated in $\mathcal{O}(M)$ operations, where M is the number of non-zeroes.

$$\mathbf{C}' \simeq \begin{pmatrix} ||c_1||^{-2} & & 0 \\ & \ddots & \\ 0 & & ||c_{N_2}||^{-2} \end{pmatrix} \mathbf{C}^{\mathrm{T}}$$

We call this method *L-Solve*. The complexity of calculating a left-inverse by this method is of

	Document Normalization	
Frequency Normalization	No	Yes
TF	0.31	0.78
Relative	0.23	0.42
TFIDF	0.21	0.63
SQRT	0.28	0.66

Table 1: Effect of Term Frequency and Document Normalization on Top-1 Precision

order $\mathcal{O}(N_1^a)$, being much more efficient than the eigenvalue methods. However, it is potentially more error-prone as it requires that a left-inverse of $\mathbf{C}$ exists. On real data this might be violated if we do not have linear independence of the rows of $\mathbf{C}$, for example if $W_2 < N_2$ or if we have even one document which has only words that are also contained in the first N_1 documents and hence there is a row in $\mathbf{C}$ that consists of zeros only. This can be solved by removing documents from the collection until $\mathbf{C}$ is row-wise linear independent.[4]

3.3 Normalization

A key factor in the effectiveness of topic-based methods is the appropriate normalization of the elements of the document matrix $\mathbf{X}$. This is even more relevant for orthonormal topics as the matrix inversion procedure can be very sensitive to small changes in the matrix. In this context, we consider two forms of normalization, term and document normalization, which can also be considered as row/column normalizations of $\mathbf{X}$.

A straightforward approach to normalization is to normalize each column of $\mathbf{X}$ to obtain a matrix as follows:

$$\mathbf{X}' = \left(\frac{x_1}{||x_1||} \cdots \frac{x_N}{||x_N||}\right)$$

If we calculate $\mathbf{X'}^{\mathrm{T}}\mathbf{X}' = \mathbf{Y}$ then we get that the (i, j)-th element of $\mathbf{Y}$ is:

$$y_{ij} = \frac{x_i^{\mathrm{T}} x_j}{||x_i|| ||x_j||}$$

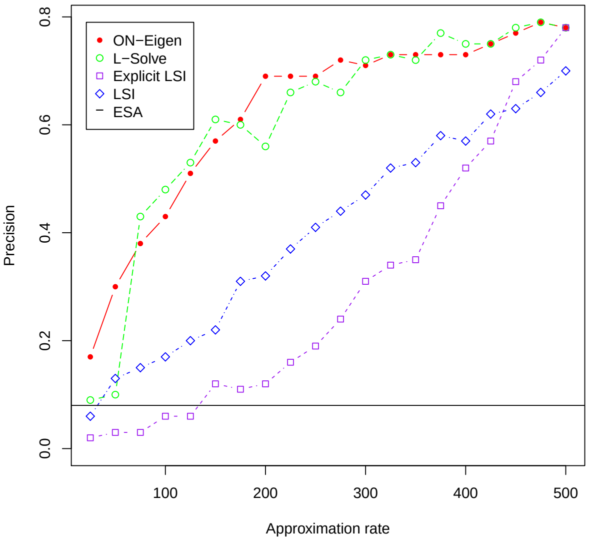

Figure 1: Effect on Top-1 Precision by various approximation method

Thus, the diagonal of $\mathbf{Y}$ consists of ones only and due to the Cauchy-Schwarz inequality we have that $|y_{ij}| \leq 1$, with the result that the matrix $\mathbf{Y}$ is already close to $\mathbf{I}$. Formally, we can use this to state a bound on $||\mathbf{X'}^{\mathrm{T}}\mathbf{X}' - \mathbf{I}||_{\mathcal{F}}$, but in practice it means that the orthonormalizing matrix has more small or zero values.

A further option for normalization is to consider some form of term frequency normalization. For term frequency normalization, we use *TF* (tf_{wn}), *Relative* ($\frac{\mathrm{tf}_{wn}}{F_w}$), *TFIDF* ($\mathrm{tf}_{wn}\log(\frac{N}{\mathrm{df}_w})$), and *SQRT* ($\frac{\mathrm{tf}_{wn}}{\sqrt{F_w}}$). Here, tf_{wn} is the term frequency of word w in document n, F_w is the total frequency of word w in the corpus, and df_w is the number of documents containing the words w. The first three of these normalizations have been chosen as they are widely used in the literature. The SQRT normalization has been shown to be effective for explicit topic methods in previous experiments not reported here.

4 Experiments and Results

For evaluation, we consider a cross-lingual mate retrieval task from English/Spanish on the basis of Wikipedia as aligned corpus. The goal is to, for each document of a test set, retrieve the aligned document or *mate*. For each test document, on the basis of

[4] In the experiments in the next section we discarded 4.2% of documents at $N_1 = 1000$ and 47% of documents at $N_1 = 5000$

Method	Top-1 Prec.	Top-5 Prec.	Top-10 Prec.	MRR	Time	Memory
ONETA L-Solve ($N_1 = 1000$)	0.290	0.501	0.596	0.390	73s	354MB
ONETA L-Solve ($N_1 = 2000$)	0.328	0.531	0.600	0.423	2m18s	508MB
ONETA L-Solve ($N_1 = 3000$)	0.462	0.662	0.716	0.551	4m12s	718MB
ONETA L-Solve ($N_1 = 4000$)	0.599	0.755	0.781	0.667	7m44s	996MB
ONETA L-Solve ($N_1 = 5000$)	0.695	0.817	0.843	0.750	12m28s	1.30GB
ONETA L-Solve ($N_1 = 6000$)	0.773	0.883	0.905	0.824	18m40s	1.69GB
ONETA L-Solve ($N_1 = 7000$)	0.841	0.928	0.937	0.881	26m31s	2.14GB
ONETA L-Solve ($N_1 = 8000$)	0.896	0.961	0.968	0.927	37m39s	2.65GB
ONETA L-Solve ($N_1 = 9000$)	0.924	0.981	0.987	0.950	52m52s	3.22GB
ONETA (No Approximation)	**0.929**	**0.987**	**0.990**	**0.956**	57m10s	3.42GB
Word Translation	0.751	0.884	0.916	0.812	n/a	n/a
ESA (SQRT Normalization)	0.498	0.769	0.835	0.621	72s	284MB
LDA (K=1000)	0.287	0.568	0.659	0.417	4h12m	8.4GB
LSI (K=4000)	0.615	0.756	0.783	0.676	13h51m	19.7GB
ONETA + Word Translation	**0.932**	**0.987**	**0.993**	**0.958**	n/a	n/a

Table 2: Result on large-scale mate-finding studies for English to Spanish matching

the similarity of the query document to all indexed documents, we compute the value rank_i indicating at which position the mate of the i^{th} document occurs. We use two metrics: *Top-k Precision*, defined as the percentage of documents for which the mate is retrieved among the first k elements, and *Minimum Reciprocal Rank*, defined as

$$\text{MRR} = \sum_{i \in \text{test}} \frac{1}{\text{rank}_i}$$

For our experiments, we first extracted a subset of documents (every 20th) from Wikipedia, filtering this set down to only those that have aligned pages in both English and Spanish with a minimum length of 100 words. This gives us 10,369 aligned documents in total, which form the background document collection B. We split this data into a training and test set of 9,332 and 1,037 documents, respectively. We then removed all words whose total frequencies were below 50. This resulted in corpus of 6.7 millions words in English and 4.2 million words in Spanish.

Normalization Methods: In order to investigate the impact of different normalization methods, we ran small-scale experiments using the first 500 documents from our dataset to train ONETA and then evaluate the resulting models on the mate-finding task on 100 unseen documents. The results are presented in Table 1, which shows the Top-1 Precision for the different normalization methods. We see that the effect of applying document normalization in all cases improves the quality of the overall result. Surprisingly, we do not see the same result for frequency normalization yielding the best result for the case where we do no normalization at all[5] . In the remaining experiments we thus employ document normalization and no term frequency normalization.

Approximation Methods: In order to evaluate the different approximation methods, we experimentally compare 4 different approximation methods: standard LSI, ON-Eigen (Equation 1), Explicit LSI (Equation 2), L-Solve (Equation 3) on the same small-scale corpus. For convenience we plot an *approximation rate* which is either K or N_1 depending on method; at $K = 500$ and $N_1 = 500$, these approximations become exact. This is shown in Figure 1. We also observe the effects of approximation and see that the performance increases steadily as we increase the computational factor. We see that the orthonormal eigenvector (Equation 1) method and the L-solve (Equation 3) method are clearly similar in approximation quality. We see that the explicit LSI method (Equation 2) and the LSI method both perform significantly worse for most of the approxi-

[5] A likely explanation for this is that low frequency terms are less evenly distributed and the effect of calculating the matrix inverse magnifies the noise from the low frequency terms

mation amounts. Explicit LSI is worse than the other approximations as it first maps the test documents into a K-dimensional LSI topic space, before mapping back into the N-dimensional explicit space. As expected this performs worse than standard LSI for all but high values of K as there is significant error in both mappings. We also see that the (CL-)ESA baseline, which is very low due to the small number of documents, is improved upon by even the least approximation of orthonormalization. In the remaining of this section, we report results using the L-Solve method as it has a very good performance and is computationally less expensive than ON-Eigen.

Evaluation and Comparison: We compare ONETA using the L-Solve method with N_1 values from 1000 to 9000 topics with (CL-)ESA (using SQRT normalization), LDA (using 1000 topics) and LSI (using 4000 topics). We choose the largest topic count for LSI and LDA we could to provide the best possible comparison. For LSI, the choice of K was determined on the basis of operating system memory limits, while for LDA we experimented with higher values for K without any performance improvement, likely due to overfitting. We also stress that for L-Solve ONETA, N_1 is not the topic count but an approximation rate of the mapping. In all settings we use N topics as with standard ESA, and so should not be considered directly comparable to the K values of these methods.

We also compare to a baseline system that relies on word-by-word translation, where we use the most likely single translation of a word as given by a phrase table generated by the Moses system (Koehn et al., 2007) on the EuroParl corpus (Koehn, 2005). Top 1, Top 5 and Top 10 Precision as well as Mean Reciprocal Rank are reported in Table 2.

Interestingly, even for a small number of documents (e.g., $N_1 = 6000$) our results improve both the word-translation baseline as well as all other topic models, ESA, LDA and LSI in particular. We note that at this level the method is still efficiently computable and calculating the inverse in practice takes less time than training the Moses system. The significance for results ($N_1 \geq 7000$) have been tested by means of a bootstrap resampling significance test, finding out that our results significantly improve on the translation base line at a 99% level.

Further, we consider a straightforward combination of our method with the translation system consisting of appending the topic vectors and the translation frequency vectors, weighted by the relative average norms of the vectors. We see that in this case the translations continue to improve the performance of the system (albeit not significantly), suggesting a clear potential for this system to help in improving machine translation results. While we have presented results for English and Spanish here, similar results were obtained for the German and French case but are not presented here due to space limitations.

In Table 2 we also include the user time and peak resident memory of each of these processes, measured on an 8 Core Intel Xeon 2.50 GHz server. We do not include the results for Word Translation as many hours were spent learning a phrase table, which includes translations for many phrases not in the test set. We see that the ONETA method significantly outperforms LSI and LDA in terms of speed and memory consumption. This is in line with the theoretical calculations presented earlier where we argued that inverting the $N \times N$ dense matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ when $W \gg N$ is computationally lighter than finding an eigendecomposition of the $W \times W$ sparse matrix $\mathbf{X}\mathbf{X}^{\mathrm{T}}$. In addition, as we do not multiply $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and $\mathbf{X}^{\mathrm{T}}$, we do not need to allocate a large $W \times K$ matrix in memory as with LSI and LDA.

The implementations of ESA, ONETA, LSI and LDA used as well as the data for the experiments are available at `http://github.com/jmccrae/oneta`.

5 Conclusion

We have presented a novel method for cross-lingual topic modelling, which combines the strengths of explicit and latent topic models and have demonstrated its application to cross-lingual document matching. We have in particular shown that the method outperforms widely used topic models such as Explicit Semantic Analysis (ESA), Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA). Further, we have shown that it outperforms a simple baseline relying on word-by-word translation of the query document into the target language, while the induction of the model takes less time than training the machine translation system from a parallel corpus. We have also presented an effective approximation method, i.e. L-Solve, which significantly reduces the computational cost associated with computing the topic models.

Acknowledgements

This work was funded by the Monnet Project and the Portdial Project under the EC Seventh Framework Programme, Grants No. 248458 and 296170. Roman Klinger has been funded by the "Its OWL" project ("Intelligent Technical Systems Ostwestfalen-Lippe", http://www.its-owl.de/), a leading-edge cluster of the German Ministry of Education and Research.

References

Dennis S Bernstein. 2005. *Matrix mathematics, 2nd Edition*. Princeton University Press Princeton.

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

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *IJCAI*, volume 9, pages 1513–1518.

Don Coppersmith and Shmuel Winograd. 1990. Matrix multiplication via arithmetic progressions. *Journal of symbolic computation*, 9(3):251–280.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Chris Ding, Tao Li, and Wei Peng. 2006. NMF and PLSI: equivalence and a hybrid algorithm. In *Proceedings of the 29th annual international ACM SIGIR*, pages 641–642. ACM.

Susan T Dumais, Todd A Letsche, Michael L Littman, and Thomas K Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval*, volume 15, page 21.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, volume 6, page 12.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference*, pages 50–57. ACM.

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

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5.

Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Martha Palmer, Owen Rambow, and Alexis Nasr. 1998. Rapid prototyping of domain-specific machine translation systems. In *Machine Translation and the Information Soup*, pages 95–102. Springer.

Philipp Sorg and Philipp Cimiano. 2008. Cross-lingual information retrieval with explicit semantic analysis. In *Proceedings of the Cross-language Evaluation Forum 2008*.

Philipp Sorg and Philipp Cimiano. 2010. An experimental comparison of explicit semantic analysis implementations for cross-language retrieval. In *Natural Language Processing and Information Systems*, pages 36–48. Springer.

Vassilis Spiliopoulos, George A Vouros, and Vangelis Karkaletsis. 2007. Mapping ontologies elements using features in a latent space. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 457–460. IEEE.

Yik-Cheung Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.