A Theoretical Analysis of Cross-lingual Semantic Relatedness in Vector Space Models

Lei Zhang
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
l.zhang@kit.edu

Thanh Tran
San Jose State University
One Washington Square, San Jose, CA 95192-0249, USA
ducthanh.tran@sjsu.edu

Achim Rettinger
Karlsruhe Institute of Technology (KIT)
76128 Karlsruhe, Germany
rettinger@kit.edu

ABSTRACT

Semantic relatedness is essential for different text processing tasks, especially in the cross-lingual setting due to the vocabulary mismatch problem. Many concept-based solutions to semantic relatedness have been proposed, which vary in the notions of concept and document representation. In our contribution, we provide a unified model that generalizes over the existing approaches to cross-lingual semantic relatedness. It shows that the main existing solutions represent different ways for constructing the concept space, which result in different document representations and implications for semantic relatedness computation. In particular, it allows us to provide theoretical justifications of existing solutions. Through the experimental evaluation, we show that the results support our theoretical findings.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory, Languages

Keywords

Semantic Relatedness, Cross-lingual, Vector Space Models

1. INTRODUCTION

Semantic relatedness has been used in many fields of natural language processing (NLP), including word sense disambiguation, text summarization and annotation, information extraction and retrieval. In this regard, understanding semantic relatedness is crucial for processing natural language texts, especially when they are composed in different languages. Cross-lingual semantic relatedness measures the strength of semantic connection between documents (or other textual units such as words, sentences and paragraphs) in different languages.

ICTIR'15, September 27–30, Northampton, MA, USA.

DOI: http://dx.doi.org/10.1145/2808194.2809450.

Approaches to semantic relatedness can be classified according to the type of used resources: (1) *dictionary-based* approaches, where entries in dictionaries can be exploited to define semantic relatedness between terms; (2) *thesaurus-based* approaches, where terms are grouped together based on different kind of relations, such as synonymy and hyponymy; (3) *corpus-based* approaches, where co-occurrences of terms are often interpreted as an estimation of semantic relatedness. While dictionary-based and thesaurus-based approaches can measure semantic relatedness in a more precise way, the advantage of corpus-based approaches lies in the large amount of available data. In this work, we focus on the corpus-based solutions [4, 9, 2, 19, 13, 6] for computing semantic relatedness.

For cross-lingual semantic relatedness, a straight-forward way is to first translate the documents into the same language using statistical machine translation (SMT) systems and then apply the monolingual semantic relatedness methods. However, the drawbacks of applying SMT systems to translate the documents on the fly is the potentially longer execution time and the requirement of parallel training corpora, which are still missing for many language pairs. Several cross-lingual extensions [5, 11, 14, 18, 12] of the corpus-based approaches to semantic relatedness have been proposed. These approaches can rely on either a parallel corpus or an aligned comparable corpus[1], which is much easier to obtain, e.g., it can be derived from Wikipedia. However, these solutions as well as existing studies comparing them (see [21, 10, 3]) do not provide a theoretical understanding of and justification for differences among existing methods.

In this work, we provide *a generalized model for cross-lingual semantic relatedness* based on the notions of (1) interlingual concept space, (2) document representation and (3) semantic relatedness measure. In our *theoretical study*, we show that the main existing solutions can be conceived as instantiations that can be mapped to components of this generalized model. In particular, they represent different ways for constructing the concept space, which result in different document representations and implications for computation of the cross-lingual semantic relatedness measure. Through the *experimental evaluation*, we then show that these differences among existing solutions translates to different performance achievements in a cross-lingual search and retrieval scenario.

The remainder of the paper is structured as follows: in Sec. 2, we provide an overview of the main approaches stud-

[1]Parallel corpus consists of translated equivalents of each document, while aligned comparable corpus contains aligned documents in different languages that address the same topics but may differ in length, detail and style.

ied in this paper and the related work. Then, we present a generalized model for computing cross-lingual semantic relatedness in Sec. 3, which is later instantiated by different approaches. Based on the generalized model, we analyze different approaches and provide theoretical justifications for these solutions in Sec. 4. Experimental results are presented in Sec. 5, followed by conclusions in Sec. 6.

2. OVERVIEW AND RELATED WORK

The vector space models (VSM) [16, 15] have been widely used for representing documents as term vectors. Using terms from the documents alone to compute their similarity, however, suffers from the *vocabulary mismatch problem*: the similarity score is small when they have few terms in common, even though they are semantically very related. This problem is more serious in the cross-lingual setting because documents in different languages rarely share common terms.

Solutions to semantic relatedness aim to address this problem. Essentially, they can be conceived as different ways of (1) mapping terms to vectors in a semantic vector space spanned by concepts (2) to produce concept-based document representations, based on which (3) documents can be compared using standard similarity measures. In the cross-lingual setting, an *interlingual concept space* is needed, which is constructed using a parallel corpus or an aligned comparable corpus. Existing solutions vary in the notions of concept and document representation. In this work, we will study the following three main models in detail.

Clustering Model. Cluster analysis is a common technique for statistical data analysis in many fields. One specific application of clustering is to derive features or concepts from documents. If such concepts need to be valid for different languages, clustering has to be performed on a language-aligned document collection. As a common used method, K-means clustering [8] is employed in this paper to group the concatenated bilingual documents into clusters, which act as concepts.

Latent Model. Various latent approaches have been proposed to identify latent dimensions or concepts inherent in the background corpus. Among these approaches, we investigate Latent Semantic Indexing (LSI) [4] in particular, which is a well-known method based on Singular Value Decomposition (SVD). LSI was originally employed for dimensionality reduction on the term-document matrix of a corpus. The reduced dimensions correspond to latent concepts. By using a parallel corpus or an aligned comparable corpus, it can be applied to cross-lingual contexts [5].

Explicit Model. Recently, explicit approaches have been proposed as alternative to latent approaches based on externally defined knowledge (e.g. Wikipedia), which is exploited to define concepts. One prominent instantiation by now is Explicit Semantic Analysis (ESA) [6]. To adopt ESA for the cross-lingual setting, cross-language links in Wikipedia has been used [14, 18]. In this work, besides Wikipedia we also use the parallel corpus to extend ESA and the experiments show its good performance.

There are some studies that compare different solutions to cross-lingual information retrieval (CLIR), which consists of providing a query in one language and searching documents in one or more different languages. In this context, the work in [21] has reported a thorough evaluation of multiple methods for CLIR, which fall into two categories: machine translation (MT) based approaches, where dictionary-based and corpus-based MT systems have been studied, and statistical information retrieval (IR) approaches including General Vector Space Model (GVSM) [20] and LSI. The comparative study shows that corpus-based MT approaches clearly surpass general-purpose dictionary-based MT approaches and the performance of LSI proves comparable to that of other corpus-based approaches including the MT ones. The work in [10] has reported a series of experiments comparing the performance of GVSM and LSI on monolingual and translingual retrieval tasks. The results show that LSI performs better but have a larger preprocessing cost. In [3], latent models of concepts, namely LSI and Latent Dirichlet Allocation (LDA) [2], have been compared to ESA on a mate retrieval task and it claimed that ESA outperforms LSI/LDA unless the latter are trained and tested on the same dataset instead of Wikipedia as the training data. However, these studies do not provide theoretical understanding of and justification for differences among existing solutions.

In this paper, we focus on three representative instantiations of the above models, namely K-means clustering, LSI and ESA. Most other approaches in each category are just variations and incremental improvements of these three approaches. For example, LDA is a probabilistic extension of LSI [9, 2] and it has shown that ESA is very close to GVSM in the recent studies [1, 7]. Different from the existing studies, we provide both theoretical justifications and empirical comparisons of these approaches.

3. GENERALIZED MODEL

In this section, we present a unified model for cross-lingual semantic relatedness. This model generalizes over the existing approaches and we will show how different approaches can be expressed as instantiations and mapped to components of this model. Firstly, we discuss the model components and their roles w.r.t. cross-lingual semantic relatedness, where documents are represented as semantic vectors in a certain interlingual concept space, which abstracts from the background parallel or aligned comparable corpus and builds on the standard cosine similarity measure to access cross-lingual semantic relatedness.

Consider two documents x and y in languages X and Y, the vocabulary sizes of which are p and q, respectively. Based on VSM, we have

$$\mathbf{x} = (x_1, x_2, \ldots, x_p)^T \tag{1}$$

$$\mathbf{y} = (y_1, y_2, \ldots, y_q)^T \tag{2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are term vectors of x and y, x_i and y_i are the weights of terms i and j. Different weighting functions can be used, such as binary, TF and TF-IDF models. In the traditional VSM, two documents in the same language can be compared based on their term vectors using the standard similarity measure. However, in the cross-lingual setting, we cannot compare the documents directly due to the vocabulary mismatch problem. As discussed, a class of concept-based approaches have been suggested to exploit the interlingual concept space. Based on a mapping function, term vectors of documents in different languages can be mapped to concept vectors in the interlingual concept space, where they can be compared using the standard similarity measure.

We first introduce some notations to facilitate the following discussion. Let $B = \binom{B'}{B''}$ be the term-document matrix of the parallel or aligned comparable corpus containing m bilingual documents, where $B' = (b'_{ij})_{p \times m}$ and

$B'' = (b''_{ij})_{q \times m}$ are matrices for documents in languages X and Y with vocabulary size p and q respectively, and each pair of the vector $\mathbf{b}'_i$ in B' and the aligned vector $\mathbf{b}''_i$ in B'' form the vector of the concatenated bilingual document $\mathbf{b}_i = \binom{\mathbf{b}'_i}{\mathbf{b}''_i}$ in B.

Interlingual Concept Space. The construction of the interlingual concept space relies on the background corpus. Given its matrix B, we apply different approaches to obtain two sets of aligned vectors of concepts for X and Y

$$U = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n) \tag{3}$$

$$V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \tag{4}$$

where each pair of aligned vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ represent the same concept. The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ can be represented as

$$\mathbf{u}_i = (u_{1i}, u_{2i}, \ldots, u_{pi})^T \tag{5}$$

$$\mathbf{v}_i = (v_{1i}, v_{2i}, \ldots, v_{qi})^T \tag{6}$$

where the entries u_{ji} in $\mathbf{u}_i$ and v_{ki} in $\mathbf{v}_i$ corresponding to terms j and k are considered as importance indicators of terms j and k in the concept. An interlingual concept space A can be formed using U and V where each dimension corresponds to a pair of aligned vectors in U and V.

Document Representation. To produce the concept-based representation, each document in languages X and Y can be mapped to a concept vector in A

$$U(\mathbf{x}) = U^T \cdot \mathbf{x} = (\langle \mathbf{u}_1, \mathbf{x} \rangle, \ldots, \langle \mathbf{u}_n, \mathbf{x} \rangle)^T \tag{7}$$

$$V(\mathbf{y}) = V^T \cdot \mathbf{y} = (\langle \mathbf{v}_1, \mathbf{y} \rangle, \ldots, \langle \mathbf{v}_n, \mathbf{y} \rangle)^T \tag{8}$$

where each entry is the inner product of term vectors of the document and the corresponding concept representing the association strength between them.

Semantic Relatedness Measure. The semantic relatedness between x and y can be calculated using cosine similarity between $U(\mathbf{x})$ and $V(\mathbf{y})$ as

$$\begin{aligned} sim(x,y) &= cos(U(\mathbf{x}), V(\mathbf{y})) = \frac{\langle U(\mathbf{x}), V(\mathbf{y}) \rangle}{|U(\mathbf{x})| \cdot |V(\mathbf{y})|} \\ &= \frac{(\mathbf{x}^T \cdot U) \cdot (V^T \cdot \mathbf{y})}{\sqrt{(\mathbf{x}^T \cdot U) \cdot (U^T \cdot \mathbf{x})} \cdot \sqrt{(\mathbf{y}^T \cdot V) \cdot (V^T \cdot \mathbf{y})}} \\ &= \frac{\sum_{j=1}^{p} \sum_{k=1}^{q} x_j \cdot y_k \cdot g_{jk}}{\sqrt{\sum_{j=1}^{p} \sum_{k=1}^{p} x_j \cdot x_k \cdot g'_{jk}} \cdot \sqrt{\sum_{j=1}^{q} \sum_{k=1}^{q} y_j \cdot y_k \cdot g''_{jk}}} \end{aligned} \tag{9}$$

where $g_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki}$ denotes the correlation between term j from document x and term k from document y, $g'_{jk} = \sum_{i=1}^{n} u_{ji} \cdot u_{ki}$ ($g''_{jk} = \sum_{i=1}^{n} v_{ji} \cdot v_{ki}$) captures the term correlation between j and k from x (y). Essentially, the term correlation between j and k is based on their associations with each concept i.

In contrast to the standard VSM, computing semantic relatedness in the concept space introduces some new factors. In the numerator of Eq. 9, we observe that the semantic relatedness between documents x and y is *proportional* to the sum of values of $x_j \cdot y_k \cdot g_{jk}$ for each pair of terms j from x and k from y. The component g_{jk}, called *term relatedness*, captures the term correlation between j and k from different documents. Obviously, when two documents have more correlated term pairs yielding more non-zero components $x_j \cdot y_k \cdot g_{jk}$, and these term pairs appear more frequently in the respective documents and have closer correlation yielding larger values of $x_j \cdot y_k \cdot g_{jk}$, the score of semantic relatedness is higher. The term relatedness factor is used to incorporate this effect.

Regarding the denominator of Eq. 9, the semantic relatedness between x and y is *inversely proportional* to the square root of the sum of values of $x_j \cdot x_k \cdot g'_{jk}$ ($y_j \cdot y_k \cdot g''_{jk}$), called normalization factor, which has two effects. Firstly, the components $x_j \cdot x_k$ and $y_j \cdot y_k$ have the effect of *document length normalization*, which is similar to that in the standard VSM. Clearly, long documents usually use the same terms repeatedly and also contain numerous different terms resulting in higher term frequencies and more terms, and thus the components $x_j \cdot y_k$ in the numerator of Eq. 9 are larger for long documents. This increases the semantic relatedness score between long documents and others. Document length normalization is used to remove the advantage of long documents over short ones. Higher term frequencies and more terms in x (y) increase the values of $x_j \cdot x_k$ ($y_j \cdot y_k$), yielding a larger normalization factor and penalizing the documents in accordance with their lengths [17].

In addition, the components g'_{jk} and g''_{jk}, called *term dependency* of documents, discard the effect of term correlation within documents on semantic relatedness. Consider documents consisting of terms that are highly dependent, in other words, many terms in them are semantically correlated. This might increase the number of correlated term pairs, thus yielding larger semantic relatedness, with other documents. The *term dependency normalization* is used to compensate for this effect. High term dependency of x and y increases the values of g'_{jk} and g''_{jk} and thus results in a larger normalization factor, thus removing the advantage of documents with high term dependency.

Although term relatedness component g_{jk} and term dependency components g'_{jk}, g''_{jk} play different roles in the computation of semantic relatedness, they all capture the term correlation between j and k. The only difference is that the terms j and k are from different documents in different languages for g_{jk}, but from the same document (thus in the same language) for g'_{jk} and g''_{jk}. We will focus our analysis on *term relatedness* g_{ij} across documents.

4. THEORETICAL ANALYSIS

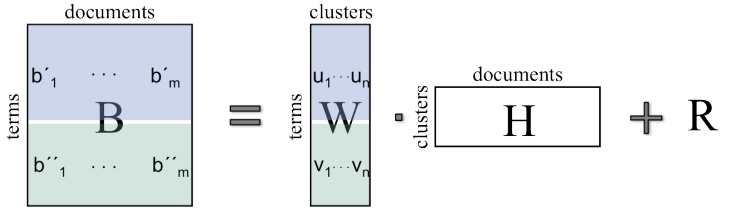
We have presented the model components and discussed the effects of specific factors on the semantic relatedness measure. Based on this, we now provide a theoretical analysis of the existing approaches, namely K-means clustering, LSI and ESA. We show how they can be mapped to the model components and in this way, make clear their differences in the semantic relatedness computation.

While all approaches exploit the *term co-occurrence* in the background corpus, the main difference between them lies in the *interlingual concept space* construction. As shown in Fig. 1, the aligned vectors of concepts $\mathbf{u}_i$ and $\mathbf{v}_i$ spanning the interlingual concept space are derived differently in these approaches. This results in different ways of computing the *cross-lingual semantic relatedness*, in particular, the term relatedness g_{jk} that can be calculated as

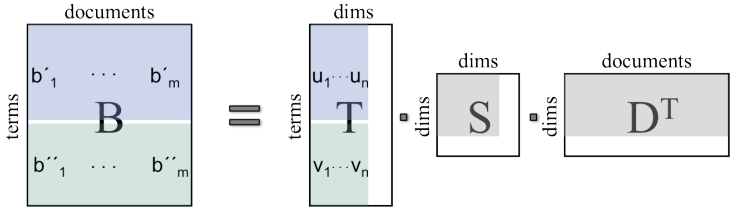
$$g_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki} \tag{10}$$

4.1 K-means Clustering based Approach

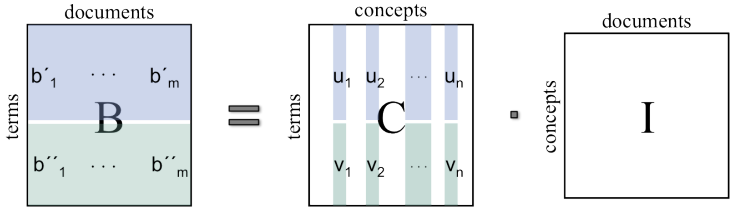
K-means clustering groups the m bilingual documents in the background corpus into n clusters and each cluster w_i corresponds to a concept, so as to minimize the within-



(a) K-means clustering based approach



(b) LSI based approach



(c) ESA based approach

Figure 1: Matrix representations of interlingual concept space construction

cluster sum of squared differences $\sum_{i=1}^{n}\sum_{b_j \in w_i} \|\mathbf{b}_j - \mathbf{w}_i\|^2$, where $\mathbf{b}_j$ is the term vector of the concatenated bilingual document b_j and the vector $\mathbf{w}_i$ corresponds to the centroid of cluster w_i. The K-means clustering based approach to interlingual concept space construction can be represented as the following matrix factorization

$$B = W \cdot H + R, \quad W = \binom{U}{V} \tag{11}$$

After the clustering performed on B, we obtain the matrix of concepts W, where the sub-matrices U and V contain the aligned vectors of concepts for languages X and Y, respectively. Each column $\mathbf{w}_i = \binom{\mathbf{u}_i}{\mathbf{v}_i}$ in W contains the term weights of a cluster centroid, which is the average of term weights for all bilingual documents in this cluster. Each column in H contains many "0"s but only one "1" indicating the membership of the document in a cluster. That means each bilingual document can only belong to one cluster. For instance, if the i-th entry h_{ij} in the column $\mathbf{h}_j$ is 1, we have $b_j \in w_i$, i.e. document b_j belongs to cluster w_i.

Given $b_j \in w_i$, the column $\mathbf{r}_j = \mathbf{b}_j - \mathbf{w}_i$ in R can be considered as noise introduced by clustering into the concept $\mathbf{w}_i$. Thus the smaller $\|\mathbf{b}_j - \mathbf{w}_i\|^2$ is, the more precisely b_j is assigned to w_i. The matrix R is called residual matrix such that clustering minimizes sum of squares of all its columns. Note that the number of clusters n is predefined and each document must belong to one of these clusters. Considering that the background corpus covers a wide range of concepts/topics, some documents might be assigned to the concepts incorrectly resulting in larger $\|\mathbf{b}_j - \mathbf{w}_i\|^2$, i.e., the concepts might contain noise.

To facilitate the following discussion, we firstly introduce some concepts about *term co-occurrence*.

Proposition 1. *Given a term-document matrix $M = (b_{ji})_{m\times n}$ with each entry b_{ji} reflecting the term frequency of term j in document i, $Z = M \cdot M^T$ is its* term co-occurrence matrix. *For each pair of terms j and k, the entry z_{jk} in Z represents the* term co-occurrence frequency *of j and k.*

Proof. Based on different weighting functions, the term frequency b_{ji} in M has different meanings. For instance, it is the raw term frequency in TF model but the normalized frequency in TF-IDF model by taking the importance of terms into account. Each entry in Z can be calculated as $z_{jk} = \sum_{i=1}^{n} b_{ji} \cdot b_{ki}$ and it reflects the term co-occurrence frequency w.r.t. all the documents in M. □

In order to derive the conclusions of the value of g_{jk} in the K-means clustering based approach, we model the *term co-occurrence* matrix $Z^{K-means}$ as

$$\begin{aligned} Z^{K-means} &= (B-R)\cdot(B-R)^T \\ &= \underbrace{B\cdot B^T}_{Z_B} \underbrace{-B\cdot R^T - R\cdot B^T + R\cdot R^T}_{N} \end{aligned} \tag{12}$$

where Z_B represents the term co-occurrence matrix of the background corpus and N stands for the noise introduced by clustering. In this regard, each entry z_{jk} in $Z^{K-means}$ captures the co-occurrence frequency of terms j and k in the background corpus with added noise.

Lemma 1. *Given the matrix of concepts $W = \binom{U}{V}$, each entry z_{jk} in $Z^{K-means}$ for terms j and k in different languages can be calculated as*

$$z_{jk} = \sum_{i=1}^{n} |w_i| \cdot u_{ji} \cdot v_{ki} \tag{13}$$

Proof. According to Eq. 11, we have $B - R = W \cdot H$ and $Z^{K-means} = (W\cdot H)\cdot(W\cdot H)^T = W\cdot H\cdot H^T\cdot W^T =$

$W \cdot \Sigma \cdot W^T$, where W contains the vectors $\mathbf{w}_i = \binom{\mathbf{u}_i}{\mathbf{v}_i}$ and $\Sigma = H \cdot H^T$ is a diagonal matrix with $\sigma_i = |w_i|$ on the diagonal. □

THEOREM 2. *The term relatedness g_{jk} in the K-means clustering based approach has a* positive correlation *with the term co-occurrence frequency z_{jk} in the* background corpus with additional noise *yielded by clustering, i.e., when $z_{jk} > 0$, we have $g_{jk} > 0$ and $\frac{z_{jk}}{\max_{1\le i\le n}(|w_i|)} \le g_{jk} \le \frac{z_{jk}}{\min_{1\le i\le n}(|w_i|)}$; otherwise $g_{jk} = 0$.*

PROOF. Following Lemma 1, when $z_{jk} > 0$, there is at least one cluster with $u_{ji} \cdot v_{ki} > 0$ such that $g_{jk} > 0$. Since $\max_{1\le i\le n}(|w_i|) \cdot g_{jk} = \sum_{i=1}^{n} \max_{1\le i\le n}(|w_i|) \cdot u_{ji} \cdot v_{ki} \ge \sum_{i=1}^{n} |w_i| \cdot u_{ji} \cdot v_{ki} = z_{jk}$ and $\min_{1\le i\le n}(|w_i|) \cdot g_{jk} = \sum_{i=1}^{n} \min_{1\le i\le n}(|w_i|) \cdot u_{ji} \cdot v_{ki} \le \sum_{i=1}^{n} |w_i| \cdot u_{ji} \cdot v_{ki} = z_{jk}$, we have $\frac{z_{jk}}{\max_{1\le i\le n}(|w_i|)} \le g_{jk} \le \frac{z_{jk}}{\min_{1\le i\le n}(|w_i|)}$. □

4.2 LSI based Approach

Given the matrix B of the background corpus, LSI [4] finds an optimal approximation X of B with low-rank at most n based on Singular Value Decomposition (SVD), so as to minimize the Frobenius norm of the matrix difference $\|B - X\|_F = \sqrt{\sum_{i=1}^{p+q}\sum_{j=1}^{m}(b_{ij} - x_{ij})}$. After SVD performed on matrix B, we obtain a loss-free factorization of the form $B = T \cdot S \cdot D^T$, where T and D, called left and right singular vectors, are two orthogonal matrices and S is a diagonal matrix with non-negative values on the diagonal usually in descending order, known as singular values of B. It is conventional to represent S as an $r \times r$ matrix, where r is the rank of B. Accordingly, T and D^T are represented as two $(p+q) \times r$ and $r \times m$ matrices, respectively. Each column $\mathbf{t}_i$ in T represents a semantic dimension corresponding to the concept t_i, which is a linear combination of vectors in B and each entry indicates how strongly a term is related to the semantic dimension. Each of singular value of B in S measures the importance of the corresponding semantic dimension. Each entry in D^T indicates how strongly a document is related to the concept represented by the semantic dimension.

Based on SVD, LSI reduces the dimensions of the matrix B by grouping the related terms to form the semantic dimensions. Nevertheless, the number of dimensions, equal to the rank of the matrix B, might be still large. In addition, the semantic dimensions with small singular values are not important such that they can be eliminated. Based on that, LSI aims to find a low-rank approximation of the matrix B by retaining only the n largest singular values. Thus, we have the matrix representation as follows

$$B_n = T_n \cdot S_n \cdot D_n^T, \quad T_n = \binom{U}{V} \tag{14}$$

where $S_n = diag(s_1, \ldots, s_n)$ contains the n largest singular values, $T_n = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$ and $D_n = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$ contain the corresponding first n vectors in T and D. The vectors in T_n will be used to span the interlingual concept space. The Eckart-Young theorem provides the fact that omitting the smallest $r - n$ singular values and their corresponding singular vectors yields the optimal approximation of matrix B with the lowest Frobenius error, namely $\min_{\mathsf{rank}(X)=n} \|B - X\|_F = \|B - B_n\|_F = \sqrt{\sum_{i=n+1}^{r} s_i^2}$. Such an approximation has the effect of preserving the important information while reducing noise in the background corpus.

In order to provide a better understanding of g_{jk} in the LSI based approach, we firstly investigate the term co-occurrence matrix $Z^{LSI} = B_n \cdot {B_n}^T$ and then discuss its relation to the l-th order term co-occurrence in the background corpus.

LEMMA 2. *Given the term co-occurrence matrix $Z_B = B \cdot B^T$ of the background corpus, the term co-occurrence matrix $Z^{LSI} = B_n \cdot {B_n}^T$ of the optimal low-rank approximation of B is a linear combination of the powers of Z_B. That is*

$$Z^{LSI} = B_n \cdot {B_n}^T = \sum_{l=1}^{r} \alpha_l \cdot Z_B^l \tag{15}$$

where r is the rank of B, α_l are constants depending on singular values of B.

PROOF. According to the factorization of $B = T \cdot S \cdot D^T$ using SVD, where T and D are orthogonal matrices and S is a diagonal matrix, we have $Z_B = (T \cdot S \cdot D^T) \cdot (T \cdot S \cdot D^T)^T = T \cdot S \cdot D^T \cdot D \cdot S \cdot T^T = T \cdot S^2 \cdot T^T$. Then, we can derive $Z_B^l = (T \cdot S^2 \cdot T^T)^l = T \cdot S^2 \cdot T^T \cdot T \cdot S^2 \cdot T^T \cdots T \cdot S^2 \cdot T^T = T \cdot S^{2l} \cdot T^T$ and each entry $z_{jk}^l = \sum_{i=1}^{r} t_{ji} \cdot t_{ki} \cdot s_i^{2l}$. With different $l \le r$, we get a linear system $A\mathbf{x} = \mathbf{b}$ with $A = \begin{bmatrix} s_1^2 & \cdots & s_r^2 \\ \vdots & \ddots & \vdots \\ s_1^{2r} & \cdots & s_r^{2r} \end{bmatrix}$, $\mathbf{x} = \begin{bmatrix} t_{j1} \cdot t_{k1} \\ \vdots \\ t_{jr} \cdot t_{kr} \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} z_{jk}^1 \\ \vdots \\ z_{jk}^r \end{bmatrix}$. The matrix A is commonly referred to as a Vandermonde matrix and its determinant is given as $det(A) = \prod_{1\le j\le r}(s_j^2) \prod_{1\le i<j\le r}(s_i^2 - s_j^2)$. Assume that the singular values s_i of B are mutually distinct[2], which is always the case in practice, we have $s_i^2 - s_j^2 \ne 0$ when $i \ne j$ and thus $det(A) \ne 0$, such that the linear system has a unique solution. Using the Cramer's rule, we have $x_i = t_{ji} \cdot t_{ki} = \frac{det(A_i)}{det(A)}$, where A_i is the matrix formed by replacing the i-th column of A by the vector $\mathbf{b}$ and $det(A_i) = \sum_{l=1}^{r} z_{jk}^l \cdot (-1)^{i+l} \cdot M_{il}$ based on the Laplace expansion, where M_{il} is the (l, i) minor of A_i, i.e. the determinant of the submatrix of A_i formed by deleting the l-th row and i-th column. Then we have $x_i = \sum_{l=1}^{r} \frac{(-1)^{i+l} \cdot M_{il}}{det(A)} \cdot z_{jk}^l$, where $det(A)$ and M_{il} are constants, which only depend on the singular values s_i of B. We conclude that $Z^{LSI} = (T_n \cdot S_n \cdot D_n^T) \cdot (T_n \cdot S_n \cdot D_n^T)^T = T_n \cdot S_n \cdot D_n^T \cdot D_n \cdot S_n \cdot T_n^T = T_n \cdot S_n^2 \cdot T_n^T$ and each entry $z_{jk} = \sum_{i=1}^{n} t_{ji} \cdot t_{ki} \cdot s_i^2 = \sum_{i=1}^{n}(\sum_{l=1}^{r} \frac{(-1)^{i+l} \cdot M_{il}}{det(A)} \cdot z_{jk}^l) \cdot s_i^2 = \sum_{l=1}^{r}(\sum_{i=1}^{n} \frac{(-1)^{i+l} \cdot M_{il}}{det(A)} \cdot s_i^2) \cdot z_{jk}^l = \sum_{i=1}^{r} \alpha_l \cdot z_{jk}^l$, where $\alpha_l = \sum_{i=1}^{n} \frac{(-1)^{i+l} \cdot M_{il}}{det(A)} \cdot s_i^2$ is a constant. Therefore, we have $Z^{LSI} = \sum_{l=1}^{r} \alpha_l \cdot Z_B^l$. □

DEFINITION 1. *Given the background corpus, its* term co-occurrence graph *is an undirected graph $G = (N, E)$, with each node $n_j \in N$ standing for term j and each edge $e(n_j, n_k) \in E$ capturing the co-occurrence of terms j and k in the background documents, where the weight of $e(n_j, n_k)$ is the co-occurrence frequency of j and k. The* l-th order co-occurrence *relation between terms j and k exists, if there is a path of length l from n_j to n_k in G, with the path weight as the product of weights of edges along the path. The* l-th order term co-occurrence frequency *of j and k is the sum of the weights of all paths from n_j to n_k in G.*

[2]When two or more singular values s_i of B are equal, one may use a generalization called confluent Vandermonde matrix, which is out of the scope of this work.

PROPOSITION 3. *Given the term co-occurrence matrix $Z_B = B \cdot B^T$ of the background corpus, the entry z^l_{jk} in Z^l_B for terms j and k, which is the l-th power of Z_B, is nonzero if the l-th order co-occurrence relation between j and k exists and the value of z^l_{jk} is the l-th order term co-occurrence frequency of j and k.*

PROOF. Let $Z_B = B \cdot B^T = (z_{jk})_{(p+q)\times(p+q)}$, where $B = \binom{B'}{B''}$ with $B' = (b'_{ij})_{p\times m}$ and $B'' = (b''_{ij})_{q\times m}$. Each entry z^l_{jk} in Z^l_B is calculated as $z^l_{jk} = \sum_{i_1=1}^{p+q}\sum_{i_2=1}^{p+q}...\sum_{i_{l-1}=1}^{p+q} z_{ji_1} \cdot z_{i_1i_2}...z_{i_{l-2}i_{l-1}} \cdot z_{i_{l-1}k}$, where $i_1, i_2, ..., i_{l-2}, i_{l-1}$ represent different terms and $z_{ji_1}, z_{i_1i_2}, ..., z_{i_{l-2}i_{l-1}}, z_{i_{l-1}k}$ denote the term co-occurrence frequencies in B. If there is a path $p = \langle e(n_j, n_{i_1}), e(n_{i_1}, n_{i_2}), ..., e(n_{i_{l-2}}, n_{i_{l-1}}), e(n_{i_{l-1}}, n_k)\rangle$ in the co-occurrence graph G, we have $z_{ji_1} \neq 0, z_{i_1i_2} \neq 0, ..., z_{i_{l-1}k} \neq 0$. The weight of path p can be computed as $w_p = z_{ji_1} \cdot z_{i_1i_2}...z_{i_{l-1}k} \neq 0$ and thus $z^l_{jk} \neq 0$. Given the set P of all paths with length l from n_j to n_k, we have $z^l_{jk} = \sum_{p\in P} w_p$. □

LEMMA 3. *Given the matrix of concepts $T_n = \binom{U}{V}$ obtained by LSI, each entry z_{jk} in Z^{LSI} for terms j and k in different languages can be calculated as*

$$z_{jk} = \sum_{i=1}^{n} s_i^2 \cdot u_{ji} \cdot v_{ki} \tag{16}$$

PROOF. According to Eq. 14, we have $Z^{LSI} = (T_n \cdot S_n \cdot D_n^T) \cdot (T_n \cdot S_n \cdot D_n^T)^T = T_n \cdot S_n \cdot D_n^T \cdot D_n \cdot S_n \cdot T_n^T = T_n \cdot \Sigma \cdot T_n^T$, where T_n contains the vectors $\mathbf{t}_i = \binom{\mathbf{u}_i}{\mathbf{v}_i}$ and $\Sigma = S_n^2$ is a diagonal matrix with $\sigma_i = s_i^2$ on the diagonal. □

THEOREM 4. *The term relatedness g_{jk} between j and k computed in the LSI based approach has a* positive correlation *with the term co-occurrence frequency in the* optimal low-rank approximation of the background corpus*, i.e. z_{jk} in Z^{LSI}, which can be represented as a linear combination of the* l-th order co-occurrence frequency *for terms k and j in the* background corpus*. When $z_{jk} > 0$, then $g_{jk} > 0$ and $\frac{z_{jk}}{\max_{1\leq i\leq n}(s_i^2)} \leq g_{jk} \leq \frac{z_{jk}}{\min_{1\leq i\leq n}(s_i^2)}$; otherwise $g_{jk} = 0$.*

For the sake of space, we omit the proof of Theorem 4, which is similar to the proof of Theorem 2.

4.3 ESA based Approach

Recent work [1] has reported that instead of Wikipedia, documents in other corpora can be employed to construct the concepts in ESA and achieve good performance. In this work, besides Wikipedia we also use the parallel corpus to extend ESA. Since ESA [6] simply uses the documents in the background corpus as concepts, we have the matrix representation $B = C \cdot I$, where I is an identity matrix such that the matrix of the concepts C is same as B. Given two documents x and y in different languages, ESA maps them into concept vectors in a high dimensional concept space constructed by the bilingual documents, where the entries in the concept vector represent the association strength between the input documents and the corresponding concepts.

Without dimension reduction in ESA, the number of concepts is equal to the number of bilingual documents in the background corpus. In order to speed up processing and yield more compact vectors, ESA considers only the top-k concepts with the highest relevance scores w.r.t. each input document to construct the interlingual concept space [18]. Given a pair of input documents x and y in languages X and Y, we first generate two top-k concept sets $S_x = \{b_s | b'_s \in kNN(x)\}$ and $S_y = \{b_s | b''_s \in kNN(y)\}$, where $kNN(x)$ ($kNN(y)$) is the set of k nearest neighbors (most highly-ranked concepts) retrieved using x (y) based on the inner product $\langle \mathbf{x}, \mathbf{b}'_s\rangle$ ($\langle \mathbf{y}, \mathbf{b}''_s\rangle$).

Clearly, the entries in the concept vector of x (y) corresponding to the concepts which are not contained in S_x (S_y) will be zero. Therefore, given the documents x and y, only the concepts contained in $S_x \cap S_y$ will play a role in the semantic relatedness computation. In this sense, x and y are transformed into vectors in a concept space constructed by such concepts in $S_x \cap S_y$, which forms the matrix $C_{xy} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$. Then we have

$$C_{xy} = C_{[S_x\cap S_y]}, \quad C_{xy} = \binom{U}{V} \tag{17}$$

where $C_{[S_x\cap S_y]}$ represents the sub-matrix defined as the columns of C corresponding to the concepts listed in $S_x \cap S_y$. The dimensionality n of the concept space is the size of $S_x \cap S_y$. In contrast to K-means clustering and LSI based approaches, n varies for different input documents and also depends on the background corpus, which will be shown in the experiments.

In the following, we model the *term co-occurrence* matrix Z^{ESA} and discuss the correlation between g_{jk} and z_{jk} in Z^{ESA}. Firstly, we define Z^{ESA} as

$$Z^{ESA} = C_{xy} \cdot C_{xy}^T \tag{18}$$

Each entry z_{jk} in Z^{ESA} captures the co-occurrence frequency of terms j and k in the top-k background documents retrieved for both input documents x and y.

LEMMA 4. *Given the matrix of concepts $C_{xy} = \binom{U}{V}$ yielded by ESA, each entry z_{jk} in Z^{ESA} for terms j and k in different languages can be calculated as*

$$z_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki} \tag{19}$$

According to Eq. 18, the proof of Lemma 4 is obvious. Following Lemma 4, it is straightforward to derive the following Theorem based on Eq. 10 and Eq. 19.

THEOREM 5. *The term relatedness g_{jk} computed in the ESA based approach is* equal to *the* term co-occurrence frequency *in the* most relevant part of the background corpus w.r.t. the input documents*, i.e. z_{jk} in Z^{ESA}.*

4.4 Summary

In this section, we summarize the different approaches w.r.t. *interlingual concept space construction* and the resulting ways of *cross-lingual semantic relatedness computation*. A summary of these approaches is shown in Table 1.

Interlingual Concept Space Construction. The concepts generated by the K-means clustering based approach are *clusters of bilingual documents* from the *entire background corpus*, where the documents are folded into these clusters and their centroids act as concepts. The number of clusters n is *predefined* and each document can only belong to a *single cluster*. When grouping the similar documents into clusters, K-means clustering might introduce *noise* (captured by the residual matrix H) into the concepts.

The LSI based approach reduces the dimensions of the term document matrix B by bringing the related terms together to form *semantic dimensions*. It uses the n most important semantic dimensions that yield the *optimal low-rank approximation* of B as concepts. Different from the K-means clustering based approach, each document could

	K-means clustering	LSI	ESA
Interlingual Concept Space Construction	$B = W \cdot H + R$ $W = \binom{U}{V}$	$B = T \cdot S \cdot D^T$ $B_n = T_n \cdot S_n \cdot D_n^T$ $T_n = \binom{U}{V}$	$B = C \cdot I$ $C_{xy} = C_{[S_x \cap S_y]}$ $C_{xy} = \binom{U}{V}$
Concept Representation	clusters of bilingual documents	semantic dimensions of related terms in different languages	bilingual documents
	based on the entire background corpus	based on the optimal approximation of the background corpus	based on the most relevant part of the background corpus w.r.t. inputs
Dimensionality	static and low	static and low	dynamic and high
Cross-lingual Semantic Relatedness Computation	$sim(x,y) = \frac{\sum_{j=1}^{p}\sum_{k=1}^{q} x_j \cdot y_k \cdot g_{jk}}{\sqrt{\sum_{j=1}^{p}\sum_{k=1}^{p} x_j \cdot x_k \cdot g'_{jk}} \cdot \sqrt{\sum_{j=1}^{q}\sum_{k=1}^{q} y_j \cdot y_k \cdot g''_{jk}}}$		
	$g_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki}$	$g_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki}$	$g_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki}$
	$u_{ji} = W[j,i] \in \mathbf{w}_i$ $v_{ki} = W[k,i] \in \mathbf{w}_i$	$u_{ji} = T_n[j,i] \in \mathbf{t}_i$ $v_{ki} = T_n[k,i] \in \mathbf{t}_i$	$u_{ji} = C_{xy}[j,i] \in \mathbf{c}_i$ $v_{ki} = C_{xy}[k,i] \in \mathbf{c}_i$
Term Co-occurrence Matrix and Implication of its Entry ($Z_B = B \cdot B^T$)	$Z^{K-means} = Z_B + N$	$Z^{LSI} = \sum_{l=0}^{r} \alpha_l \cdot Z_B^l$	$Z^{ESA} = C_{xy} \cdot C_{xy}^T$
	$z_{jk} = \sum_{i=1}^{n} \lvert w_i \rvert \cdot u_{ji} \cdot v_{ki}$	$z_{jk} = \sum_{i=1}^{n} s_i^2 \cdot u_{ji} \cdot v_{ki}$	$z_{jk} = \sum_{i=1}^{n} u_{ji} \cdot v_{ki}$
	term co-occurrence frequency in the background corpus with added noise	a linear combination of the high-order term co-occurr frequency in the background corpus	term co-occurrence frequency in the most relevant part of the background corpus
Correlation of g_{jk} with z_{jk}	g_{jk} has a positive correlation with z_{jk}	g_{jk} has a positive correlation with z_{jk}	g_{jk} is equal to z_{jk}

Table 1: Concept-based approaches to cross-lingual semantic relatedness

be folded into *more than one semantic dimension* and the weight of a document on a semantic dimension reflects its fractional membership, which can be viewed as a soft clustering *without noise* generated.

In contrast to K-means clustering and LSI based approaches, ESA based approach considers the *bilingual documents* in the background corpus as concepts. For each pair of input documents, it constructs the interlingual concept space dynamically using the intersection of two sets of top-k background documents that are *most relevant* to the respective input documents. Instead of a *fixed and relatively low* dimensionality n of the concept space in K-means clustering and LSI based approaches, n in ESA based approach is determined *dynamically*.

Cross-lingual Semantic Relatedness Computation. While all these approaches are based on term co-occurrence derived from the background corpus, the difference of interlingual concept space construction between these approaches results in different ways of computing cross-lingual semantic relatedness and term relatedness g_{jk} in particular.

In the K-means clustering based approach, g_{jk} is *coarse-grained* due to the term co-occurrence captured at *cluster* level ($u_{ji}, v_{ki} \in \mathbf{w}_i$) and *sensitive to the noise* yielded by K-means clustering. This results in a *positive correlation* of g_{jk} with the term co-occurrence frequency z_{jk} in the background corpus with added *noise*.

In the LSI based approach, g_{jk} is also *coarse-grained* due to the term co-occurrence captured at *semantic dimension* level ($u_{ji}, v_{ki} \in \mathbf{t}_i$), but *not sensitive to any noise*. This leads to a *positive correlation* of g_{jk} with the term co-occurrence frequency z_{jk} in the *optimal low-rank approximation* of the background corpus, which has been proved to be a linear combination of the *high-order term co-occurrence frequency* in the background corpus.

In the ESA based approach, g_{jk} is *fine-grained* due to the term co-occurrence captured at *document* level ($u_{ji}, v_{ki} \in \mathbf{c}_i$) and *sensitive to the dimensionality* n of the interlingual concept space constructed dynamically. And g_{jk} is equal to the term co-occurrence frequency z_{jk} in the *most relevant part* of the background corpus w.r.t. the input documents.

	English	German	Spanish
#Wikipedia articles	4,014,643	1,438,325	896,691

(a) Number of Wikipedia articles

	English-German	English-Spanish	German-Spanish
#Cross-language links (→)	721,878	568,210	295,415
#Cross-language links (←)	718,401	581,978	302,502
#Cross-language links (merged)	722,069	593,571	307,130

(b) Number of cross-language links

Table 2: Statistics about Wikipedia dataset used in our experiments

5. EVALUATION

In order to investigate the performance of different approaches to cross-lingual semantic relatedness, we carried out the experiments, similar to [3], in a cross-lingual search and retrieval scenario using a standard mate retrieval setup for different language pairs covering English, German and Spanish.

5.1 Data and Methodology

To provide the background corpus, we extracted large collections from the parallel corpus JRC-Acquis[3] and the aligned comparable corpus Wikipedia[4] (henceforth also denoted by JRC and Wiki). The JRC-Acquis corpus comprises of approximately 23,000 legislative documents from European Union in each of 22 European languages. We used a random sample of 90% of parallel documents in English, German and Spanish from JRC-Acquis corpus as the background corpus and the remaining 10% parallel documents in these languages for testing. For constructing the aligned Wikipedia comparable corpus as the additional background corpus, we analyzed cross-language links between Wikipedia articles for each pair of supported languages in both directions and keep articles for which aligned versions exist at least in one direction. For instance, we extracted $721,878$ cross-language links from English Wikipedia to German Wikipedia, and $718,401$ cross-language links from German to English. By merging them, we obtain $722,069$ cross-language links, which are used to construct the aligned Wikipedia comparable corpus of the English-German language pair. Table 2 shows some statistics of the Wikipedia dataset used in our experiments.

For mate retrieval evaluation, we take the document in one language as query and retrieve the relevant documents in another language. We assumed that only the translated version (mate) is considered as relevant to the query document. In this experimental setup, we are concerned about whether the translation can appear on top of the ranked result list and the observed position of the mate is also used as a comparison factor. Based on such observation, we consider recall at cutoff rank k (R@k) and Mean Reciprocal Rank (MRR) as quality criteria. Recall defines the number of relevant documents that are retrieved in relation to the total number of relevant documents. R@k is defined by only considering the top-k results. In the mate retrieval setting, R@k defines the number of queries for which the mate document was found in the top-k results. MRR measures the average reciprocal ranks of the mate documents. Different from R@k, MRR also takes into account the position of the mate document, resulting in higher value when the position of the mate in the ranked result list is higher.

[3]http://langtech.jrc.it/JRC-Acquis.html

[4]http://dumps.wikimedia.org

5.2 Results

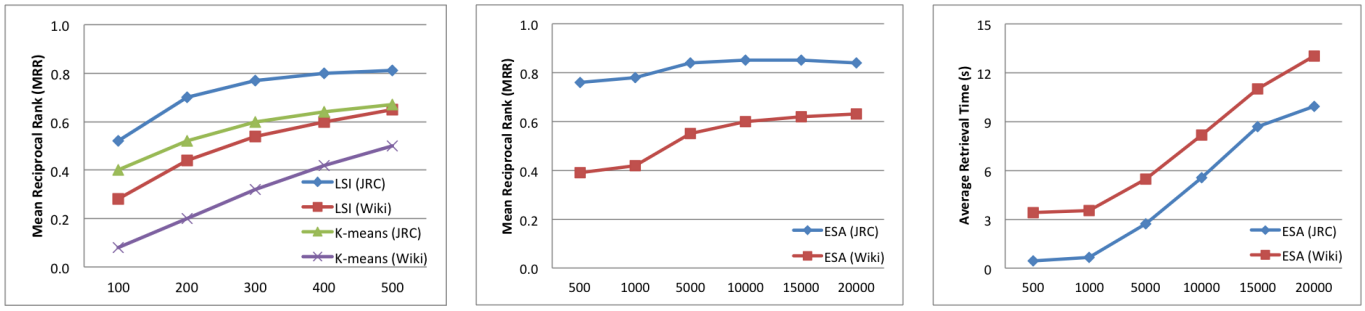
In the experiments, we observed that the performance of all these approaches varied with the parameters, namely the number n of the concepts used by K-means clustering and LSI based approaches and the number k of the most highly-ranked concepts under consideration for each input document in ESA based approach. Figs. 2(a-b) show the MRR results of three approaches using JRC-Acquis and Wikipedia as the background corpora averaged on different language pairs (i.e. English-German, English-Spanish and German-Spanish).

As shown in Fig. 2(a), the performance of K-means clustering and LSI based approaches tends to increase from 100 to 500[5] concepts using both JRC-Acquis and Wikipedia as the background corpora. For K-means clustering, the reason is that less noise will be introduced during clustering when n increases because the background documents can be assigned to more concepts, especially for Wikipedia which covers a wide range of topics. This explains why the performance of K-means clustering using Wikipedia as the background corpus improves significantly when n increases. For LSI with larger n, more semantic dimensions are involved to capture the term co-occurrence for semantic relatedness computation.

In ESA based approach, the dimensionality n of the interlingual concept space changes dynamically for specific input documents and it also depends on the top-k concepts under consideration for each input document. As shown in Fig. 2(b), ESA reaches its peak performance at $k = 10,000$ when using JRC-Acquis as the background corpus and the performance tends to slightly increase after $k > 10,000$ using Wikipedia as the background corpus. In general, ESA needs a minimum number of concepts to perform reasonably, but also reaches a point where further concepts will not help and may start introducing noise. After the top-$10,000$ concepts for each input document are considered, the number of the overlapped concepts for both input documents, i.e. the dimensionality n of the concept space, is large enough to capture the term co-occurrence when using JRC-Acquis as the background corpus, while there is still room to increase n when using Wikipedia as the background corpus. We will discuss this issue later.

It is observed that JRC-Acquis as the background corpus leads to much better results than Wikipedia for all approaches. That is due to the large vocabulary overlap be-

[5]The exploration of the concept space in the K-means clustering and LSI based approaches ends with 500 dimensions due to computational limitations of our servers.



(a) MRR for different n (b) MRR for different k (c) Retrieval Time for different k

Figure 2: Evaluation results for different parameters

Method	R@1	R@10	R@100	MRR	R@1	R@10	R@100	MRR	R@1	R@10	R@100	MRR
	English-German				English-Spanish				German-Spanish			
K-means (JRC)	0.49	0.84	0.98	0.61	0.70	0.94	0.99	0.78	0.50	0.85	0.98	0.61
LSI (JRC)	0.67	0.93	**1.00**	0.77	0.83	0.97	**1.00**	0.89	0.68	**0.94**	**1.00**	0.77
ESA (JRC)	**0.75**	**0.94**	**1.00**	**0.82**	**0.86**	**0.98**	**1.00**	**0.91**	**0.76**	**0.94**	**1.00**	**0.82**
K-means (Wiki)	0.32	0.61	0.85	0.42	0.57	0.82	0.94	0.66	0.34	0.62	0.85	0.43
LSI (Wiki)	0.52	0.81	**0.98**	**0.62**	**0.66**	**0.90**	**0.99**	**0.74**	**0.49**	**0.81**	**0.98**	**0.60**
ESA (Wiki)	**0.53**	**0.82**	0.96	0.59	0.60	0.86	**0.99**	0.67	0.47	0.76	0.95	0.55

Table 3: Evaluation results using the optimal settings

Language Pair	Min. dim.	Max. dim.	Avg. dim.	Min. dim.	Max. dim.	Avg. dim.	Fixed dim.
	ESA(JRC)			ESA(Wiki)			K-means/LSI
English-German	1,678	8,569	4,499	23	5,749	929	500
English-Spanish	1,550	8,830	4,519	10	6,407	1,131	500
German-Spanish	1,643	8,373	4,502	93	5,660	1,299	500
Average	1,624	8,591	4,489	42	5,939	1,120	500

Table 4: Dimensionality of the interlingual concept space

tween the test collection and the background corpus since the semantic relatedness computed in these approaches are all based on term co-occurrence derived from the background corpus. Moreover, in contrast to a parallel corpus, Wikipedia is a comparable corpus where the aligned articles may vary in size, quality and vocabulary. In other words, the term co-occurrence frequency in JRC-Acquis reflects more reliable term relatedness than that in Wikipedia.

In K-means clustering and LSI based approaches, each query document can be processed in less than 1 second on average. This is because the time complexity of both approaches only depends on the dimensionality n of the concept space, which is relatively small ($n \leq 500$) in both approaches. For the sake of space, we omit the results because individual times exhibit only minor differences.

Fig. 2(c) shows the average retrieval time of ESA based approach for each query document. We observe that ESA takes significantly more time than the other two approaches, because it has to compare each candidate document to retrieve with the background documents to yield the top-k concepts. This results in a much higher time complexity depending on m, which is the total number of all background documents. In practice, the inverted index and top-k query processing techniques can be employed such that the time complexity is much smaller than the worst case. However, compared with K-means clustering and LSI based approaches, the computation in ESA based approach is still more expensive, especially when k is large.

While K-means clustering and LSI are more efficient than ESA for online retrieval process, ESA does not require comprehensive computation for offline preprocessing, which is needed for K-means clustering and LSI. In our experiments, the preprocessing of ESA can be performed within 1 hour for any desired number of concepts. In contrast, the preprocessing of K-means clustering and LSI takes from several hours to several days with the increasing dimensionality.

For the reported evaluation results, we used these settings ($n = 500$ for K-means clustering and LSI based approaches and $k = 10,000$ for ESA based approach) to achieve the trade-off between the effectiveness (MRR) and the efficiency (retrieval time). Table 3 shows the R@k and MRR results of all three approaches using JRC-Acquis and Wikipedia as the background corpora for different language pairs, where the best results are formatted in bold.

LSI and ESA outperform K-means clustering in all the cases. With the previous theoretical analysis, we can explain this with the differences in the semantic relatedness computation: the term relatedness in K-means clustering is related to the term co-occurrence frequency in the background corpus with *noise* yielded by clustering, which leads to distortion of the computed term relatedness.

ESA outperforms LSI in most cases using JRC-Acquis as the background corpus. This is because the term relatedness captured in ESA is *fine-grained*, i.e. at the *bilingual document* level, while LSI captures a relatively *coarse-grained* term relatedness at the *semantic dimension* level. In ad-

dition, ESA calculates the term relatedness based on the term co-occurrence frequency in the *most relevant part* of the background corpus w.r.t. the input documents, which reduces the noise, i.e. the irrelevant background documents, while the term relatedness computed in LSI is related to a linear combination of the high-order term co-occurrence frequency in the background corpus without considering the input documents.

Interestingly, when using Wikipedia as the background corpus, LSI achieves slightly better results than ESA. The reason is that the term relatedness in ESA is sensitive to the dimensionality n, which is determined dynamically based on the input documents and the background corpus. As shown in Table 4, we investigate the values of n yielding the results in Table 3. While K-means clustering and LSI have a fixed n, it varies significantly in ESA when the background corpus changes from JRC-Acquis to Wikipedia. This is due to the large size and wide range of covered topics of Wikipedia, such that the overlap between the top-k concepts for the input documents and thus the dimensionality n is much smaller compared with the case when using JRC-Acquis as the background corpus. Since ESA needs a large number of concepts spanning the concept space to perform reasonably, we can generate more top-k concepts from Wikipedia for each input document to increase n. However, this will also result in more retrieval time as shown in Fig. 2(c).

6. CONCLUSIONS

In this paper, we study the foundation of cross-lingual semantic relatedness in vector space models. We investigate three fundamental solutions: the clustering model instantiated by K-means clustering, the latent model instantiated by LSI and the explicit model instantiated by ESA. Most approaches proposed earlier or later are variations and incremental improvements of these three approaches. As the main contribution, we establish a generalized model, which subsumes and helps to analyze the differences among the three existing approaches. In particular, we elaborate on differences in interlingual concept space construction and cross-lingual semantic relatedness computation based on concepts. We perform a theoretical analysis of these approaches and validate them in the experiments in a cross-lingual search and retrieval scenario.

The merit of our work is twofold. Firstly, it helps to obtain a better understanding of existing approaches; while the work is carried out in the more general cross-lingual context, the results are transferable to semantic relatedness in the monolingual case. Secondly, it can be used as a guide to choose among existing approaches and to design future semantic relatedness solutions for the particular type of data and tasks at hand.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

8. REFERENCES

[1] M. Anderka and B. Stein. The esa retrieval model revisited. In *SIGIR*, pages 670–671, 2009.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] P. Cimiano, A. Schultz, S. Sizov, P. Sorg, and S. Staab. Explicit versus latent concept models for cross-language information retrieval. In *IJCAI*, pages 1513–1518, 2009.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[5] S. Dumais, T. Letsche, M. Littman, and T. Landauer. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*, 1997.

[6] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611, 2007.

[7] T. Gottron, M. Anderka, and B. Stein. Insights into explicit semantic analysis. In *CIKM*, pages 1961–1964, 2011.

[8] J. Hartigan. *Clustering algorithms*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1975.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[10] M. Littman and F. Jiang. A comparison of two corpus-based methods for translingual information retrieval. Technical report, 1998.

[11] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *EMNLP*, pages 880–889, 2009.

[12] C. Müller and I. Gurevych. Using wikipedia and wiktionary in domain-specific information retrieval. In *Working Notes of the Annual CLEF Meeting*, 2008.

[13] C. Müller, I. Gurevych, and M. Mühlhäuser. Integrating semantic knowledge into text similarity and information retrieval. In *ICSC*, pages 257–264, 2007.

[14] M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *ECIR*, pages 522–530, 2008.

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[16] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[17] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR*, pages 21–29, 1996.

[18] P. Sorg and P. Cimiano. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes of the Annual CLEF Meeting*, 2008.

[19] M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, pages 1419–1424, 2006.

[20] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalized vector space model in information retrieval. In *SIGIR*, pages 18–25, 1985.

[21] Y. Yang, J. G. Carbonell, R. D. Brown, and R. E. Frederking. Translingual information retrieval: Learning from bilingual corpora. *Artif. Intell.*, 103(1-2):323–345, 1998.