Semantic-Based Multilingual Document Clustering via Tensor Modeling

Salvatore Romeo, Andrea Tagarelli
DIMES, University of Calabria
Arcavacata di Rende, Italy
sromeo@dimes.unical.it
tagarelli@dimes.unical.it

Dino Ienco
IRSTEA, UMR TETIS
Montpellier, France
LIRMM
Montpellier, France
dino.ienco@irstea.fr

Abstract

A major challenge in document clustering research arises from the growing amount of text data written in different languages. Previous approaches depend on language-specific solutions (e.g., bilingual dictionaries, sequential machine translation) to evaluate document similarities, and the required transformations may alter the original document semantics. To cope with this issue we propose a new document clustering approach for multilingual corpora that (i) exploits a large-scale multilingual knowledge base, (ii) takes advantage of the multi-topic nature of the text documents, and (iii) employs a tensor-based model to deal with high dimensionality and sparseness. Results have shown the significance of our approach and its better performance w.r.t. classic document clustering approaches, in both a balanced and an unbalanced corpus evaluation.

1 Introduction

Document clustering research was initially focused on the development of general purpose strategies to group unstructured text data. Recent studies have started developing new methodologies and algorithms that take into account both linguistic and topical characteristics, where the former include the size of the text and the type of language, and the latter focus on the communicative function and targets of the documents.

A major challenge in document clustering research arises from the growing amount of text data that are written in different languages, also due to the increased popularity of a number of tools for collaboratively editing through contributors across the world. *Multilingual document clustering* (MDC) aims to detect clusters in a collection of texts written in different languages. This can aid a variety of applications in cross-lingual information retrieval, including statistical machine translation and corpora alignment.

Existing approaches to MDC can be divided in two broad categories, depending on whether a parallel corpus rather than a comparable corpus is used (Kumar et al., 2011c). A *parallel corpus* is typically comprised of documents with their related translations (Kim et al., 2010). These translations are usually obtained through machine translation techniques based on a selected anchor language. Conversely, a *comparable corpus* is a collection of multilingual documents written over the same set of classes (Ni et al., 2011; Yogatama and Tanaka-Ishii, 2009) without any restriction about translation or perfect correspondence between documents. To mine this kind of corpus, external knowledge is employed to map concepts or terms from a language to another (Kumar et al., 2011c; Kumar et al., 2011a), which enables the extraction of cross-lingual document correlations. In this case, a major issue lies in the definition of a cross-lingual similarity measure that can fit the extracted cross-lingual correlations. Also, from a semi-supervised perspective, other works attempt to define must-link constraints to detect cross-lingual clusters (Yogatama and Tanaka-Ishii, 2009). This implies that, for each different dataset, the set of constraints needs to be redefined; in general, the final results can be negatively affected by the quantity and the quality of involved constraints (Davidson et al., 2006).

To the best of our knowledge, existing clustering approaches for comparable corpora are customized for a small set (two or three) of languages (Montalvo et al., 2007). Most of them are not generalizable to many languages as they employ bilingual dictionaries and the translation is performed sequentially considering only pairs of languages. Therefore, the order in which this process is done can seriously impact the results. Another common drawback concerns the way most of the recent approaches perform their analysis: the various languages are analyzed independently of each other (possibly by exploiting external knowledge like Wikipedia to enrich documents (Kumar et al., 2011c; Kumar et al., 2011a)), and then the language-specific results are merged. This two-step analysis however may fail in profitably exploiting cross-language information from the multilingual corpus.

Contributions. We address the problem of MDC by proposing a framework that features three key elements, namely: (1) to model documents over a unified conceptual space, with the support of a large-scale multilingual knowledge base; (2) to decompose the multilingual documents into topically-cohesive segments; and (3) to describe the multilingual corpus under a multi-dimensional data structure.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 600–609,
October 25-29, 2014, Doha, Qatar.

The first key element prevents loss of information due to the translation of documents from different languages to a target one. It enables a conceptual representation of the documents in a language-independent way preserving the content semantics. BabelNet (Navigli and Ponzetto, 2012a) is used as multilingual knowledge base. To the extent of our knowledge, this is the first work in MDC that exploits BabelNet.

The second key element, document segmentation, enables us to simplify the document representation according to their multi-topic nature. Previous research has demonstrated that a segment-based approach can significantly improve document clustering performance (Tagarelli and Karypis, 2013). Moreover, the conceptual representation of the document segments enables the grouping of linguistically different (portions of) documents into topically coherent clusters.

The latter aspect is leveraged by the third key element of our proposal, which relies on a tensor-based model (Kolda and Bader, 2009) to effectively handle the high dimensionality and sparseness in text. Tensors are considered as a multi-linear generalization of matrix factorizations, since all dimensions or modes are retained thanks to multi-linear structures which can produce meaningful components. The applicability of tensor analysis has recently attracted growing attention in information retrieval and data mining, including document clustering (e.g., (Liu et al., 2011; Romeo et al., 2013)) and cross-lingual information retrieval (e.g., (Chew et al., 2007)).

The rest of the paper is organized as follows. Section 2 provides an overview of BabelNet and basic notions on tensors. We describe our proposal in Section 3. Data and experimental settings are described in Section 4, while results are presented in Section 5. We summarize our main findings in Section 6, finally Section 7 concludes the paper.

2 Background

2.1 BabelNet

BabelNet (Navigli and Ponzetto, 2012a) is a multilingual semantic network obtained by linking Wikipedia with WordNet, that is, the largest multilingual Web encyclopedia and the most popular computational lexicon. The linking of the two knowledge bases was performed through an automatic mapping of WordNet synsets and Wikipages, harvesting multilingual lexicalization of the available concepts through human-generated translations provided by the Wikipedia interlanguage links or through machine translation techniques. The result is an encyclopedic dictionary containing concepts and named entities lexicalized in 50 different languages.

Multilingual knowledge in BabelNet is represented as a labeled directed graph in which nodes are concepts or named entities and edges connect pairs of nodes through a semantic relation. Each edge is labeled with a relation type (is-a, part-of, etc.), while each node corresponds to a *BabelNet synset*, i.e., a set of lexicalizations of a concept in different languages.

BabelNet can be accessed and easily integrated into applications by means of a Java API provided by the toolkit described in (Navigli and Ponzetto, 2012b). The toolkit also provides functionalities for *graph-based WSD in a multilingual context*. Given an input set of words, a semantic graph is built by looking for related synset paths and by merging all them in a unique graph. Once the semantic graph is built, the graph nodes can be scored with a variety of algorithms. Finally, this graph with scored nodes is used to rank the input word senses by a graph-based approach.

2.2 Tensor model representation

A tensor is a multi-dimensional array $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times \cdots \times I_M}$. The number of dimensions M, also known as *ways* or *modes*, is called *order* of the tensor, so that a tensor with order M is also said a M-way or M-order tensor. A *higher-order* tensor (i.e., a tensor with order three or higher) is denoted by boldface calligraphic letters, e.g., $\boldsymbol{\mathcal{T}}$; a matrix (2-way tensor) is denoted by boldface capital letters, e.g., $\mathbf{U}$; a vector (1-way tensor) is denoted by boldface lowercase letters, e.g., $\mathbf{v}$. The generic entry (i_1, i_2, i_3) of a third-order tensor $\boldsymbol{\mathcal{T}}$ is denoted by $t_{i_1 i_2 i_3}$, with $i_1 \in [1..I_1], i_2 \in [1..I_2], i_3 \in [1..I_3]$.

A one-dimensional fragment of tensor, defined by varying one index and keeping the others fixed, is a 1-way tensor called *fiber*. A third-order tensor has column, row and tube fibers. Analogously, a two-dimensional fragment of tensor, defined by varying two indices and keeping the rest fixed, is a 2-way tensor called *slice*. A third-order tensor has horizontal, lateral and frontal slices.

The *mode-m matricization* of a tensor $\boldsymbol{\mathcal{T}}$, denoted by $\mathbf{T}_{(\mathrm{m})}$, is obtained by arranging the mode-m fibers as columns of a matrix. A third-order tensor $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times I_3}$ is *all-orthogonal* if $\sum_{i_1 i_2} t_{i_1 i_2 \alpha} t_{i_1 i_2 \beta} = \sum_{i_1 i_3} t_{i_1 \alpha i_3} t_{i_1 \beta i_3} = \sum_{i_2 i_3} t_{\alpha i_2 i_3} t_{\beta i_2 i_3} = 0$ whenever $\alpha \neq \beta$. The *mode-m product* of a tensor $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times \cdots \times I_M}$ with a matrix $\mathbf{U} \in \Re^{J \times I_m}$, denoted by $\boldsymbol{\mathcal{T}} \times_m \mathbf{U}$, is a tensor of dimension $I_1 \times \ldots I_{m-1} \times J \times I_{m+1} \times \cdots \times I_M$ and can be expressed in terms of matrix product as $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{T}} \times_m \mathbf{U}$, whose mode-$m$ matricization is $\mathbf{Y}_{(\mathrm{m})} = \mathbf{U}\mathbf{T}_{(\mathrm{m})}$.

3 Our Proposal

3.1 Multilingual Document Clustering framework

We are given a collection of multilingual documents $\mathcal{D} = \bigcup_{l=1}^{L} \mathcal{D}_l$, where each $\mathcal{D}_l = \{d_{l_i}\}_{i=1}^{N_l}$ represents a subset of documents written in the same language, with $N = \sum_{l=1}^{L} N_l = |\mathcal{D}|$. Our framework can be applied to any multilingual document collection regardless of the languages, and can deal with balanced as well as

Algorithm 1 *SeMDocT* (**Se**gment-based **M**ultiLingual **Doc**ument Clustering via **T**ensor Modeling)

Input: A collection of multilingual documents $\mathcal{D}$, the number k of segment clusters, the number of tensorial components r.
Output: A document clustering solution $\mathcal{C}$ over $\mathcal{D}$.

1: Apply a text segmentation algorithm over each of the documents in $\mathcal{D}$ to produce a collection of document segments $\mathcal{S}$. */* Section 3.1.1 */*
2: Represent $\mathcal{S}$ in either a bag-of-words (BoW) or a bag-of-synsets (BoS) space. */* Section 3.1.2 */*
3: Apply any document clustering algorithm on $\mathcal{S}$ to obtain a segment clustering $\mathcal{C}^{\mathcal{S}} = \{C_i^s\}_{i=1}^k$. */* Section 3.1.2 */*
4: Represent $\mathcal{C}^{\mathcal{S}}$ in either a bag-of-words (BoW) or a bag-of-synsets (BoS) space. */* Section 3.1.3 */*
5: Model $\mathcal{S}$ as a third-order tensor $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times I_3}$, with $I_1 = |\mathcal{D}|$, $I_2 = |\mathcal{F}|$, and $I_3 = k$. */* Section 3.1.4 */*
6: Decompose the tensor using a Truncated HOSVD. */* Section 3.1.4 */*
7: Apply a document clustering algorithm on the mode-1 factor matrix to obtain the final clusters of documents $\mathcal{C} = \{C_i\}_{i=1}^K$. */* Section 3.1.5 */*

unbalanced corpora. Therefore, no restriction is given on both the number L of languages and the distribution of documents over the languages (i.e., $N_i \lesseqgtr N_j$, with $i, j = 1..L, i \neq j$).

Real-world documents often span multiple topics. We assume that each document in $\mathcal{D}$ is relatively long to be comprised of smaller textual units, or *segments*, each of which can be considered cohesive w.r.t. a topic over the document. This represents a key aspect in our framework as it enables the use of a *tensor model* to conveniently address the multi-faceted nature of the documents.

Our overall framework, named *SeMDocT* (**Se**gment-based **M**ultiLingual **Doc**ument Clustering via **T**ensor Modeling), is shown in Algorithm 1. In the following, we shall describe in details each of the steps involved in *SeMDocT*.

3.1.1 Computing within-document segments

Text segmentation is concerned with the fragmentation of an input text into multi-paragraph, contiguous and disjoint blocks that represent subtopics. Regardless of the presence of logical structure clues in the document, linguistic criteria (Beeferman et al., 1999) and statistical similarity measures (Hearst, 1997; Choi et al., 2001; Cristianini et al., 2001) have been mainly used to detect subtopic boundaries between segments. A common assumption is that terms that discuss a subtopic tend to co-occur locally, and a switch to a new subtopic is detected by the ending of co-occurrence of a given set of terms and the beginning of the co-occurrence of another set of terms.

Our *SeMDocT* does not depend on a specific algorithmic choice to perform text segmentation; in this work, we refer to the classic *TextTiling* (Hearst, 1997), which is the exemplary similarity-block-based method for text segmentation.

3.1.2 Inducing document segment clusters

The result of the previous step is a collection of document segments, henceforth denoted as $\mathcal{S}$. Each segment in $\mathcal{S}$ is represented as a vector of feature occurrences, where a feature can be either *lexical* or *semantic*. This corresponds to two alternative representation models: the standard *bag-of-words* (henceforth *BoW*), whereby features correspond to lemmatized, non-stopword terms, and the obtained feature space results from the union of the vocabularies of the different languages; and *bag-of-synsets* (henceforth *BoS*), whereby features correspond to BabelNet synsets. We shall devote Section 3.2 to a detailed description of our proposed BoS representation.

The segment collection $\mathcal{S}$ is given in input to a document clustering algorithm to produce a clustering of the segments $\mathcal{C}^{\mathcal{S}} = \{C_i^s\}_{i=1}^k$. The obtained clusters of segments can be disjoint or overlapping. Again, our *SeMDocT* is parametric to the clustering algorithm as well; here, we resort to a state-of-the-art clustering algorithm, namely *Bisecting K-Means* (Steinbach et al., 2000), which is widely known to produce high-quality (hard) clustering solutions in high-dimensional, large datasets (Zhao and Karypis, 2004). Note however that it requires as input the number of clusters. To cope with this issue, we adopt the method described in (Salvador and Chan, 2004), which explores how the within-cluster cohesion changes by varying the number of clusters. The number of clusters for which the slope of the plot changes drastically is chosen as a suitable value for the clustering algorithm.

3.1.3 Segment-cluster based representation

Upon the segment clustering, each document is represented by its segments assigned to possibly multiple segment clusters. Therefore, we derive a document-feature matrix for each of the k segment clusters. The features correspond either to the BoW or BoS model, according to the choice made for the segment representation.

Let us denote with $\mathcal{F}$ the feature space for all segments in $\mathcal{S}$. Given a segment cluster C^s, the corresponding document-feature matrix is constructed as follows. The representation of each document $d \in \mathcal{D}$ w.r.t. C^s is a vector of length $|\mathcal{F}|$ that results from the sum of the feature vectors of the d's segments belonging to C^s. Moreover, in order to weight the appearance of a document in a cluster based on its segment-based portion covered in the cluster, the document vector of d w.r.t. C^s is finally obtained by multiplying the sum of the segment-vectors by a scalar representing the portion of d's features that appear in the segments belonging to C^s. The document-feature matrix of C^s resulting from the previous step is finally normalized by column.

3.1.4 Tensor model and decomposition

The document-feature matrices corresponding to the k segment-clusters are used to form a third-order tensor.

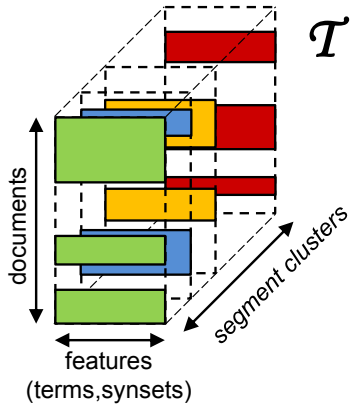


Figure 1: The third-order tensor model for the representation of a multilingual document collection based on segment clusters.

Our third-order tensor model is built by arranging as frontal slices the k segment-cluster matrices. The resulting tensor will be of the form $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times I_3}$, with $I_1 = |\mathcal{D}|$, $I_2 = |\mathcal{F}|$, and $I_3 = k$. The proposed tensor model is sketched in Fig. 1.

The resulting tensor is decomposed through a Truncated Higher Order SVD (T-HOSVD) (Lathauwer et al., 2000) in order to obtain a low-dimensional representation of the segment-cluster-based representation of the document collection. The T-HOSVD can be considered as an extension of the Truncated Singular Value Decomposition (T-SVD) to the case of three or more dimensions. For a third-order tensor $\boldsymbol{\mathcal{T}} \in \Re^{I_1 \times I_2 \times I_3}$ the T-HOSVD is expressed as

$$\boldsymbol{\mathcal{T}} \approx \boldsymbol{\mathcal{X}} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \mathbf{U}^{(3)}$$

where $\mathbf{U}^{(m)} = [\mathbf{u}_1^{(m)} \mathbf{u}_2^{(m)} \dots \mathbf{u}_{r_m}^{(m)}] \in \Re^{I_m \times r_m}$ $(m = 1, 2, 3)$ are orthogonal matrices, $r_m \ll I_m$, and the core tensor $\boldsymbol{\mathcal{X}} \in \Re^{r_1 \times r_2 \times r_3}$ is an all-orthogonal and ordered tensor. T-HOSVD can be computed in two steps:

1. For $m \in \{1, 2, 3\}$, compute the unfolded matrices $\mathbf{T}_{(\mathrm{m})}$ from $\boldsymbol{\mathcal{T}}$ and related standard SVD: $\mathbf{T}_{(\mathrm{m})} = \mathbf{U}^{(\mathrm{m})} \mathbf{S}^{(\mathrm{m})} \mathbf{V}^{(\mathrm{m})}$. The orthogonal matrix $\mathbf{U}^{(\mathrm{m})}$ contains the leading left singular vectors of $\mathbf{T}_{(\mathrm{m})}$.

2. Compute the core tensor $\boldsymbol{\mathcal{X}}$ using the inversion formula: $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}} \times_1 {\mathbf{U}^{(1)}}^T \times_2 {\mathbf{U}^{(2)}}^T \times_3 {\mathbf{U}^{(3)}}^T$.

Note that, since T-HOSVD is computed by means of 3 standard matrix T-SVDs, its computational cost can be reduced by using fast and efficient SVD algorithms. Moreover, the ability of T-HOSVD in effectively capturing the variation in each of the modes independently from the other ones, is particularly important to alleviate the problem of concentration of distances, thus making T-HOSVD well-suited to clustering purposes. In this work, in order to obtain a final clustering solution of the documents, we will consider the mode-1 factor matrix $\mathbf{U}^{(1)}$ of the T-HOSVD.

3.1.5 Document clustering

The mode-1 factor matrix is provided in input to a clustering method to obtain a final organization of the documents into K clusters, i.e., $\mathcal{C} = \{C_i\}_{i=1}^{K}$. Note that there is no principled relation between the number K of final document clusters and k. However, K is expected to reflect the number of topics of interest for the document collection. Also, possibly but not necessarily, the same clustering algorithm used for the segment clustering step (i.e., Bisecting K-Means) can be employed for this step.

3.2 Bag-of-synset representation

In the BoS model, our objective is to represent the document segments in a conceptual feature space instead of the traditional term space. Since we deal with multilingual documents, this task clearly relies on the multilingual lexical knowledge base functionalities of BabelNet. Conceptual features will hence correspond to BabelNet synsets.

The segment collection $\mathcal{S}$ is subject to a two-step processing phase. In the first step, each segment is broken down into a set of lemmatized and POS-tagged sentences, in which each word is replaced with related lemma and associated POS-tag. Let us denote with $\langle w, POS(w) \rangle$ a lemma and associated POS-tag occurring in any sentence sen of the segment. In the second step, a WSD method is applied to each pair $\langle w, POS(w) \rangle$ to detect the most appropriate BabelNet synset σ_w for $\langle w, POS(w) \rangle$ contextually to sen. The WSD algorithm is carried out in such a way that all words from all languages are disambiguated over the same concept inventory, producing a language-independent feature space for the whole multilingual corpus. Each segment is finally modeled as a $|\mathcal{BS}|$-dimensional vector of BabelNet synset frequencies, being $\mathcal{BS}$ the set of retrieved BabelNet synsets.

As previously discussed in Section 2.1, BabelNet provides WSD algorithms for multilingual corpora. More specifically, the authors in (Navigli and Ponzetto, 2012b) suggest to use the Degree algorithm (Navigli and Lapata, 2010), as it showed to yield highly competitive performance in a multilingual context as well. Note that the Degree algorithm, given a semantic graph for the input context, simply selects the sense of the target word with the highest vertex degree. Clearly, other graph-based methods for (unsupervised) WSD, particularly PageRank-style methods (e.g., (Mihalcea et al., 2004; Agirre and Soroa, 2009; Yeh et al., 2009; Tsatsaronis et al., 2010)), can be plugged in to address the multilingual WSD task based on BabelNet. An investigation of the performance of existing WSD algorithms for a multilingual context is however out of the scope of this paper.

4 Evaluation Methodology

In order to evaluate our proposal we need a multilingual comparable document collection with annotated

RCV2 Topics	*English*	*French*	*Italian*
Balanced Corpus			
C15 - PERFORMANCE	850	850	850
C18 - OWNERSHIP CHANGES	850	850	850
E11 - ECONOMIC PERFORMANCE	850	850	850
E12 - MONETARY/ECONOMIC	850	850	850
M11 - EQUITY MARKETS	850	850	850
M13 - MONEY MARKETS	850	850	850
Total	5 100	5 100	5 100
Unbalanced Corpus			
C15 - PERFORMANCE	850	850	0
C18 - OWNERSHIP CHANGES	850	850	0
E11 - ECONOMIC PERFORMANCE	0	850	850
E12 - MONETARY/ECONOMIC	850	0	850
M11 - EQUITY MARKETS	0	850	850
M13 - MONEY MARKETS	850	0	850
Total	3 400	3 400	3 400

Table 1: Number of documents for each topic and language.

Statistics	*Balanced Corpus*	*Unbalanced Corpus*
# of docs	15 300	10 200
# of terms	58 825	44 535
# of synsets	16 395	14 339
BoW Density	1.5×10^{-3}	2.0×10^{-3}
BoS Density	2.6×10^{-3}	3.1×10^{-3}

Table 2: Main characteristics of the corpora.

topics. For this reason, we used *Reuters Corpus Volume 2* (RCV2), a multilingual corpus containing news articles in thirteen language.[1] In the following, we present the corpus characteristics and competing methods used in our analysis.

4.1 Data preparation

We consider a subset of the RCV2 corpus corresponding to three languages: *English*, *French* and *Italian*. It covers six different topics, i.e., different labels of the RCV2 TOPICS field. Topics are chosen according with their coverage in the different languages. The language-specific documents were lemmatized and POS-tagged through the Freeling library (Padró and Stanilovsky, 2012) in order to obtain a suitable representation for the WSD process.

To assess the robustness of our proposal, we design two different scenarios. The first (*Balanced Corpus*) is characterized by a completely balanced dataset. Each language covers all topics and for each pair language/topic the same number of documents is selected. The second scenario corresponds to an *Unbalanced Corpus*. Starting from the balanced corpus, we removed for each topic all the documents belonging to one language. In this way, we obtained a corpus in which each topic is covered by only two of the three languages.

Main characteristics of both evaluation corpora are reported in Table 1 and Table 2. In the latter table, we report the number of documents, number of terms, number of synsets and the dataset density for both representations. To quantify the density of each corpus/representation combination, we counted the non-zero entries of the induced document-synset matrix (alternatively, document-term matrix) and we divided this value by the size of such matrix. This number provides an estimate about the density/sparseness of each dataset. Lower values indicate more sparse data. We can note that BoS model yields more dense datasets for both *Balanced Corpus* and *Unbalanced Corpus*.

[1] http://trec.nist.gov/data/reuters/reuters.html

RCV2 Topics	*English*	*French*	*Italian*
C15 - PERFORMANCE	3.41	3.67	3.27
C18 - OWNERSHIP CHANGES	3.20	3.32	2.40
E11 - ECONOMIC PERFORMANCE	4.89	3.17	2.07
E12 - MONETARY/ECONOMIC	5.22	3.69	2.05
M11 - EQUITY MARKETS	4.29	2.94	2.15
M13 - MONEY MARKETS	3.31	3.12	2.10

Table 3: Average number of document segments, for each topic and language.

	English		*French*		*Italian*	
RCV2 Topics	avg *BoS* seg. leng.	avg *BoW* seg. leng.	avg *BoS* seg. leng.	avg *BoW* seg. leng.	avg *BoS* seg. leng.	avg *BoW* seg. leng.
C15	21.76	36.32	11.54	34.92	10.58	37.75
C18	20.94	36.87	10.94	35.62	11.24	41.20
E11	22.90	37.24	11.47	34.73	11.96	38.60
E12	22.70	37.70	11.50	37.44	12.59	43.63
M11	22.04	36.83	10.91	32.76	11.57	42.39
M13	22.22	36.97	11.34	34.75	11.72	39.36

Table 4: Average length of document segment in the BoW and BoS spaces, for each topic and language.

As our proposal explicitly models document segments, we also report statistics, considering both topics and languages, related to the average number of segments per document (Table 3), and the average length of segments per document (Table 4). The latter statistic is computed separately for BoW and BoS representations. We made this distinction because a term cannot have a mapping to a synset, or it can be mapped to more than one synset in the BoS space during the WSD process (Section 3.2).

Looking at the average number of segments per document in Table 3, it can be noted that English documents contain, for all topics, a larger number of segments. This means that English documents are generally richer than the ones in the other languages. Italian language corresponds to the smallest documents, each of them containing between 2 and 3.2 segments on average. A sharper difference appears in the *MONETARY/ECONOMIC* topic for which English documents contain 5.2 segments, while the Italian ones are composed, on average, by only 2 segments.

Table 4 shows the average length of segments per document for both space representations. Generally, segments in the BoS representation are smaller than the corresponding segments in the BoW space. More in detail, if we consider the ratio between the segment length in BoS and the one in BoW, this ratio is around 2/3 for the English language, while for both French and Italian it varies between 1/4 and 1/3. This disequilibrium is induced by the multilingual concept coverage of BabelNet, as stated by its authors (Navigli and Ponzetto,

2012a), (Navigli and Ponzetto, 2012b). In particular, the WSD process tightly depends from the concept coverage supplied from the language-specific knowledge base.

4.2 Competing methods and settings

We compare our *SeMDocT* with two standard approaches, namely *Bisecting K-Means* (Steinbach et al., 2000), and *Latent Semantic Analysis (LSA)*-based document clustering (for short, *LSA*). Given a number K of desired clusters, *Bisecting K-Means* produces a K-way clustering solution by performing a sequence of K-1 repeated bisections based on standard K-Means algorithm. This process continues until the number K of clusters is found. *LSA* performs a decomposition of the document collection matrix through Singular Value Decomposition in order to extract a more concise and descriptive representation of the documents. After this step, *Bisecting K-Means* is applied over the new document space to get the final document clustering.

All the three methods, *SeMDocT*, *Bisecting K-Means* and *LSA* are coupled with either BoS or BoW representation models. The comparison between BoS and BoW representations allows us to evaluate the presumed benefits that can be derived by exploiting synsets instead of terms for the multilingual document clustering task.

Both *SeMDocT* and *LSA* require the number of components as input; as concerns specifically *SeMDocT*, we varied r_1 (cf. Section 3.1.4) from 2 to 30, with increments of 2. To determine the number of segment clusters k, we employed an automatic way as discussed in Section 3.1.2. By varying k from 2 to 40, for *Balanced Corpus* and *Unbalanced Corpus*, respectively, the values of k obtained were 22 and 23 under BoS, and 25 and 11 under BoW.

As concerns the step of text segmentation, TextTiling requires the setting of some interdependent parameters, particularly the size of the text unit to be compared and the number of words in a token sequence. We used the setting suggested in (Hearst, 1997) and also confirmed in (Tagarelli and Karypis, 2013), i.e., 10 for the text unit size and 20 for the token-sequence size.

4.3 Assessment criteria

Performance of the different methods are evaluated using two standard clustering validation criteria, namely *F-Measure* and *Rand Index*.

Given a document collection $\mathcal{D}$, let $\mathbf{\Gamma} = \{\Gamma_j\}_{j=1}^{H}$ and $\mathcal{C} = \{C_i\}_{i=1}^{K}$ denote a reference classification and a clustering solution for $\mathcal{D}$, respectively. The local precision and the local recall of a cluster C_i w.r.t. a class Γ_j are defined as $P_{ij} = |C_i \cap \Gamma_j|/|C_i|$ and $R_{ij} = |C_i \cap \Gamma_j|/|\Gamma_j|$, respectively. F-Measure (FM) is computed as follows (Steinbach et al., 2000):

$$F = \sum_{j=1}^{H} \frac{|\Gamma_j|}{|\mathcal{D}|} \max_{i=1\ldots K}\{F_{ij}\}$$

where $F_{ij} = 2P_{ij}R_{ij}/(P_{ij} + R_{ij})$.

Rand Index (RI) (Rand, 1971) measures the percentage of decisions that are correct, penalizing false positive and false negative decisions during clustering. It takes into account the following quantities: *TP*, i.e., the number of pairs of documents that are in the same cluster in $\mathcal{C}$ and in the same class in $\mathbf{\Gamma}$; *TN*, i.e., the number of pairs of documents that are in different clusters in $\mathcal{C}$ and in different classes in $\mathbf{\Gamma}$; *FN*, i.e., the number of pairs of documents that are in different clusters in $\mathcal{C}$ and in the same class in $\mathbf{\Gamma}$; and *FP*, i.e., the number of pairs of documents that are in the same cluster in $\mathcal{C}$ and in different classes in $\mathbf{\Gamma}$. Rand Index is hence defined as:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

Note that for each method, results were averaged over 30 runs and the number of final document clusters K was set equal to the number of topics in the document collection (i.e., 6).

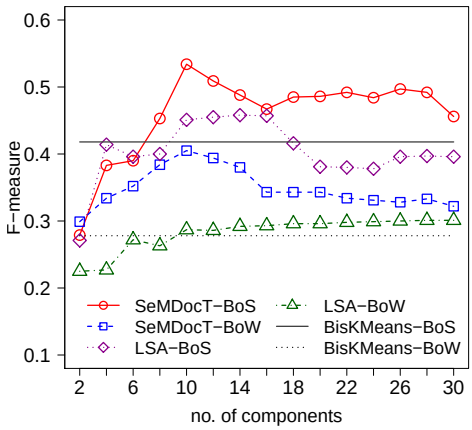
5 Results

We present here our main experimental results. We first provide a comparative evaluation of our *SeMDocT* with the competing methods, on both balanced and unbalanced corpus evaluation cases. Then we provide a per language analysis focusing on *SeMDocT*.

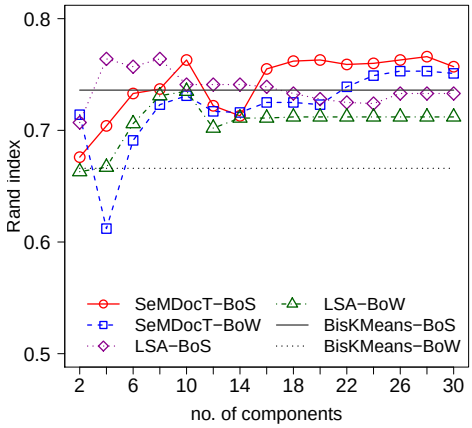
5.1 Evaluation with competing methods

Evaluation on balanced corpus. Figure 2 shows FM and RI results obtained by the various methods coupled with the two document representations on the *Balanced Corpus*. Several remarks stand out. First, the BoS space positively influences the performance of all the employed approaches. This is particularly evident for *Bisecting K-Means* and *LSA* that clearly benefit from this kind of representation. The former almost doubles its performance in terms of FM and significantly improves its result w.r.t. RI. *LSA* shows improvements in both cases. *SeMDocT*-BoS generally outperforms all the competitors for both FM and RI when the number of components is greater than 16. Note that, under the BoW model, *SeMDocT*-BoW still outperforms the other methods.

Evaluation on unbalanced corpus. Figure 3 reports results for the *Unbalanced Corpus*. Also in this evaluation, the best performances for all the methods are reached using the BoS representation. *SeMDocT*-BoS shows similar behavior according to the two measures. It always outperforms the competitors considering a number of components greater than or equal to 12. More precisely, *SeMDocT*-BoW obtains a gain of 0.047 and 0.103 in terms of FM and 0.006 and 0.058 in terms of RI, w.r.t. *LSA*-BoW and *Bisecting K-Means*-BoW, respectively. Similarly, *SeMDocT*-BoS obtains improvements of 0.05 in terms of FM w.r.t. both BoS

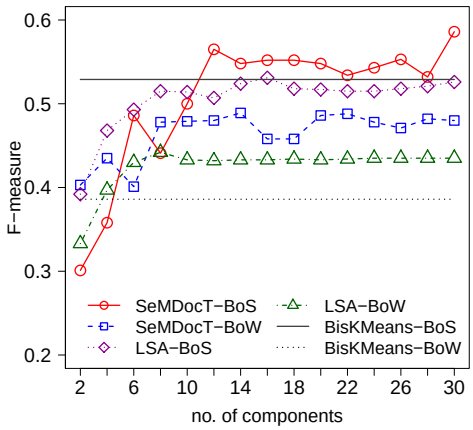


(a)

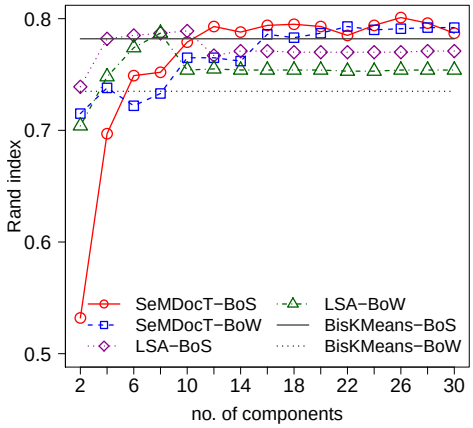


(b)

Figure 2: Average F-Measure (a) and Rand Index (b) on the *Balanced Corpus* using BoW and BoS document representation and varying the number of components for both *SeMDocT* and *LSA*.



(a)



(b)

Figure 3: Average F-Measure (a) and Rand Index (b) on the *Unbalanced Corpus* using BoW and BoS document representation and varying the number of components for both *SeMDocT* and *LSA*.

competitors, while in terms of RI the differences in performance are 0.012 and 0.019 for *LSA*-BoS and *Bisecting K-Means*-BoS, respectively.

5.2 Per language evaluation of *SeMDocT*-BoS

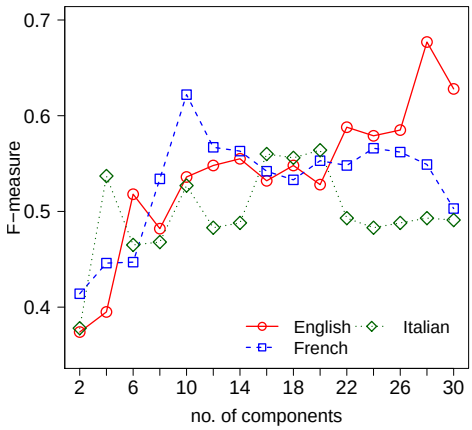
Starting from the clustering solutions produced by *SeMDocT*-BoS in both balanced and unbalanced cases, for each language we extracted a language-specific projection of the clustering. After that, we computed the clustering validation criteria according to language specific solutions to quantify how well the clustering result fits each specific language. The results of this experiment are reported in Fig. 4 and Fig. 5.

On the *Balanced Corpus*, *SeMDocT*-BoS shows comparable performance for English and French documents, while it behaves slightly worse for Italian texts. This trend is highlighted for both clustering evaluation criteria. Inspecting the results for the *Unbalanced Corpus*, we observe a different trend. Results obtained for the English texts are generally better than the results for the French and Italian documents. For this benchmark, *SeMDocT*-BoS obtains similar results for documents written in French and in Italian.

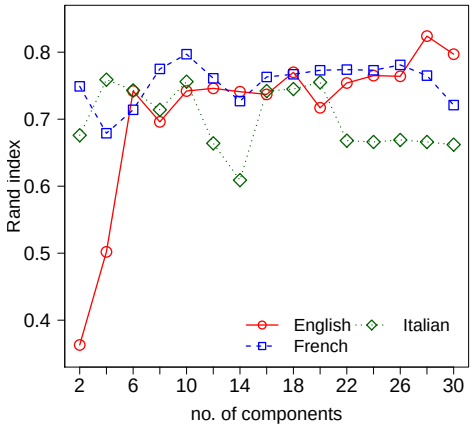
Dataset	*Language*	*BoW size*	*BoS size*	*avg # synsets per term* (β)
Balanced	English	29 999	12 065	0.4021
	French	17 826	5 310	0.2978
	Italian	16 951	4 471	0.2637
Unbalanced	English	19 432	10 387	0.5345
	French	14 439	4 431	0.3068
	Italian	14 743	4 012	0.2721

Table 5: Balanced corpus: language statistics.

We gained an insight into the above discussed performance behaviors by computing some additional statistics that we report in Table 5: for each language and each dataset, the size of the term and synset dictionaries and the average number of synsets per lemma (β) we retrieved with BabelNet according to the related corpus. More in detail, β is the ratio between the BoS and the BoW dictionaries. This quantity roughly evaluates how many synsets are produced per term during the multilingual WSD process (Section 3.2). As we can observe, this value is always smaller than one, which means that not all the terms have a corresponding mapping to a synset. The β ratio can explain the discrep-

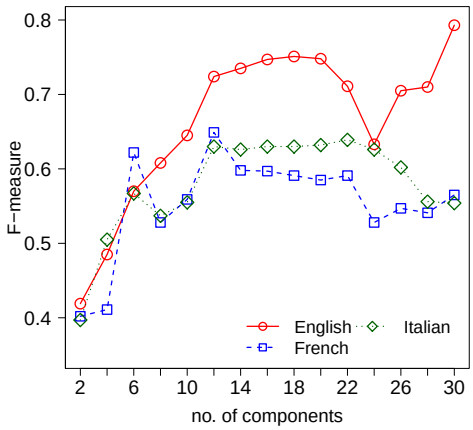


(a)

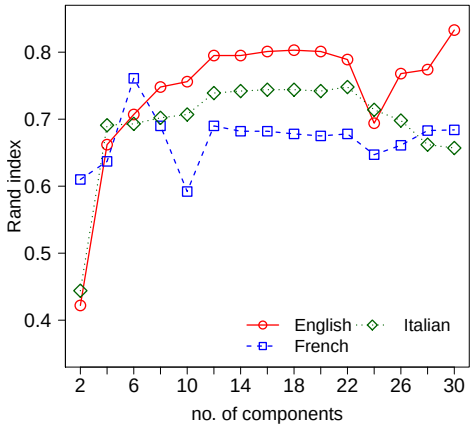


(b)

Figure 4: Average F-Measure (a) and Rand Index (b) for language specific solutions on the *Balanced Corpus* obtained by *SeMDocT*-BoS.



(a)



(b)

Figure 5: Average F-Measure (a) and Rand Index (b) for language specific solutions on the *Unbalanced Corpus* obtained by *SeMDocT*-BoS.

ancy in (language-specific) performances in the two scenarios. In particular, the difference in the β statistic between English and the other languages is more evident for the *Unbalanced Corpus* (i.e., 0.23 between English and French), while it is lower for the *Balanced Corpus* (around 0.1). The relatively large gap in β between the first and the second language (respectively, English and French) for the *Unbalanced Corpus* reduces the relative gap between the second and the third languages (respectively, French and Italian) while this trend is less marked for the *Balanced Corpus* as β range is narrower. In summary, we can state that our framework works well if BabelNet knowledge base provides a good coverage of the terms in the analyzed language. Experimental evidence shows that, if this condition is met, *SeMDocT*-BoS provides better clustering results w.r.t. the competing approaches.

5.3 Runtime of tensor decomposition

As previously discussed, T-HOSVD of a third-order tensor can be computed through three standard SVDs. Furthermore, for clustering purposes, we considered only the mode-1 factor matrix of the decomposition. To compute the SVD, we used the *svds()* function of MATLAB R2012b, which is based on an iterative algorithm.[2] Experiments were carried out on an Intel Core I7-3610QM platform with 16GB DDR RAM.

Figure 6 shows the execution time of the SVD over the mode-1 matricization of our tensor for the *Balanced Corpus*, by varying the number of components, for both BoW and BoS representation models. As it can be observed, in both cases the runtime is linear in the number of components. However, the SVD computation in the BoS setting is one order of magnitude faster than time performance in the BoW setting. This is mainly due to a large difference in size between the feature spaces of BoW and BoS (cf. Table 2), since the selected number of segment clusters (k) was nearly the same (25 for BoW, and 22 for BoS). Therefore, by providing a more compact feature space, BoS clearly allows for a much less expensive SVD computation for our tensor decomposition.

[2] http://www.mathworks.it/it/help/matlab/ref/svds.html

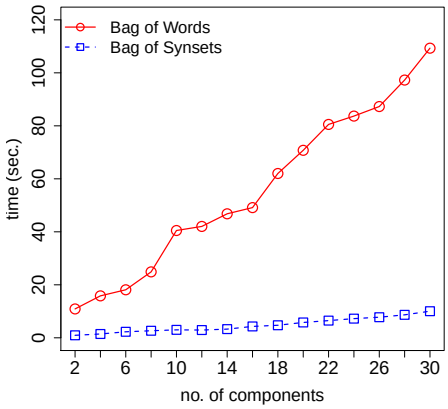


Figure 6: Time performance of SVD over the mode-1 matricization of the *Balanced Corpus* tensor.

6 Discussion

Our work paves the way for the use of a multilingual knowledge base to deal with the multilingual document clustering task. Here we sum up our main findings.

***SeMDocT* vs. *LSA*.** *LSA* achieved its best results for a number of components generally smaller than the one for which *SeMDocT* obtained its maximum. This is due to the initial information that the two methods summarize. *LSA* tries to capture the variation of the initial document-term (alternatively, document-synset) matrix representing the texts in a lower space, whereas *SeMDocT* does the same starting from a richer representation of the documents (i.e., a third-order tensor model). For this reason, *SeMDocT* tends to employ relatively more components in order to summarize the documents content; however, a number of components between 16 and 30 is generally enough to ensure good performance of *SeMDocT*. Moreover, in most cases, the highest performance results by *SeMDocT* are better than the highest performances of *LSA*. for

BoS vs. BoW. Our results have highlighted the better quality in multilingual clustering supplied by synsets compared with the one provided by terms. BoS produces a smaller representation space over which documents are projected, but it is enough rich to well capture the documents content. In particular, BoS benefits from the WSD process that is able to discriminate the same term w.r.t. the context in which it appears.

BabelNet. BabelNet is a recent project that supports many different languages. As the intention of the authors is to enrich this resource, in the future our framework will benefit of this fact. Moreover, our framework can deal with documents written in many different languages as they are represented through the same space; the only constraint is related to the available language support in BabelNet. On the other hand, we point out that any other multilingual knowledge base and WSD tools can in principle be integrated in our framework.

7 Conclusion

In this paper we proposed a new approach for multilingual document clustering. Our key idea lies in the combination of a tensor-based model with a bag-of-synsets description, which enables a common space to project multilingual document collections. We evaluated our approach w.r.t. standard document clustering methods, using both term and synset representations. Results have shown the benefits deriving from the use of a multilingual knowledge base in the analysis of comparable corpora, and also shown the significance of our approach in both a balanced and an unbalanced corpus evaluation. Our tensor-based representation of topically-segmented multilingual documents can also be applied to cross-lingual information retrieval or multilingual document categorization.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proc. of the International Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–41.

Doug Beeferman, Adam L. Berger, and John D. Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210.

Peter A. Chew, Brett W. Bader, Tamara G. Kolda, and Ahmed Abdelali. 2007. Cross-language information retrieval using PARAFAC2. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 143–152.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for Text Segmentation. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 109–117.

Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. 2001. Latent Semantic Kernels. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 66–73.

Ian Davidson, Kiri Wagstaff, and Sugato Basu. 2006. Measuring constraint-set utility for partitional clustering algorithms. In *Proc. of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 115–126.

Dino Ienco, Céline Robardet, Ruggero G. Pensa, and Rosa Meo. 2013. Parameter-less co-clustering for star-structured heterogeneous data. *Data Mining and Knowledge Discovery*, 26(2):217–254.

Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

Young-Min Kim, Massih-Reza Amini, Cyril Goutte, and Patrick Gallinari. 2010. Multi-view clustering

of multilingual documents. In *Proc. of the ACM SIGIR International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 821–822.

Tamara G. Kolda and Brett W. Bader. 2009. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500.

N. Kiran Kumar, G. S. K. Santosh, and Vasudeva Varma. 2011. A language-independent approach to identify the named entities in under-resourced languages and clustering multilingual documents. In *Proc. of the International Conference of the Cross-Language Evaluation Forum (CLEF)*, pages 74–82.

N. Kiran Kumar, G. S. K. Santosh, and Vasudeva Varma. 2011. Effectively mining Wikipedia for clustering multilingual documents. In *Proc. of the International Conference on Applications of Natural Language to Information Systems (NLDB)*, pages 254–257.

N. Kiran Kumar, G. S. K. Santosh, and Vasudeva Varma. 2011. Multilingual document clustering using Wikipedia as external knowledge. In *Proc. of the Information Retrieval Facility Conference (IRFC)*, pages 108–117.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A Multilinear Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Xinhai Liu, Wolfgang Glänzel, and Bart De Moor. 2011. Hybrid clustering of multi-view data via Tucker-2 model and its application. *Scientometrics*, 88(3):819–839.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. PageRank on Semantic Networks, with Application to Word Sense Disambiguation. In *Proc. of the International Conference on Computational Linguistics (COLING)*.

Soto Montalvo, Raquel Martínez-Unanue, Arantza Casillas, and Víctor Fresno. 2007. Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters*, 28(16):2305–2311.

Roberto Navigli and Mirella Lapata. 2010. An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Roberto Navigli and Simone P. Ponzetto. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217-250.

Roberto Navigli and Simone P. Ponzetto. 2012. Multilingual WSD with Just a Few Lines of Code: The BabelNet API. In *Proc. of the System Demonstrations of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 67–72.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2011. Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proc. of the ACM International Conference on Web Search and Web Data Mining (WSDM)*, pages 375–384.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proc. of the Language Resources and Evaluation Conference (LREC)*.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850.

Salvatore Romeo, Andrea Tagarelli, Francesco Gullo, and Sergio Greco. 2013. A Tensor-based Clustering Approach for Multiple Document Classifications. In *Proc. of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 200–205.

Stan Salvador and Philip Chan. 2004. Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms. In *Proc. of the International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 576–584.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A Comparison of Document Clustering Techniques. In *Proc. of the KDD Workshop on Text Mining*.

Andrea Tagarelli and George Karypis. 2013. A segment-based approach to clustering multi-topic documents. *Knowledge and Information Systems*, 34(3):563–595.

George Tsatsaronis, Iraklis Varlamis, and Kjetil Nørvåg. 2010. SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs. In *Proc. of the International Conference on Computational Linguistics (COLING)*, pages 1074–1082.

Chih-Ping Wei, Christopher C. Yang, and Chia-Min Lin. 2008. A Latent Semantic Indexing-based Approach to Multilingual Document Clustering. *Decision Support Systems*, 45(3):606–620.

Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. 2009. WikiWalk: Random walks on Wikipedia for Semantic Relatedness. In *Proc. of the ACL Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49.

Dani Yogatama and Kumiko Tanaka-Ishii. 2009. Multilingual spectral clustering using document similarity propagation. In *Proc. of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 871–879.

Ying Zhao and George Karypis. 2004. Empirical and Theoretical Comparison of Selected Criterion Functions for Document Clustering. *Machine Learning*, 55(3):311–331.