FP7-ICT-2011.4.2
Contract no.: 288342
www.xlike.org

Deliverable D4.1.1

Cross-lingual document linking prototype

Editor:	Achim Rettinger, KIT
Author(s):	Achim Rettinger, KIT; Lei Zhang, KIT; Jan Rupnik, JSI; Andrej Muhič, JSI;
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M12
Actual Delivery Date:	M12
Suggested Readers:	All partners using XLike Toolkit
Version:	1.0
Keywords:	Factor analysis, Canonical correlation analysis, cross-lingual information retrieval

Disclaimer

This document contains material, which is the copyright of certain XLike consortium parties, and may not be reproduced or copied without permission.

All XLike consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the XLike consortium as a whole, nor a certain party of the XLike consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full Project Title:	XLike – Cross-lingual Knowledge Extraction
Short Project Title:	XLike
Number and Title of Work package:	WP4 – Cross-lingual Semantic Integration
Document Title:	D4.1.1 Cross-lingual document linking prototype
Editor (Name, Affiliation)	Achim Rettinger, KIT
Work package Leader (Name, affiliation)	Marko Grobelnik, JSI
Estimation of PM spent on the deliverable:	11 PM

Copyright notice

Executive Summary

One aspect of the XLike project is to monitor and aggregate knowledge from web text contents from around the globe. In order to achieve this goal it is necessary to compare and link documents (e.g. contents of Wikipedia articles) together that address the same topics or are about the same event type. However, one problem in today's society is that not all documents are composed in the same language. Thus, it is crucial to measure the similarity of multilingual documents to be able to group documents, which deal with the same topic. Techniques from cross-lingual information retrieval, such as Explicit Semantic Analysis and Canonical Correlation Analysis, are used to determine the similarity of documents. Press agencies could for example examine whether their lead story has already been published in another language or vice versa continuously search for stories that have not yet been published in their language. This deliverable provides a matrix of cross-lingual similarity functions for XLike languages.

Generally speaking, cross-lingual document linking compares pairs of documents using a background corpus. The background corpus used for this task can be parallel, such as JRC-ACQUIS Corpus, or comparable, such as Wikipedia articles with language links. During corpus creation it is also necessary to apply standard pre-processing techniques such as stop word removal and stemming. A problem of cross-lingual document linking is vocabulary mismatch. Although two documents might contain information about the same topic, the vocabulary used to describe the information might be different, especially in the multilingual/cross-lingual setting. Thus, it is critical not to rely on a simple overlap of words to link documents but rather identify statistical links.

In our experimental evaluation, we address two different scenarios based on the use cases: 1) cross-lingual plagiarism detection and 2) cross-lingual recommendation. The task of cross-lingual plagiarism detection is to track the republishing of its articles that are translated into other languages and the task of cross-lingual recommendation is to provide a list of recommended articles from a multilingual news stream. Since cross-lingual plagiarism detection and cross-lingual recommendation require different similarity criteria, we carried out two separate experiments: one which searches specifically for translation, and one which generally searches for related documents. In the experiments, we compared the K-means clustering, LSI, ESA and CCA based approaches. We draw the following conclusions. The results show that all the approaches achieve similar performance for cross-lingual recommendation scenario. However, for cross-lingual plagiarism detection, LSI, ESA and CCA outperform K-means clustering significantly. Among LSI, ESA and CCA, CCA achieves the best results. LSI and ESA yield comparable results that are not far behind CCA.

When in this deliverable we use the term "XLike languages" we will refer to English, German, Spanish, Chinese, Catalan and Slovenian.

Table of Contents

List of Figures and/or List of Tables

Abbreviations

ESA	Explicit Semantic Analysis
LSI	Latent Semantic Indexing
CCA	Canonical Correlation Analysis
TFIDF	Term Frequency–Inverse Document Frequency
MRR	Mean Reciprocal Rank
MAP	Mean Average Precision

Definitions

Parallel Corpus — Parallel corpus consists of documents that are translated directly into different languages.

Comparable Corpus — Comparable corpus contains, unlike parallel corpora, no direct translations. Overall they may address the same topic but can differ significantly in length, detail and style.

1 Introduction

1.1 Motivation

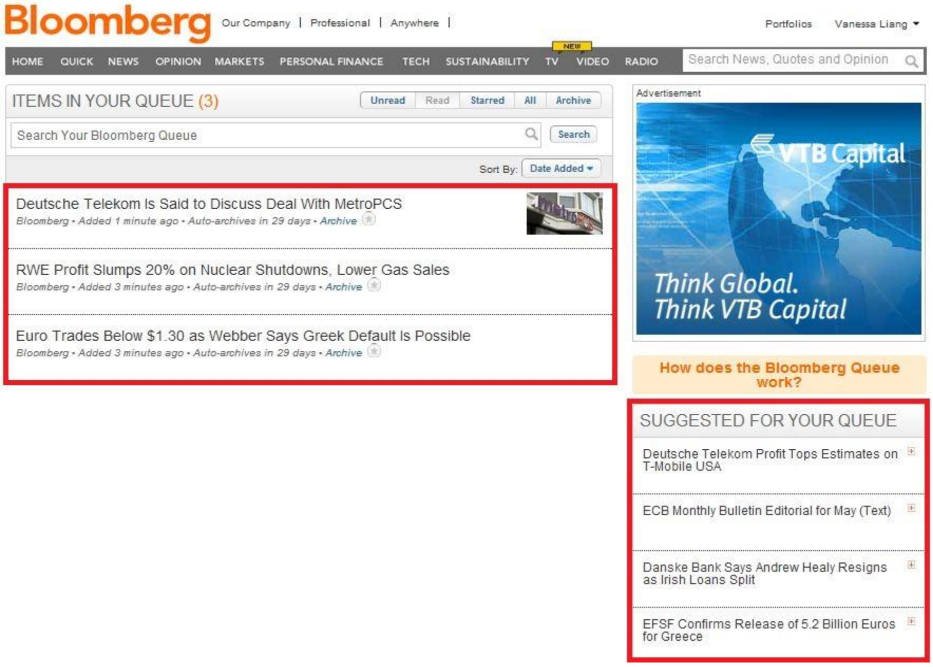
Over the past decades the volume of information available online has been exploding very fast. Never has so much information been available over networks and accessed by so many people independent of their location. Therefore the consequence has been unfortunately brought on: it has become very difficult for users to find the desired, relevant information. Scientists are interested in those articles from the domain on which they focus, like biologists who search for articles about biology, or financial professionals who daily read financial news on the internet. As very many articles about the same topics are available everywhere it’s important to choose only reliable providers of articles and to use comfortable services from information providers. Users don’t want to spend too much time on complicated or repeated searches.

Document to document linking allows users to link related documents together that have similar topics thus allowing users to quickly see a listing of all interesting documents with minimal clicks. To implement document linking there are several technologies available. We will adopt two techniques from the latent topic model (LSI and CCA), one technique from the explicit model (ESA) as well as one technique from clustering (K-means clustering). After the four techniques are implemented and tested we will evaluate the results and thus compare the techniques according to several criteria.

We now live in the age of globalization. The same or related news to news at Bloomberg.com, academic achievements, or general information can be spread instantly across the world through the internet, television, or other media. People in the modern world possess more language skills than before. This phenomenon demands that the techniques for document linking also make corresponding changes because users need and can read articles about the same topics in several languages. Cross-lingual statistical document linking techniques offer the opportunity to allow users to get desired articles about similar topics which are written in many major languages, such as English, German, Chinese etc.

1.2 Bloomberg use case

Bloomberg’s business is the delivery of financial information. The core of their business is based on Bloomberg Terminals, a specialized platform for financial professionals. Besides this, they also maintain more mainstream-oriented news portals at Bloomberg.com. The Bloomberg use case in project XLike will focus on the website part, by evaluating techniques for cross-lingual integration of news articles.


Bloomberg
Our Company | Professional | Anywhere |
Portfolios
Vanessa Liang ▾
HOME QUICK NEWS OPINION MARKETS PERSONAL FINANCE TECH SUSTAINABILITY TV NEW VIDEO RADIO
Search News, Quotes and Opinion
Unread Read Starred All Archive
Search Your Bloomberg Queue
Search
Sort By: Date Added ▾
Advertisement
VTB Capital
Think Global.
Think VTB Capital

ITEMS IN YOUR QUEUE (3)

Deutsche Telekom Is Said to Discuss Deal With MetroPCS
Bloomberg • Added 1 minute ago • Auto-archives in 29 days • Archive

RWE Profit Slumps 20% on Nuclear Shutdowns, Lower Gas Sales
Bloomberg • Added 3 minutes ago • Auto-archives in 29 days • Archive

Euro Trades Below $1.30 as Webber Says Greek Default Is Possible
Bloomberg • Added 3 minutes ago • Auto-archives in 29 days • Archive

How does the Bloomberg Queue work?

SUGGESTED FOR YOUR QUEUE

Deutsche Telekom Profit Tops Estimates on T-Mobile USA

ECB Monthly Bulletin Editorial for May (Text)

Danske Bank Says Andrew Healy Resigns as Irish Loans Split

EFSF Confirms Release of 5.2 Billion Euros for Greece

Figure 1. Stored articles in the queue and the recommendation list of related articles at Bloomberg.com

Bloomberg.com provides personalized lists of suggested articles alongside each article. The list is assembled from the recent Bloomberg articles and custom fitted for the specific user, based on his/her history. In Figure 1, the user's read stories are stored in their queue; based on that a suggested list for their queue is generated. The second task in Bloomberg use case is to extend the suggested articles by including external mainstream sources across more languages. Formally, the task is defined as follows. Given a user u, with visit history H(u), identify relevant recent articles from multi-lingual news stream. All articles from H(u) are in English and published by Bloomberg. Assembling a relevant recent articles list requires cross-lingual integration with Bloomberg.com articles.

1.3 STA use case

Like in the Bloomberg use case here are also two tasks in the STA use case: article tracking as well as topic and entity tracking. Article tracking is meaningful for STA because of two business cases. First, the main income of agency is licensing its content, and publish unlicensed material requires their attention. Second, knowing which articles are republished by their subscribers helps the agency at better understanding their market, and to provide better coverage for the events relevant for them.

STA covers topics related to Slovenia or Slovenian entities (E.g. companies, athletes). As such, tracking relevant news is an important part of editors’ daily routine. Technologies developed within XLike project can improve this process by providing tools for detecting relevant articles across languages and media (mainstream, social media).

STA covers domestic and international events by producing and selling copyrighted content. However, protection and tracking of copyrighted material on the Web is an open problem especially when the material involved is translated in the process. To solve the above problem an application should be developed which uses the components for cross-lingual document linking and information flow to detect articles, which have significant overlap with the source article.

2 Theories about techniques for statistical document linking

Cross-lingual document linking is defined as follows: given one document in any language, identify relevant documents based on their similarities to the input document from a multi-lingual document collection. Cross-lingual document linking deals with retrieving information written in a language different from the language of the target-document. For example, we have to compare a target-document written in English and a source-document written in German in order to find out whether the source-document is relevant to the target-document. With cross-lingual document linking technique we can link a certain article to a corresponding article in a certain website like 'Bloomberg.com' in another language. Two kinds of approaches for cross-lingual document linking are available: translation-based approaches and concept-based approaches. We will focus on the latter ones.

In a general framework we will explain basic knowledge about cross-lingual document linking techniques and show how the approaches are developed step by step. First at all, our research objects are corpora, documents and terms. We represent documents and terms in numerical and geometric framework in order to form models and execute algorithms. Several measures are introduced based on the representations of documents and terms, such as similarity and tf-idf measure.

2.1 Translation-based approach

Translation-based approaches translate two documents into a certain language. Several translation techniques are available for it and they differ in the choice of that into which language documents are translated. We can either translate source-document into the language of target-document, or translate target-document into the language of source-document. As alternative of above techniques we can also find a pivot language into which both documents are translated. The pivot language should be available for translation systems from many languages. Hence, English is mostly adopted as pivot language because of wide applicability in translation systems.

Translations can be either gained by professional manual translators or through the application of Machine translation. Since the manual translation of documents does not scale to large corpora and requires some processing time the manual translation would be replaced by machine translation. Two popular approaches are mostly employed to machine translation: dictionary-based translation and statistical machine translation. Dictionary-based translation use bilingual dictionaries for term-by-term translation. The approach is straight-forward, but not suitable to big corpus. For big corpus the statistical machine translation can be used instead of dictionary-based translation. Most current SMT systems are developed based on the IBM Models which is introduced by Brown. These models are iteratively induced for language pairs on the basis of a training multilingual corpus. Translation can be improved by using additional background knowledge, for example by using language models derived from large monolingual training corpora.

Through translation of documents the multilingual document linking problem will be reduced to monolingual document linking problem. Therefore the weaknesses of monolingual information retrieval are also exposed, like vocabulary mismatch, term sense disambiguation etc. In addition, a new problem of term mismatch is formed through translation. As well know, many terms in a target language correspond to one term in source language, it depends on meanings of terms, context in document and preference of professional translators or machine translation systems. For this reason one term in a source language can be translated as several synonymous terms in target language. In sentence environment one paragraph in source language is translated according to context in the whole document, rules and conventions in target language. So the paragraphs are not translated term-by-term. Furthermore, the machine translation systems mostly show poor accuracy. Hence, in source document and in target document the terms cannot be perfectly matched because of the flexibility of terms.

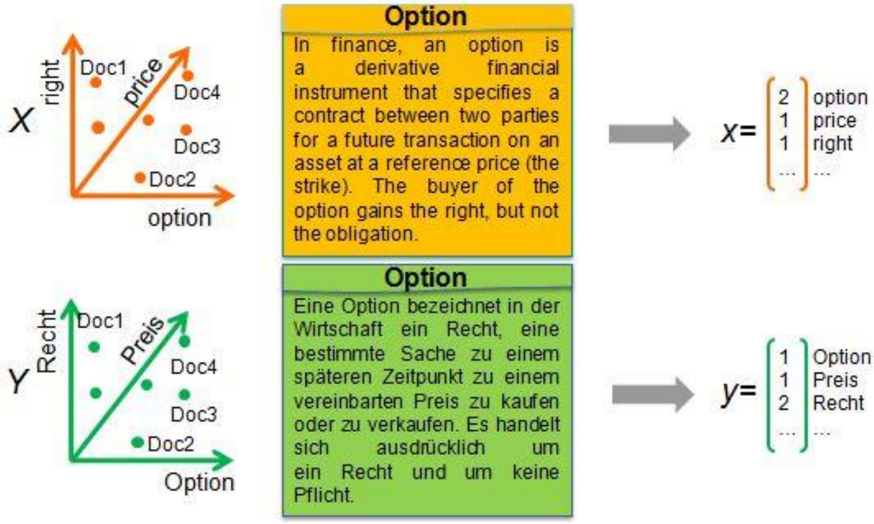
2.2 Inter-lingual Representation based Approach

Due to the problem of translation-based approaches, we employ the concept-based approaches. We employ 4 theories ‘K-means Clustering’, ‘ESA’, ‘LSI’ and ‘CCA’ as bases for cross-lingual document linking approaches, thus the approaches can be interpreted differently with cross-lingual generalized vector space model. Based on the analysis with inter-lingual representation based extension of generalized vector space model, these approaches can be compared easily and comprehensibly. The monolingual generalized vector space model is a standard model, but here we need the derived version of standard model for our use case, that is inter-lingual representation based extension of generalized vector space model.

2.2.1 General Framework

In concept-based approaches, documents are transformed into a multilingual space. To realize the transformation the aligned concept vectors are introduced. At starting point documents in each language are represented as vectors of term weights in each corresponding vector space. The term vectors in different languages (i.e. different term vector spaces) should be transformed with help of aligned concept vectors into multilingual concept space. The aligned concept vectors can be seen as descriptions of concepts in different languages and the concepts might be explicit and external defined or data-derived, for example, in the case of ESA. We group relevant terms together to a topic and form a concept. The concepts can be identical although they are in different languages, they can also differ in variant term descriptions or they can be overlapped in some terms. With help of aligned concept vectors we can get the new document representations in a multilingual concept space where no more terms exist, just multilingual concepts are built up. The multilingual concept space is a set of concepts, which are multilingual, and is spanned by aligned concept vectors. A multilingual concept represents unit of thoughts and contains all terms related to a topic in several source languages. Many concepts can be separated or overlapped in a concept space. Thus, the concept-based representations of documents are language-independent because the possibilities of terms occurring in several source languages are already considered in concepts. Under the common multilingual concept space the similarity between two documents which are represented as new vectors can be calculated again.

With an example we see before the transformation how the documents would be represented under Vector space model (VSM).



Option

In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.

Option

Eine Option bezeichnet in der Wirtschaft ein Recht, eine bestimmte Sache zu einem späteren Zeitpunkt zu einem vereinbarten Preis zu kaufen oder zu verkaufen. Es handelt sich ausdrücklich um ein Recht und um keine Pflicht.

Figure 2. VSM-based document representation of two documents in different languages

Here parameters have following meanings:

- Two vector spaces X and Y for two languages L_X and L_Y;
- Representation of documents as vectors x and y in X and Y;

$$x^T = (x_1, x_2, \dots, x_p), \quad y^T = (y_1, y_2, \dots, y_q)$$

- x_i and y_j are weights of term i and j in x and y respectively ;

- p and q are sizes of vocabularies in the different languages LX and LY.

Each document can be represented as a real-valued vector of term weights using vector space model. Each language can be modeled as a vector space. The dimension of the vector space is the size of the vocabulary, i.e. the number of terms that make up the language. Terms are axes of the vector space. Documents are points in this space.

In order to compare the documents (representations) in different languages (vector spaces), we need to find aligned concept vectors for these vector spaces, where each pair of corresponding vectors in these aligned concept vectors are supposed to represent the same topic/concept. Each document (representations) in different languages (vector spaces) can be represented by these aligned concept vectors, where each entry corresponds to the inner product of the document term vector and concept vector.

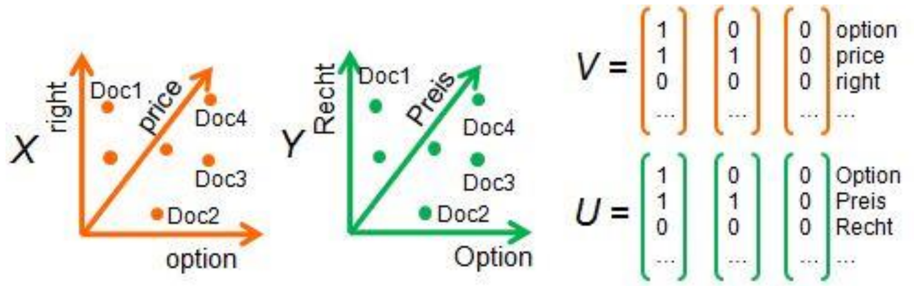


Figure 3. Step 1: Find aligned concept vectors V and U

In order to transform documents vector x and y into a multilingual concept space we should at first find out two aligned concept vectors V for X and U for Y, where vi should represent the same topic/concept as u_i.

$$\mathbf{V} = (\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}), \quad \mathbf{U} = (\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_n})$$

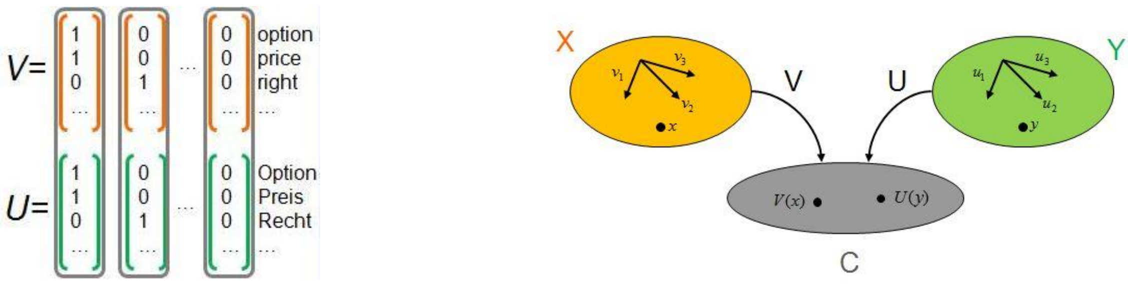
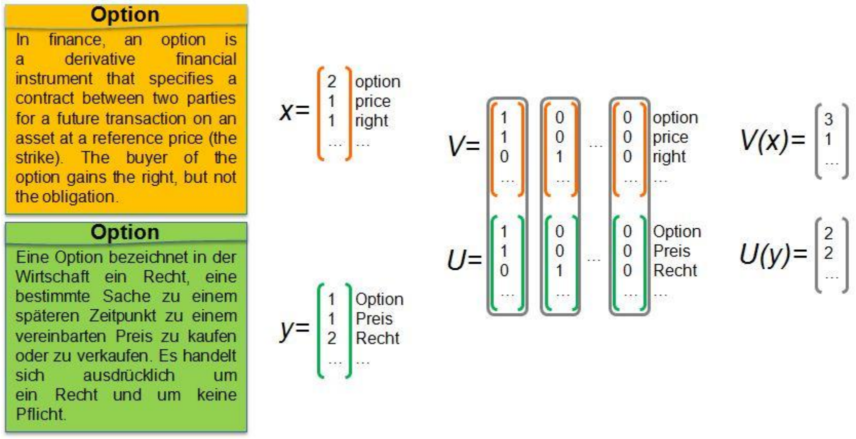


Figure 4. Step 2: Find aligned topic/concept space C

An aligned topic/concept space C is spanned by V and U such that each dimension corresponds to a pair of aligned vectors in V and U.



Option

In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.

Option

Eine Option bezeichnet in der Wirtschaft ein Recht, eine bestimmte Sache zu einem späteren Zeitpunkt zu einem vereinbarten Preis zu kaufen oder zu verkaufen. Es handelt sich ausdrücklich um ein Recht und um keine Pflicht.

Figure 5. Step 3: Represent documents as V(x) and U(y)

Representation of documents as vector V(x) and U(y) in C:

$$\mathbf{V(x)^T} = (<\mathbf{v_1}, \mathbf{x}>, <\mathbf{v_2}, \mathbf{x}>, \dots, <\mathbf{v_n}, \mathbf{x}>),$$

$$\mathbf{U(y)}^{\mathbf{T}} = (<\mathbf{u_1}, \mathbf{y}>, <\mathbf{u_2}, \mathbf{y}>, \dots, <\mathbf{u_n}, \mathbf{y}>),$$

The similarity between document vectors x in X and y in Y in Figure 5 is calculated as follow:

$$\mathbf{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{cos}\big(\mathbf{V}(\mathbf{x}), \mathbf{U}(\mathbf{y})\big) = \frac{<\mathbf{V}(\mathbf{x}), \mathbf{U}(\mathbf{y})>}{|\mathbf{V}(\mathbf{x})| \cdot |\mathbf{U}(\mathbf{y})|}$$

Further it's derived:

$$<\boldsymbol{V}(\mathbf{x}), \mathbf{U}(\mathbf{y})> = \mathbf{V}(\mathbf{x})^{\mathbf{T}} \cdot \mathbf{U}(\mathbf{y}) = \left(\mathbf{x}^{\mathbf{T}} \cdot \mathbf{V}\right) \cdot \left(\mathbf{U}^{\mathbf{T}} \cdot \mathbf{y}\right) = \mathbf{x}^{\mathbf{T}} \cdot \left(\mathbf{V} \cdot \mathbf{U}^{\mathbf{T}}\right) \cdot \mathbf{y} = \mathbf{x}^{\mathbf{T}} \cdot \mathbf{G} \cdot \mathbf{y}$$

$$|\mathbf{V}(\mathbf{x})| = \sqrt{\mathbf{V}(\mathbf{x})^{\mathbf{T}} \cdot \mathbf{V}(\mathbf{x})} = \sqrt{(\mathbf{x}^{\mathbf{T}} \cdot \mathbf{V}) \cdot (\mathbf{V}^{\mathbf{T}} \cdot \mathbf{x})} = \sqrt{\mathbf{x}^{\mathbf{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathbf{T}}) \cdot \mathbf{x}} = \sqrt{\mathbf{x}^{\mathbf{T}} \cdot \mathbf{G}' \cdot \mathbf{x}}$$

$$|\mathbf{U}(\mathbf{y})| = \sqrt{\mathbf{U}(\mathbf{y})^{\mathbf{T}} \cdot \mathbf{U}(\mathbf{y})} = \sqrt{(\mathbf{y}^{\mathbf{T}} \cdot \mathbf{U}) \cdot (\mathbf{U}^{\mathbf{T}} \cdot \mathbf{y})} = \sqrt{\mathbf{y}^{\mathbf{T}} \cdot (\mathbf{U} \cdot \mathbf{U}^{\mathbf{T}}) \cdot \mathbf{y}} = \sqrt{\mathbf{y}^{\mathbf{T}} \cdot \mathbf{G}'' \cdot \mathbf{y}}$$

Here $\mathbf{G}, \mathbf{G}', \mathbf{and}\ \mathbf{G}''$ are term correlation matrices.

Hence, the similarity between document vectors x and y can be calculated as follows:

$$\mathbf{sim}(\mathbf{x}, \mathbf{y}) = \mathbf{cos}\big(\mathbf{V}(\mathbf{x}), \mathbf{U}(\mathbf{y})\big) = \frac{<\boldsymbol{V}(\mathbf{x}), \mathbf{U}(\mathbf{y})>}{|\mathbf{V}(\mathbf{x})| \cdot |\mathbf{U}(\mathbf{y})|} = \frac{\mathbf{x}^{\mathbf{T}} \cdot \left(\mathbf{V} \cdot \mathbf{U}^{\mathbf{T}}\right) \cdot \mathbf{y}}{\sqrt{\mathbf{x}^{\mathbf{T}} \cdot (\mathbf{V} \cdot \mathbf{V}^{\mathbf{T}}) \cdot \mathbf{x}} \cdot \sqrt{\mathbf{y}^{\mathbf{T}} \cdot (\mathbf{U} \cdot \mathbf{U}^{\mathbf{T}}) \cdot \mathbf{y}}}$$

$$= \frac{\mathbf{x}^{\mathbf{T}} \cdot \mathbf{G} \cdot \mathbf{y}}{\sqrt{\mathbf{x}^{\mathbf{T}} \cdot \mathbf{G}' \cdot \mathbf{x}} \cdot \sqrt{\mathbf{y}^{\mathbf{T}} \cdot \mathbf{G}'' \cdot \mathbf{y}}}$$

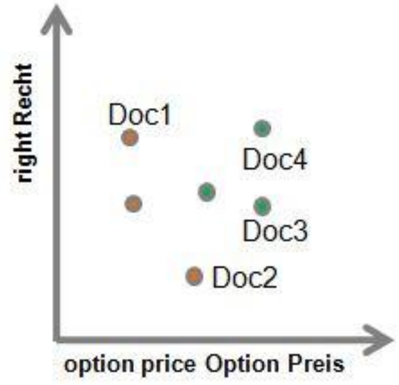


Figure 6. Step 4: Represent documents in low-dimensional concept space

We combine two dimensions 'option', 'price' of vector space X and two dimensions 'Option', 'Preis' of vector space Y together as a new dimension 'option price option preis'. And we combine dimension 'right' of vector space X and dimensions 'Recht' of vector space Y together as a new dimension 'right Recht'. The new dimensions are multilingual concepts. So we can represent documents in a two-dimensional vector space after transformation from term representation to concept representation.

2.2.2 Background data corpora

We develop four approaches for cross-lingual document linking: K-means based approach, ESA based approach, LSI based approach and CCA based approach. With each of those approaches we can determine concepts from given corpus and further execute document linking tasks in a generalized vector space model. After we built up models and developed approaches we will execute experiments on these four approaches and evaluate them. For the evaluation we use parallel and comparable multilingual corpora.

From parallel multilingual corpora we take JRC-Acquis corpus for experiments. From comparable multilingual corpora we use Wikipedia corpus. Parallel corpora are suitable to almost all types of cross-lingual research. The larger the size of a parallel corpus and the larger the number of languages, for which translations exist, the greater is the value of a parallel corpus.

Joint Research Centre Collection of the Acquis Communautaire, abbreviated as JRC-Acquis, is a multilingual parallel corpus extracted from Acquis Communautaire. Acquis Communautaire (AC) is a French term that means ‘the EU as it is’. This is a body of common rights and obligations which bind all the Member States together within the European Union. The JRC-Acquis corpus is the biggest parallel corpus in existence. The corpus is available in 22 languages, consists of almost 8000 documents per language, with average size of nearly 9 million words per language. The corpus is available in TEI-compliant XML format, and consists of two parts, the marked-up texts and the bilingual alignment information for all the 190 language pairs. With help of Eurovoc thesaurus the JRC-corpus manually is classified into subject domains. At following webpage the JRC-Acquis corpus is available for download: http://langtech.jrc.it/JRC-Acquis.html

Wikipedia is currently the largest knowledge treasure on the web in the world and it is developed constantly by diverse editors, therefore its breadth and depth also are expanded continually. The documents at Wikipedia are available in 260 languages and they are linked to each other cross languages in case they describe the same topic. The most documents at Wikipedia are available in English they are currently more than 4 million documents. The German, Spanish, French, Italian and other 5 Wikipedias contain more than 750 thousand documents. The Chinese, Slovenian and other 18 Wikipedias contain more than 150 thousand documents. The Greek, Norwegian and other 8 Wikipedias have more than 50 thousands documents. The documents not only are available in large number of different languages also available in diverse domains. That is why we use Wikipedia corpus as comparable corpus. The Wikipedia database dumps you can download at http://dumps.wikimedia.org/ . There page content, page-to-page link list, image metadata and misc bits are available for download in different XML wrapper formats. The download tool for Wikipedia database dumps you can find at https://github.com/babilen/wp-download/ .

2.3 K-means Clustering based approach

The term ‚K-means‘ was first used by James MacQueen in 1967. K-means is a partitioning approach with exact assignment, which uses cluster centres to form clusters. Therefore K-means also belongs to centroid-based approaches.

Given a set of observations (x1, x2, ⋯, xn), where each observation is a d-dimensional real vector, K-means clustering aims to partition the n observations into k sets (k ⩽ n) S = {S1, S2, ⋯, Sk} so as to minimize the within-cluster sum of squares.

$$\sum_{\mathbf{k}}\sum_{\mathbf{g}\in\mathbf{k}}\sum_{\mathbf{j}}(\mathbf{x}_{\mathbf{gj}}-\bar{\mathbf{x}}_{\mathbf{kj}})^{2} \quad \rightarrow \mathbf{min}\ (*)$$

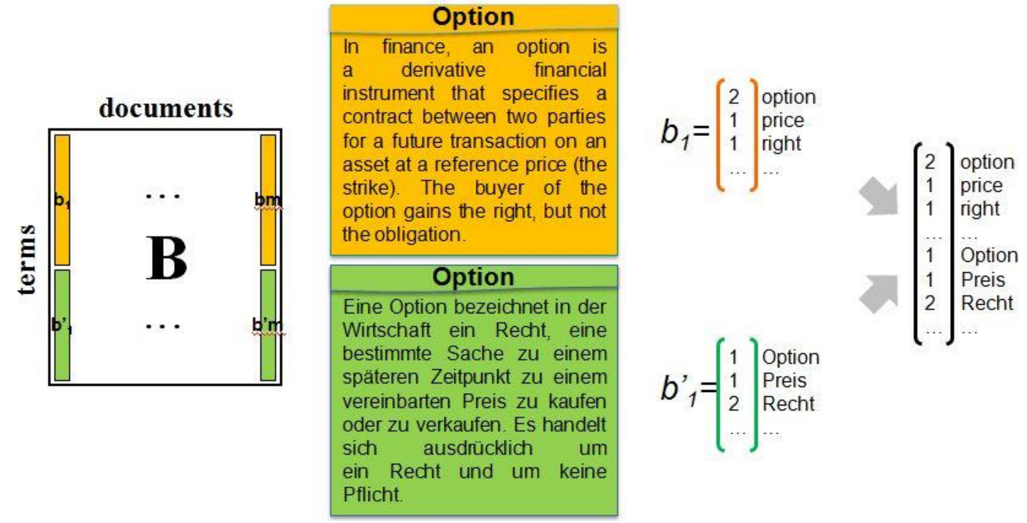
where K = number of clusters (k=1,…, K)

m = number of variables (j = 1, ... , m)

$\mathbf{x}_{\mathbf{gj}}$ = value of variable j for object g

$\bar{\mathbf{x}}_{\mathbf{kj}}$ = cluster center for variable j in cluster k

In the K-means clustering based approach K-means clustering technique is used to find out concepts from given documents collection. At the end of process documents are represented as vectors, where each dimension is a cluster which is derived from the given documents collection.



In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.

Eine Option bezeichnet in der Wirtschaft ein Recht, eine bestimmte Sache zu einem späteren Zeitpunkt zu einem vereinbarten Preis zu kaufen oder zu verkaufen. Es handelt sich ausdrücklich um ein Recht und um keine Pflicht.

Figure 7. Matrix representation of input documents

In Figure 7, each document in one language is represented as a vector of term weights using Vector Space Model (VSM). Two documents 'Option' in English and 'Option' in German are represented as vectors b_1 and b_1'. They are integrated as one multilingual vector in the term-document matrix B, we call it object. Through similar way other documents couple in English and in German can also be bound together. So we got the term-document matrix in which corpora in English and in German as input documents are represented through correlations with terms in that.

This approach first concatenates all vectors of the parallel or comparable documents in different languages and finds the best cluster vectors, after that decouple them to obtain the aligned concept vectors.

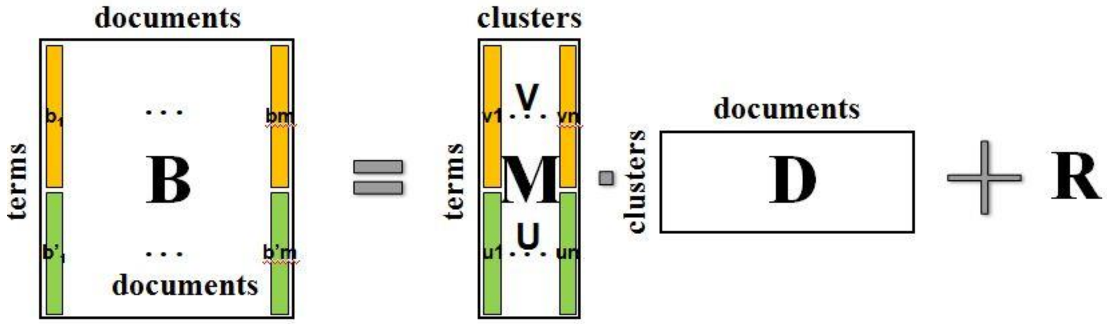


Figure 8. Matrix representation of K-means clustering based approach

In Figure 8, we will show how a term-document matrix is decomposed into a linear combination of matrices. In the term-document matrix B each column is a vector of term weights for a concatenated multilingual document. Now the matrix B is described through term-cluster matrix M, cluster-document matrix D and a residual matrix R. In the matrix M each column is a vector of term weights for a cluster centroid, each of which is average of term weights for all documents in the cluster. In the matrix D each column contains some '0's and a cluster indicator '1' indicating the membership of one document in a cluster. This R is the residual matrix such that clustering minimizes sum of squares of all columns. Each cluster corresponds to a concept. If we find out the clusters through running K-means clustering algorithm we also get the aligned concepts.

2.4 Latent Semantic Indexing based approach

Latent Semantic Indexing (LSI) is a prominent representative of latent models based on Singular Value Decomposition (SVD), and at first outlined by Deerwester in 1990, that is also called Latent Semantic Analysis (LSA). Latent model computes latent concepts or dimensions in documents and builds latent topic structure over terms. Compared to explicit model in latent model apart from observed vectors some hidden characters exist.

The vector space representation cannot deal with two in natural languages arising classic problems: synonymy and polysemy. Synonymy is a kind of semantic relation which describes the situation that two terms have the same meaning, thus we say the two terms are synonyms, for example, bank and financial institute. As vector space representation cannot capture the relationship between synonymous terms such as bank and financial institute because each term has a separate dimension in the vector space, the problem can be solved with LSI since through dimension reduction the similar terms can be assigned in a same dimension. Polysemy refers to a term that has two or more similar meanings, for example, 'Foot' refers to the bottom part of the mountains or the bottom part of the leg. Unfortunately LSI overcomes the problem just partially not completely because of its linear nature.

In LSI Singular Value Decomposition is applied in order to compute the approximation matrix to term-document matrix by decomposition the approximation matrix into three matrices. Low-rank approximation to term-document matrix is also used and combined in Singular Value Decomposition due to yield a new representation of term-document matrix.

In the approach of SVD the data set with high dimension and many variables is reduced to a lower dimensional vector space where substructure of the original data is built and amount of variations is also reduced. SVD has many characters, like transforming correlated variables into a set of uncorrelated ones; identifying and ordering the dimensions along which data points exhibit the most variation; reducing dimensions of data points.

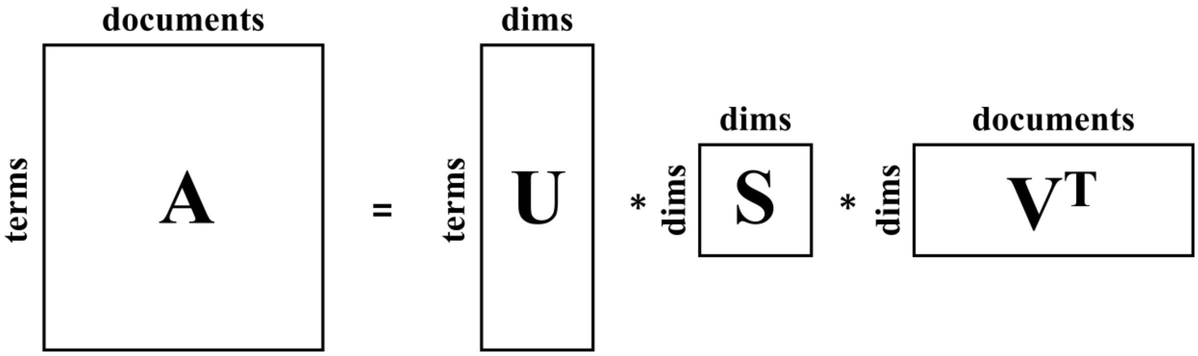


Figure 9. Singular value decomposition of term-document matrix A

In Figure 9 an original term-document matrix A is broken down into the product of three matrices U, S and V. The theorem is represented as follow:

$$A_{td} = U_{tm} * S_{mm} * V_{md}^{T}$$

where A_{td} is the original rectangular t x d matrix of terms and documents. Each column is a vector of term weights for a document. U_{tm} is an t x m term matrix whose columns are the orthogonal eigenvectors of AA^T, and whose each row vector is represented for each term. Each entry indicates how strongly a term is related to the semantic dimension. S_{mm} is an m x m singular value matrix with the singular values on the diagonals, since all its entries outside this sub-matrix are zeros. Each singular value reflects importance of the corresponding semantic dimension. V_{md}^{T} is the transpose of V_{dm}. V_{dm} is an d x m document matrix whose columns are the orthogonal eigenvectors of A^TA, and whose each row vector is represented for each document. Each entry in this matrix indicates how strongly a document is related to the topic represented by the semantic dimension.

Formally, we represent term-document matrix A_{td} by tf.idf-representing single entry of the matrix: $a_{ij} = tf(w_j\,,\ d_i) \cdot idf(w_j)$, where d_i (1 ≤ i ≤ d) refers to the i-th document and w_j (1 ≤ j ≤ t) refers to the j-th term. U_{tm} and V_{dm} denote matrices such that $U_{mt}^{T} \cdot U_{tm} = V_{md}^{T} \cdot V_{dm} = I_m$. Here I_m is an identity matrix: $\mathrm{I_m} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & 1 \end{bmatrix}$. $\mathrm{S_{mm}}$ is a diagonal matrix which contains the square roots of eigenvalues of $\mathrm{AA^T}$ or $\mathrm{A^TA}$. $\mathrm{S_{mm}} = \begin{bmatrix} \mathrm{s_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{s_m} \end{bmatrix}$. The single entry of matrix $\mathrm{S_{mm}} : \mathrm{s_i} = \sqrt{\lambda_\mathrm{i}}$. Here $\lambda_1, \lambda_2, \ldots, \lambda_\mathrm{m}$ are eigenvalues of $\mathrm{AA^T}$ or eigenvalues of $\mathrm{A^TA}$ (They are the same.). The rank of original term-document matrix $\mathrm{A_{td}}$ is m, here m ≤ min (t , d).

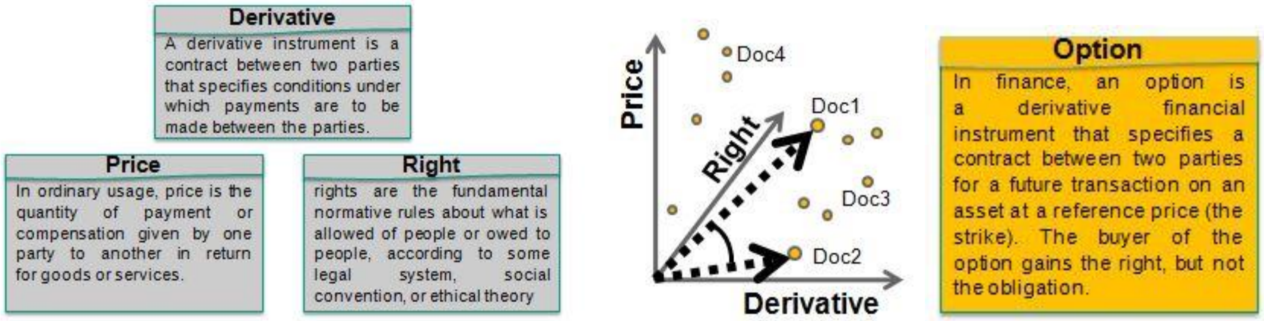
In LSI approach documents are represented as vectors in a concept/topic space, where topics are mined from the given text. Through decomposition of term-document matrix some dimensions are determined. The dimensions are also concepts. If we execute the LSI approach then we will get dimensions i.e. concepts.

2.5 Explicit Semantic Analysis based approach

Explicit Semantic Analysis (ESA) is a recent prominent example of explicit models, and developed by Gabrilovich and Markovitch in 2007. Explicit model is a concept based retrieval model, allows us to explicitly represent the meaning of documents based on concepts. In an explicit model external defined concepts are given, we manipulate manifest concepts grounded in the human cognition. The ESA method represents semantics of natural language texts using natural concepts, is easy to explain to human users.

Given external defined concepts C = { $\mathrm{c_1, c_2, \ldots, c_n}$ } classical monolingual Explicit Semantic Analysis takes a document x represented by a term vector as input and maps it to a concept vector. This concept vector space is spanned by a given collection of documents like $\mathrm{D_k} = \{\mathrm{D_1, D_2, \ldots, D_n}\}$ in language $\mathrm{L_x}$ such that each dimension corresponds to a document. For standard vector space, synonyms contribute nothing to document similarity. Documents that are semantically similar, i.e. talking about the same topics, might be not similar in the vector space if they use different words. ESA and LSI address the problems of synonym and semantic relatedness.

The process how we transform term-vectors into concept space is like in GVSM. The particular setting in ESA is that each document of corpus corresponds to a concept. With an example we can see how a document term-vector is transformed into concept-vector through ESA approach.



Derivative

A derivative instrument is a contract between two parties that specifies conditions under which payments are to be made between the parties.

Price

In ordinary usage, price is the quantity of payment or compensation given by one party to another in return for goods or services.

Right

rights are the fundamental normative rules about what is allowed of people or owed to people, according to some legal system, social convention, or ethical theory

Option

In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.

Figure 10. Document representation in concept space

The three documents ‘Derivative’, ‘Price’ and ‘Right’ are seen as three concepts and are described as three dimensions in a concept space. An input document ‘Option’ is represented in the three-dimensional concept space. How the document ‘Option’ is represented by the three concepts we will explain through Figure 10.

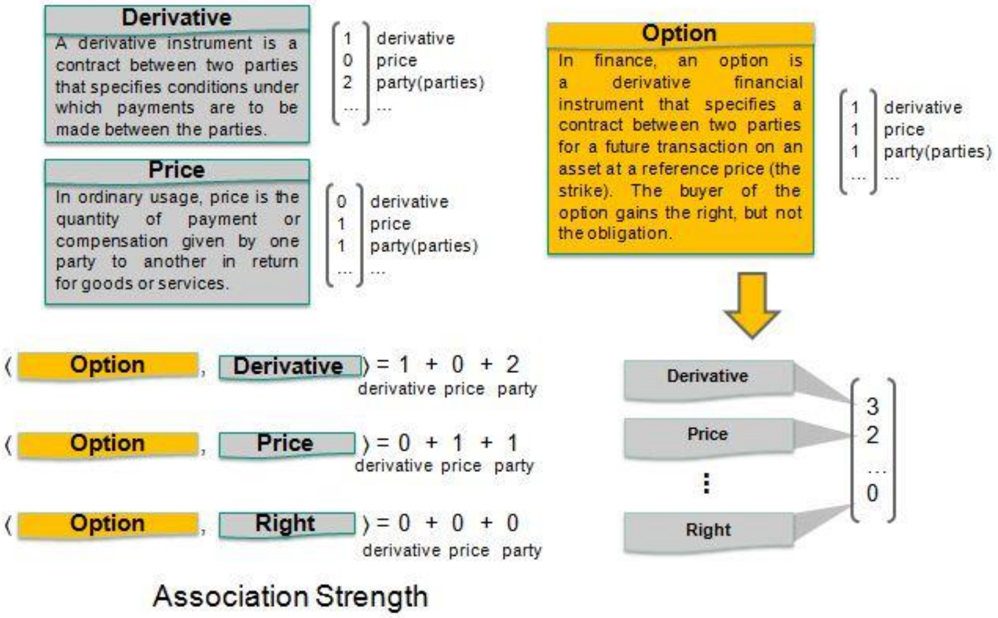
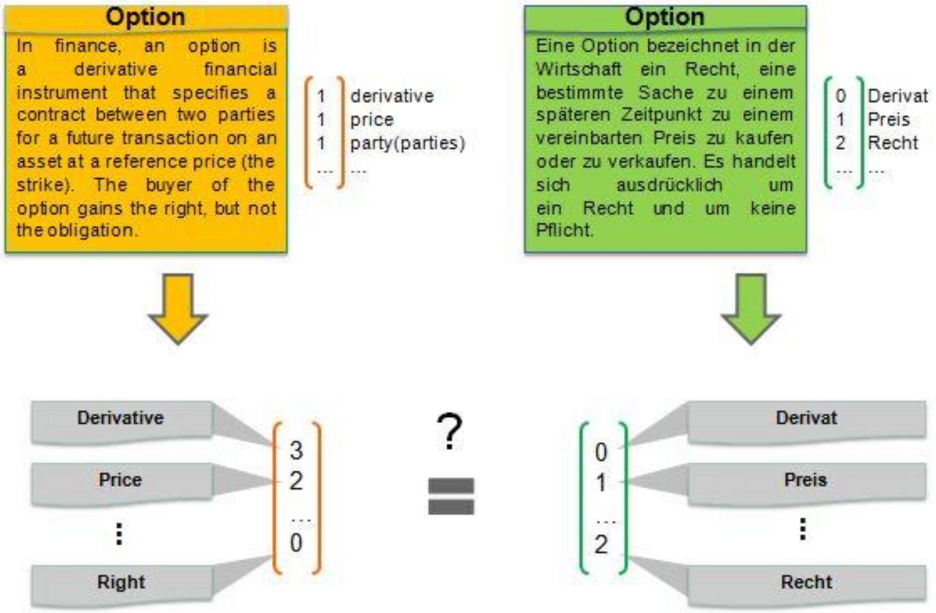

Derivative
A derivative instrument is a contract between two parties that specifies conditions under which payments are to be made between the parties.
1 0 2 …
derivative
price
party(parties)
…
Price
In ordinary usage, price is the quantity of payment or compensation given by one party to another in return for goods or services.
0 1 1 …
derivative
price
party(parties)
…
Option
In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.
1 1 1 …
derivative
price
party(parties)
…
(Option , Derivative) = 1 + 0 + 2
derivative price party
(Option , Price) = 0 + 1 + 1
derivative price party
(Option , Right) = 0 + 0 + 0
derivative price party
Association Strength
Derivative
Price
⋮
Right
3
2
…
0

Figure 11. Transforming term-vector into concept-vector through association strengths

We take terms ‘derivative’, ‘price’ and ‘party(parties)’ in order to build up term vector representation of concepts. The input document ‘Option’ can also be represented as term vector with the three terms in a vector space. After both concepts and input document are represented as term vectors we can calculate the association strengths between the input document and concepts. Association strength expresses the strength of association between terms of an input document and a concept. The association strengths between ‘Option’ and ‘Derivative’, between ‘Option’ and ‘Price’ as well as between ‘Option’ and ‘Right’ are calculated as in Figure 11. Each value expresses the strength of association between document "Option" and each background document. The association strengths can also be seen as the weights of concepts in input document. Through the association strengths the input document is represented as concept vector in concept space. In order to speed up processing and yield more compact vectors, we consider only the top n dimensions of the ESA vectors by using projection of the vectors and selecting only the n dimensions with the highest values.

Cross-lingual ESA is a generalization of monolingual ESA. Like monolingual ESA the concepts are also given by C = $\{ c_1, c_2, \dots, c_n \}$. Apart from that a set of languages is given by L = $\{ L_1, L_2, \dots, L_m \}$, accordingly a set of document collection is given by $D = \{ D_1, D_2, \dots, D_m \}$, where each D_i contains documents of language L_i.

As above two figures described the representation of the input document ‘Option’ in English language is from term vector transformed into concept vector. Similarly the representation of an input document ‘Option’ in German language can also be transformed into concept vector. Please see the results in Figure 12 which are brought by application of ESA approach. Each concept in English corresponds to a concept in German. We assume that the two concepts in two languages are identical. But actually another suitable concept in German exists for a concept in English. So we can only say that the concepts in English and in German approximate to each other. Based on that the documents in both languages are represented in concept space and the concepts in English and in German are comparable the similarity between two documents can be calculated.



Option

In finance, an option is a derivative financial instrument that specifies a contract between two parties for a future transaction on an asset at a reference price (the strike). The buyer of the option gains the right, but not the obligation.

Option

Eine Option bezeichnet in der Wirtschaft ein Recht, eine bestimmte Sache zu einem späteren Zeitpunkt zu einem vereinbarten Preis zu kaufen oder zu verkaufen. Es handelt sich ausdrücklich um ein Recht und um keine Pflicht.

Figure 12. Comparing two documents in two languages as concept vectors

In ESA approach documents are represented as vectors in a concept space, where each dimension corresponds to a given textual description. ESA is explicit in the sense that the concept space corresponds exactly to the background document space.

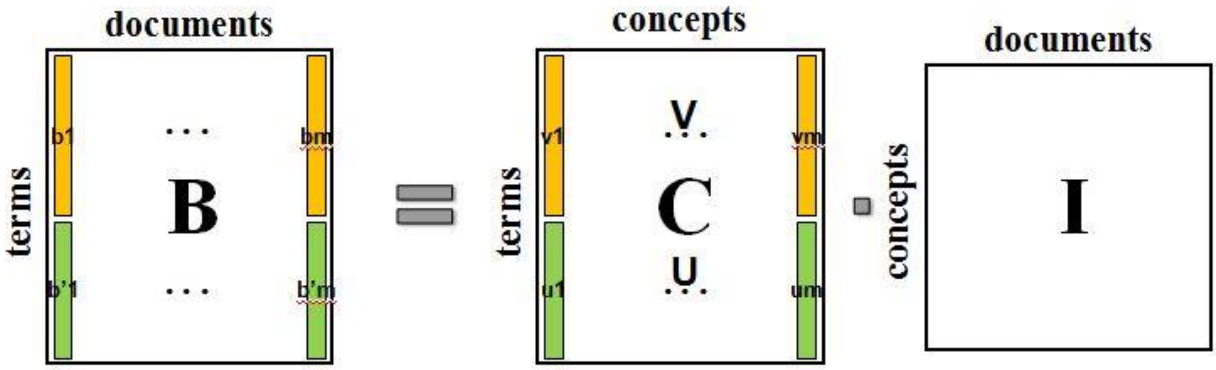


Figure 13. Matrix representation of ESA based approach

In the term-document matrix B each column is a vector of term weights for a concatenated multi-lingual document. This term-concepts matrix C is the same matrix as matrix B because each document corresponds to a concept. This matrix I is an identity matrix.

2.6 Canonical correlation analysis based approach

The canonical correlation approach is motivated by the following observations:

- When on increases the number of languages in the analysis, the aligned document set size decreases (possibly to an empty set) – the intersection of Wikipedia pages available in all the languages becomes prohibitively small. One option is to treat the missing documents as empty documents. This biases the U matrix towards zero, which is undesirable.
- Even though U is orthogonal, the block components may not be. Using them as projection matrices is not as justified as in the monolingual case.

The approach described in [6] consists of two steps: the first one is dimensionality reduction closely related to LSI, where the difference is that we perform low rank decompositions of the cross-covariance operator which is better suited to deal with missing documents. After this step we refine the projection operators using a canonical correlation based formulation. When considering the structure of links between documents in Wikipedia, we observe that if a document written in a language other than English is linked to another language, it is very likely linked to English. For this reason we refer to English as the hub language and that fact can be exploited in the approach in several places.

Let $L_1 = \mathbb{R}^{N_1}, \dots, L_m = \mathbb{R}^{N_m}$ denote the vector spaces corresponding to m languages and let L_1 denote the hub language (English). Let $C_{i,j} \in \mathbb{R}^{N_i \times N_j}$ denote the empirical estimate of the cross-covariance matrix between language i and language j, which is computed by using the aligned documents between L_1 and L_2 (see [6] for more details). We now use the hub-language assumption and only consider cross-covariances related to the hub language: $C_{1,2}, \dots, C_{1,m}$ and compute the following low-rank decomposition:

$$\left[C_{1,2} \cdots C_{1,m}\right] = U\,S\,V',$$

where $U \in \mathbb{R}^{N_1 \times k}, S \in \mathbb{R}^{k \times k}$ and $V \in \mathbb{R}^{(N_2+\cdots+N_m)\times k}$. We split the matrix V according to the dimensions $N_2, \dots, N_m$: $V = [V_2' \cdots V_m{'}]'$ and define: $U_1 = U$ and $U_i = V_i$. We proceed by projecting the cross-covariance matrices $C_{i,j}$ to:

$$C_{i,j} \leftarrow U_i' C_{i,j} U_j.$$

We now search for concept vectors that maximize linear dependence between the languages by solving a generalized version of canonical correlation analysis (sum of squared correlations [6][7]) where similar to the dimensionality reduction step we only consider cross-covariance matrices related to the hub language. The search problem can be posed as constrained optimization problem:

$$\begin{array}{ll} \max\limits_{w_1,\dots,w_m \in \mathbb{R}^k} & \sum\limits_{i=2}^{m} \left(w_1' C_{1,i} w_i\right)^2 \\ s.t. & w_i' C_{i,i} w_i = 1, \qquad \forall i \end{array},$$

which can be expressed as an eigenvalue problem (due to ignoring the non-hub language cross-covariances):

$$\left(\sum_{i=2}^{m} G_i G_i'\right) \cdot V = \Lambda \cdot V,$$

where $G_i = H_1' C_{1,i} H_i$ and $H_i = Chol\left(C_{i,i}\right)^{-1}$. The solutions are obtained from the maximal eigenvector denoted by v as:

$$\begin{aligned} w_1 &= H_1 v \\ w_i &= H_i G_i' v, \qquad i = 2, \dots, m. \end{aligned}$$

So far we described how a single concept represented as $w_1, \dots, w_m$ is discovered. By using deflation techniques we can find a second set of vectors $u_1, \dots, u_m$ that maximize the sum of squared correlations and are orthogonal (uncorrelated) to the first set:

$$\begin{array}{ll} \max\limits_{u_1,\dots,u_m \in \mathbb{R}^k} & \sum\limits_{i=2}^{m} \left(u_1' C_{1,i} u_i\right)^2 \\ s.t. & u_i' C_{i,i} u_i = 1, \qquad \forall i \\ & u_i' C_{i,i} w_i = 0 \end{array}.$$

The orthogonality constraints can be automatically enforced by deflating the matrices:

$$H_i \leftarrow H_i - H_i w_{i+1} w_{i+1}' - w_1 w_1' H_i + w_1 w_1' H_i w_{i+1} w_{i+1}', \quad \forall i = 2, \dots, m,$$

And solving the same eigenvalue problem defined above. The same deflation technique is used to find more than two concept vectors. Let $W_1 \in \mathbb{R}^{k \times r}, \dots, W_m \in \mathbb{R}^{k \times r}$ denote r sets of solutions over m languages.

The final r concept vectors (in the basis of the original vector spaces L_i, before the dimensionality reduction step) can be expressed as the columns of the matrix:

$$P_i = U_i W_i.$$

3 Cross-lingual Document Linking Web Service

3.1 KIT Cross-lingual Similarity Service

The cross-lingual document linking prototype consists of two functionalities and thus also has two output formats. The first service (Cross-lingual Similarity) determines the cross-lingual similarity between two documents. The second service (Cross-lingual Analysis) retrieves related Wikipedia articles in a specified language given an input document. The XML format is used to represent the output of the prototype.

3.1.1 Cross-lingual Similarity output format

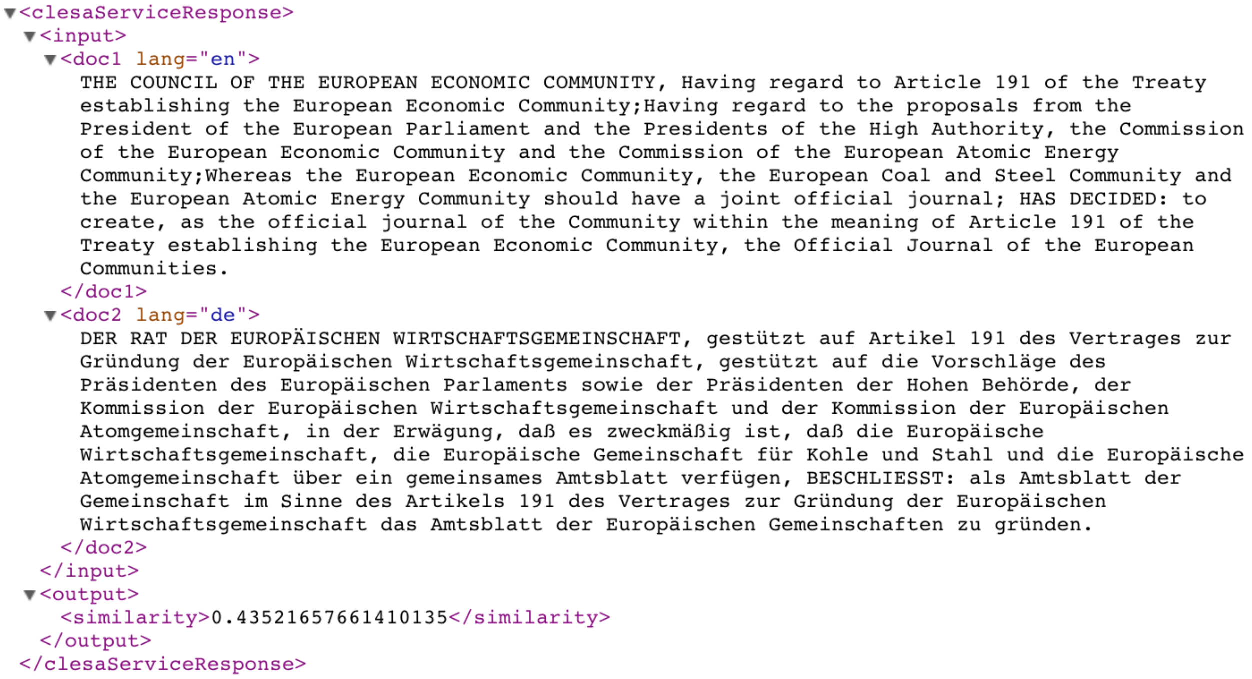
This web service is based on Cross-lingual extension of Explicit Semantic Analysis (ESA) and uses Wikipedia dumps from Mai 2012 as knowledge source. It supports English, German, Spanish, French, Catalan and Slovenian. The service can be called by using POST or GET request to the following URL address and input parameters.

Service URL: http://km.aifb.kit.edu/services/clesa/similarity

Input Parameters:

- **doc1**: the first input document
- **lang1**: language of doc1
- **doc2**: the second input document
- **lang2**: language of doc2

The response of the service consists of an input and output element. The input element contains two doc elements with a language attribute. The output element contains the similarity score between the two input documents. **Error! Reference source not found.**14 shows the similarity of two example documents in English and German. Note that the two example documents are in this case from a parallel corpus.



```
▼<clesaServiceResponse>
  ▼<input>
    ▼<doc1 lang="en">
       THE COUNCIL OF THE EUROPEAN ECONOMIC COMMUNITY, Having regard to Article 191 of the Treaty
       establishing the European Economic Community;Having regard to the proposals from the
       President of the European Parliament and the Presidents of the High Authority, the Commission
       of the European Economic Community and the Commission of the European Atomic Energy
       Community;Whereas the European Economic Community, the European Coal and Steel Community and
       the European Atomic Energy Community should have a joint official journal; HAS DECIDED: to
       create, as the official journal of the Community within the meaning of Article 191 of the
       Treaty establishing the European Economic Community, the Official Journal of the European
       Communities.
     </doc1>
    ▼<doc2 lang="de">
       DER RAT DER EUROPÄISCHEN WIRTSCHAFTSGEMEINSCHAFT, gestützt auf Artikel 191 des Vertrages zur
       Gründung der Europäischen Wirtschaftsgemeinschaft, gestützt auf die Vorschläge des
       Präsidenten des Europäischen Parlaments sowie der Präsidenten der Hohen Behörde, der
       Kommission der Europäischen Wirtschaftsgemeinschaft und der Kommission der Europäischen
       Atomgemeinschaft, in der Erwägung, daß es zweckmäßig ist, daß die Europäische
       Wirtschaftsgemeinschaft, die Europäische Gemeinschaft für Kohle und Stahl und die Europäische
       Atomgemeinschaft über ein gemeinsames Amtsblatt verfügen, BESCHLIESST: als Amtsblatt der
       Gemeinschaft im Sinne des Artikels 191 des Vertrages zur Gründung der Europäischen
       Wirtschaftsgemeinschaft das Amtsblatt der Europäischen Gemeinschaften zu gründen.
     </doc2>
   </input>
  ▼<output>
     <similarity>0.43521657661410135</similarity>
   </output>
 </clesaServiceResponse>
```

Figure 14. Example output of KIT cross-lingual similarity service

3.1.2 Cross-lingual Analysis output format

This web service is based Explicit Semantic Analysis (ESA) and language links in Wikipedia. It uses Wikipedia dumps from Mai 2012 and supports English, German, Spanish, French, Catalan and Slovenian. The service can be called by using POST or GET request to the following URL address and input parameters.

Service URL: http://km.aifb.kit.edu/services/clesa/analyzer

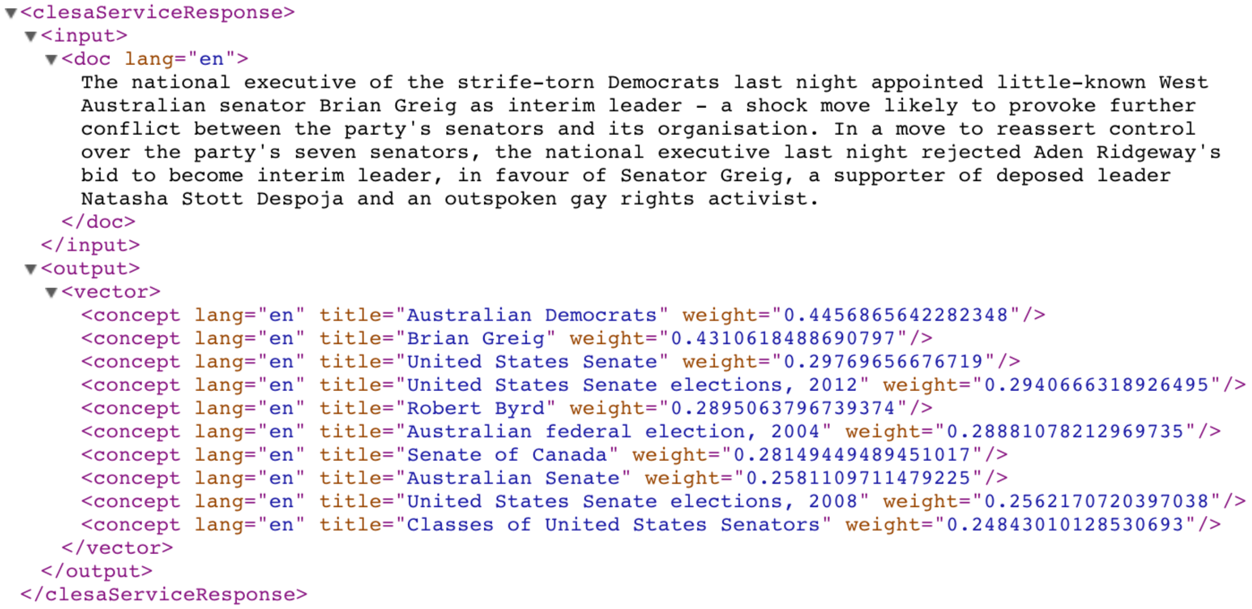
Input Parameters:

- **doc**: the input document
- **lang1**: language of doc
- **lang2**: language of Wikipedia articles to retrieve
- **retrieve**: number of Wikipedia articles to retrieve

The response of the service consists of the xml elements **`input`** and **`output`**. The **`input`** element consists of a **`doc`** element, which contains the raw input document before pre-processing. The **`output`** element consists of a vector, which in turn contains the related concepts. The **`concept`** element has the following attributes:

- **lang**: the language of the related Wikipedia article
- **title**: the title of the related Wikipedia article
- **weigth**: measure of the similarity between input document and the Wikipedia article

Error! Reference source not found. 15 shows the output format and returns the top ten related English Wikipedia articles.



```
▼<clesaServiceResponse>
  ▼<input>
    ▼<doc lang="en">
       The national executive of the strife-torn Democrats last night appointed little-known West
       Australian senator Brian Greig as interim leader - a shock move likely to provoke further
       conflict between the party's senators and its organisation. In a move to reassert control
       over the party's seven senators, the national executive last night rejected Aden Ridgeway's
       bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader
       Natasha Stott Despoja and an outspoken gay rights activist.
     </doc>
   </input>
  ▼<output>
    ▼<vector>
       <concept lang="en" title="Australian Democrats" weight="0.4456865642282348"/>
       <concept lang="en" title="Brian Greig" weight="0.4310618488690797"/>
       <concept lang="en" title="United States Senate" weight="0.29769656676719"/>
       <concept lang="en" title="United States Senate elections, 2012" weight="0.2940666318926495"/>
       <concept lang="en" title="Robert Byrd" weight="0.2895063796739374"/>
       <concept lang="en" title="Australian federal election, 2004" weight="0.28881078212969735"/>
       <concept lang="en" title="Senate of Canada" weight="0.28149449489451017"/>
       <concept lang="en" title="Australian Senate" weight="0.2581109711479225"/>
       <concept lang="en" title="United States Senate elections, 2008" weight="0.2562170720397038"/>
       <concept lang="en" title="Classes of United States Senators" weight="0.24843010128530693"/>
     </vector>
   </output>
 </clesaServiceResponse>
```

Figure 15. Example output of KIT cross-lingual analysis service

3.2 JSI Cross-lingual Similarity Service

The JSI Cross-lingual Similarity Service consists of two main components and web interface. First one (**CLSI**) can be used to compute similarity between two documents **doc1** and **doc2** in two languages, **lang1** and **lang2**. Second one (**REVEAL**) enables the insight in how the similarity is computed. It returns words in language **lang1** and language **lang2** that add the most to the similarity.

Both services can be called by using POST or GET request to the appropriate address. Using POST is preferable and the query needs to be URL escaped. The ISO 639-1 language codes are used for the specification of document's language.

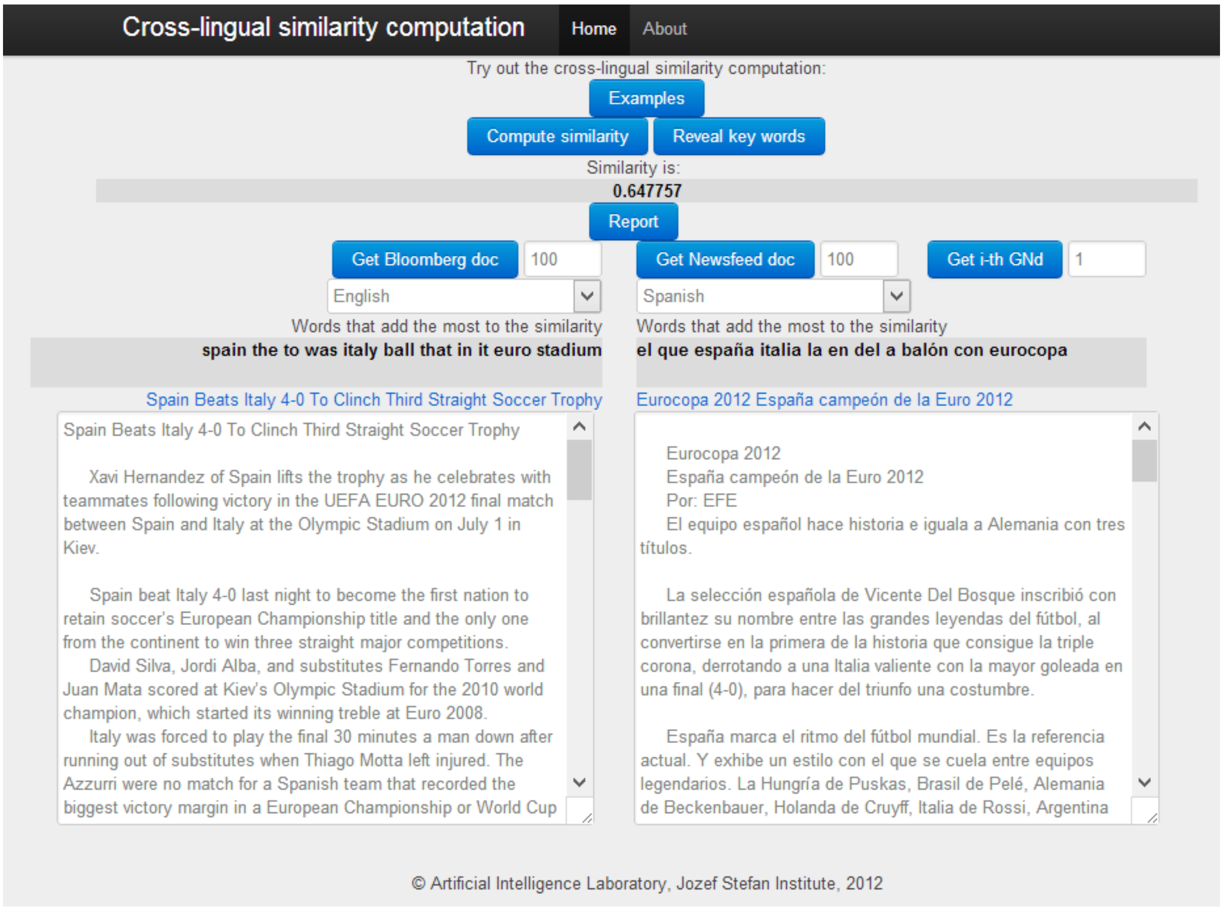
The **CLSI** component call example is:

- Request: http://xling.ijs.si:2222/clsi?doc1=Car%20is%20the%20most%20commonly%20used%20vehicle%20in%20Slovenia&lang1=en&doc2=Como%20es%20el%20veh%C3%ADculo%20m%C3%A1s%20utilizado%20en%20Eslovenia&lang2=es
- Response: *0.906426*

The two input documents are URL escaped:

- Car is the most commonly used vehicle in Slovenia becomes Car%20is%20the%20most%20commonly%20used%20vehicle%20in%20Slovenia
- Como es el vehículo más utilizado en Eslovenia is transformed to Como%20es%20el%20veh%C3%ADculo%20m%C3%A1s%20utilizado%20en%20Eslovenia
- The cosine similarity between this two documents is *0.906426.*

To illustrate the second component two somewhat related news articles about Eurocup are compared.
The request is posted on http://xling.ijs.si:2222/clsi in the same format as for **CLSI** component.
The resulting similarity is *0.647757*. They can be found at http://www.bloomberg.com/news/2012-07-01/spain-defeats-italy-4-0-to-become-first-to-retain-european-title.html and http://www.elespectador.com/deportes/futbolinternacional/articulo-356545-espana-campeon-de-euro-2012.
The resulting JSON output is array of two strings with top 10 words that add the most to similarity, *{" spain the to was italy ball that in it euro stadium", " el que españa italia la en del a balón con eurocopa"}.* Retrieved words clearly show that the similarity computation was successful.
Both components and some additional functionality related to Bloomberg's use case are available in the form of the web interface at http://xling.ijs.si:2222/wikipedia.html that can be used for demo purposes.



Try out the cross-lingual similarity computation:

Similarity is:

0.647757

Get Bloomberg doc: 100 | Get Newsfeed doc: 100 | Get i-th GNd: 1

English | Spanish

Words that add the most to the similarity

spain the to was italy ball that in it euro stadium

Words that add the most to the similarity

el que españa italia la en del a balón con eurocopa

Spain Beats Italy 4-0 To Clinch Third Straight Soccer Trophy

Spain Beats Italy 4-0 To Clinch Third Straight Soccer Trophy

Xavi Hernandez of Spain lifts the trophy as he celebrates with teammates following victory in the UEFA EURO 2012 final match between Spain and Italy at the Olympic Stadium on July 1 in Kiev.

Spain beat Italy 4-0 last night to become the first nation to retain soccer's European Championship title and the only one from the continent to win three straight major competitions.

David Silva, Jordi Alba, and substitutes Fernando Torres and Juan Mata scored at Kiev's Olympic Stadium for the 2010 world champion, which started its winning treble at Euro 2008.

Italy was forced to play the final 30 minutes a man down after running out of substitutes when Thiago Motta left injured. The Azzurri were no match for a Spanish team that recorded the biggest victory margin in a European Championship or World Cup

Eurocopa 2012 España campeón de la Euro 2012

Eurocopa 2012
España campeón de la Euro 2012
Por: EFE

El equipo español hace historia e iguala a Alemania con tres títulos.

La selección española de Vicente Del Bosque inscribió con brillantez su nombre entre las grandes leyendas del fútbol, al convertirse en la primera de la historia que consigue la triple corona, derrotando a una Italia valiente con la mayor goleada en una final (4-0), para hacer del triunfo una costumbre.

España marca el ritmo del fútbol mundial. Es la referencia actual. Y exhibe un estilo con el que se cuela entre equipos legendarios. La Hungría de Puskas, Brasil de Pelé, Alemania de Beckenbauer, Holanda de Cruyff, Italia de Rossi, Argentina

© Artificial Intelligence Laboratory, Jozef Stefan Institute, 2012

Figure 16. Example output of JSI cross-lingual similarity service

4 Experimental Evaluation

In our experiments, we compared the K-means clustering, LSI and ESA based approaches for both cross-lingual plagiarism detection and cross-lingual recommendation scenarios.

4.1 Evaluation of Cross-lingual Plagiarism Detection

4.1.1 Experimental Setting

For cross-lingual plagiarism detection, the task is to track the republishing of its articles especially if the articles are translated. We will investigate the performance using a standard mate retrieval setup, which has already been used to evaluate cross-lingual plagiarism detection and also general cross-lingual IR.

To provide the background and testing data, we extracted large collections from the parallel corpus JRC-Acquis and the comparable corpus Wikipedia. The JRC-Acquis parallel corpus comprises of approx. 23000 legislative documents from European Union in each of 22 European languages. We randomly select a sample of 10% of parallel documents in English, German, Spanish and Slovenian from JRC-Acquis corpus for testing and use the remaining 90% parallel documents in these languages as background data. For constructing the comparable corpus Wikipedia as additional background data, we analysed cross-language links between Wikipedia articles for each pair of supported languages and used only articles for which aligned versions exist.

For mate retrieval evaluation, we take the document in one language from the testing collection as a query and retrieve the relevant documents in another language from the testing collection. We assumed that only the translated version (mate) is considered as relevant to the document itself. This experimental setup simulates the cross-lingual plagiarism detection scenario, where we are concerned about whether the translations appear on top of the ranked result lists. In addition, the observed position of the mate can also be used as a comparison yardstick.

Based on the above observation, we consider recall at cutoff rank k (R@k) and the Mean Reciprocal Rank (MRR) as quality criteria [12]. 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. In other words, it measures how many of all translations have been found. MRR measures the average reciprocal ranks of the mate documents. In contrast to R@k, MRR also takes into account the position of the mate document, resulting in higher values the higher the position of the mate in the ranked result list is. Note that the results for each language pair are averages in both directions (e.g. English-German and German-English).

4.1.2 Evaluation Results

Figure 17 shows MRR results for mate retrieval experiments on JRC-Acquis dataset.

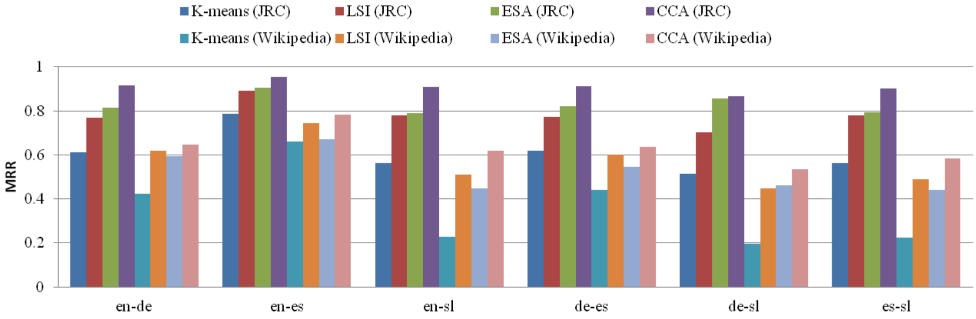


Figure 177. Mean Reciprocal Rank for JRC-Acquis Test Dataset

Figure 18 shows R@k reuslts for mate retrieval experiments on JRC-Acquis dataset.

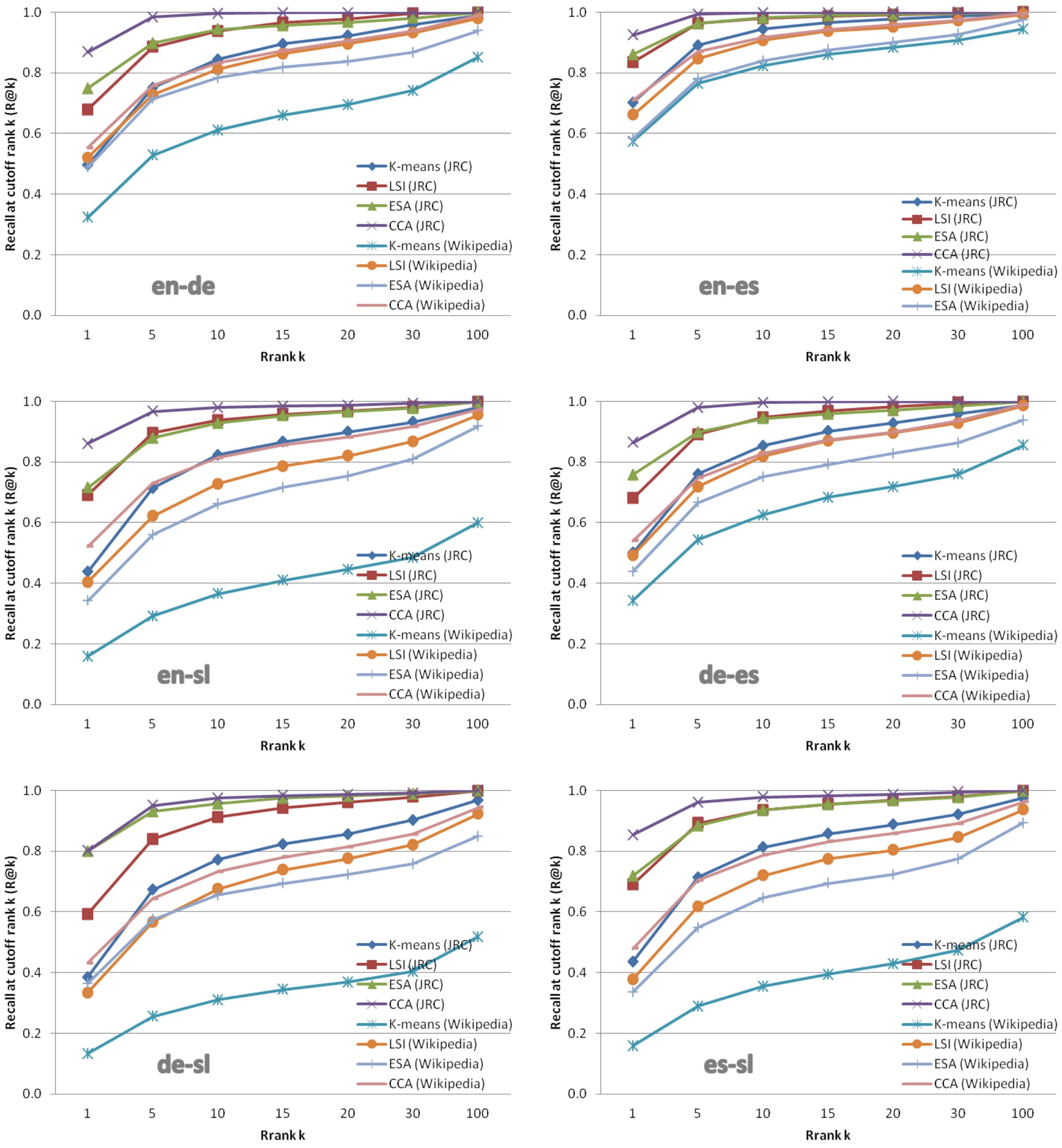


Figure 188. Recall at cutoff rank k for JRC-Acquis Test Dataset

It is observed that the use of JRC-Acquis itself as background data leads to significantly better results than the use of Wikipedia as background data for all approaches. That is due to the large vocabulary overlap with the test collection when using JRC-Acquis itself as background data. Moreover, in contrast to a parallel corpus, Wikipedia is a comparable corpus where the aligned articles may vary in size, quality and vocabulary wildly.

Clearly, LSI and ESA based approaches outperform K-means based approach in all cases. The explanation is that the semantic relatedness computed by K-means based approach is based on the term co-occurrence at the cluster level, which is more coarse-grained and thus not suitable for cross-lingual plagiarism detection scenario. While ESA outperforms LSI slightly when using JRC-Acquis itself as background data, LSI based approach obtains better results than ESA when using Wikipedia as background data. We conjecture that is because the wide topic range in Wikipedia may introduce noise through the homonyms from different topics and thus distort the correlation information for ESA, especially when addressing a particular topic domain where the number of different meanings of a term is likely low. By contrast, LSI brings related terms together and forms concepts by capturing the frequent term co-occurrence patterns and thus to some extent reduce noise in the original background data.

By performing the experiments on the same dataset, our results are different from those reported in [1], especially for LSI based approach. We believe that is because we use a better implementation of LSI. And the results of ESA using the dataset itself as background data is missing in [1].

4.2 Evaluation of Cross-lingual Recommendation

4.2.1 Experimental Setting

For cross-lingual recommendation, the task is to provide a list of recommended articles from a multilingual news stream. We will evaluate the performance using news stories from Reuters Corpus Volume 1 (RCV1) and Volume 2 (RCV2). The RCV1 contains about 810,000 Reuters, English language news stories, which have been used to evaluate various recommendation systems. The RCV2 contains over 487,000 Reuters news stories in thirteen languages which are contemporaneous with RCV1. The corpus contains 117 topics organized in a topic hierarchy with 4 top-level categories: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), MCAT (Market). For cross-lingual recommendation evaluation, we will use the news items contained in some sub-categories in one language as the context and create a stream of news with news items in another language, and then we use different approaches for recommending news items that are close to the context. As a ground truth, we consider as relevant all the news items that are in the context categories. We randomly select about 2500 documents for each language of English, German, Spanish and Chinese, a total of about 10000 documents, for testing and also use the parallel corpus JRC-Acquis and the comparable corpus Wikipedia as background data.

As quality criteria, we consider the Mean Average Precision (MAP) and the Mean Reciprocal Rank (MRR) as well as the 11-point interpolated average precision [12]. Precision defines the number of relevant documents that are retrieved in relation to the total number of retrieved documents. For each information need, the interpolated precision is measured at the 11 recall levels of 0.0, 0.1, 0.2, . . . , 1.0. For each recall level, we then calculate the arithmetic mean of the interpolated precision at that recall level for each information need in the test collection. MAP is another standard measure used in IR that is also sensitive to the rank of relevant documents. It averages precision measured at the rank of each relevant document. Similarly, the results for each language pair are averages in both directions (e.g. English-German and German-English).

4.2.2 Evaluation Results

Figure 19 shows MAP results for experiments on Reuters test dataset.

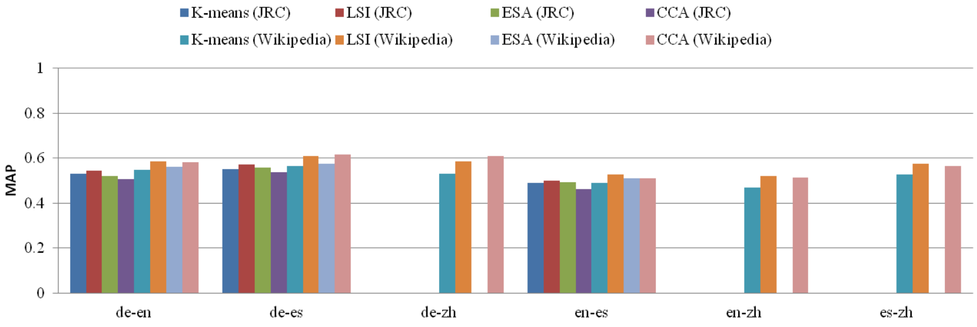


Figure 19. Mean Average Precision for Reuters Test Dataset

Figure 20 shows MRR results for experiments on Reuters test dataset.

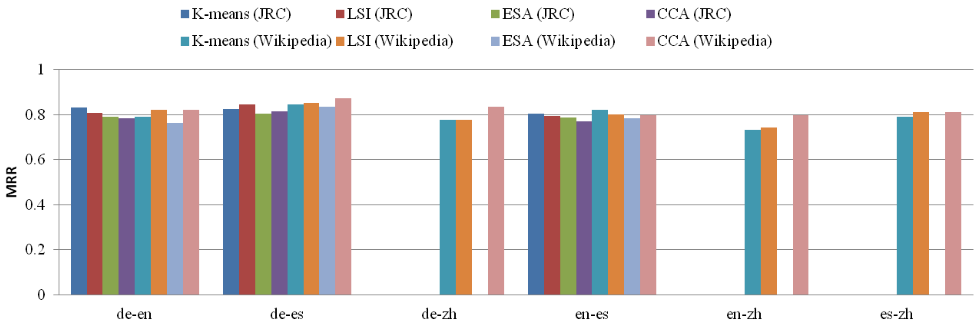


Figure 20. Mean Reciprocal Rank for Reuters Test Dataset

Figure 21 shows averaged 11-point precision-recall curve for experiments on Reuters test dataset.

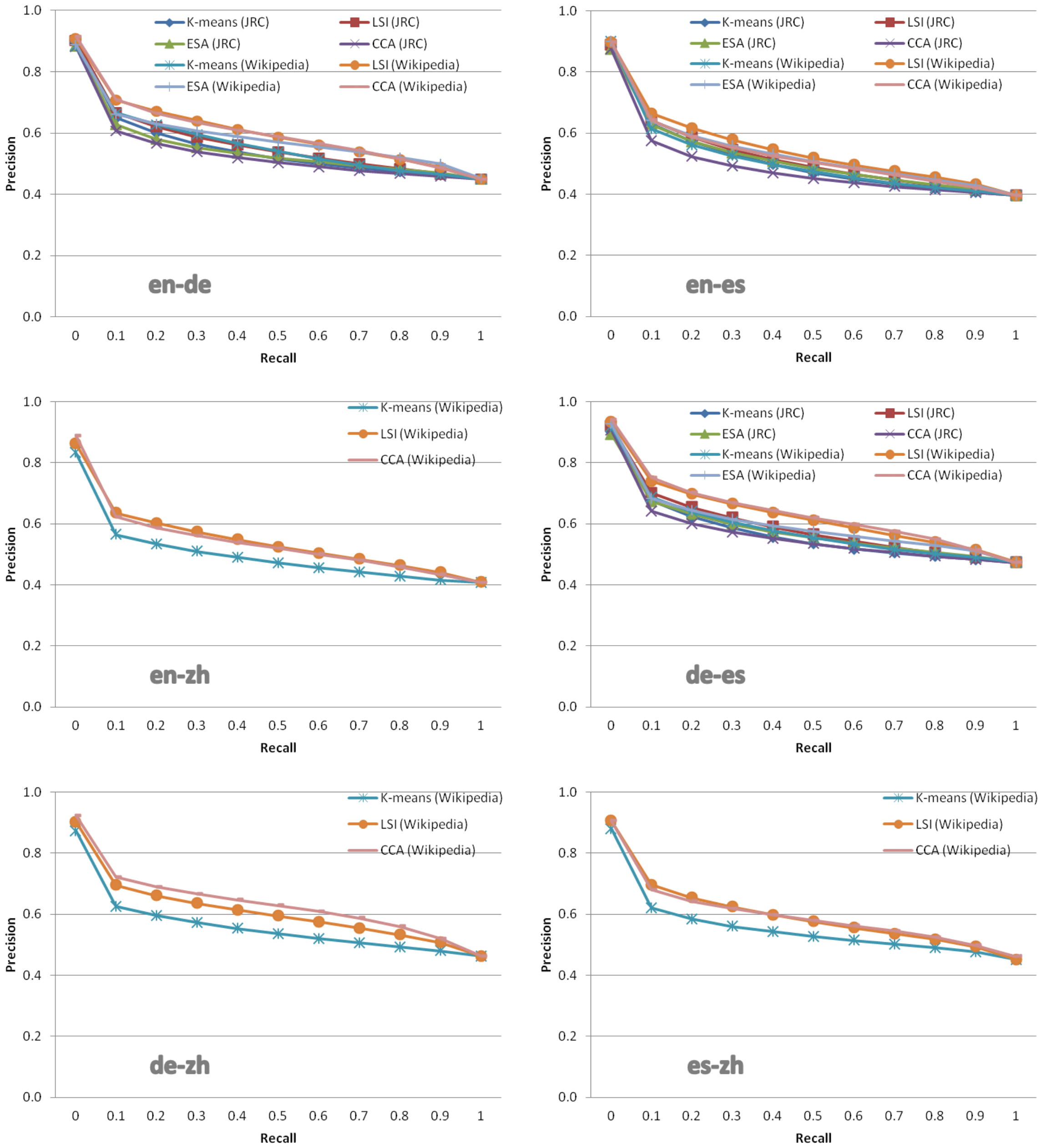


Figure 21. Averaged 11-point Precision-Recall Curve for Reuters Test Dataset

Different from the results of cross-lingual plagiarism detection evaluation, we observe that all approaches achieve similar performance for each language pair. The reason is that cross-lingual recommendation requires loose semantic relatedness that all approaches can capture. K-means based approach even yields better performance in some cases because the coarse-grained term co-occurrence information used by it may more appropriately reflect the requirement of loose semantic relatedness.

The results yielded using Wikipedia as background data are comparable or even slightly better than those using JRC-Acquis. In spite of the disadvantage of Wikipedia as background data mentioned previously, the wide topic range in Wikipedia may play a positive role here due to the large vocabulary overlap.

4.3 Conclusions

In our experimental evaluation, we address two different scenarios based on the use cases: 1) cross-lingual plagiarism detection and 2) cross-lingual recommendation. The task of cross-lingual plagiarism detection is to track the republishing of its articles that are translated into other languages and the task of cross-lingual recommendation is to provide a list of recommended articles from a multilingual news stream. Since cross-lingual plagiarism detection and cross-lingual recommendation require different similarity criteria, we carried out two separate experiments: one which searches specifically for translation, and one which generally searches for related documents. In the experiments, we compared the K-means clustering, LSI, ESA and CCA based approaches. The results show that all the approaches achieve similar performance for cross-lingual recommendation scenario. However, for cross-lingual plagiarism detection, LSI, ESA and CCA outperform K-means clustering significantly. Among LSI, ESA and CCA, CCA achieves the best results. LSI and ESA yield comparable results that are not far behind CCA.

References

[1] Cimiano, P., Schulz, A., Sizov, S., Sorg, P., Staab, S. **Explicit vs. Latent Concept Models for Cross-Language Information Retrieval**. Proceedings of the International Joint Conference on Artificial Intelligence, 2009.

[2] Furnas, G., Landauer, T., Gomez, L., Dumais, S. **The Vocabulary Problem in Human-System Communication**. Communications of the ACM, 30(1), 964-971, 1987.

[3] Gabrilovich, E., Markovitch, S. (2007) **Computing semantic relatedness using Wikipedia-based explicit semantic analysis**. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 1606-1611, 2007.

[4] Sorg, P., Cimiano, P. **Cross-lingual Information Retrieval with Explicit Semantic Analysis**. Working Notes of the Annual CLEF Meeting, 2008.

[5] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D. **The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages**. Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006.

[6] Andrej Muhic, Jan Rupnik and Primoz Skraba. **Cross-Lingual Document Retrieval through Hub Languages**, xLiTe: Cross-Lingual Technologies, NIPS 2012 WORKSHOP

[7] J. R. Kettenring. **Canonical analysis of several sets of variables**. Biometrika, 1971.

[8] 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.

[9] S. Dumais, T. Letsche, M. Littman, and T. Landauer. **Automatic Cross-language Retrieval using Latent Semantic Indexing.** In Proceedings of the AAAI Symposium on Cross-Language Text and Speech Retrieval, 1997.

[10] E. Gabrilovich and S. Markovitch. **Wikipedia-based semantic interpretation for natural language processing.** J. Artif. Intell. Res. (JAIR), 34:443-498, 2009.

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

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. **Introduction to Information Retrieval.** Cambridge University Press, New York, NY, USA, 2008.