UNIVERSITAT JAUME I
Departament de Llenguatges i Sistemes Informàtics

Discovering and Describing Coherent and Meaningful Topics from Document Collections

Ph. D. dissertation
Henry ANAYA SÁNCHEZ

Supervisors
Dr. Rafael BERLANGA LLAVORI
Dr. Anselmo PEÑAS PADILLA

Honorary Supervisor
Dr. Aurora PONS PORRATA

Castellón, November 2015

To my parents.
To Lisette, Sandra and Eric.

Acknowledgements

First of all, I would like to express my deepest gratitude and acknowledgement to Rafael Berlanga and Anselmo Peñas for their extreme patience, support and permanent motivation. Without them, this work would never have been possible.

I would also like to acknowledge the instruction and assistance I have received during all my studies in Cuba from many people. It would be difficult to mention all of them, but these names are mandatory: Mirtha L. Fernández Venero, Aurora Pons Porrata (RIP) and Hebert Pérez Rosés.

Thanks to the people in TKBG and the Computer Vision Group at Universitat Jaume I for their important support and friendship. It has been a pleasure to meet you all.

This thesis is also dedicated to Tamara and Wilbe, who provided me with much help and care when I moved to Madrid.

Last but not least, I want to take this opportunity to thank my family. Special thanks to my parents and grandmas for the immense dedication and support they have provided me with since ever. Thanks to Sandra and Eric for being my inspiration. Thanks to Lisette for her guidance, encouragement, and for always being there.

This thesis has been partially funded by MINECO (PCIN-2013-002-C02-01) and EPSRC (EP/K017845/1) in the framework of CHIST-ERA READERS project and by the projects TIN2005-09098-C05-04 and TIN2008-01825/TIN.

Abstract

The main motivation behind this thesis is the problem of automatically discovering and describing coherent and meaningful topics underlying a target collection of text documents; where a topic is a theme that runs through documents in the collection.

In this work, discovering topics means to (automatically) produce a processable representation for each of the individual topics in the collection despite they are unobserved data (e.g. using clusters of documents or probability distributions of words); whereas describing a topic aims to generate a summary of the representation of the topic that allows users to identify and discriminate the topic in the context of the target collection.

By semantically coherent topics, we refer to topics that can be easily interpreted by humans, bearing an intelligible (underlying) subject or matter; whereas meaningful topics are meant to represent and summarize the main (vs. background or supporting) themes addressed by each of the individual documents in the target collection.

Discovering and describing topics with these two features can be shown useful to exploratory browsing, but also to obtain semantic decompositions of document collections that bring support to many information accessing and processing tasks. Notice that these topics and their descriptions can be directly applied to provide ostensible end-users with a summary of the main contents included in a target collection of texts.

There are two major trends to discover topics from a collection of text documents. These are clustering-based approaches and the approaches based on *Probabilistic Topic Modeling* (PTM). The first ones represent each topic using a cluster of documents; whereas the second ones employ a probability distribution of words to define each topic.

Nevertheless, as far as we know, none of the existing approaches simultaneously address the issues of ensuring coherence and meaningfulness on the discovered topics as defined in this work. Indeed, only a few existing approaches have been focused on the problem of discovering coherent topics, whereas the issue of providing meaningful topics has not been addressed so far.

In this context, this thesis firstly proposes an abstract framework for dis-

covering and describing topics. Then, from the proposed framework we derive and evaluate two general methodologies, one producing clusters of documents and the other one obtaining probability distributions of words, both aimed to discover and describe topics deemed to satisfy the requirements of coherence and meaningfulness. The main novelty of these methodologies is the combination of both:

- modeling topics from sets of lexically related words in the context of the collection, so that these sets of words determine the *aboutness* of each topic and hence topic coherence is deemed to be satisfied.

- assessing topic meaningfulness by means of probabilistic criteria that penalize topics with an underlying content close to the random contents underlying the target text collection (e.g., topics determined by abstract concepts such as "death victims of murder or accidents", that can merge topics about specific accidents or crimes, etc.).

In the framework and, consequently, in the two derived methodologies, the topic discovery process is implemented as an iterative search in which topics are successively discovered, in a fully unsupervised manner, until all the documents in the target collection are considered to be covered by at least one topic.

No prior knowledge about the topics is utilized, and the number of topics is not needed to be prescribed beforehand. The latter is one of the strongest points of our proposal, since many approaches –most based on PTM– require from setting a priori the number of topics to be discovered from the collection, which is very difficult to know in practice (mainly, if we are indeed interested in obtaining data that describe the collection).

The experiments carried out over target collections of news stories and collections of tweets about different entities in a given domain (e.g., *music/artists* and *carmakers*) show that the proposed methodologies achieves a higher performance in terms of coherence scores and meaningfulness than state-of-the-art related approaches. The latter is based on the agreement (i.e., comparison) with human annotations.

Resumen, principales contribuciones y resultados

La presente tesis trata el problema no supervisado del descubrimiento y la descripción de tópicos coherentes y significativos a partir de una colección de textos, donde un *tópico* es una abstracción que representa una temática que fluye a través de los documentos de la colección.

Descubrir tópicos en esta tesis significa entonces producir de manera automática una representación procesable mediante ordenador de cada uno de los tópicos de la colección (por ejemplo, mediante grupos de documentos o distribuciones de probabilidad sobre un vocabulario), a pesar de ser el conjunto de tópicos una variable no observada en los datos de entrada, es decir, en la colección de documentos.

Mediante tópicos coherentes nos referimos a tópicos que pueden ser fácilmente interpretable por personas, a partir de los cuales se puede inferir un asunto. El calificativo de significativos se refiere a tópicos que representan el tema principal tratado por documentos individuales de la colección.

El descubrimiento y la descripción de tópicos con estas características resulta de gran utilidad en tareas tales como la exploración por parte de usuarios de un grandes colecciones de documentos; pero se puede emplear, además, en la obtención de descomposiciones semánticas de colecciones de documentos que puedan dar soporte a muchas tareas de procesamiento y acceso a la información. Nótese que estos tópicos y sus descripciones pueden ser usados para proporcionar a usuarios finales un resumen de los principales contenidos de la colección de entrada.

En la actualidad existen dos grandes tendencias en el descubrimiento de tópicos a partir de una colección de textos. Estas tendencias son: el agrupamiento de documentos y el modelado probabilíctico de tópicos (en inglés, Probabilistic Topic Modeling). En la primera, cada tópico se representa mediante un grupo o clúster de documentos, mientras que en la segunda se emplea una distribución de probabilidad sobre un vocabulario para representar cada tópico.

Sin embargo, hasta donde conocemos, no existen en la actualidad aproxi-

maciones que traten la cuestión de descubrir al mismo tiempo tópicos coherentes y significativos tal y como se definen en esta tesis. Sólo algunas aproximaciones han tratado en solitario el problema de obtener tópicos interpretables. La cuestión de descubrir tópicos significativos no ha sido tratada hasta ahora.

En este contexto, esta tesis se centra en proponer nuevas methodologías generales para descubrir y describir simultáneamente y de manera automática los tópicos coherentes y significativos de una colección de documentos de texto que es dada como entrada.

Se toman como punto de partida dos hipótesis principales que se corresponden con las propiedades de coherencia y significatividad de los tópicos. Estas son:

1. *Hipótesis de descubrimiento de tópicos coherentes*: Cada tópico coherente puede ser descubierto o aprendido a partir de un conjunto de palabras relacionadas léxicamente en el contexto de la colección.

2. *Hipótesis de descubrimiento de tópicos significativos*: Partiendo de que cada tópico tratado en un documento se descubre a partir de un conjunto de palabras relacionadas de manera léxica, se asume que un tópico significativo no debe ser tan general o abstracto que esté demasiado próximo a contenidos selecionados al azar en la colección. Tampoco debe ser tan específico como para que las palabras que permiten definirlo estén presentes muy probablemente en otros tópicos.

El problema de descubrir y describir tópicos se trata de manera totalmente no supervisada. No se consideran muestras de los tópicos a descubir y el número de éstos tampoco se conoce de antemano.

Principales contribuciones

La principales contribuciones de esta tesis son las siguientes:

1. Primeramente, se realiza una revisión de los principales métodos existentes para el descubrimiento de tópicos que al mismo tiempo proporcionan una descripción de los mismos, realizándose todo de manera no supervisada. Se incluyen tanto métodos basados en agrupamiento (específicamente, métodos basados en el minado de conjuntos frecuentes de palabras) y métodos basados en modelos probabilísticos de tópicos. Se describen las principales limitaciones de los métodos en cuanto a la obtención de tópicos coherentes y significativos.

2. En línea con las principales hipótesis de la tesis, y teniendo en cuenta el estudio de las principales limitaciones encontradas en los métodos existentes, se propone un marco general y abstracto para el descubrimiento y la descripción de tópicos. El marco implementa el proceso de

descubrimiento de tópicos en térrminos de una búqueda basada en una definición abstracta de signaturas léxicas (conjuntos de palabras relacionadas de manera léxica en el contexto de la coleccion de documentos). Además, se basa en un conjunto de componentes abstractas que se emplean para dar soporte al descubrimiento de tópicos con las características de coherencia y significatividad.

3. A partir del marco general, primeramente se deriva un método nuevo para descubrir y describir tópicos representados por una agrupación de documentos. El método propuesto implementa el concepto de signaturas léxicas por medio de pares de palabras, que intentan representar de manera breve y precisa el asunto tratado por cada uno de los tópicos. Estos pares guían directamente el proceso de búsqueda de los tópicos en la colección.

4. Como parte de este método, se introduce el criterio de homogeneidad de un conjunto soporte de documentos para evaluar la significatividad de un tópico a partir del par de palabras que lo define. El criterio de homogeneidad necesita de un umbral de semejanza entre documentos que se calcula de manera automática a partir de la colección de documentos de entrada.

5. Usando el umbral de semejanza anterior, se propone un nuevo método para definir un tópico coherente a partir de un par de palabras. El método se define a partir del concepto de grafo de máxima β-semejanza.

6. Se propone también un nuevo mecanismo para obtener descripciones extendidas de grupos de documentos que se basa en una prueba de razón de verosimilitud.

7. A partir del marco general se deriva, además, otro nuevo método para descubrir y describir tópicos que se expresan en términos de distribuciones de probabilidad sobre el vocabulario de la colección de documentos. El método se basa tanto en modelos estadísticos de lenguajes como en el modelado probabilístico de tópicos para aprender tópicos coherentes y significativos de manera no supervisada. La búsqueda de tópicos en este caso se basa en signaturas léxicas que se obtienen de documentos individuales de la colección.

8. Como parte de esta instancia del marco general, se propone el método SLM para aprender y describir tópicos a partir de una signatura léxica. SLM se define mediante la combinación de modelos de probabilidades condicionadas entre palabras y de un procedimiento de refinamiento de modelos de lenguajes.

9. Tomando como base el modelado probabilístico de tópicos, se introduce SDM como un mecanismo para modelar la significatividad semántica

de un conjunto de tópicos a partir de un nuevo modelo probabilístico basado en urnas, cuyo objetivo es modelar el grado de significatividad de los tópicos en un documento.

10. Se realiza la evaluación de los métodos propuestos haciendo uso de colecciones de documentos de registros diferentes. Específicamente, se consideran dos colecciones de noticias (una en español y otra en inglés) y dos colecciones de mensajes de Twitter acerca de entidades específicas de dos dominios: automoción y artistas musicales. Los documentos de las colecciones han sido etiquetados previamente con tópicos por parte de expertos. Se comparan los resultados obtenidos por cada método con los obtenidos por métodos relacionados. Además, se comparan los resultados obtenidos por los métodos propuestos entre sí en términos de información mutua con respecto a los tópicos manuales.

Resultados

A partir del marco general propuesto y de los dos métodos que se derivan, se obtienen los principales resultados de esta tesis que consisten en corroborar las hipótesis planteadas. Además, se corrobora:

- el impacto positivo de combinar el criterio de homogeneidad y el método basado en el grafo de máxima β-semejanza en el método de agrupamiento para obtener tópicos de gran calidad en cuanto a coherencia y significatividad semántica. Los tópicos obtenidos se acercan más a los definidos por los expertos que aquellos obtenidos por los algoritmos de agrupamiento que se basan en conjuntos frecuentes de palabras tales como FIHC y similares (que son los métodos del estado de la cuestión que guardan más similitud con el propuesto).

- la validez del método SLM que se propone para el aprendizaje de tópicos coherentes y sus descripciones a partir de conjuntos de palabras relacionadas léxicamente. Los tópicos aprendidos por medio de SLM resultan ser más coherentes que aquellos obtenidos por LDA y métodos similares en general. También resultan ser más coherentes que los inferidos por el método Quad-Reg en colecciones de mensajes de Twitter, a pesar de que Quad-Reg se centra en modelar tópicos coherentes por medio de mecanismos de regularización (que no se emplean en nuestra propuesta pero que pudieran ser incorporados directamente).

- la utilidad del método SDM propuesto para modelar la significatividad semántica de los tópicos. Mediante SDM, el método de descubrimiento y descripción de tópicos propuesto que se basa en modelos de lenguajes y en el modelado probabilístico de tópicos es capaz de aprender de manera no supervisada tópicos con mayor significancia semántica que

los obtenidos por métodos tradicionales basados en el modelado probabilístico de tópicos.

- que el método derivado del marco general que se basa en modelos de lenguajes y en el modelado probabilístico de tópicos supera al método basado en agrupamiento de documentos en términos de información mutua respecto a los tópicos manuales. Además, este método descubre un número de tópicos que se aproxima mejor al número de tópicos manuales etiquetados por los expertos.

- el marco general propuesto puede ser usado para derivar métodos que de manera satisfactoria descubran tópicos coherentes y significativos desde el punto de vista semántico, sin tener en cuenta supervisión alguna y sin necesidad de proporcionar de antemano el número de tópicos a descubrir.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The ever-increasing availability of text documents has led to a growing challenge for information systems to effectively manage and retrieve the information comprised in large collections of texts according to the users' information needs.

As previously pointed out in (Cutting et al., 1992), the standard formulation of the information access problem presumes a query, which is the user's expression of an information need. The task is then to search a target collection for documents that match this need and retrieve them for the user.

However, it is not always easy or even possible for users to formulate such needs precisely. For example, users may not be familiar with the vocabulary that defines the themes of their interest, or simply they wish to get a broad summary of the collection in order to guide their searches.

For this reason, there exists a great interest to develop methodologies and tools for analyzing and summarizing these collections according to their main topics; i.e., the main themes addressed by the collection documents.

Traditionally, two major approaches have been applied to organize a text collection according to their main topics. These are clustering-based approaches and the approaches based on *Probabilistic Topic Modeling*.

Clustering is an unsupervised learning technique that has been widely used in the process of topic discovery from documents. Basically, clustering methods are aimed at generating document groups or clusters, each one representing a different topic. Clusters are often generated in such a way that documents belonging to the same cluster are very similar to each other while exhibit some differences with respect to documents in other clusters.

On the other hand, Probabilistic Topic Modeling (PTM) (Blei, 2012; Blei et al., 2010; Steyvers and Griffiths, 2007; Blei et al., 2003; Griffiths and Steyvers, 2004; Mimno et al., 2011) has been proposed to discover –also in an unsuper-

vised manner– different distributions of words from a text collection in such a way that these distributions "jointly" model the generation of individual documents as a mixture model. Each of these distributions of words is also expected to capture a salient theme that runs through the documents in the text collection, and therefore, they are considered to be topics.

Clearly, both clustering and PTM approaches can be applied to organize and summarize large collections of text documents in terms of a relative small number of topics represented by clusters and word distributions respectively. Unfortunately, a major limitation of these approaches is the quality of the discovered topics.

As pointed out in previous work (Boyd-Graber et al., 2009; Newman et al., 2009; Mimno et al., 2011; Newman et al., 2011), traditional PTM approaches do not always correlate with human judgments so as to always provide ostensible end-users (beyond Machine Learning practitioners, in a fully unsupervised scenario) with semantically coherent and meaningful topics. By semantically coherent topics, we refer to topics that can be easily interpreted by humans, bearing an intelligible (underlying) subject or theme; whereas meaningful topics are meant to represent and summarize the main (vs. background or supporting) themes addressed by each of the individual documents in the target collection.

In traditional PTM approaches, topics are modeled as latent (hidden) variables representing word distributions, and despite their values are statistically significant, they are sometimes difficult to interpret and explain by humans since the information they convey in many cases is not at all directly related to a subject heading (i.e., they are not coherent). In other cases, the topics correspond to either background (abstract) or supporting (very specific) themes in the collection (e.g., a very abstract concept or a specific event from a striking news).

A similar criticism has been applied to common clustering-based approaches. Since experimental results have shown that document clusters often tend to merge documents from different topics as manually labeled by human annotators, it has been claimed that the obtained clusters do not always correspond to actual coherent and meaningful topics (Anaya-Sánchez et al., 2010; Fung et al., 2003; Pons-Porrata et al., 2007a).

Besides, the application of topic discovery techniques to summarize the contents of text collections needs from a mechanism that summarizes or describes each topic for the users, in order to let them determining at a glance those topics of their interest. In this regard, common clustering-based approaches do not use to provide a "built-in" mechanism that summarize the clusters' contents; whereas a word distribution from a PTM approach can be considered as a topic description for ostensible end-users up to some degree of interpretability of the topic.

In this context, this thesis addresses the following issues:

(i) How to discover the semantically coherent and meaningful topics un-

derlying a target collection of text documents?

(ii) How to simultaneously provide an appropriate description for each topic so that humans can easily judge its relevance?

Here, *semantically coherent topics* refers to topics that can be easily interpreted by humans, bearing an intelligible (underlying) subject or matter that we refer to as *aboutness*. *Meaningful topics* are meant to represent and summarize the main (vs. background/abstract or supporting) themes addressed by each of the individual documents in the target collection.

In particular, we are interested in providing new methodologies for discovering and describing topics to cope with the issues mentioned above, while overcome other technical limitations of existing approaches such as:

- *Use of user-specific parameters*: Some topic discovery approaches require from defining a priori a set of parameters by the user. Often, these parameters have an important impact on the performance of the approaches, and they are also difficult to be defined by expert users. For example, most PTM approaches require from prescribing the number of topics to be discovered in advance (which is unfeasible to predict by ostensible end-users in a fully unsupervised manner). In the same vein, some clustering-based approaches that rely on frequent word sets requires from setting up a minimum support threshold for mining these word sets, and this threshold directly determines a minimum bound on the size of the topics to be discovered.

- *Generating a large number of redundant topics*: Existing approaches –most in the class of clustering-based approaches– tend to produce a very large number of topics; which, sometimes, entails a large degree of overlapping/redundancy between pairs of discovered topics. Approaches based on PTM do not systematically address the issue of obtaining non-redundant topics.

- *Failure to detect/discover small-size topics*: Despite the possible unbalance between topic sizes in a target collection, the approaches should not only attempt to discover large-size topics but also medium- and small-size topics whenever these topics actually represent the main themes addressed by existing documents in the target collection. [1]

1.2 Goals

The main goal of this thesis is to develop new methodologies to effectively discover and describe topics from a target collection of texts. Specifically, the aim is to:

[1] In this thesis, a topic's size is measured by the number of documents in the collection that are addressed by the topic.

1. Contribute with general methodologies to approach the problem of simultaneously discovering and describing semantically coherent and meaningful topics from both perspectives: (a) representing topics using clusters of documents and (b) defining each topic as a probability distribution of words. The proposed methodologies are required to:

   A) be fully unsupervised approaches,

   B) obtain topics covering all of the documents in the target collection of texts,

   C) discover topics of any size despite the possible unbalance of topic sizes in the target collection,

   D) provide illustrative and discriminating descriptions of the discovered topics (descriptions with vague or very ambiguous words, such as *thing*, *today*, *person*, etc., should be avoided),

   E) do not need to know a priori the number of topics to be discovered (to predict the number of topics addressed by a collection of documents is currently a very hard problem),

   F) rely on the smallest number of user-defined parameters as possible,

   G) discover actual semantically coherent and meaningful topics (i.e., discover topics with a clear, subject-heading like interpretation, which must be the main non-abstract theme of a non-empty subset of documents in the collection).

2. Evaluate the performance of the proposed approaches using manually-labeled target collections of different document registers (e.g., news stories of medium and large size documents, and tweets –i.e. shorts texts of up to 140 characters that are posted using a nonstandard language with similarities to SMS style–). Broadly, the evaluation should include both: (i) validating the adequacy of each component of the approaches (i.e., the engineering of the solutions) and (ii) comparing the performance of the approaches to state-of-the-art methods by mainly regarding topic coherence and meaningfulness.

1.3 Hypothesis

There are two main (general) hypotheses underlying this thesis. These hypotheses correspond to the quality features of coherence and meaningfulness demanded for the topic to be discovered, namely:

(H1) *Coherent topic discovery hypothesis*: Each coherent topic can be learned as an explanation of a set of lexically related words in the context of the target text collection.

(H2) *Topic meaningfulness hypothesis*: Assuming that each topic in a document is discovered from a set of lexically related words, we claim that a meaningful topic should not be too general/abstract so that it is too close to random contents (or concepts) underlying the target text collection; neither too specific/concrete so that the set of lexically related words from which it is discovered be likely generated from other (more general) topics in the collection.

Two primary abstractions can be used to define these hypotheses: *sets of lexically related words* and *topic meaningfulness*. Thus, different instantiations of these abstractions (that is, different ways of interpreting or representing lexical relations between words and different ways to assess topic meaningfulness) will lead to different implementations of these hypotheses and, finally, to different approaches to our problem of simultaneously discovering and describing topics.

1.4 Methodology

Aligned with the main goal and hypotheses of this thesis, the operational methodology devised can be summarized as follows:

1) Perform a thorough analysis of the problem of simultaneously discovering and describing topics in the context of the state-of-the-art of topic discovery approaches. The aim is to elucidate the issues that need to be solved. Since there are two main perspectives for discovering topics, the study needs to be partitioned; however, the issues must be lined up with the quality features they affect in order to be successfully addressed by the proposed methods.

2) Build upon and generalize existing approaches in line with the quality features/hypotheses to devise an abstract framework for discovering and describing topics regardless the topic representation perspective.

3) Derive concrete methods from the abstract framework to discover and describe topics from the perspectives of clustering and PTM by separate. To implement in each case the necessary components to deal with the issues identified in Step 1 of this methodology.

4) Evaluate each concrete method regarding traditional quality measures from the corresponding perspective. Validate the engineering solutions (i.e., the adequacy of each component in the approaches).

5) Make general conclusions about the strengths and limitations of the different approached perspectives for discovering and describing high quality topics.

1.5 Organization

The rest of this document is organized as follows.

- Firstly, Chapter 2 provides a background on document clustering and PTM (Section 2.2 and Section 2.3), which are currently the main techniques employed to discover topics in a fully unsupervised manner. This chapter also provides a more formal definition of the research problem (Section 2.1), summarizes the different evaluation approaches (Section 2.4), and briefly describe related tasks (Section 2.5).

- Then, Chapter 3 reviews the current state-of-the-art of the research problem, which mainly includes those topic discovery approaches that simultaneously discover and describe topics. Overall, the set of clustering-based approaches that rely on frequent word-based itemsets (Section 3.1) and a variety of extensions and modifications to LDA aimed at improving the quality of the topics (Section 3.2) is surveyed. Some conclusions summarizing the main issues that concern topic quality in the reviewed approaches are provided (Section 3.3).

- Chapter 4 is devoted to introduce a general abstract framework for discovering and describing topics. In this chapter, the main framework components are also outlined and the most related approaches in the state-of-the-art are contextualized within the framework.

- In Chapter 5, a novel clustering-based approach derived from the abstract framework is presented for discovering and describing topics.

- Chapter 6 introduces a new method, also derived from abstract framework, that discover and describe coherent and meaningful topics based on PTM.

- Finally, Chapter 7 concludes this thesis by summarizing the contributions and results obtained. Some discussion comparing the different perspectives approached is also provided in this chapter. Finally, the chapter outlines and describes feasible directions for further research relying on the results of this thesis.

Chapter 2

Background

2.1 The problem of discovering and describing coherent and meaningful topics

The problem of discovering and describing the coherent and meaningful topics comprised in a given collection of text documents consists of both determining the main themes addressed by the collection documents and providing an illustrative yet discriminant description for each one; where each theme is referred to as a topic.

More formally, this problem can be expressed as that one of determining a set of pairs $\{(T_1, \delta_1), \ldots, (T_K, \delta_K)\}$ from a collection of text documents $D = \{d_1, \ldots, d_N\}$, in such a way that:

i. $\forall k \in \{1, \ldots, K\}$, T_k represents the main topic (i.e., the main concrete theme) addressed by each of the documents in a non-empty subset of documents $G(T_k) \subseteq D$,

ii. δ_k is a word-based description for T_k in the context of D, and

iii. there exists an intelligible (latent but concrete) subject or matter s_k conveyed by both T_k and δ_k that represents the "aboutness" of the topic and makes it interpretable; that is, s_k uniquely determines both T_k and its description δ_k in the given text collection.

In this problem, no prior knowledge about the collection or the topics is considered to be known in advance (e.g., domain information of the collection documents, topic samples, etc.). Even, the number of topics to be discovered is a priori unknown.

The above problem statement leads to a topic definition that is in accordance to that followed by the Topic Detection and Tracking (TDT) research program, in which a topic is defined in the domain of news stories to be a

seminal event or activity, plus all its derivative (directly related) facts, events or activities Doddington (1998); being an event defined as something that happens at some specific time and place (e.g., a specific airplane crash; whereas *airplane crash* in general is not).

Thus, according to TDT, a news story (i.e., a news report of any length, usually presented in a straightforward style and without editorial comment) is considered to address a given topic whenever the story is directly connected to the associated event. [1] Besides, as part of an effort to broaden the notion of topic, in TDT a topic is also a set of news with a coherent focus on a concrete theme, even when there is no a clear underlying event.

However, our definition of topic is not limited to news stories. Instead, we consider arbitrary collections of text documents, where each document coherently focuses on one or more concrete themes.

To address the issue of discovering topics from a collection of text documents immediately implies to give a representation for topics. From our problem statement two alternative representation can be straightforward realized: (1) to represent a topic as a group or cluster of documents from the collection and (2) representing a topic as a combination of words (e.g., a probability distribution of words) from the vocabulary of the collection. Indeed, most approaches to the problem of topic discovery are based on document clustering or Probabilistic Topic Modeling (PTM).

2.2 Document clustering

The aim of clustering text documents is to sort out a collection of documents $D = \{d_1, \ldots, d_N\}$ into a set of document groups or clusters $G = \{G_1, \ldots, G_K\}$ $(1 \leq K \leq 2^N - 1)$, each one representing a document category $G_k \subseteq G$ $(\forall k \in \{1, \ldots, K\})$.

It can be assumed that for each clustering G of the documents in D there exists a function $f : D \times G \to \mathbb{R}$ satisfying the following propositions:

i. $\forall(d \in D)[\forall(G_k \in G)[d \in G_k \Rightarrow \forall(G_{k'} \in G)[f(d, G_k) \geq f(d, G_{k'})]]]$

ii. $\forall(d \in D)[\forall(G_k \in G)[d \notin G_k \Rightarrow \exists(G_{k'} \in G)[f(d, G_{k'}) > f(d, G_k)]]]$

where function f is frequently determined by an optimization problem referred to as clustering scheme. Thus, solving function f is equivalent to solve the target text clustering problem.

In practice, the objective function of the clustering scheme depends on both the function f to be solved and some data objects (i.e., constants and/or vari-

[1] For example, a story on the search for survivors of an airplane crash, or on the funeral of the crash victims, will be considered to be a story on the crash event topic. Obviously, there are limits to this inclusiveness. Thus, stories on FAA repair directives that derive from a crash investigation probably would not be considered to be stories on the crash event Doddington (1998).

ables) representing the documents and/or clusters. For example, a prototypical objective function may take the form:

$$J(\Phi, \Psi) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathrm{I}(d_i, G_k)\delta(\phi_i, \psi_k) \tag{2.1}$$

where ϕ_i is a representation for document d_i, ψ_k is an internal representation for cluster G_k, δ is an expression that specifies some estimated or desired relationship between a document representation and an internal cluster representation (e.g., $\delta(\phi_i, \psi_k) = \|\phi_i - \psi_k\|$ if both document d_i and cluster G_k are represented by vectors), and $\mathrm{I}(d_i, G_k)$ is a binary variable that represents the membership of document d_i to cluster G_k.

In this way, a key problem to face for specifying a clustering algorithm is that one of specifying a representation for the collection documents and the clusters in order to allow defining the clustering scheme in a suitable manner.

2.2.1 Document representation models

In the literature, there are several models that are frequently used as a general framework for representing documents in a text collection. The following sections present three of the most commonly used models.

2.2.1.1 Bag-of-words model

The bag-of-words model (Manning and Schütze, 1999) is perhaps the simplest model for representing documents. In this model, a text document is represented as the bag (multiset) of its indexing terms –commonly, words in the vocabulary of the collection or their lemmas– disregarding grammar and even term order but keeping term multiplicity; i.e., the number of times a term occurs in the document.

2.2.1.2 Vector Space Model

In the Vector Space Model (VSM) (Grossman and Frieder, 2004), each document d_i in the collection is represented by means of an r-dimensional vector (r is the number of different indexing terms chosen for the collection), in which each component represents the weight of the term associated to that dimension.

The weight associated to a term t_j in a document d_i –denoted by w_j^i– uses to represent a statistical estimate of the importance of the term for describing the document; that is, the usefulness of the term for distinguishing the document from among the other documents in the collection.

Similar to the bag-of-words model, indexing terms may correspond to words in the vocabulary of the collection or their lemmas, but also they can be keywords or phrases extracted from the collection documents.

In this model, a term is often weighted with value 0 in the representation of every document in which it does not appear. Both very frequent and rare terms appearing in the collection are usually disregarded as indexing terms. Commonly, document vectors are normalized so that the length of each document affect weight of its terms.

There are different techniques referred to as *weighting schemes* aimed to assign the weight to each term in a document. Some of these schemes include the following.

- *Boolean or binary scheme*. Weights take values in the set $\{0, 1\}$. If the term occurs in the document it is weighted with value 1; otherwise, the term's weight is 0.

- *Term Frequency (TF) scheme*. Each term is assigned with a weight proportional to the number of times it occurs in the document. In this scheme, the term weight is typically denoted by the expression $tf(t_j, d_i)$. Often, frequencies are normalized to mitigate high frequency phenomena caused by very large documents. In this regard, the standard L_1 normalization is frequently applied.

- *TF×IDF*. It considers both the frequency with which the term occurs in a document (i.e., the TF factor) and how frequently the term occurs in the collection documents (i.e., the IDF factor). The weight of term t_j in document d_i is defined as $w_j^i = tf\_idf(t_j, d_i) = tf(t_j, d_i) \cdot idf(t_j)$, where $idf(t_j) = \log \frac{N}{df(t_j)}$ and $df(t_j)$ is the number of documents in the collection containing the term t_j. This combination favors those terms with both a high term frequency in a document and a sparse presence among the documents in the collection.

- *SMART ltc*. It is a variant of the TF×IDF scheme firstly implemented in SMART system (Buckley et al., 1995a) that defines the weight of term t_j in document d_i as follows:

$$w_j^i = (1 + \log tf(t_j, d_i)) \cdot \log \frac{N}{df(t_j)} \tag{2.2}$$

Usually, text clustering methods that employ the VSM for representing documents rely on a similarity function between documents to define its clustering scheme.

In the literature, the most widely used similarity function to relate vectors representing documents is the *cosine* function; which defines the similarity between two documents d_{i_1} and d_{i_2} as the cosine of the angle determined by

their corresponding vectors as follows:

$$\cos(d_{i_1}, d_{i_2}) = \frac{d_{i_1} \cdot d_{i_2}}{||d_{i_1}|| \cdot ||d_{i_2}||} = \frac{\sum_{j=1}^{r} w_j^{i_1} \cdot w_j^{i_2}}{\sqrt{\sum_{j=1}^{r} \left(w_j^{i_1}\right)^2} \cdot \sqrt{\sum_{j=1}^{r} \left(w_j^{i_2}\right)^2}} \tag{2.3}$$

where $w_j^{i_1}$ and $w_j^{i_2}$ represent the weight of the jth dimension in the representation of documents d_{i_1} and d_{i_2} respectively.

2.2.1.3 Statistical Language Models

Statistical Language Models constitutes another common way to represent text documents.

Formally, a statistical language model is a function that defines a probability distribution over the elements in a language; where a language is a set of word sequences over a vocabulary.

In this document representation scheme, each document is represented by means of a statistical language model, that is often a model from which the document is the sample of maximum likelihood.

Usually, the language underlying the model is defined in terms of all possible sequences of fixed length composed over the vocabulary of the document collection. These sequences are called n-grams, and the statistical language model is referred to as n-gram language model; where n is the length of the sequences ($n \geq 1$).

In an n-gram language model, the probability distribution that represents a text document is estimated from all the sequences of length n included in the document (i.e., the n-grams of the document). Some estimation methods of the probability distribution that defines a model for a document d_i are the following:

- *Maximum Likelihood Estimator* (MLE):

$$\mathrm{p}_{MLE}(s|d_i) = \frac{tf(s, d_i)}{\sum_{s' \in d_i} tf(s', d_i)} \tag{2.4}$$

- *Laplace or adding one smoothing*

$$\mathrm{p}(s|d_i) = \frac{tf(s, d_i) + 1}{\sum_{s' \in d_i} tf(s', d_i) + |S|} \tag{2.5}$$

- *Jelinek-Mercer smoothing*

$$\mathrm{p}(s|d_i) = (1 - \lambda)\, \mathrm{p}_{MLE}(s|d_i) + \lambda\, \mathrm{p}(s|D) \tag{2.6}$$

- *Dirichlet smoothing*

$$\mathrm{p}(s|d_i) = \frac{tf(s,d_i) + \mu\, \mathrm{p}(s|D)}{\sum_{s' \in d_i} tf(s',d_i) + \mu} \tag{2.7}$$

where $s \in S$ is an n-gram over the vocabulary of the collection, $tf(s,d_i)$ accounts for the number of times n-gram s is included document d_i, $\mathrm{p}(s|D)$ is an estimated probability value for s under the document collection, and both λ and μ represent smoothing factors ($0 < \lambda < 1$, $\mu > 0$).

Commonly, an n-gram language model representing a document d_i is chosen to be a stochastic language model; i.e, a model $\{\mathrm{p}(s|d_i)\}_{s \in S}$ such that $\sum_{s \in S} \mathrm{p}(s|d_i) = 1$.

Particular cases of n-gram language models frequently used for representing text documents in practice are the *unigram* and *bigram* language models, which are defined by setting $n = 1$ and $n = 2$ respectively.

Representing documents using statistical language models allows to relate text documents in a variety of ways. For example, it can be estimated the probability of generating an arbitrary document or phrase (over the vocabulary of the collection) from the statistical language model representing a given document in a document collection. Besides, distance metrics between documents, such as the geodesic distance between distributions (Lafferty et al., 2005; Dillon et al., 2012), can be employed to set up a clustering scheme. The geodesic distance between distributions $\mathrm{p}_i = \{\mathrm{p}_i(s)\}_{s \in S}$ and $\mathrm{p}_j = \{\mathrm{p}_j(s)\}_{s \in S}$ is defined as follows:

$$g(\mathrm{p}_i, \mathrm{p}_j) = 2 \arccos\left(\sum_{s \in S} \sqrt{\mathrm{p}_i(s)}\sqrt{\mathrm{p}_j(s)}\right) \tag{2.8}$$

Other measures traditionally employed to relate/compare probability distributions are the following:

- *Hellinger distance*:

$$h(\mathrm{p}_i, \mathrm{p}_j) = \left(\frac{1}{2}\sum_{s \in S}\left(\sqrt{\mathrm{p}_i(s)} - \sqrt{\mathrm{p}_j(s)}\right)\right)^{\frac{1}{2}} \tag{2.9}$$

- *Kullback-Leibler divergence*:

$$KLD(\mathrm{p}_i \,||\, \mathrm{p}_j) = \sum_{s \in S} \mathrm{p}_i(s) \log \frac{\mathrm{p}_i(s)}{\mathrm{p}_j(s)} \tag{2.10}$$

2.2.2 Internal representation for document clusters

Generally, internal representations of a set of documents in a cluster result from combining the representation of the individual documents in the cluster, or from relying on more complex structures (e.g., graphs) that relate a set of individuals in their definition. Thus, examples of non-trivial cluster representations utilized by existing document clustering approaches are the following:

- **Term vectors**. In this representation, each cluster is represented using a vector that results from an algebraic combination of the vectors representing the individual documents in the cluster (e.g., as the median of the vectors that represent the documents in the cluster (Hartigan and Wong, 1979; Arora et al., 1998)).

- **Weighted graphs**. Given a similarity or distance function between document representations, a document cluster can be represented by means of a connected digraph whose set of nodes is given by the documents in the cluster, and each edge in the graph is weighted using the similarity or distance value between the representation of its documents (Aslam et al., 2004; Pons-Porrata et al., 2002).

- **Term sets**. Sets of indexing terms that frequently co-occur in the documents comprising a cluster and that do not frequently co-occur in the complement of the cluster (w.r.t. the entire collection of documents) have been also used to represent a cluster of documents (Beil et al., 2002; Fung et al., 2003; Yu et al., 2004; Malik and Kender, 2006).

2.2.3 Classifying text clustering algorithms

Document clustering methods can be classified into two broad classes: strict partitioning clusterings and overlapping clusterings. Strict partitioning clusterings obtain pairwise disjoint document clusters; whereas in overlapping clusterings each document may belong to more than one cluster.

A distinguished subclass of overlapping clusterings is the class of hierarchical clusterings. The document groups obtained from a hierarchical clustering can be arranged into a document inclusion hierarchy. As a categorization tool, the hierarchical clustering provides a categorization schema composed by multiple levels of abstraction. Clusters belonging to higher levels in the hierarchy are considered to be more general categories than those ones in lower levels.

2.3 Probabilistic Topic Modeling

PTM proposes to model each document d in a collection of text documents from a mixture of K topics (i.e., distributions of words) $\beta_1, \ldots, \beta_K$:

$$\mathrm{p}_d(w) = \sum_{k=1}^{K} \mathrm{p}(w|z = \beta_k)\theta_{d,k} \tag{2.11}$$

where, for all $k \in \{1, \ldots, K\}$, $\theta_{d,k}$ is the prior probability (or proportion) of topic β_k used to model document d, $\mathrm{p}(w|z = \beta_k)$ is the probability of word w under distribution β_k, and both the topics and the priors are (hidden) latent variables from a (stochastic) generative model of documents that allows the generation of all the documents in the collection.

PTM approaches can be easily described by means of the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is the simplest generative model of documents that falls into the PTM framework.

2.3.1 The LDA model

LDA is a generative PTM approach that applies to a collection of text documents $D = \{d_1, \ldots, d_N\}$ represented using the bag-of-words scheme.

Similar to most PTM approaches, LDA assumes that a generative process is responsible for creating the target collection of documents. Then, applying LDA to collecttion D consists of doing inference to "invert" the generative process and recover the latent (unobserved) topics from the observed documents.

Thus, LDA is mainly described by its generative process; that is, the random process by which the model assumes the documents in a collection arose from a mixture of topics.

2.3.1.1 The generative process

Assuming that there are K topics $\beta_1, \ldots, \beta_K$ that have been specified before any data has been generated, the generative model of LDA proposes to generate the collection of documents D by individually (independently) generating each document d in D from the following two-stage process:

i. Firstly, randomly choose a distribution $\theta_d = \langle \theta_{d,1}, \ldots, \theta_{d,K} \rangle \in \mathbb{P}_K$ over the K topics. [2]

ii. Then, generate each word occurrence $w_{d,n}$ in d ($1 \leq n \leq N_d$, N_d being the number of word occurrences in document d) as follows:

[2] $\mathbb{P}_K$ denotes the $(K-1)$-simplex, defined as $\mathbb{P}_K = \{\theta \in \mathbb{R}^K : \forall i \in \{1, \ldots, K\}, \theta_i > 0, \sum_{i=1}^{K} \theta_i = 1\}$.

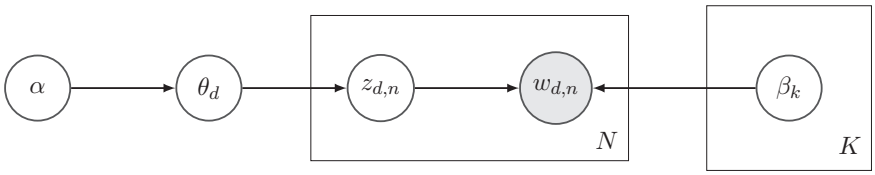


Figure 2.1: Generative model for a document d according to LDA.

a) Randomly choose a topic index $z_{d,n}$ ($z_{d,n} \in \{1, \ldots, K\}$) from the distribution of topics θ_d.
b) Randomly choose $w_{d,n}$ from the topic (i.e., distribution of words) $\beta_{z_{d,n}}$.

This process supports the idea of generating documents with multiple topics. Specifically, each document d in a collection "addresses" topics in different proportions (according to distribution θ_d) as each word occurrence $w_{d,n}$ in d is randomly chosen from one of the topics in the second stage of the process.

Figure 2.1 graphically depicts using plate notation the generative process for a document d according to LDA. In the model, α is the (constant) hyperparameter of a uniform Dirichlet distribution from which topic proportions are sampled, β_k represents the kth topic in the model, θ_d represents the topic proportions for document d, $w_{d,n}$ represents the nth word in document d, and $z_{d,n}$ is the topic assignment for word $w_{d,n}$. The variables representing the topics, the topic proportions and the assignment of topics to words are all latent variables; whereas word occurrences are observed in the corpus of documents. These variables are considered to be distributed as follows:

$$\theta_d \sim \text{Dirichlet}_K(\alpha) \tag{2.12}$$

$$z_{d,n} \sim \text{Discrete}(\theta_d) \tag{2.13}$$

$$w_{d,n}|z_{d,n}, \beta_1, \ldots, \beta_K \sim \text{Discrete}(\beta_{z_{d,n}}) \tag{2.14}$$

where $\text{Dirichlet}_K(\alpha)$ represents a symmetric Dirichlet distribution with hyperparameter α.

2.3.1.2 Inference in LDA

The central computational problem of a probabilistic topic model is that of inferring the values of the hidden variables in order to use the model. This problem is typically addressed by approximating the posterior distribution of the hidden variables given the actual data (i.e., observations, hyperparameters, model constants, etc.). In LDA, this posterior has the form $\text{p}(\beta, \theta, z|w, \alpha)$, where $\beta = \{\beta_1, \ldots, \beta_K\}$, $\theta = \{\theta_d\}_d$, $z = \{z_{d,n}\}_{d,n}$ and $w = \{w_{d,n}\}_{d,n}$. Thus, the topic discovery process in LDA consists in estimating the values of variables

β, θ and z that maximize their joint posterior conditioned on both the word observations and the hyperparameter α.

In PTM, the full inference problem (i.e., the assignment of the most likely values to all latent variables) can be thought of as "reversing" the current generative process, since the generative model leads to a joint posterior probability distribution for variables (θ, z, w) defined as:

$$\mathrm{p}(\theta, z, w|\alpha, \beta) \propto \left(\prod_{d \in D} \mathrm{p}(\theta_d|\alpha) \right) \left(\prod_{\substack{d \in D, \\ 1 \leq n \leq N_d}} \mathrm{p}(z_{d,n}|\theta_d)\, \mathrm{p}(w_{d,n}|\beta, z_{d,n}) \right) \quad (2.15)$$

Techniques such as Gibbs sampling (Geman and Geman, 1984) or variational methods (Blei et al., 2003; Teh et al., 2007) are usually employed to perform the inference. It has been shown that the choice between this two inference methods has negligible effect on the probability of held-out documents or inferred topics (Asuncion et al., 2009).

In the case of LDA, Griffiths and Steyvers (Griffiths and Steyvers, 2004) proposed a (collapsed) Gibbs sampling procedure that considers each word occurrence $w_{d,n}$ observed in the text collection in turn, and estimates the probability of individually assigning topics $\beta_1, \ldots, \beta_K$ to $w_{d,n}$, conditioned on the topic assignment to all other word occurrences $z_{-d,n}$. That is, for each $w_{d,n}$ the value of $\mathrm{p}(z_{d,n} = k|z_{-d,n}, w_{d,n}, \beta, \theta, \alpha)$ is estimated for each $k \in \{1, \ldots, K\}$. From this posterior distribution, a topic is sampled and stored as the topic assignment to word $w_{d,n}$ (i.e., the estimated value of $z_{d,n}$). The posterior distribution of topic assignments is calculated as follows:

$$\mathrm{p}(z_{d,n} = k|z_{-d,n}, w, \beta, \theta, \alpha) \propto \frac{N^{(\theta)}_{-d,k} + \alpha}{\sum_{i=1}^{K} N^{(\theta)}_{-d,i} + \alpha K} \frac{N^{(\beta)}_{-w_{d,n},k} + \mu}{\sum_{w' \in V} N^{(\beta)}_{-w',k} + \mu|V|} \quad (2.16)$$

where V is the vocabulary of the document collection, $z_{-d,n}$ represents all the topic assignment to word occurrences except for $w_{d,n}$, μ is a (constant) parameter suitably introduced to smooth the distribution of words that represent each topic, $N^{(\theta)}_{-d,k}$ is the number of times topic β_k is assigned to a word occurrence in document d (excluding $w_{d,n}$), and $N^{(\beta)}_{-w',k}$ is the number of times that word w' in the collection is assigned to topic β_k (excluding current occurrence $w_{d,n}$ from the count).

This sampling procedure provides direct estimates for θ and β based on

the following equations:

$$\theta_{d,k} = \frac{N_{d,k}^{(\theta)} + \alpha}{\sum_{k'=1}^{K} N_{d,k'}^{(\theta)} + \alpha K} \tag{2.17}$$

$$\beta_{k,w} = \frac{N_{w,k}^{(\beta)} + \mu}{\sum_{w' \in V} N_{w',k}^{(\beta)} + \mu |V|} \tag{2.18}$$

where $\theta_{d,k}$ represents the probability of sampling topic k from distribution θ_d, $\beta_{k,w}$ is the probability of word w in topic β_k, and $N_{d,k}^{(\theta)}$ accounts for the number of times topic β_k is assigned to a word in d, and $N_{w,k}^{(\beta)}$ is the number of times that word w is assigned to topic β_k in the whole collection.

In the Gibbs sampling procedure, the initial state assigned to z should not matter in theory, since the Markov chain eventually converges to the true distribution of the data after many sampling iterations. However, in this thesis we perform "on-line" initialization, a heuristic procedure aimed to speed convergence in inference problems. On-line initialization begins with an empty z, which is increased in each iteration with a sample $z_{d,n}$ according to Equation 2.16. The very first $z_{d,n}$ is based on the hyperparameters only, whereas the final one is a true Gibbs sample conditioned on all other $z_{-d,n}$ (Andrzejewski, 2010).

The hyperparameters in the model can be either empirically set or automatically learned from the data (e.g., by coupling an Expectation Maximization procedure with the Gibbs sampling to optimize them, or by including them in the Gibbs sampling procedure assuming they are generated from certain distributions). In (Griffiths and Steyvers, 2004), it is said that a reasonable starting point is to set $\beta = 0.1$ and $\alpha = 50/K$; where K is the number of topics modeled.

Currently, fully unsupervised approaches to automatically determine the number of topics to model a collection rely on nonparametric Bayesian statistics. For example, Dirichlet processes have been applied in (Teh et al., 2006) to encode uncertainty about the number of topics in the PTM approach known as Hierarchical Dirichlet Process (HDP), since a Dirichlet process is a distribution over multinomial distributions with potentially infinitely many components. Thus, inference under these models automatically sets the number of topics based on the observed data and given hyperparameters.

2.4 Evaluating topic discovery approaches

Traditionally, approaches to the problem of topic discovery have been empirically evaluated in experimental environments using benchmark test collections. Broadly, one or several quality measures are employed to assess the

performance of individual approaches on the benchmark test collection. Then, it follows a comparison between the different approaches in terms of the obtained values for the different measures.

Benchmarking collections in topic discovery are generally provided with a explicit structure that organizes documents into topics. All these collections and the topic structures are typically built in by human experts (often in a semi-automatic manner) using documents extracted from diverse information sources (e.g., newspapers, archives of historical documents, databases of scientific articles, etc.).

The topics in a benchmark test collection are also referred to as either gold-standard topics, manual topics or classes. They are used as the reference from which to compare the topics generated by the individual approaches (also called peer topics).

Some examples of widely used benchmark test collection in topic discovery are the different versions of TDT2 document collection from the tracks of *Topic Detection and Tracking*, and the collections of document retrieval from TREC (*Text REtrieval Conference*).

There are several measures for evaluating the quality of the topics discovered by an approach. However, the specific measures to be used in each case obviously depends on the perspective of the approach since the obtained topics are modeled by different kinds of abstractions. Thus, the measures employed to evaluate clustering-based approaches cannot be directly applied to evaluate PTM-based approaches and *vise versa*.

Next subsections briefly summarize the most widely used measures to evaluate the quality of topic discovery approaches.

2.4.1 Quality measures for clustering-based approaches

The evaluation of clustering-based approaches to topic discovery has been mainly based on extrinsic measures that compare the peer clusters to the gold-standard produced by human annotators.

2.4.1.1 Micro- and macro-F1 measures

Two of the most widely used measures are the measures of micro- and macro-F1 measures. These measures compare the clusters generated by the approaches to the manual topics by combining both precision and recall factors. Whereas micro-averaging gives equal weight to every document, macro-averaging gives equal weight to each topic. Overall, the higher the values of these measures the better the clustering is.

The definition of these measures involves the calculation of the F1 value between each obtained cluster and each manual topic. The F1 value of the ith topic in the gold-standard with respect to the jth obtained cluster is defined as:

$$\begin{aligned} \mathrm{F1}(i,j) &= \frac{2 \cdot Precision(i,j) \cdot Recall(i,j)}{Precision(i,j) + Recall(i,j)} && (2.19) \\ &= \frac{2 \cdot N_{ij}}{N_i + N_j} && (2.20) \end{aligned}$$

where N_{ij} is the number of common members in the ith manual topic and jth cluster, N_i and N_j are the respective cardinalities of ith manual topic and the jth cluster. The measures of recall and precision *Recall* and *Precision* of jth cluster with respect to the ith manual topic are defined as follows:

$$\begin{aligned} Precision(i,j) &= \frac{N_{ij}}{N_j} && (2.21) \\ Recall(i,j) &= \frac{N_{ij}}{N_i} && (2.22) \end{aligned}$$

Then, the macro- and micro-averaged F1 measures are calculated as follows:

$$\begin{aligned} \text{F1-}macro &= \frac{1}{m} \sum_{i=1}^{m} \mathrm{F1}(i, \sigma(i)) && (2.23) \\ \text{F1-}micro &= \frac{2 \cdot microP \cdot microR}{microP + microR} && (2.24) \end{aligned}$$

where m is the number of manual topics, and

$$\begin{aligned} \sigma(i) &= \arg\max_{j} \{\mathrm{F1}(i,j)\} && (2.25) \\ microP &= \frac{1}{m} \sum_{i=1}^{m} Precision(i, \sigma(i)) && (2.26) \\ microR &= \frac{1}{m} \sum_{i=1}^{m} Recall(i, \sigma(i)) && (2.27) \end{aligned}$$

It is worth noting that these measures do not explicitly take into account the number of peer topics generated by an approach. Indeed, these measures fail to provide good quality estimates for approaches that produce massive overlapping clusters. In such cases, the closer the obtained clustering is to the power set of the target document collection, the larger the values of these measures.

Nevertheless, the obtained values for these both measures are a good indicator of the overall quality of the discovered topics (namely, in terms of topic coherence and meaningfulness) in the case that a small or no overlapping is produced. This is because they provide a comparison of the obtained clusters to topics that have been defined by human annotators and so they can be regarded as coherent and meaningful.

2.4.1.2 Mutual Information between sets of clusters

Another measure that has been often employed to compare a peer clustering to a gold-standard is the Mutual Information (MI). Broadly, this measure is define as follows:

$$\mathrm{MI} = \sum_{i,j} \mathrm{p}(i,j) \log \frac{\mathrm{p}(i,j)}{\mathrm{p}_{gold}(i)\,\mathrm{p}_{peer}(j)} \tag{2.28}$$

where $\mathrm{p}(i,j)$, $\mathrm{p}_{gold}(i)$ and $\mathrm{p}_{peer}(j)$ respectively represent estimates of:

- the joint probability between the ith cluster in the gold-standard and the jth peer cluster,
- the marginal probability of the ith cluster in the gold-standard, and
- the marginal probability of the jth peer cluster.

MI measures how much knowing one of these sets of topics reduces the uncertainty about the other; which intuitively measures the amount of information that the peer and the gold-standard share.

The main concern for applying MI to compare two sets of topics represented by document clusters is that it would need "good" estimates for the joint probability distribution between the clusters in both sets. It is worth mentioning that in the case that the topics were represented by probability distributions of words, this joint probability distribution would be easier to estimate (also in a more natural way) than in the case of clusters.

Nevertheless, the measure of MI opens the door to compare topic discovery approaches from different perspectives (e.g., one based in clustering to one based in PTM), whenever we can properly estimate the joint probability distribution between the topics' representations.

Other measures such as the *clustering cohesiveness* have been applied to intrinsically evaluate the quality of the discovered topics under the clustering perspective. However, such measures have not been shown to correlate with human judgments of actual topic coherence and meaningfulness. They simply relies on heuristics about the closeness and separation of intra- and inter-cluster documents.

So far, no methodology nor quality measure has been applied to explicitly evaluate the quality of topic descriptions in the case of clustering-based approaches.

2.4.2 Quality measures for PTM-based approaches

Unlike the approaches based on clustering, the quality of the topics discovered by means of PTM has been mainly evaluated using intrinsic measures; where the averaged value of log-likelihood (that is obtained in the generation of held-aside data) has been perhaps the most widely used evaluation measure.

However, as shown in (Boyd-Graber et al., 2009; Newman et al., 2009; Mimno et al., 2011; Newman et al., 2011) such a measure does not correlate with human judgments of what are actually coherent topics.

2.4.2.1 The UMass measure

The recent work by Mimno et al. (2011) and posteriorly the work by Stevens et al. (2012) have corroborated the use of a new intrinsic measure called UMass to evaluate the coherence of individual topics modeled by means of word distributions. Specifically, (Mimno et al., 2011; Stevens et al., 2012) have shown that the values of UMass measure correlate with human judgments of topic coherence.

The definition of the UMass measure given by Stevens et al. (2012) regards word co-occurrence frequencies and a positive real value ϵ to measure the coherence of a peer topic β_i as follows:

$$\text{UMass}(\beta_i; n) = \sum_{r=1}^{n} \sum_{\substack{l=1 \\ l \neq r}}^{n} \log \frac{S(w_r^{(i)}, w_l^{(i)}) + \epsilon}{S(w_l^{(i)})} \tag{2.29}$$

where $(w_1^{(i)}, .., w_n^{(i)})$ is the list of the n most probable words under topic t_i and $S(w_r^{(i)}, w_l^{(i)})$ is the number of documents in the collection containing both words $w_r^{(i)}$ and $w_l^{(i)}$. Similarly, $S(w_l^{(i)})$ represents the number of documents in which word $w_l^{(i)}$ occurs. The parameter ϵ is employed to penalize the inclusion of words that do not co-occur with other words in the top n. Thus, values of $\epsilon \in (0, 1)$ are used to help distinguishing between topics that are semantically interpretable and topics that are artifacts of statistical inference. The larger the values of UMass(β_i, n) the more coherent the topic β_i is.

The UMass measure significantly computes its counts from the original corpus used to train the topic models, instead of using an external corpus Stevens et al. (2012). So that, it attempts to confirm that the models learned data known to be in the corpus.

The values of UMass provides a topical score indicating to what extent there is a meaning underlying a topic. However, such a score does not offer hints about how much abstract or specific the meaning is, and therefore it cannot be applied to assess the meaningfulness of a topic.

2.4.2.2 Mutual Information between sets of distributions

To our better knowledge, the problem of evaluating the meaningfulness of a topic has not been addressed before (at least in an explicit manner). Nevertheless, the overall quality of the inferred topics can be evaluated by measuring the correspondence between the word distributions that represents the peer topics and the word distributions that correspond to the gold-standard.

Thus, MI can be adopted to measure such a correspondence as follows:

$$\mathrm{MI} = \sum_{i=1}^{K_{peer}} \sum_{j=1}^{K_{gold}} \mathrm{p}(t_i, t_j^*) \cdot \mathrm{PMI}(t_i, t_j^*) \tag{2.30}$$

where:

$$\mathrm{PMI}(t_i, t_j^*) = \log \left(\frac{\mathrm{p}(t_i, t_j^*)}{\mathrm{p}(t_i)\,\mathrm{p}(t_j^*)} \right) \tag{2.31}$$

$$\mathrm{p}(t_i, t_j^*) \propto \frac{1}{K_{gold} K_{peer}} \sum_{w \in \mathcal{V}} \frac{\mathrm{p}(w|t_i)\,\mathrm{p}(w|t_j^*)}{\mathrm{p}(w)} \tag{2.32}$$

$$\tag{2.33}$$

Here K_{gold} is the number of topics in the gold-standard, K_{peer} is the number of topics modeled by the approach under evaluation, and $\{\mathrm{p}(w|t_i)\}_{w \in \mathcal{V}}$ and $\{\mathrm{p}(w|t_j^*)\}_{w \in \mathcal{V}}$ represent the probability distributions of words that define the peer topic t_i and the topic t_j^* in gold-standard, respectively. Since the aim is to obtain a qualitative evaluation, topics can be regarded as equally probable in this comparison (and so, $\mathrm{p}(t_i) = 1/K_{peer}$, $\mathrm{p}(t_j^*) = 1/K_{gold}$).

The distribution of words corresponding to a manual topic can be estimated from the averaged MLE models of its documents.

2.5 Related tasks

The problem of discovering and describing topics can be inscribed in the broad area of Text Mining (Feldman and Dagan, 1995), which refers to the process of extracting interesting and non-trivial information and knowledge from unstructured texts. In particular, it can be contextualized into the set of problems concerning the organization of text collections that aims to obtain a structure in which the different contents that arise from the collection are represented.

In this way, some closely related tasks to the problem of discovering and describing topics are the following:

- **Text classification**. Given a collection of text document D and a predefined set of document classes or categories $\mathbb{C} = \{C_1, ..., C_{|\mathbb{C}|}\}$, this task consists in finding a relation $\mathcal{R} \subseteq D \times \mathbb{C}$ such that each document $d \in D$ be in correspondence with the classes in $\mathbb{C}$ that it is intended to belong to, given a set of class samples from D.

  If topic samples are provided in the topic discovery problem, the problem of classifying collection documents into their respective topics could be seen as a task of text classification.

- **Automatic summarization**. The goal of automatic summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs (Mani, 2001).

  A summary can be generated from a single document (*Single Document Summarization*) or, alternatively, from a collection of documents (*Multi-Document Summarization*).

  The result of combining the descriptions of the topics discovered from a collection of text documents can be seen as a generic multi-document summary built from the documents in the collection.

- **Subtopic retrieval**. In the context of Information Retrieval, the problem of subtopic retrieval has to do with finding documents that cover as many different subtopics of a general topic as possible (Zhai et al., 2003; Zhai and Lafferty, 2006). In this problem a topic is usually formulated in terms of a user query, and the aim is to obtain a ranking of relevant documents for the query in such a way that the top ranked documents cover all possible subtopics (i.e. aspects, interpretations, etc.) underlying the query.

  Subtopic retrieval approaches can be categorized into implicit or explicit approaches (Santos et al., 2010). The implicit approaches assume the similar documents will contain similar subtopics; which leads to redundant information; whereas the explicit ones directly model the subtopics of queries using generally a topic discovery approach, and then search the retrieved documents to maximize the coverage of the subtopics.

2.6 Summary

This chapter has described the main research problem addressed by this thesis together with the main techniques –namely, document clustering and PTM– on which the different existing approaches rely. The main aspects of these techniques have been reviewed (e.g., the issue of document representation in the case of clustering-based approaches, and the problem of inference in PTM), as well as the methodology concerning their evaluation on the problem of topic discovery. Finally, we have outlined some of the most related tasks.

Chapter 3

Related work

The goal of this chapter is to review the current state-of-the-art of our research problem, which mainly includes those topic discovery approaches that simultaneously discover and provide some description of the topics. Overall, this set of approaches subsumes both the set of clustering-based approaches that rely on frequent word-based itemsets and the set of approaches based on PTM.

3.1 Clustering approaches based on frequent word-based itemsets

Unlike traditional clustering-based approaches (that do not center on providing topic descriptions and disregard the notion of topic aboutness to define the topics), the series of works such as FIHC (Fung et al., 2003), CFWS (Li et al., 2008) and the method proposed by Malik and Kender (Malik and Kender, 2006), aim at obtaining simultaneously both the coverage of a topic and its description by means of a new clustering criterion based on the concept of frequent word-based itemsets (e.g., sets of words that co-occur in at least a minimum number of documents in the text collection). In these approaches the topics correspond to either clusters of documents that share frequent word-based itemsets or mixtures of these itemsets if they are similar.

Given a document collection $\mathcal{D}$ and a minimum support threshold μ_0 ($0 \leq \mu_0 \leq 1$), a word-based itemset τ is said to be frequent if the number of documents in $\mathcal{D}$ where τ occurs (e.g., the number of documents that simultaneously contain all the words in τ) is greater or equal than $\mu_0 \cdot |\mathcal{D}|$; being $\mathcal{D}$ the total number of documents in the collection.

The set of all documents in a collection $\mathcal{D}$ where a frequent word-based itemset τ occurs is called support set of τ in $\mathcal{D}$, and will be denoted by $\mathcal{D}|_\tau$ hereafter.

Thus, approaches based on frequent word-based itemsets rely on these

concepts to approach the topics in a collection together with their respective descriptions. The topics are mainly determined by the support sets of the frequent itemsets and the descriptions are given by the own frequent itemsets. A summary of these approaches is presented in the next subsections.

3.1.1 FTC and HFTC

In (Beil et al., 2002), two greedy methods that operate in an iterative manner for clustering documents are presented: the method *FTC* (Frequent Term-based Clustering) that obtains a partition of the target collection, and a hierarchical version of FTC called *HFTC*. These methods rely on the concept of frequent term sets as their frequent itemsets to represent both the document clusters and the clusters' descriptions.

In each iteration, FTC generates a cluster and its description as follows. Firstly, it is obtained the set of all frequent word sets from the target document collection by relying on a minimum support threshold that is given by the user. Each frequent word set determines a candidate cluster and its description represented by the support set and the set of frequent words respectively. Then, the candidate cluster with minimum overlap with respect to the other candidates is selected as the cluster generated by the iteration, and a new iteration is carried out in order to generate new clusters using as target collection the set of documents that have not been included in a cluster yet. The iterative process –and hence, the generation of new clusters– finishes when the target collection is empty.

The following entropy-based measure is employed to measure the overlapping of each cluster G:

$$EO(G) = -\sum_{d \in G} \mathrm{P}(d) \log \mathrm{P}(d) \tag{3.1}$$

where,

$$\mathrm{P}(d) = \frac{1}{|\{\tau \mid \tau \text{ is a frequent word set in } \mathcal{D} \wedge d \in \mathcal{D}|_\tau\}|} \tag{3.2}$$

and $\mathcal{D}$ represents the current target collection of documents for the iteration.

Different from FTC, the hierarchical version HFTC generates an entire level of the hierarchy in each iteration. The first level is obtained by applying FTC to the whole document collection using only frequent word sets of cardinality 1. The clusters in the ith level ($i \geq 2$) are generated from the clusters in level $i-1$ by applying FTC on each individual cluster, using only frequent word sets of cardinality i. HFTC adds to the hierarchy as many levels and clusters as possible.

3.1.2 FIHC

The clustering approach *FIHC* (Frequent Itemset Hierarchical Clustering) (Fung et al., 2003) also obtains a hierarchy of groups determined by the frequent word sets in the collection.

This method relies on the vector space model to represent each document using the TF×IDF weighting scheme. Besides, it employs two user-defined minimum support thresholds: one for mining frequent word sets and the other one for obtaining the most frequent words in a group of documents. The latter sets up a minimum bound on the ratio between the size of the support of a frequent word in a group and the cardinality of the group in order to regard the word among the most frequent ones in the group.

This approach starts by mining the frequent word sets in the target collection. Each frequent word set is employed as a label for the group consisting of its support. Then, the overlapping between groups is eliminated by assigning each document d_j to the group G_i that maximizes the following function:

$$\begin{aligned} Score(G_i \leftarrow d_j) \quad = \quad & [\sum_{\{t\}} \omega(w, d_j) * |\,(G_i|_{\{t\}})\,|] - \\ & - [\sum_{\{w'\}} \omega(w', d_j) * |\,(\mathcal{D}|_{\{w'\}})\,|] \end{aligned} \tag{3.3}$$

where G_i is a document group that contains d_j, $\{w\}$ represents a frequent word set in the collection that is also frequent in G_i, $\{w'\}$ represents a frequent word set in the collection such that word w' is not a frequent one in G_i, and $\omega(w, d_j)$ and $\omega(w', d_j)$ represent the weights of w and w' in d_j respectively.

Afterwards, the groups are organized in a hierarchy by levels, from the deepest one to level 0 containing the root. The root joins all the groups in level 1 together with the singletons that correspond to each document that has not been clustered by means of the frequent word sets mechanism. The root is labeled as $\emptyset$.

The $k$the level ($k \geq 1$) in the hierarchy consists of the groups labeled with frequent word sets of size k. For each group G_i in this level, a parent is selected from among the groups labeled with subsets of the label of G_i having size $k-1$. The selection of the parent follows a similar criterion to that of selecting the group for a document when eliminating the overlapping between groups (see Equation 3.3). In this case, all documents in the sub-hierarchy of G_i are merged together in a single conceptual document and the value of function *Score* for this document is calculated with respect to each possible parent.

Once the hierarchy has been created, its branches are bounded with the aim of obtaining a more "natural" and "accurate" hierarchy to be employed by users for browsing. The branch bounding consists of: (1) removing descendant groups if the parent and descendant are similar enough (only applied to

parent and descendant at level 2 or deeper), and (2) merging of groups level 1 regarding their similarity.

The process of removing descendant groups is carried out in a bottom-up approach; whereas merging groups at level 1 is performed greedily by merging the most similar group pairs each time if its similarity is above a predefined threshold or a predefined number of clusters is obtained.

The similarity between two groups is defined by means of function $Inter\_Sim$ as follows:

$$Inter\_Sim(G_i \leftrightarrow G_j) = [Sim(G_i \leftarrow G_j) * Sim(G_j \leftarrow G_i)]^{\frac{1}{2}} \quad (3.4)$$

$$Sim(G_i \leftarrow G_j) = \frac{Score(G_i \leftarrow doc(G_j))}{\sum_{\{w\}} \omega(w, doc(G_j)) + \sum_{\{w'\}} \omega(w', doc(G_j))} + 1 \quad (3.5)$$

where $doc(G_j)$ represents the conceptual single document obtained by merging all the documents in the sub-hierarchy G_j, $\{w\}$ represents a frequent word set in the collection that is also frequent in G_i, $\{w'\}$ represents a frequent word set in the collection such that t' is not frequent in G_i, and $\omega(w, doc(G_j))$ and $\omega(w', doc(G_j))$ are the weights of w and w' in $doc(G_j)$. The similarity threshold proposed by authors in the algorithm is 1.

3.1.3 TDC

In (Yu et al., 2004), the clustering algorithm *TDC* (Topic Directory Construction) is proposed to generate a topic directory from a collection of documents by relying on the concept of *closed term set*. A closed term set is a frequent frequent word set such that its support set is a strict superset of all of the support sets that correspond to their strict word subsets.

TDC employs the vector space model with TF×IDF weighting scheme to represents the documents, and different from other approaches based on frequent itemsets, TDC does not need a user-defined minimum support threshold to mine the frequent sets. Instead, this threshold is automatically calculated in order to ensure that the union of all support sets of the closed term sets constitutes a cover of the document collection. Thus, all of the documents in the collection are considered in the process of generating the topic directory.

The algorithm starts from an initial set of document groups obtained from the support sets of the closed term sets (each closed term set identifies an labels a group consisting in its support set). Then, the overlapping between the groups is minimized by applying the following branching criteria:

i. *Removing inner term sets*. If multiple nodes in the same path in a directory contain the same documents, to minimize the document redundancy, we only leave the one in the lowest node and remove the others. This is done by removing inner term sets – among frequent closed term sets, the termsets whose superset exists in the same document (Yu et al., 2004).

ii. *Constraining the maximal number of document duplication*. The user is allowed to specify a maximum number of duplication for the documents in the directory. According to that number, each document d is only assigned to those groups whose labels maximize the following function:

$$score(d, \tau) = \sum_{w \in \tau} tf\_idf(w, d) \tag{3.6}$$

where τ represents a closed term set in whose support set d occurs, and $tf\_idf(w, d)$ represents the weight of w in d.

The directory's hierarchy is built by relying on the subsumption relationship between the labels of the groups. Starting from an abstract root that is place at level 0, level 1 is built using the groups labeled with closed term sets of size (i.e., cardinality) 1. These groups are placed as direct descendants of the root. Each group labeled with term set $\tau = \{w_1, \ldots, w_k\}$ ($k \geq 2$) is placed at level k as direct descendant of all groups at level $k-1$ whose labels are subsets of τ.

Finally, in a similar way to FIHC, groups at level 1 are merged according to their similarities to reach a number of groups less or equal than a user-defined threshold. The well-known Jaccard coefficient (i.e., the ratio between the cardinality of the intersection and union of the groups) is used as similarity function between groups.

3.1.4 Method by Malik and Kender

Arguing in favor of the closeness property of term sets from (Yu et al., 2004) and against the use of minimum support thresholds, the use of "closed interesting" itemsets is proposed in (Malik and Kender, 2006) to obtain a hierarchical document clustering. The notion of *close interesting itemset* refers to closed term sets (Yu et al., 2004) that replace the minimum threshold property of closed term sets by a property of high interestingness, which is implemented by putting a threshold on an association measure referred to as *interestingness measure*. Association measures such as *Chi cuadrado*, *Jaccard coefficient*, *Mutual Information* and the correlation coefficient have been used as interestingness measures.

The method by Malik and Kender (Malik and Kender, 2006) operates in a similar way to TDC. The main differences are:

1) Relying on closed interesting itemsets to obtain the initial groups instead of using closed term sets.

2) The criterion of removing inner term sets is redefined as follows:

   - If a document is contained in multiple clusters that are based on itemsets of varying sizes, this document duplication is reduced by pruning the document from all but the clusters based on the largest sized itemsets (Malik and Kender, 2006).

3) When building the hierarchical structure, each group different from the root labeled as $\tau = \{w_1, \ldots, w_k\}$ is assigned with one and only one direct ancestor, which is selected as the group labeled as τ' ($|\tau'| = k - 1$, $\tau' \subset \tau$) that maximizes the association between the partitions $\{\mathcal{D}|_{\tau'}, \mathcal{D} \setminus \mathcal{D}|_{\tau'}\}$ and $\{\mathcal{D}|_{\tau \setminus \tau'}, \mathcal{D} \setminus \mathcal{D}|_{\tau \setminus \tau'}\}$ with respect to the interestingness measure.

3.1.5 STC

The clustering algorithm *STC* (Suffix Tree Clustering) (Zamir and Etzioni, 1998) is a non-hierarchical document clustering algorithm that relies on a suffix tree to discover and describe document clusters from a target collection of text documents.

STC represents each document in the collection as a set of (lemmatized) word sequences, each one representing a sentence in the document.

Basically, STC operates in two phases. Firstly, it obtains a set of initial clusters from the set of all (word-based) substrings of the sentences that are shared by the documents in the collection. Specifically, the substrings are obtained by building a suffix tree from the sentences. Each substring determines an group and its description in such a way that the substring is the description of the group and the documents in the group are those one containing the substring.

In a second phase, similar initial groups are merged together to obtain the document clustering. In (Zamir and Etzioni, 1998), two initial groups G_i and G_j are considered to be similar if both $|G_i \cap G_j|/|G_i| > 0.5$ and $|G_i \cap G_j|/|G_j| > 0.5$. Each document cluster is obtained by merging together the documents belonging to the connected components of the similarity graph determined by the similar initial groups. This graph is an undirected one whose vertices are the initial groups and there is an edge between each pair of similar initial groups.

STC allows to obtain overlapping clusters and can build the document clustering in an incremental manner. However, it is extremely expensive in terms of memory usage since it has to store all substrings of words shared by the documents in the target collection. Therefore, STC can be applied in practice only to small collections comprised of small size documents (e.g., *snippets*).

3.1.6 CFWS and CFWMS

In (Li et al., 2008), the document clustering algorithms *CFWS* (Clustering based on Frequent Word Sequences) and *CFWMS* (Clustering based on Frequent Word Meaning Sequences) have been proposed based on the concepts of *frequent word sequences* and *frequent word meaning sequences* respectively.

Both CFWS and CFWMS simultaneously obtain the document clusters and their respective descriptions by relying on a similar idea to STC. That is, theses algorithms firstly obtain a set of initial clusters by mining the suffix tree in which the document representations have been inserted, and then merge

together similar groups to obtain the final clustering. However, different from STC, CFWS and CFWMS represent each document using a single sequence of frequent words and concepts respectively (arranged in the order they "occur" in the text) instead of a set of sequences with all the lemmatized words from the sentences.

In CFWS and CFWMS, frequent words (concepts) correspond to words (concepts) belonging to frequent sets of cardinality 2 in the collection. Thus, both CFWS and CFWMS obtain an important save of memory usage with respect to STC. [1]

Concepts in CFWMS are based on sysnsets from WordNet (Miller, 1995) that correspond to nouns and verbs in the texts. Specifically, each concept is build as a combination of the two most frequent synsets from each word senses. Thus, despite CFWMS obtains a conceptual description of the obtained document clusters, it can introduce some errors and inconsistencies caused by possible word ambiguities.

3.1.7 Main limitations

One of the claims of these works is that they outperform classical document clustering algorithms such as *Bisecting K-Means* (Steinbach et al., 2000) and *UPGMA* (Jain and Dubes, 1988) at the same time that they provide a description for the clusters relying on word sets. However, several issues still remain open in order to apply such algorithms. For example:

- *Minimum support threshold*. Clustering algorithms based on frequent itemsets need to set up a minimum support for mining frequent word-based itemsets. Determining this value is one of the most critical aspects of all these algorithms. High values for the support threshold produce a handful set of word-based itemsets, but these ones only cover the broadest topics (i.e. many documents will not be assigned to a topic). Instead, low support values produce either a very large set of term sets or a combinatorial explosion, mainly in large and heterogeneous document collections. To alleviate this problem some approaches rely on closed and interesting word-based itemsets to remove those ones that make little contribution. However, from our point of view there is no minimum support threshold able to simultaneously capture the underlying topics in a collection and generate a treatable number of frequent itemsets at the same time. Related to this issue is the fact that in general these algorithms rely on several user-defined parameters whose values are difficult to determine in advance. In addition to the minimum support threshold, these algorithms rely on user-defined parameters such as the

[1] Frequent words and concepts are obtained using a minimum support threshold that authors propose to be in the interval [0.005, 0.15].

overlapping or minimum similarity threshold between groups, a minimum term frequency threshold in a group of documents, lower bounds on interestingness measures, a maximum number of groups in the first level of the hierarchy, etc.

- *Topic coverage*. Overall, the mechanism of building a document cluster using only documents in the support set of a frequent itemset is too constrained. For example, there can be topics in a document collection not covered by a single frequent itemset but by multiple ones in such a way that each frequent itemset is a lexical variation of the other ones. Such topics could not be determined by one but more frequent itemsets. Several approaches attempt to address this issue by joining together similar support sets. On the other hand, it is also possible that the minimum support threshold employed does not entail covering the entire target collection by means of the frequent itemsets. In this case, the actual topics containing documents not included in the support sets could not be fully discovered by means of these clustering criteria.

- *Topic Redundancy*. The collection of support sets corresponding to frequent word-based itemsets usually determines a high overlapping cover of the target text collection. Even when specific strategies are applied to reduce the overlapping between document groups or to decrease the number of support sets to be considered, the number of topics finally produced by these approaches is much greater than the number of actual topics in the collection.

- *Topic meaningfulness*. The selection of a word-based itemset from which a topic is produced in these approaches is mainly based on the number of documents in the collection that simultaneously contain all its words. Thus, despite each topic is built around a set of words which can be though of as the topic meaning, the relative importance of these words in a document is not regarded at all and therefore its meaningfulness is either. Notice that association measures used as interestingness measure in (Malik and Kender, 2006) do not measure meaningfulness but correlation among words. In practice, if we randomly choose a frequent word set based on (frequency- based) language statistics, the chosen word set is more likely to be a frequent domain pattern or a language collocation than a true topic descriptor. For example, in a collection addressing sport topics, possible frequent word sets in a collection like {*sport*, *athlete*} (regardless stopwords) are frequent correlations between possible frequent words in a domain, and they are more likely to be generated (i.e., to be more frequent) than {*Sochi*, *Olympics*}, which would be more likely to be a topic-based co-occurrence.

3.2 Approaches based on PTM

As explained in Chapter 2, topic discovery approaches based on PTM (from which LDA Blei et al. (2003) is the simplest generative approach that falls in this category) simultaneously discover and describe topics by inferring a set of word distributions that jointly model the generation of each individual document in the target text collection; each distribution being the representation and description of a topic.

Since LDA, the class of PTM-based approaches has been growing with LDA extensions aimed at improving the quality of the discovered topics. In this section, we describe a set of these extensions that are closely related to the approach proposed in this thesis. To sum up, these extensions focus on:

- Using non-parametric Bayesian statistic to automatically infer the number of topics (Teh et al., 2006).
- Using asymmetric Dirichlet priors to model the topic proportions that generate each document (Wallach et al., 2009).
- Using regularization factors to improve topic coherence Newman et al. (2011).

3.2.1 Hierarchical Dirichlet Processes

Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006) assumes that each document is modeled as a Dirichlet Process Mixture Model (DPMM) (Antoniak et al., 1974); i.e., as a mixture model with an infinite number of distributions representing the topics, though only finitely distributions are used (i.e., the topics).

3.2.1.1 The generative process in HDP

The DPMM aimed to generate a document d supposes that each word $w_{d,n}$ in d arises as follows:

$$G_d \sim \mathrm{DP}(G_0, \alpha) \tag{3.7}$$

$$\phi_{d,n}|G_d \sim G_d \tag{3.8}$$

$$w_{d,n}|\phi_{d,n} \sim F(\phi_{d,n}) \tag{3.9}$$

where $F(\phi_{d,n})$ denotes the topic employed to generate word $w_{d,n}$. The factors $\{\phi_{d,n}\}_{n\in\{1,\ldots,N\}}$ are conditionally independent given G_d (factors belong to an infinite set of factors Φ such that for each factor $\phi \in \Phi$ there is a topic $F(\phi)$), and the observation $w_{d,n}$ is conditionally independent on the other observations given factor $\phi_{d,n}$. Finally, G_d is distributed according to a Dirichlet process (Ferguson, 1973) with concentration parameter α and base distribution G_0 (G_0 is a distribution over factor in Φ).

One way to characterize a distribution drawn from a Dirichlet process with concentration parameter α and base distribution G_0 is by means of a modified urn model in which initially there are α balls labeled with factor ϕ_0 ($\phi_0 \notin \Phi$). Each time, one ball is drawn randomly from the urn and its label is inspected. If the label is in the set of factors Φ, the ball is placed back in the urn together with an additional ball with the same label. Otherwise, the ball is placed back in the urn together with a new ball labeled with a new factor ϕ from Φ that is randomly drawn according to base distribution G_0.

Thus, a drawn G from the Dirichlet process $\mathrm{DP}(G_0, \alpha)$ is a distribution of factors that correspond to the distribution of labels from Φ in the urn.

Figure 3.1 graphically depicts both the DPMM aimed at generating a single document (left) and the HDP focused on generating a collection of documents (right). In the HDP, the variables in the model are distributed as follows: [2]

$$G_0 \sim \mathrm{DP}(H, \gamma) \tag{3.10}$$
$$G_d \sim \mathrm{DP}(G_0, \alpha) \tag{3.11}$$
$$\phi_{d,n}|G_d \sim G_d \tag{3.12}$$
$$w_{d,n}|\phi_{d,n} \sim F(\phi_{d,n}) \tag{3.13}$$

where γ is a concentration parameter and H is a base distribution for the Dirichlet process G_0 that is employed to generate the distribution of factors for each document.

Typically, H is chosen to be a conjugate prior for the family of distributions $F(.)$; e.g., a Dirichlet distribution.

3.2.1.2 Inference in HDP

Inference in HDP consists in repeatedly performing the following two sampling steps until convergence over a variable state that includes factors and topics associated to word observations:

- For all $w_{d,n}$: If the present value of $\phi_{d,n}$ is associated with no other observation, remove the factor and the associated topic from the state. Draw a new value for $\phi_{d,n}$ from $\phi_{d,n}|\phi_{-d,n}, w_{d,n}$ according to the following posterior:

$$\mathrm{p}(\phi_{d,n} = \phi|\phi_{-d,n}, w_{d,n}) \propto \begin{cases} (n_{-d,n}^{d,\phi} + \alpha m_\phi) f(w_{d,n}|\phi) & \text{if } \phi \text{ is in the state} \\ \alpha\gamma f(w_{d,n}) & \text{otherwise} \end{cases} \tag{3.14}$$

  where $f(w_{d,n}|\phi)$ is the probability of $w_{d,n}$ under topic $F(\phi)$, $f(w_{d,n})$ is the prior of word $w_{d,n}$ ($f(w) = \int f(w|\phi)h(\phi)d\phi$, being $h(\phi)$ the density

[2] The term hierarchical refers to the hierarchy of variables distributed as a DP in the generative process.

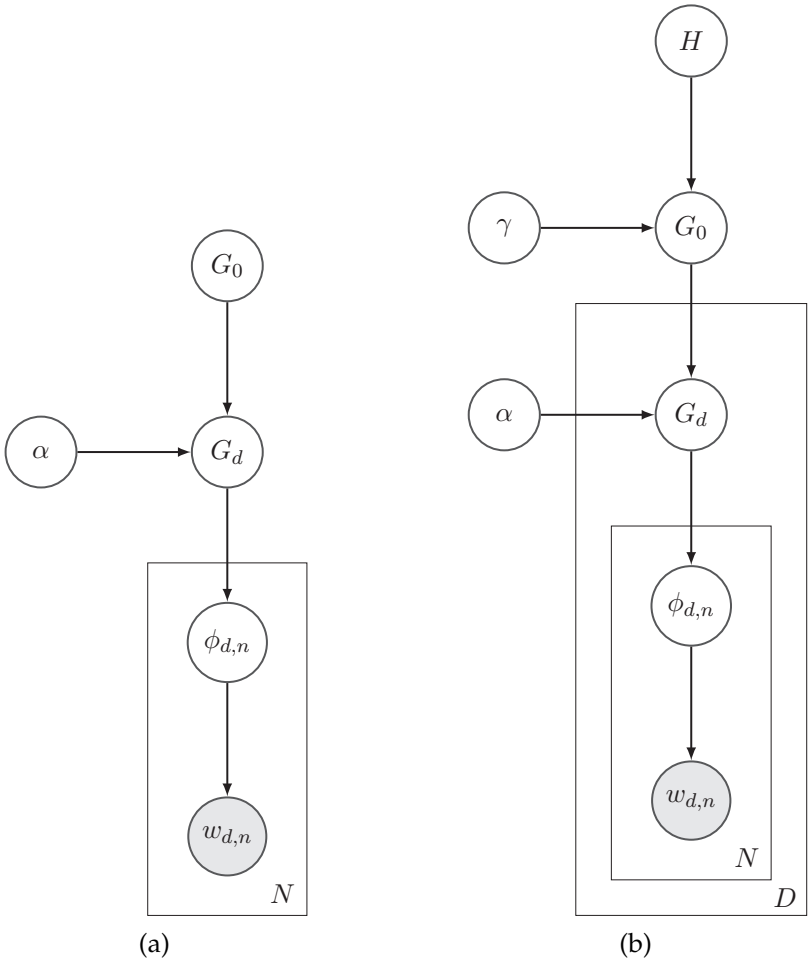


Figure 3.1: (a) A DPMM modeling the generation of a document. (b) A HDP modeling the generation of a document collection (Teh et al., 2006).

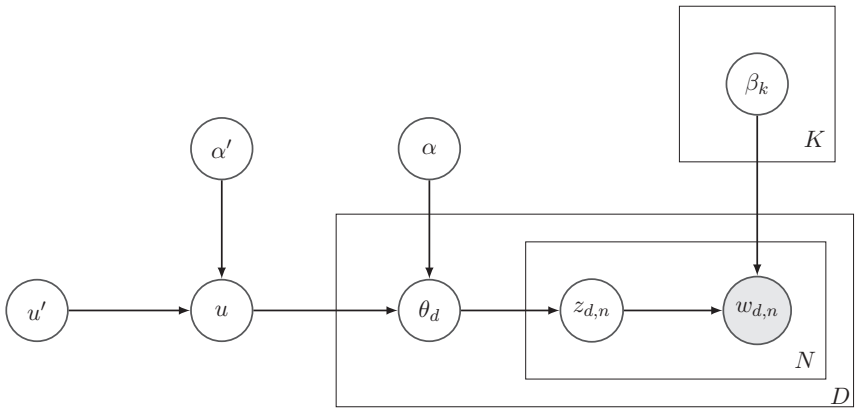


Figure 3.2: LDA with symmetric priors over $\Theta = \{\theta_d\}_{d \in D}$ representing the model proposed by Wallach et al. (2009).

of H), m_ϕ is the number of times factor ϕ has been selected as a new factor in all G_d for an observation excluding $w_{d,n}$ in the actual factor assignment, and $n^{d,\phi}_{-d,n}$ similarly accounts for the number of times ϕ has been selected as a particular previously used value for an observation in d excluding $w_{d,n}$ in the actual factor assignment. If the new $\phi_{d,n}$ is not associated with any other observation, draw a new topic $F(\phi_{d,n})$ from H and add both the new factor and the topic to the state.

- For all ϕ in the set of factors in the state: Draw a new value $F(\phi)$ from $F(\phi)|$ prior H and all $w_{d,n}$ for which $\phi_{d,n} = \phi$.

After performing inference, the finite set of topics in the state corresponds to the set of topics discovered by the HDP approach. Thus, the number of topics in the collection is automatically inferred instead of being prescribed by the users.

Additionally, some implementations include a third sampling step in which hyperparameters are learned as samples from predetermined distributions.

3.2.2 LDA with asymmetric priors

In (Wallach et al., 2009), authors study the performance of using asymmetric Dirichlet priors over the document-topic distributions in LDA. They empirically find that an asymmetric Dirichlet prior has substantial advantages over a symmetric one in the generation of held-out documents.

3.2.2.1 The generative process

Figure 3.2 represents the extended generative model of LDA that uses an asymmetric prior αm over the document-topic distributions as in (Wallach et al., 2009). In this extended model, variables m and θ_d are distributed as follows:

$$u \sim \text{Dirichlet}(\alpha' u'_1, \ldots, \alpha' u'_K) \tag{3.15}$$

$$\theta_d \sim \text{Dirichlet}(\alpha u_1, \ldots, \alpha u_K) \tag{3.16}$$

where $u' = \langle u'_1, \ldots, u'_K \rangle$ is a constant vector with $u'_i = 1/K$ for all $i \in \{1, \ldots, K\}$.

In this case, one way to approach the topic proportions in θ_d is by means of a urn model with two urns u and u' that operates as follows. Initially, there are α and α' balls in u and u' respectively, each one labeled with category c_0.

Each time, one ball is randomly drawn from u and its label is inspected. If the label is in the set $\{c_1, \ldots, c_K\}$, the ball is placed back in the urn u together with an additional ball with the same label. Otherwise (i.e., if the label is c_0), the ball is placed back in the urn u, and a new ball is randomly drawn from u'. In this case, if the label of the ball is in the set $\{c_1, \ldots, c_K\}$ the ball is placed in u', and two new balls are included in the urns (one in u and the other one in u'). The two new balls are labeled with the label of the drawn ball.

In the case that the ball drawn from u' is labeled with c_0, the ball is placed back in the urn u', and again two new balls (one in u and the other in u') are included the urns. The two balls are labeled with label c randomly chosen from the set $\{c_1, \ldots, c_K\}$.

Topic proportions are distributed in the same way that balls with labels in $\{c_1, \ldots, c_K\}$ are distributed in urn u.

3.2.2.2 Inference

In (Wallach et al., 2009), inference is performed by means of a Gibbs sampling procedure in which the sampling path of each topic assignment is maintained; that is, it is known if the topic drawn comes from a draw from u or from a draw from u'.

Overall, sampling a topic assignment for a word occurrence $w_{d,n}$ is based on the posterior distribution defined as follows:

$$\begin{aligned} \mathrm{p}(z_{d,n} = j | z_{-d,n}, w_{d,n}, \beta, \theta\alpha, \alpha', u') \propto \\ \frac{N^{(\theta)}_{-d,j} + \alpha \frac{\hat{N}_j + \frac{\alpha'}{K}}{\sum_{i=1}^{K} \hat{N}_i + \alpha'}}{\sum_{i=1}^{K} N^{(\theta)}_{-d,i} + \alpha} \frac{N^{(\beta)}_{-w_{d,n},j} + \frac{\mu}{|V|}}{\sum_{w' \in V} N^{(\beta)}_{-w',j} + \mu} \end{aligned} \tag{3.17}$$

where $\hat{N}_k$ is the number of observations (different from $w_{d,n}$) that has been assigned with topic j and this topic is sampled from u'. The rest of variables in the form are defined as in Equation 2.16.

3.2.3 Conv-Reg and Quad-Reg

Based on the method GPUM-LDA Mimno et al. (2011), which relies on MI scores to perform a kind of regularization in the definitions of the topics during inference, the approaches Conv-Reg and Quad-Reg Newman et al. (2011) formally add a regularization factor to the target function that seeks to maximize the likelihood of the latent variables in LDA (in the case of Quad-Reg it is a quadratic term that is added, whereas Conv-Reg define each topic from a convolution to include a spread of related words). The aim is to enhance the coherence of LDA topics by means of a MI-based regularization, so that the burstiness of a word in a topic entails the burstiness of its related words via MI.

The main concern with this approach is that frequent words usually have frequent co-occurrences with many other words, and therefore it does not produce coherent enough topics. Frequent words are ranked top in many topics.

To alleviate this problem, the regularized topic models in Newman et al. (2011) rely on more sophisticated correlation matrices between words to set up their models. However, these correlation matrices are expected to be built from external knowledge, so that these methods can only be useful to discover topics from collections of specific, well-characterized domains.

Conv-Reg has been shown to outperform Quad-reg in terms of topic coherence Newman et al. (2011).

3.2.4 Topic Signature Language Models

Topic Signature Language Models (TSLM) have been recently introduced in (Zhou et al., 2007) to provide an internal representation of documents in terms of a word distribution to be applied to ad-hoc document retrieval.

Given a set of topic signatures $\{t_1, \ldots, t_K\}$ from a document collection $\mathcal{D}$, where topic signatures correspond to frequently occurring word sets in $\mathcal{D}$ or, alternatively, frequent concepts drawn from an existing ontology, TSLM manages to represent a document $d \in \mathcal{D}$ as the distribution of words:

$$\mathrm{p}_t(w|d) = \sum_{k=1}^{K} \mathrm{p}(w|t_k)\, \mathrm{p}_{mle}(t_k|d) \tag{3.18}$$

where $\mathrm{p}_{mle}(t_k|d)$ represents the likelihood of generating topic signature t_k from document d, which is estimated as follows:

$$\mathrm{p}_{mle}(t_k|d) = \frac{c(t_k, d)}{\sum_{i=1}^{K} c(t_i, d)} \tag{3.19}$$

being $c(t_i, d)$ the frequency of occurrence of topic signature t_i in d.

The topic signature model $\{\mathrm{p}(w|t_k)\}$ is estimated from the set of documents containing t_k, by assuming that words in this set are generated by a

mixture model that interpolates the topic signature model with a background collection model as follows:

$$\mathrm{p}(w|t_k, \mathcal{D}) = (1-\alpha)\,\mathrm{p}(w|t_k) + \alpha\,\mathrm{p}(w|\mathcal{D}) \tag{3.20}$$

Here, the coefficient α is accounting for the background noise and is set to 0.5 in (Zhou et al., 2007). Under this mixture language model, the log-likelihood of generating the document set D_k containing topic signature t_k is:

$$\log \mathrm{p}(D_k|t_k, \mathcal{D}) = \sum_w c(w, D_k) \log \mathrm{p}(w|t_k, \mathcal{D}) \tag{3.21}$$

where, $c(w, D_k)$ is the frequency of word w in D_k. Thus, the topic signature language model for t_k is estimated by means of an Expectation-Maximization algorithm with the following update formulas:

$$Z(w) = \frac{(1-\alpha)\,\mathrm{p}(w|t_k)}{(1-\alpha)\,\mathrm{p}(w|t_k) + \alpha\,\mathrm{p}(w|\mathcal{D})} \tag{3.22}$$

$$\mathrm{p}(w|t_k) = \frac{c(w, D_k)Z(w)}{\sum_{w'} c(w', D_k)Z(w')} \tag{3.23}$$

The main concern with this approach in order to be applied to learn coherent and meaningful topics is that, despite each topic signature language model is contextually learned from a topic signature, topics signatures correspond to frequent patterns observed in the context of the document collection, which hardly correlate with the actual topics underlying the collection (Anaya-Sánchez et al., 2008; Anaya-Sánchez et al., 2010).

3.2.5 Main limitations

Overall, topic discovery approaches based on PTM do not guarantee to obtain high quality topics in terms of topic coherence and meaningfulness.

On the one hand, except in the case of TSLM, each topic learned by these approaches corresponds to a pure latent word distribution that can be hardly associated to a short topic description or summary from which to display a topic as discussed in (Blei, 2012).

On the other hand, the proposal of TSLM builds a similar scenario for discovering and describing topics to that set up by the clustering approaches based on frequent word-based itemsets, in which many topic quality issues (mainly, topic meaningfulness concerns) are still open.

So far, from our point of view the attempts to obtain true better quality topics are mainly focused on applying correlation measures between words such as MI-based coefficients to improve topic coherence. These methods has been centered on regularizing topic definitions (see Section 3.2.3); which is not enough to assess topic meaningfulness in arbitrary document collections, and not enough to obtain completely coherent topics neither.

Both the problem of automatically determining the number of topics in a collection and the one of minimizing the redundancy between the generated topics have received relatively little attention in the perspective of PTM.

3.3 Conclusions

This chapter has reviewed the class of approaches that simultaneously discover and describe topics, which is the class that better fits in with our research problem of discovering and describing high quality topics from a target text collection.

This class of approaches mainly consists of two major sub-classes of approaches: the clustering approaches based on frequent (word-based) itemsets and the approaches based on PTM.

In each case, the reviewed approaches have attempted to step forward in the topic discovery performance. Nevertheless, the quality of the discovered topics by these approaches is still far from satisfying the quality requirements of coherence and meaningfulness as stated in this thesis. This is mainly due to:

- *Topic redundancy*: this quality property of topics has not been well addressed in both kinds of approaches. Topic redundancy has not been systematically regarded in the PTM-based approaches. In the case of clustering-based approaches (that often produce "comprehensive" topic hierarchies with high overlapping between groups), the attempts to reduce redundancy have been carried out mainly in an ad-hoc manner during late stages of the approaches, such as in post processing steps by merging similar clusters.

- *Topic meaningfulness*: Addressing topic coherence (i.e., ensuring to obtain interpretable topics) is not enough to produce an effective set of topics for end-users. Topic meaningfulness is still demanded. However, his quality property has not been fully taken into account in none of the approaches. To our better knowledge, no mechanism has been previously implemented to avoid obtaining too abstract topics. In the case of the clustering-based approaches, the similarity-based mechanisms already used to assess topic cohesion can be hardly applied to arbitrary document collections (at least in a direct manner) in order to avoid obtaining abstract topics. This is because of the possible variability on the granularity (i.e., broadness or coverage) of the actual topics.

- *Quality of generated descriptions*. The quality of the generated topic descriptions is another issue of existing approaches. In both the clustering-based approaches based on frequent word sets and the PTM-based approaches, the quality of the topic descriptions directly depends on the

quality of discovered topics and vice versa. The methods that employs the most frequent words to produce the topic descriptions (which are the majority in the approaches based on clustering) cause the generation of vague and irrelevant descriptions with a very low discriminative power for the users.

Hence, we can conclude that there is still enough room to develop new methodologies focused on discovering and describing high quality topics.

Chapter 4

An abstract framework to discover and describe topics

This chapter introduces a general framework for discovering and describing topics with two important properties not simultaneously addressed before in a completely unsupervised manner; namely, coherence and meaningfulness.

The framework is entirely abstract and relies on the concept of lexical signatures, which broadly refers to sets of lexically related words. These signatures are intended to represent the aboutness of the topics; that is, the basic elements from which each topic can be accurately discovered and described.

The aim of the framework is to provide a general and open enough methodology to discover high quality user-interesting topics from a concrete, yet arbitrary, definition of lexical signatures.

The framework does not constrain all of the topic discovering approaches derived from it to be exclusively included in one of the pre-existing topic discovery perspectives (i.e., clustering vs PTM-based approaches); neither necessarily entails a strict partition of the documents regarding their topic coverage, though no topic hierarchy is explicitly constructed.

4.1 The proposed framework

In line with the hypotheses of this work, the proposed framework mainly relies on the following abstract components:

C1) a concrete representation and implementation of lexical signatures, which will represent the possible aboutness of the topics (see Section 2.1) and hence they will determine the topics and their respective descriptions,

C2) a method to discover/learn the possible topic underlying a lexical signature in the context of the target collection of texts,

C3) a method to assess topic meaningfulness, which should assign each lexical signature with a score of meaningfulness, useful to decide whether the underlying topic is not too abstract nor specific, and also

C4) a method for generating a topic description; which can be properly defined by default as the lexical signature, or can be learned as an enhanced version of the signature from either the own signature or the underlying (discovered or learned) topic in the context of the text collection.

From these components, we formulate the framework in terms of an iterative search in which each iteration is focused on discovering a new coherent and meaningful topic from the target document collection as follows:

- Firstly, a (finite) set of lexical signatures S is chosen to find topic diversity regarding the set of previously discovered topics; where the lexical signatures in S follow C1.

- Then, the meaningfulness of the topic underlying each lexical signature in S is assessed by assigning a scoring of meaningfulness to the signature regarding the abstract component C3 (the aim is to determine if a lexical signature is the aboutness of a meaningful topic).

- A filtering process based on the meaningfulness score is then applied to filter out non-meaningful topics (i.e., disregard the lexical signatures that are not deemed to represent meaningful topics).

- Finally, a new topic is obtained from each remaining lexical signature by applying component C2 to discover/learn the topic underlying the signature; and also is so its description by means of C4.

At the end of each iteration, the new topics are stored together with their respective descriptions. Then, if all the documents in the collection are covered by the discovered topics, the iterative search is stopped and the set of stored topics (and their respective descriptions) is returned. Otherwise, the topic discovery search proceed with a new iteration in order to find new topics. The general steps of the proposed methodology are shown in Algorithm 1.

In the algorithm, functions *topic-definition*, *topic-meaningfulness* and *topic-description* respectively represent the abstract components C2, C3 and C4 from the framework; whereas *sample-lexical-signatures* and *stop-discovering-criterion* are other abstract components in the framework that respectively represent:

- A mechanism to generate a set of lexical signatures focused on discovering new topics from the target text collection by regarding the set of discovered topics in previous iterations. In the algorithm, $\bar{T}$ denotes an abstract data element representing the complement of the discovered topics. The signatures in S are assumed to follow C1.

Algorithm 1 A general framework for discovering and describing coherent and meaningful topics from a target collection of texts.

Entrada: A target collection of texts $\mathcal{D} = \{d_1, .., d_N\}$.

Salida: The set of pairs $T = \{(T_i, \delta_i)\}_i$ containing the topics and their respective descriptions.

1: $T \leftarrow \emptyset$
2: **repeat**
3: $S \leftarrow$ *sample-lexical-signatures*$(\bar{T})$
4: Let $\omega_i \leftarrow$ *topic-meaningfulness*(s_i) for all $s_i \in S$.
5: $S' \leftarrow \{s_i | s_i \in S \wedge \omega_i > \omega_0\}$
6: $T \leftarrow T \cup \{(T_i, \delta_i) | s_i \in S' \wedge T_i =$ *topic-definition*$(s_i) \wedge \delta_i =$ *topic-description*$(s_i)\}$
7: **until** *stop-discovering-criterion*

- The general condition that is satisfied when all of the documents in the collection are covered by at least one of the topics discovered, or no lexical signatures can be obtained for generating new topics.

The value ω_0 is intended to express a lower bound on the meaningfulness score that performs like a decision boundary for filtering meaningful topics. As it will be shown later, this element can be defined as a global constant value or can be contextually determined by the elements in S.

By specifying different definitions for all of the above abstract components, different approaches can be derived for discovering and describing topics from this abstract framework as performed in the next chapters.

4.2 Specifying related approaches

The proposed framework is general enough so as to be employed to specify the state-of-the-art approaches reviewed in Chapter 3 as instances of the framework. By doing this, we will be able to further analyze the framework components and decide where to put the focus of our research in order to obtain significant improvements over the state-of-the-art related approaches.

To sum up, the related approaches can be broadly specified as follows:

- *Clustering approaches based on frequent itemsets*: Overall, these approaches can be expressed as framework instances by defining lexical signatures to be the corresponding frequent word-based itemsets mined from the entire document collection in each case, and also by defining function *sample-lexical-signatures* to return all these frequent itemsets at the same time. In the case of methods that do not rely on interestingness measures (e.g., FTC, HFTC and FIHC), the meaningfulness scoring function *topic-meaningfulness* can be defined as a constant, so that every lexical signature generated is considered to be the aboutness of a meaningful topic.

In the case of other methods, this function can be fulfilled with the corresponding interestingness measure to only regard the interesting itemsets as meaningful enough to determine the topics. Only one framework iteration is needed to specify all these approaches; thus, *stop-discovering-criterion* can be defined as propositional constant '*True*'. Function *topic-definition* can be defined from the support sets of the lexical signatures according to the topic definition of each specific approach.

- *PTM-based approaches*: Traditional PTM-based approaches can be expressed as framework instances in a similar way to that of clustering-based approaches. The main difference stems from the use of multisets of words to define the lexical signatures. In this case, lexical signatures can be defined so that there is exactly one lexical signature for each topic, defined as the multiset of word occurrences that have been labeled with the topic at inference time. The topic definition from each lexical signature is straightforward from the way in which the specific PTM-based approach define its topics from the labeled word occurrences (see for example Equation 2.18). The main issue here is that these lexical signatures do not necessarily correspond to actual sets of lexically related words.

4.2.1 Main observations

Several observations can be made from specifying existing approaches as framework instances:

- *Lexical signatures and topic coherence*: Regarding the clustering proposals based on frequent itemsets, we support the idea that lexical signatures should not be directly based on frequent word sets from the target text collection. Not all of the actual topics in a text collection (mainly, the smallest ones) can be covered by these kinds of lexically related words.

  Neither should we directly base topic definitions on the support sets of lexical signatures, whatever these signatures may be.

  To directly base lexical signatures on labeled words from latent topics is actually not a good idea. These multisets do not correspond to actual sets of lexically related words, and therefore the coherence of the topics may be compromised.

- *Topic meaningfulness*: We cannot directly base the meaningfulness score of a lexical signature on word correlations. It might lead to obtain false-positive meaningfulness assessments such as validating lexical signatures that cover very specific events (mainly, in case of target collections comprising very separated topics) or too abstract signatures entailing several concrete topics in case of domain-specific document collections. Topic meaningfulness assessments have not been implemented beyond

correlation measures, so there is enough room to explore different alternatives despite the complexity of the task.

- *Topic redundancy*: The above specification of existing approaches in terms of the framework components is based on a single iteration to find the topics, and the issue of obtaining redundant topics (that is not systematically addressed by the approaches) cannot be addressed by the iterative search mechanism aimed at finding topic diversity. This makes extremely important to accurately define function *sample-lexical-signatures*. The aim should be to attempt discovering just one meaningful topic in each iteration to minimize the risk of discovering redundant topics.

4.3 Framework evaluation

Since the proposed framework is an abstract one, we consider evaluating it by means of performing empirical evaluations (see Section 2.4) on concrete approaches derived from the framework as instances.

Specifically, we consider to evaluate the performance of the different approaches on different benchmark text collections, namely:

- TDT2 English corpus (version 4.0) of news stories from the TDT research campaign.
- AFP Spanish collection of news stories from TREC. [1]
- The collection of tweets (RL-MA) about entities in the domain of MUSIC/ARTISTS from the training set of RepLab 2013 evaluation campaign on Online Reputation Management. [2]
- The collection of tweets (RL-CARS) about entities in the automotive domain from the training set of RepLab 2013 evaluation campaign.

These collections correspond to different document registers, and they have been manually labeled with topical information by human annotators at the document level. A description of these collections is shown in Table 4.1.

In the case of TDT2 and AFP, topics correspond to the main events addressed by news stories; whereas in the case of RL-M/A and RL-CARS topics correspond to opinion aspects submitted by users in the form of tweets about different entities in a domain.

The aim of using these document collections is to assess the performance of the derived methods on very different input collections. It is worth mentioning that in addition to the differences in the type of documents between the collections of news and the collections of tweets, the topics based on tweets

[1] http://trec.nist.gov

[2] http://www.limosine-project.eu/events/replab2013

Table 4.1: Description of the benchmark text collections that will be used in the evaluation of the framework instances.

Feature	TDT2	AFP	RL-M/A	RL-CARS
Number of documents	8042	2384	11956	11121
Vocabulary size	38847	27549	20895	29620
Number of topics	96	25	1018[3]	1237[3]
Document register	news	news	tweets	tweets
Source language	English	Spanish	English	English

[3]One of the topics labeled as "Other topics" groups together a broad set of tweets that do not belong to the rest of the topics.

have a significant vocabulary overlapping between each other because they correspond to opinion aspects of entities in the same domain. Thus, to some extent these topics can be considered as subtopics of a broad topic that involve opinions about a certain type of entities.

These collections were preprocessed to consider only word lemmas as the documents' features. Stopwords were removed in the case of news stories, but they were kept in the case of the collections based on tweets because in many cases they corresponded to meaningful words in the message being transmitted.

4.3.1 Experimental targets

In first place, our goal is to validate the reliability of the instances of the main abstract component in each derived approach; that is, to check for the effectiveness of the components in the solution.

We will be interested in evaluating the performance of the approaches in terms of the coherence and meaningfulness of the discovered topics. To do so, we will consider the assumption:

> *The closer the obtained set of topics to the gold standard, the more coherent and meaningful the individual topics are, and vice versa.*

The aim is to assess topic meaningfulness in absence of intrinsic evaluation metrics for this quality property.

Finally, we will be focused on comparing the performance of our proposals to those of the main state-of-the-art methods reviewed in Chapter 3.

4.4 Conclusions

In this chapter, a new abstract framework for discovering and describing topics have been introduced. The framework has been devised as an iterative

topic search from a set of abstract components that seek to verify the main hypotheses of this work. The aim is to derive concrete approaches from the proposed framework to successfully address the problem of discovering and describing coherent and meaningful topics.

The proposed framework is general enough so as to be used to specify related approaches from the state-of-the-art as instances of the framework. This allows to contextualize the framework with these approaches and, in turn, to know where to put the focus of our research in order to obtain significant improvements.

To evaluate the proposed abstract framework we will consider to carry out empirical evaluations on concrete framework instances that will be mainly based on the comparison of the obtained topics to gold-standard produced by human annotators.

The evaluation will be based on very different benchmark text collections (in terms of document register and vocabulary overlapping between the topics).

Chapter 5

A clustering-based framework instance

5.1 Introduction

In this chapter, we introduce a new clustering algorithm that is derived from the framework proposed in the previous chapter. The concrete hypothesis is that high quality topics (i.e., the coherent and meaningful ones) can be identified from highly probable word pairs that co-occur across the documents in the target collection and that are also likely to represent homogeneous (i.e., non-abstract) contents.

The method assumes that no prior knowledge about the collection exists, and therefore no training samples are available to supervise neither the discovery nor the description processes.

The approach is an extension of the methods proposed in (Anaya-Sánchez et al., 2008) and (Anaya-Sánchez et al., 2010) Firstly, based on the general framework, we provide a new formalization of the main concepts on which the approach relies. Secondly, a comprehensive set of experiments is carried out on benchmark text collections of different document register. That is, in addition to collections of news stories we apply our approach to discover opinion topics from collections of tweets.

Despite the method is based on word pairs, it is able to provide larger topic descriptions than just a pair of words.

The method represents the documents in the VSM (see Section 2.2.1). However, it avoids similarity threshold tuning by automatically estimating a similarity value from the collection, which produces near-optimal results.

5.2 Overview and notation

Given a document collection $\mathcal{D} = \{d_1, \ldots, d_N\}$, the proposed method aims to obtain a clustering $\mathcal{G} = \{(G_1, \delta_1), \ldots, (G_K, \delta_K)\}$; where, for all $1 \leq i \leq K$, each cluster G_i represents a topic in the collection ($G_i \subseteq \mathcal{D}$), δ_i being its description.

Assuming that each topic can be represented by a pair of words from the vocabulary of the collection, in this clustering instance we implement the concept of *lexical signatures* (i.e., the abstraction that represents the aboutness of each topic) by means of word pairs. Then, the approach relies on a probabilistic model of word pairs from the collection to guide the search for a "good" partition of the data in terms of coherence and meaningfulness as follows.

Starting from the most probable word pair in the collection, a homogeneity criterion is applied to the pair in order to test whether the content underlying the meaning of the pair in the collection is meaningful or not (see Section 5.4). If the homogeneity criterion holds, a coherent cluster consisting of the set of relevant documents for the content represented by the pair is created (see Section 5.5). Otherwise, the pair is discarded. Thus, word pairs representing abstract contents are disregarded for representing a topic.

Once a cluster has been built, its documents are removed from the collection. Then, this process is repeated again (regarding only the remaining documents and the pairs not evaluated yet) until either the set of remaining documents is empty or no more relevant pairs can be found. Finally, if there are documents not clustered yet, a singleton is created for each one considering its most probable word pair as its description.

Next sections are respectively focused on giving details about:

- the probabilistic model of word pairs,
- the homogeneity criterion employed to assess the meaningfulness of a word pair to define a topic, and
- the definition of coherent topics

Then, Section 5.6 focuses on specifying how the different framework components defined in Chapter 4 are instantiated to derive the entire methodology.

To illustrate the concepts introduced in the next sections we mainly rely on the following example of target text collection.

Example Collection. Consider the text collection composed of the seven documents shown in Table 5.1. These documents have been built from TDT2 collection, and their document names (specifically, the prefix before the period) indicate their topics. For each term, the table includes the number of its occurrences in the documents. The last row in the table indicates the total number of terms in each document.

Table 5.2 summarizes the notation adopted in this chapter.

Table 5.1: Example document set.

Term	**T20011.1**	**T20002.2**	**T20096.3**	**T20001.4**	**T20001.5**	**T20001.6**	**T20001.7**
clinton	4	2	4	-	-	-	-
president	2	4	2	-	-	-	-
state	2	-	-	-	-	-	-
nation	2	-	-	-	-	-	-
white_house	2	-	-	-	-	-	-
address	2	-	-	-	-	-	-
monica_lewinsky	-	2	-	-	-	-	-
affair	-	2	-	-	-	-	-
sexual	-	2	-	-	-	-	-
scandal	-	3	-	-	-	-	-
china	-	-	2	-	-	-	-
beijing	-	-	2	-	-	-	-
tiannamen	-	-	2	-	-	-	-
jiang_zemin	-	-	2	-	-	-	-
crisis	-	-	-	4	2	4	2
asia	-	-	-	4	4	2	-
market	-	-	-	2	-	2	-
⋮							
Total	100	92	90	138	156	120	100

Table 5.2: Main notation used in the proposed clustering-based methodology to discover and describe coherent and meaningful topics.

Variable	Description
$\mathcal{D} = \{d_1, \ldots, d_N\}$	Set of documents defining the *target document collection*.
$\mathcal{G} = \{(G_1, \delta_1), \ldots, (G_K, \delta_K)\}$	The document clustering that we seek to find; where each G_i is a cluster of documents that represents a topic in the collection ($G_i \subseteq \mathcal{D}$), δ_i being its description.
$\mathcal{V} = \{w_1, .., w_{\lvert\mathcal{V}\rvert}\}$	*Vocabulary* of the target document collection.
$\mathcal{P}$	The set of all term pairs that co-occur in at least one document in $\mathcal{D}$. This set represents the set of all possible lexical signatures.
π, $\{w_i, w_j\}$	An arbitrary term pair representing a lexical signature from $\mathcal{P}$.
$\mathrm{p}(\pi\lvert\mathcal{D})$, $\mathrm{p}(\{w_i, w_j\}\lvert\mathcal{D})$	The probability of a word pair that is used to model the probabilistic model of term pairs.
$\mathcal{D}\lvert_\pi$	The support set of a word pair in the collection $\mathcal{D}$; that it, the set of documents from $\mathcal{D}$ that simultaneously contain the words in π.

Table 5.3: Top five most probable term pairs.

Term pair	**Probability proportional to**
$\{clinton,\ president\}$	0.000390
$\{asia,\ crisis\}$	0.000246
$\{president,\ scandal\}$	0.000202
$\{beijing,\ clinton\}$	0.000141
$\{clinton,\ jiang\_zeming\}$	0.000141

5.3 A probabilistic model of word pairs

Let $\mathcal{P}$ be the set of all word pairs that co-occur in at least one document in the collection $\mathcal{D}$. For a given $\pi \in \mathcal{P}$, we denote by $\mathcal{D}|_\pi$ the *support set* of π in $\mathcal{D}$, i.e. the set of documents in $\mathcal{D}$ that simultaneously contain both words in π.

We define the probability of a word pair $\{w_i, w_j\} \in \mathcal{P}$ from $\mathcal{D}$ as:

$$\mathrm{p}(\{w_i, w_j\}|\mathcal{D}) \propto \sum_{d \in \mathcal{D}} \mathrm{p}(w_i|d)\mathrm{P}(w_j|d)\mathrm{p}(d|\mathcal{D}) \tag{5.1}$$

where $\mathrm{p}(d|\mathcal{D})$ is the probability of selecting document d from among all documents in $\mathcal{D}$, and $\mathrm{p}(w|d)$ represents the conditional probability of word w given d. As we consider each document in $\mathcal{D}$ to be equally probable, we estimate $\mathrm{p}(d|\mathcal{D})$ as $1/|\mathcal{D}|$ for all $d \in \mathcal{D}$.

In this work, the conditional probability of a word w given a document d is estimated using MLE as the fraction:

$$\mathrm{p}(w|d) = \frac{TF(w,d)}{\sum\limits_{w' \in d} TF(w',d)} \tag{5.2}$$

where $TF(w,d)$ is the number of occurrences of word w in d. Note that Formula 5.1 weights a word pair not only by considering the number of documents that contain the words, but also by regarding the frequency of each word in the documents.

Example In Table 5.3, we show the top five most probable word pairs generated from the example collection. As it can be appreciated, the most probable word pair is $\{clinton, president\}$, whose probability is $\mathrm{p}(\{clinton, president\}|\mathcal{D})$ $\propto 4/100 \cdot 2/100 \cdot 1/7 + 2/92 \cdot 4/92 \cdot 1/7 + 4/90 \cdot 2/90 \cdot 1/7 = 0.000390$.

Table 5.4: Entropies for some TDT2 topics.

Source topics	20002 ∪ 20096	20002 *Monica Lewinsky Case*	20096 *Clinton-Jiang Debate*
Vocabulary entropy of the sample	9.81	9.18 (-6.4%)	9.38 (-4.4%)

5.4 A homogeneity criterion to assess topic meaningfulness

We consider a homogeneity criterion to test whether a word pair represents a meaningful content in the target document collection. Our intuition is that highly probable word pairs in a collection are likely to represent a meaningful content only if such a content is homogeneous; that is, if it comprises a single, cohesive content instead of several or many. The task is then to devise a boolean criterion that expresses the homogeneity of a document set that represents the content underlying a word pair.

Information entropy has been often used as a characterization of the information content comprised in a data source. For example, it has been used as a measure for feature selection, lossless data compression methods, or for evaluating the quality of clustering partitions. In order to measure the homogeneity of a document collection $\mathcal{D}'$, entropy has been usually applied over the vocabulary of the collection at hand. That is,

$$H(V) = -\sum_{w \in V} \mathrm{p}(w|\mathcal{D}') \log_2 \mathrm{p}(w|\mathcal{D}') \tag{5.3}$$

where V is the vocabulary of the collection $\mathcal{D}'$ and $\mathrm{p}(w|\mathcal{D}')$ represents the probability of word w in $\mathcal{D}'$.

However, this value gives us little information about the number of topics a collection is actually covering. To show this let us consider, for example, two TDT2 topics (20002 and 20096) that share some vocabulary (they are about events related to President Clinton). For each topic, we select a random sample of 85 documents and then we prepare a uniform topic mixture by completely merging the two topic samples. The individual topic samples are considered to be homogeneous, whereas the topic mixture is considered to be heterogeneous. In Table 5.4, we show the vocabulary entropies obtained for these document sets. Notice that the differences in the vocabulary entropy of the homogeneous document sets with respect to the heterogeneous mixture are negligible (the relative differences are shown in parenthesis). Thus, we cannot easily define a threshold for the vocabulary entropy that determines the homogeneity of these document sets.

In this way, we propose an alternative manner to estimate the homogeneity of a set of documents (specifically, for the support set of a given word pair) by analyzing its possible content coverage. The intuition behind our proposal is

that for a homogeneous document set there must be a content-based partition in which a component prevails. This component would represent the (main) theme of the document set. Thus, given the support set of a word pair, the proposed method estimates a content-based partition for this set, and by using the notion of entropy it analyzes the possible existence of such a component.

As a content-based partition for the support set of a word pair π, we use that one induced by the connected components of the *β-similarity graph* that corresponds to the support set. Given a similarity function $\mathcal{S} : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$, the *β-similarity graph* of the support set $\mathcal{D}|_\pi$ is an undirected graph whose vertices are the documents in $\mathcal{D}|_\pi$, and there is an edge between documents d_i and d_j if they are β-similar. Two documents d_i and d_j are β-similar if $\mathcal{S}(d_i, d_j) \geq \beta$, where β is a minimum similarity threshold (Pons-Porrata et al., 2007b). This partition is equivalent to that one produced from the β-level of the document hierarchy obtained by applying the standard *single-link* clustering method (Sibson, 1973).

For instantiating function $\mathcal{S}$ we use the cosine similarity function (see Equation 2.3). The features in this case are the documents' words, which are weighted using SMART *ltc* (Buckley et al., 1995b).

We estimate the minimum similarity threshold β for a document collection $\mathcal{D}$ by averaging the similarities between each document in this collection and its k-most similar documents, that is:

$$\beta = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{1}{k} \sum_{d' \in msd(k,d)} \mathcal{S}(d, d') \tag{5.4}$$

where $msd(k, d)$ represents the collection of the k-most similar documents of d in $\mathcal{C}$. It is worth mentioning that we calculate the β value from the document collection at once, before the iterative process of topic discovery begins.

For estimating the value of k, we rely on the *k-nearest neighbor* estimation approach. Thus, we fix the value of k to be $\lfloor \sqrt{|\mathcal{D}|} \rfloor$ in the case of the news stories, and we use $k = \lfloor \log |\mathcal{D}| \rfloor$ in the case of RL-M/A and RL-CARS; that is, we define k in terms of a sublinear function on the size of the document collection, as it is suggested across the literature on *density estimation* (Loftsgaarden and Quesenberry, 1965; Duda et al., 2001; Zhang et al., 2007). In a previous work (Anaya-Sánchez et al., 2008), we have shown that the quality results obtained by varying the β threshold are stable. In the collections tested here, the proposed estimation of β thresholds produces result values that are not statistically different from estimations of the optimal ones (obtained by varying the value of β in the range $[0, 1]$).[1]

[1] In previous experiments performed on news stories (Anaya-Sánchez et al., 2010), we have shown that $k = \lfloor \sqrt{|\mathcal{D}|} \rfloor$ is a good value to estimate the threshold β. However, we have recently found that in document collections such as RL-M/A and RL-CARS that have a significant overlapping in the vocabulary of their documents, a good value of k should be estimated from a lower-order function such as the logarithmic one.

Table 5.5: Entropies for some TDT2 topics.

Source topics	20002 ∪ 20096	20002 *Monica Lewinsky Case*	20096 *Clinton-Jiang Debate*
Most probable term pair of the sample	{*clinton, president*}	{*monica, lewinsky*}	{*clinton, president*}
Vocabulary entropy of the support set	9.85	9.02 (-8.4%)	9.36 (-5.0%)
Content-based entropy of the support set	1.69	0.42 (-75.1%)	0.37 (-78.1%)

In this way, we define that the content underlying a word pair π is *homogeneous in content* if the "pure" entropy of the partition induced by the connected components of the β-similarity graph of $\mathcal{D}|_\pi$ is less than 1. The pure entropy of a partition $\Theta = \{\Theta_1, \ldots, \Theta_q\}$ is calculated as follows:

$$H(\Theta) = -\sum_{i=1}^{q} \mathrm{p}(\Theta_i|\Theta) \log_2 \mathrm{p}(\Theta_i|\Theta) \tag{5.5}$$

where $\mathrm{p}(\Theta_i|\Theta)$ can be estimated as $|\Theta_i| / \sum_{j=1}^{q} |\Theta_j|$.

As entropy expresses the number of units on the average required to describe some information (in this case the partition), this definition stems from the following fact: if less than one unit is needed to encode the contents comprised in a support set, then such a set includes a predominant content which makes it homogeneous enough. We will call *core of the support set* $\mathcal{D}|_\pi$, denoted as *core*$(\mathcal{D}|_\pi)$, to the largest connected component, which represents this predominant content.

In Table 5.5, we show both the vocabulary entropy and the content-based entropy of the support sets of the most probable term pair generated from each of the TDT2 samples that had been used above. As it can be noticed, unlike the vocabulary entropy, the content-based entropy values of single topics are well distinguished from that of the mixture. Also, we can corroborate that value 1 is a good estimate for the homogeneity decision boundary.

Example Consider the similarity matrix for the documents of the example collection (see Table 5.6). From these values, we estimate $k = \lfloor\sqrt{7}\rfloor = 2$ and β=1/7· (1/2·(0.00388+0.00513) + 1/2·(0.00388+0.00420) + 1/2·(0.00513 + 0.00420) + 1/2·(0.01111+0.01205) + 1/2·(0.00726+0.01076) + 1/2·(0.01111+0.01189) + 1/2·(0.01205+0.01189)) = 0.00818. Regarding the pair {*clinton*, *president*} and its support set {T20011.1, T20002.2, T20096.3}, it can be appreciated that the documents are not β-similar to each other. Thus, the content-based partition induced by the connected components is {{T20011.1}, {T20002.2}, {T20096.3}}, which has the entropy value: $H(\mathcal{D}|_{\{clinton,president\}})$ = -(1/3· $\log_2$(1/3)+1/3· $\log_2$(1/3)+1/3· $\log_2$(1/3)) = 1.585. Therefore, the content represented by the

Table 5.6: Similarity matrix from the example collection.

	T20011.1	**T20002.2**	**T20096.3**	**T20001.4**	**T20001.5**	**T20001.6**	**T20001.7**
T20011.1	-	0.00388	0.00513	0.00	0.00	0.00	0.0
T20002.2	0.00388	-	0.00420	0.00	0.00	0.00	0.0
T20096.3	0.00513	0.00420	-	0.00	0.00	0.00	0.0
T20001.4	0.00	0.00	0.00	-	0.00726	0.01111	0.01205
T20001.5	0.00	0.00	0.00	0.00726	-	0.01076	0.00710
T20001.6	0.00	0.00	0.00	0.01111	0.01076	-	0.01189
T20001.7	0.00	0.00	0.00	0.01205	0.00710	0.01189	-

pair $\{clinton, president\}$ is not homogeneous and therefore the pair is not regarded for generating a topic. This was expected because this pair clearly merges three topics. The next most probable word pair is $\{asia, crisis\}$, whose support set is {T20001.4, T20001.5, T20001.6}. For this pair, the generated partition includes a single connected component. Thus, the content-based entropy is 0, and therefore the pair is regarded for generating a topic. In this case, the core coincides with the support set of the pair.

5.5 Building coherent topics

Let $\pi \in \mathcal{P}$ be a pair of words that represents a homogeneous content. Let also $core(\mathcal{D}|_\pi)$ be the core of its support set. We define the *set of relevant documents for the content represented by* π as:

$$Rel(\pi) = core(\mathcal{D}|_\pi) \cup \left\{ d \in \mathcal{D} \,\middle|\, \exists d' \in core(\mathcal{D}|_\pi)[\mathcal{S}(d,d') = \max_{\substack{d'' \in \mathcal{D}_0 \setminus \{d\} \\ \mathcal{S}(d,d'') \geq \beta}} \mathcal{S}(d,d'')] \right\} \quad (5.6)$$

where $\mathcal{D}_0$ denotes the original document collection (i.e. the current collection $\mathcal{D}$ plus all the documents included in the previously identified topics).

That is, we consider as relevant documents for the content represented by a pair of words all of the documents in the core of its support set, together with those documents in the collection $\mathcal{D}$ whose most β-similar document belongs to the core.

We consider that this set of relevant documents constitutes a topic. Notice that this way of generating topics accepts documents about a topic in which the word pair does not occur. Also, documents belonging to the support set that are not relevant to its predominant content can be disregarded to build a topic.

Example $Rel(\{asia, crisis\})$={T20001.4, T20001.5, T20001.6} $\cup$ {T20001.7}, because the most β-similar document of T20001.7 is T20001.4.

5.5.1 Generating topic descriptions

The word pair π may be insufficient to describe the topic $Rel(\pi)$, because the pair may be an arbitrary term correlation under a topic or it may represent a single entity (e.g. *President Clinton*). In order to give a more adequate context for topic interpretation, we propose a method to determine a larger description extracted from the words occurring in the topic.

Assuming we have preclassified documents into a set of relevant documents $Rel(\pi)$ and a set of non-relevant documents $\mathcal{C} \setminus Rel(\pi)$, we define a *word w to be descriptive for the content labeled by a word pair π* if w occurs in $Rel(\pi)$ and also if it is highly correlated to $Rel(\pi)$ in the context $\mathcal{C}$. Let $\delta(\pi)$ denote the set of all words that are descriptive for the content labeled by π. In our method, we consider that $\delta(\pi)$ is the description of the topic generated by π.

For calculating the correlation between a word w and a topic $Rel(\pi)$, we apply the likelihood ratio score (Dunning, 1993) to the contingency tables between the topic and the word w. This score has been widely used for estimating the correlation of words with respect to a target topic (e.g., to define topic signatures (Lin and Hovy, 2000; Harabagiu and Lacatusu, 2005)).

Given the following contingency table:

	$Rel(\pi)$	$\mathcal{D} \setminus Rel(\pi)$
w	O_{11}	O_{12}
$\neg w$	O_{21}	O_{22}

where $\neg w$ represents the category of documents that do not contain word w, the likelihood ratio score of word w is defined as:

$$-2log\lambda(w) = 2\sum_{i,j} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}} \tag{5.7}$$

where O_{11} is the number of documents in topic $Rel(\pi)$ that contain word w, O_{12} is the number of collection documents that do not belong to the topic but contain w, O_{21} is the number of documents in the topic that do not contain word w, O_{22} is the number of documents that do not contain w and belong to $\mathcal{D} \setminus Rel(\pi)$, and E_{ij} is the expected value for cell i, j.

For generating the description, we regard as highly correlated words all of the words w occurring in $Rel(\pi)$ such that $-2log\lambda(w)/-2log\lambda \geq p$, where $-2log\lambda$ represents the perfect score for the topic $Rel(\pi)$, and p is a given correlation ratio ($0 < p < 1$). The perfect score, $-2log\lambda$, is obtained from the values $O_{11} = |Rel(\pi)|$, $O_{12} = O_{21} = 0$, and $O_{22} = |\mathcal{D} \setminus Rel(\pi)|$.

In the experiment carried out in Section 5.8, we use $p = 0.25$. Notice that, except for this parameter, the proposed method provides a parameter-less algorithm for topic discovery. The value of p affects the length of the topics' descriptions. However, it can be noticed that this value does not affect the process of topic generation (see for example Equation 5.6).

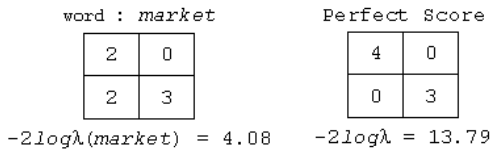


Figure 5.1: Likelihood ratio scores.

Example Consider the topic $Rel(\{asia, crisis\})$={T20001.4, T20001.5, T20001.6, T20001.7}. Figure 1 depicts the contingency tables and the likelihood ratio scores corresponding to the word *market* and the perfect score value for this topic. As it can be appreciated, for $p = 0.25$ the word *market* is included in $\delta(\{asia, crisis\})$.

5.6 Instantiating the abstract framework

To set up the topic discovery approach, we instantiate the components of the abstract framework from the above definitions as follows:

- Lexical signatures (component C1): We define lexical signatures as word pairs that co-occur in at least on document in the target collection $\mathcal{D}$. That is, the set of lexical signatures is given by the set of word pairs $\mathcal{P}$.

  In the context of the iterative search for the topics, let $\mathcal{D}'$ be the set of documents that have not been clustered yet as part of an already discovered topic. Then, we define component *sample-lexical-signatures*($\bar{T}$) as the result of sampling a word pair as follows:

  $$\textit{sample-lexical-signatures}(\bar{T}) = \{\pi\} \tag{5.8}$$

  where $\pi = \underset{\{w_i,w_j\}\in\mathcal{P}}{\arg\max}\ \mathrm{p}(\{w_i, w_j\}|\mathcal{D}')$. This means that in each iteration, we process the most probable word pair in order to search for a topic; which would be defined from the content represented by the pair in $\mathcal{D}'$.

- Topic definition/learning (Component C2) and the topic's description: As previously stated in Section 5.5, for a given lexical signature π we define:

  $$\textit{topic-definition}(\pi) = Rel(\pi) \tag{5.9}$$

  $$\tag{5.10}$$

  Accordingly, the topic description will be given by the set of words:

  $$\textit{topic-description}(\pi) = \{w|w \text{ occurs in } Rel(\pi) \wedge \frac{-2log\lambda(w)}{-2log\lambda} \geq p\} \tag{5.11}$$

- Topic meaningfulness (Component C3): Based on the meaningfulness criterion presented in Section 5.4, we define *topic-meaningfulness*(π) as:

$$\textit{topic-meaningfulness}(\pi) = -H(\Theta) \tag{5.12}$$

where $H(\Theta)$ is defined as in Equation 5.5, and Θ represents the partition induced by the connected components of the β-similarity graph of $\mathcal{D}'|_\pi$. In this case, the framework parameter w_0 is defined to regard only those pairs having -$H(\Theta)$ greater than -1.

- Stopping criterion of the search:

  In our search for the cluster-based topics, two conditions can determine the stop for the search. The first one is that all of the documents be clustered into already discovered topics. In this case, the search must be immediately stopped and the set of clusters together with their respective descriptions must be given as result.

  The second case holds when there are still documents that remain unclustered yet and no pair of words from these documents satisfy the condition of representing an homogeneous content. In such a case, the topic search must be stopped but we also propose to include a singleton cluster ...

The general steps of the proposed clustering-based approach is summarized in Algorithm 2. This algorithm is equivalent to Algorithm 1 under the substitution of the abstract components with the concrete ones as defined earlier in this section.

The general steps of the proposed method are shown in Algorithm 2.

5.7 Time Complexity

For giving an expression of the time complexity of the proposed method, we analyze the time complexity of the following operations:

i. *The computation of both the minimum similarity threshold β and the most β-similar document for each document in the collection*. The latter is used for calculating the part after the union symbol in Formula 5.6. These two computations are carried out before the iterative process of topic discovery begins. The similarity between documents d_i and d_j (see Formula 2.3) can be computed in time $O(l_i + l_j)$, where, $\forall n \in \{1, \ldots, N\}$, l_n denotes the length of the collection document d_n (i.e. the number of different words contained in d_n). As we need to calculate all the pairwise similarities between documents, the complexity of these calculations is $O(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} l_i + l_j) = O((N-1)\sum_{i=1}^{N} l_i) = O((N-1)NL_1) = O(N^2 L_1)$, where N is the number of documents in the collection and L_1 is the arithmetic mean of the document lengths.

Algorithm 2 A clustering algorithm for discovering and describing coherent and meaningful topics.

Entrada: A set of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$.
Salida: The set of topics generated from $\mathcal{D}$ together with their descriptions, $\mathcal{G} = \{(G_1, \delta_1), \ldots, (G_K, \delta_K)\}$.

1. Build the set of term pairs $\mathcal{P}$.
2. Let $\mathcal{G} = \emptyset$.
3. $\pi = \underset{\{t_i, t_j\} \in \mathcal{P}}{\arg\max} \, \mathrm{P}(\{t_i, t_j\} | \mathcal{D})$
4. If $\mathcal{D}|_\pi$ is homogeneous in content then
   (a) $G = Rel(\pi)$
   (b) $\delta = \delta(\pi)$
   (c) $\mathcal{G} = \mathcal{G} \cup \{(G, \delta,)\}$
   (d) $\mathcal{D} = \mathcal{D} \setminus G$
5. $\mathcal{P} = \mathcal{P} \setminus \{\pi\}$
6. If $\mathcal{D} \neq \emptyset \wedge \mathcal{P} \neq \emptyset$ then go to Step 3.
7. If $\mathcal{D} \neq \emptyset$ then
   (a) $\mathcal{G} = \mathcal{G} \cup \{(\{d\}, \delta) | \, d \in \mathcal{D} \wedge \delta$ is the most probable term pair in $d\}$
8. Return $\mathcal{G}$.

ii. *The iterative process for discovering and describing the topics*. The time complexity of the i-th iteration depends on the time complexity of:

a) *The computation of the probabilities of generating all term pairs that co-occur in the collection documents*. This calculation is performed by accumulating progressively the pairs' probabilities over the documents of the collection. In this way, the complexity of this step is $O(l_1^2 + \ldots + l_N^2) = O(NL_2^2)$, where L_2 is the quadratic mean of the document lengths.

b) *The computation of the homogeneity test for the support set of the most probable term pair*. As this step requires to compute the connected components in the β-similarity graph of the support set, its time complexity is $O(s_i^2)$, where s_i is the cardinality of the support set of the term pair (the pairwise similarities between the documents have been previously computed).

c) *The calculation of the topic's documents*. This step is performed by firstly retrieving the core from the connected components of the β-similarity graph, and then incorporating the documents from the collection whose most β-similar documents are in the core. This can be carried out in $O(s_i + N)$.

d) *The generation of the description*. This operation involves the calculation of the likelihood ratio score for the terms occurring in the topic, and therefore its time complexity is $O(l_1 + \ldots + l_N) = O(NL_1)$.

Thus, we can reduce the time complexity of the i-th iteration to $O(NL_2^2 + s_i^2)$. In this way, t iterations are performed in $O(tNL_2^2 + \sum_{i=1}^{t} s_i^2)$. An upper bound for the number of iterations of our method may be the number of word pairs, which is $O(NL_2^2)$. Hence, the complexity of the iterative process is $O(N^2L_2^4 + NL_2^2S_2^2)$, where S_2 represents the quadratic mean of the cardinalities of the support sets of the term pairs that co-occur in the collection documents.

The previous analysis suggests that the time complexity of our method is dominated by the iterative process for discovering and describing the topics. However, L_1, L_2 and S_2 can be considered as constant values when $N \to \infty$, because they become population means. Therefore, the overall time complexity of our method is $O(N^2)$.

Experimentally, we have calculated the values of L_1, L_2 and S_2 in three document collections (Anaya-Sánchez et al., 2010). In these collections, the averages of L_1, L_2 and S_2 are around 85.32 words, 104.34 words and 8.98 documents respectively.

Based on a similar analysis, it can be easily shown that the space complexity of the proposed method is $O(N^2)$.

5.8 Evaluation

For evaluating the proposed approach, we use the four benchmark collections described in Chapter 4; namely, TDT2, AFP, RL-M/A, and RL-CARS.

These collections are different in terms of number of topics, topic sizes, number of dimensions and document register. All of the documents in these collections have been manually labeled with topics by human annotators.

Since there are no measures that directly evaluate topic coherence and/or meaningfulness on document clusters (at least in a manner that correlates with human judgments), we rely on the assumption stated in Section 4.3.1 to evaluate the overall quality of the discovered topics. That is, we assume that the closer the topics to the gold standard produced by humans, the more coherent and meaningful the individual topics are.

Thus, we compare the obtained topics to the gold-standard in terms of macro- and micro-averaged F1 measure. We firstly focus on evaluating the impact of both components: the one aimed at producing coherent topics and the other one centered on ensuring topic meaningfulness in the performance of the proposed approach. Then, we compare our method to related approaches in the state-of-the-art; that is we compare to approaches that produce document clusters based on frequent word sets.

5.8.1 Performance of the main components

In the first experiment we consider two versions of our method in order to evaluate the impact of: (1) using only term pairs whose support sets are homogeneous in content, and (2) adding those documents whose most β-similar document is included in the core of the support set of the pairs (see Formula 5.6).

For the first version (Version 1), we disregard the homogeneity constraint imposed to the most probable term pairs, i.e. we remove the conditional part of Step 4 in Algorithm 2. Also, in this version clusters simply consist of all the documents that contain the pair, that is, for a given term pair $\pi \in \mathcal{P}$, $Rel(\pi)$ is defined as $\mathcal{D}|_\pi$.

For the second version (Version 2), we test the homogeneity condition on the term pairs (i.e. we regard the conditional part of Step 4 in the algorithm), but we only consider the cores for creating the clusters, i.e. $Rel(\pi)$ is defined as $core(\mathcal{D}|_\pi)$.

The aim of the first version is to validate the performance of the instantiated component C3 to ensure topic meaningfulness; whereas the second version aims at validating the performance of the instantiated component C2 to produce coherent topics.

For each test collection, we also consider a baseline directly built from its manual topics. Each cluster in the baseline coincides with the support set of the most probable term pair generated by a manual topic. The probability of

Table 5.7: Micro- and macro-averaged F1 values obtained for the test collections.

Data	Algorithm	*Macro*-F1	*Micro*-F1
TDT2	Baseline	0.727	0.787
	Version 1	0.461	0.588
	Version 2	0.828	0.861
	Our approach	**0.868**	**0.901**
AFP	Baseline	0.714	0.750
	Version 1	0.677	0.725
	Version 2	0.651	0.715
	Our approach	**0.719**	**0.766**
RL-M/A	Baseline	0.235	0.297
	Version 1	0.099	0.213
	Version 2	0.241	0.452
	Our approach	**0.424**	**0.526**
RL-CARS	Baseline	0.351	0.413
	Version 1	0.169	0.326
	Version 2	0.360	0.579
	Our approach	**0.565**	**0.685**

the pairs is calculated by using Formula 5.1 but constraining $\mathcal{D}$ in each case to be the manual topic (i.e., the set of documents labeled with the topic).

Table 5.7 shows both macro- and micro-averaged F1 values obtained for each test collection. Several observations can be made by analyzing these results.

Firstly, it can be appreciated that our proposal obtains very good results for both macro- and micro-averaged F1 measures in the case of the collections based on news stories; whereas in the collections of tweets the performance was poorer. The reason might be twofold:

- Unlike TDT2 and AFP, there is a noisy topic included in RL-M/A and RL-CARS (i.e., a topic labeled as "Other topics"). Such a topic comprises a significant number of documents and this might alter the meaning of some word pairs in the collection. Notice also that our method depends upon an estimated threshold β, whose value might be affected by the "noisy" tweets.

- The method relies on the VSM model to represents the documents. This model might not be the most appropriate one to represent the tweets.

Secondly, we can see that the results obtained by Version 1 are very poor, except for the AFP collection. This was expected since the manually labeled topics in AFP are more distinguishable than the other ones in terms of the main vocabulary that defines the topics. For example, in TDT2 dataset the

most probable word pair in the collection is $\{clinton, president\}$, which frequently occurs in documents about the topics 20002 (Monica Lewinsky Case), 20011 (State of the Union Address), 20096 (Clinton-Jiang Debate) and 20099 (Oregon bomb for Clinton?). Obviously, the support set of this pair merges these three topics, and therefore it can decrease the quality of the results. A similar situation is observed in RL-MA/ and RL-CARS, in which topics are actually different aspects of a set of entities in a domain.

We can also appreciate that the proposed method outperforms both the baseline and the two versions of our approach defined above. This indicates that using only word pairs is not enough for discovering topics, even though the pairs be the most probable ones generated from the manual topics. Moreover, these results corroborate the positive impact of both (i) filtering out word pairs by considering the homogeneity of their support sets and (ii) adding β-similar documents to the core of the support sets of the pairs in order to define a topic.

The second experiment was focused on validating the proposed estimation method of the minimum similarity threshold β for a document collection. With this aim, we consider a third version (Version 3) of our method that disregards the automatic calculation of β as defined in Formula 5.4.

In this new version, the minimum threshold β is defined as an additional input parameter of the method. Thus, we apply this third version of the algorithm over the four test collections using different values for β that try to uniformly cover its entire domain (i.e. the range [0, 1]). Specifically, we vary β from 0 to 1 with an increment of 0.01. In previous work (Anaya-Sánchez et al., 2008), we have shown that the results obtained by varying threshold β are stable.

In Table 5.8, we compare the best results of macro- and micro-averaged values of F1 obtained using the third version to those results obtained by our approach in the four test collections. As it can be appreciated, the proposed estimate for β approximates well the values that produce the best results for macro- and micro-averaged F1 in TDT2, but it is also close to the optimal values for macro-averaged F1 in AFP and micro-averaged F1 in RL-M/A and RL-CARS.

The values of both macro- and micro-averaged F1 obtained from the estimated β are close to the optimal ones in TDT2, AFP and RL-M/A collections. In this way, we can say that, overall, the proposed estimation method for threshold β produces result values that are near optimal.

5.8.2 Comparison to state-of-the-art approaches

In a third experiment, we compare the results obtained by our proposal to those ones produced by approaches based on frequent term sets in the state-of-the-art. In particular, we use FIHC (Fung et al., 2003) version 1.0 [2] and our own

[2] http://www.cs.sfu.edu.ca/~ddm/dmsoft/Clustering/products/fihcDistribution.zip

Table 5.8: Micro- and macro-averaged F1 values obtained for different β thresholds.

Data	Algorithm	β	*Macro*-F1	*Micro*-F1
TDT2	Version 3, β best macro-F1	0.19	0.870	0.903
	Version 3, β best micro-F1	0.19	0.870	0.903
	Our approach	0.197	0.867	0.901
AFP	Version 3, β best macro-F1	0.14	0.719	0.766
	Version 3, β best micro-F1	0.10	0.717	0.784
	Our approach	0.139	0.719	0.766
RL-M/A	Version 3, β best macro-F1	0.58	0.440	0.523
	Version 3, β best micro-F1	0.51	0.429	0.529
	Our approach	0.499	0.424	0.526
RL-CARS	Version 3, β best macro-F1	0.47	0.609	0.705
	Version 3, β best micro-F1	0.47	0.609	0.705
	Our approach	0.399	0.565	0.685

implementations of both CFWS (Li et al., 2008) and the method proposed by Malik and Kender in (Malik and Kender, 2006). For the latter, we use Mutual Information as interestingness measure with threshold 0.1. To ensure a fair comparison, we use the parameter values recommended by the authors. We also tuned support thresholds for each dataset and reported the best results. In the case of hierarchical algorithms (FIHC and the method by Malik and Kender), we report the values obtained by evaluating the whole hierarchy.

Table 5.9 shows the obtained values of micro- and macro-averaged F1 in the four test collections. For each approach, the table also includes the overlapping degree of the discovered topics (i.e. the number of clusters in which a document is included on the average) and the number of generated itemsets. Itemsets correspond to the number of frequent term sets, closed interesting itemsets and frequent word sequences generated by FIHC, the method by Malik and Kender and CFWS respectively. In the case of our method, we report the total number of probable word pairs on which the homogeneity criterion was satisfied.

As it can be appreciated, our algorithm significantly outperforms the other approaches. The obtained macro- and micro-averaged F1 values corroborate the dependence of these algorithms with respect to the minimum support, which rejects all the topics whose size is below this threshold. To make an understanding of this, it is worth mentioning that for a support of 5% there are only 8 manual topics in AFP and 5 in TDT2 collections, whose respective sizes are above this threshold (which is the recommended one for CFWS).

Also, we can observe that, unlike our method and FIHC, the other approaches obtain a high overlapping in their generated topics that does not correspond to the actual topic labeling. As previously mentioned, this overlapping produces an effect of boosting in the calculation of the F1 values that

Table 5.9: Comparison w.r.t. approaches based on frequent term sets.

Data	Algorithm	*Macro*-F1	*Micro*-F1	Overlapping	Itemsets
AFP	FIHC	0.537	0.642	1.0	48084
	CFWS	0.401	0.463	32.5	53417
	Malik	0.609	0.661	14.6	3047
	Our approach	**0.719**	**0.766**	1.0	134
TDT2	FIHC	0.404	0.515	1.5	40630
	CFWS	0.095	0.135	27.5	508246
	Malik	0.684	0.748	14.7	5811
	Our approach	**0.868**	**0.901**	1.0	979
RL-M/A	FIHC	0.153	0.235	1.02	208074
	CFWS	0.128	0.160	11.0	6442
	Malik	0.284	0.336	6.02	8244
	Our approach	**0.424**	**0.526**	1.0	6945
RL-CARS	FIHC	0.208	0.292	1.05	22696
	CFWS	0.119	0.147	11.06	6032
	Malik	0.378	0.412	5.70	4021
	Our approach	**0.565**	**0.685**	1.0	6547

actually hides the performance of these approaches.

Regarding efficiency, we can also see that our approach has the best performance by large in the case of the collections of news stories if we consider the number of generated itemsets as a measure of cost; whereas it is beaten only by CFWS in both collections of tweets (but only by a very narrow margin).

5.8.3 Descriptions

Finally, for illustrating how the generated descriptions are representative enough of the topics they describe, we respectively show in tables 5.10 and 5.11 the obtained descriptions for some topics in TDT2 and RL-M/A together with the corresponding topic titles (as provided by the human annotators).

For each manual topic i, the tables include the F1 value (i.e., the value $\mathrm{F1}(i, \sigma(i))$, see Section 2.4.1), the label and the description (top terms) obtained for the best matched cluster $\sigma(i)$. In the case of FIHC, the labels correspond to the frequent term sets, whereas the descriptions are the frequent terms in the clusters. In our method, the labels coincide with the term pairs that generate the topics.

It can be seen that unlike FIHC, our method is not only able to properly identify the topics, but also to provide meaningful descriptions for each one. See for example the description given by FIHC for the topic "Fossett's Balloon Ride" in TDT2 or the topic "WH performance in Cinderella movie" in RL-M/A. The words in these descriptions obviously do not identify these topics. Notice also the close correspondence between the topic titles and the obtained descriptions in the case of our method.

Table 5.10: Descriptions and F1 values obtained for some topics in TDT2.

Manual topics title / size	Method	Best F1 matching clusters F1 / label / description
Monica Lewinsky Case / 969	FIHC	0.87 / *house* / *white, president, lewinsky, clinton, monica*
	Our approach	0.92 / {*lewinsky, monica*} / *lewinsky, monica, starr, counsel, grand*
Fossett's Balloon Ride / 15	FIHC	0.18 / *problem* / *day, make, year, man, high*
	Our approach	1.00 / {*balloon, world*} / *fossett, steve, balloon, balloonist, louis*
Current Conflict with Iraq/1486	FIHC	0.92 / *council* / *security, u.n., iraq, weapon, inspector*
	Our approach	0.85 / {*u.n., iraq*} / *iraq, u.n., inspector, weapon, iraqi*
Cable Car Crash / 110	FIHC	0.23 / *force* / *military, official, death, kill, people*
	Our approach	0.97 / {*cable, car*} / *cable, marine, italian, car, italy*
Tornado in Florida / 53	FIHC	0.22 / *central* / *people, continue, hit, home, kill*
	Our approach	0.95 / {*tornado, florida*} / *tornado, florida, central, twister, storm*
Oprah Lawsuit / 70	FIHC	0.47 / *show* / *talk, make, time, bring, u.s.*
	Our approach	1.00 / {*winfrey, show*} / *winfrey, oprah, beef, cattle, cow*
LaSalle Boat FOUND! / 1	FIHC	1.00 / *find, year* / *authority, expect, large, run*
	Our approach	1.00 / {*lasalle, ship*} / *divers, lasalle, aimable, explorer, artefact*
Asteroid Coming?? / 31	FIHC	0.37 / *close* / *pass, year, chance, base, call*
	Our approach	0.98 / {*earth, astroid*} / *asteroid, earth, astronomer, scientist, orbit*
India, A Nuclear Power? / 475	FIHC	0.88 / *pakistan* / *nuclear, test, india, minister, country*
	Our approach	0.93 / {*test, nuclear*} / *nuclear, test, pakistan, india, pakistani*
Puerto Rico Phone Strike / 13	FIHC	0.13 / *friday, president* / *day, attack, bank, buy, close*
	Our approach	1.00 / {*telephone, puerto*} / *puerto, rico, gte, telephone, consortium*

Table 5.11: Descriptions and F1 values obtained for some topics in RL-M/A.

Manual topics title / size	Method	Best F1 matching clusters F1 / label / description
Follow the leader / 30	FIHC	0.33 / *jam* / *ft., i, jennifer, video, a, you*
	Our approach	0.79 / {*the, leader*} / *leader, wisin, yandel, follow, jennifer*
"Gangnam Style" spotlighted on CNN / 13	FIHC	0.20 / *charger* / *psy, style, gangnam, on, allkpop*
	Our approach	0.75 / {*style, cnn*} / *spotlight, cnn, http://www.allkpop.com/2012/08/psys -gangnam-style-spotlighted-on-cnn, gangnam, style*
WH performance in Cinderella movie / 11	FIHC	0.10 / *nbc* / *late, the, be, in, with, at*
	Our approach	0.76 / {*with, brandy*} / *brandy, conderella, houston, whitney, with*

The topic descriptions generated by the other methods were less representative of their topics, and so these methods were not included in this comparison.

5.9 Conclusions

In this chapter, a new clustering algorithm for discovering and describing the topics comprised in a text collection has been presented. The proposed algorithm provides a novel parameter-less method for the user in order to discover the topics in the collection, at the same time that it attaches suitable descriptions to the discovered topics.

The method relies on word pairs as lexical signatures to represent the topic aboutness from which topics are discovered. Within the method, a homogeneity criterion based on entropy has successfully been introduced to assess topic meaningfulness. Coherent topics are produced by relying on the maximum β similarity relation of documents in the collection.

The experiments carried out over TDT2 English corpus, AFP Spanish collection and the collections of tweets RL-M/A and RL-CARS validate our proposal and show significant improvements over state-of-the-art-methods (namely, FIHC (Fung et al., 2003), CFWS (Li et al., 2008) and the method proposed by Malik and Kender in (Malik and Kender, 2006)) in terms of the standard macro- and micro-averaged F1 measures.

The approach does not require to know the number of topics to be discovered a priori, and it can be applied to collections of documents of arbitrary

register, though we have experimentally found that it has a better performance on news stories than in tweets in terms of macro- and micro-averaged F1.

Chapter 6

A methodology based on statistical modeling of language and topics

6.1 Introduction

This chapter presents a novel methodology derived from the proposed framework to discover and describe coherent and meaningful topics. The methodology mainly relies on the statistical modeling frameworks of LM and PTM (see Chapter 2) to implement the framework components.

Specifically, two new modeling methods are proposed: Signature Language Modeling (SLM) and Signature Document Modeling (SDM). The former relies on language modeling techniques and is employed to set up the framework components related to both the modeling of lexical signatures and the learning and description of coherent topics behind them; whereas the latter is aimed to assess topic meaningfulness by means of a new PTM approach centered on modeling the main contents of individual documents from the target collection.

SLM is a new method aimed at obtaining a distribution of words that models the language of the main contents underlying a set of lexically related words; whereas SDM is introduced as a novel method focused on modeling the meaningful contents addressed by a document taking as context a backgound document collection.

The rest of this chapter is organized as follows. Firstly, Section 6.2 summarizes the main conceptual notions and the notation employed in this chapter. Then, Section 6.3 and 6.4 present SLM and SDM respectively. Section 6.5 shows how the abstract framework components are implemented from the proposed modeling techniques. In Section 6.6, the experiments carried out to

validate our proposal as well as the obtained results are presented. Finally, Section 6.7 presents some conclusions.

6.2 Overview and notation

Different from the methodology presented in Chapter 5, in which the topic search is guided by lexical signatures based on word pairs, the one introduced here directly relies on documents from the target collection to guide the search for the coherent and meaningful topics.

Thus, the abstract component *sample-lexical-signatures* is implemented by composing two operations: one in which a document d is chosen to find new topics according a topic diversity model, and another operation in which a set of lexical signatures deemed to represent the main contents in d is calculated.

Nevertheless, *lexical-signatures* remains to be the framework component from which the abstract and meaningful topics are defined. In this case, lexical signatures will correspond to subsets of words (from the vocabulary of the target collection) that attempt to describe the main contents addressed by the documents in the collection.

Words in a given lexical signature will share some lexical relationship that is realized by means of spectral clustering. No bounds or size constraints are imposed to a set of (lexically related) words to be a lexical signature.

The calculation of lexical signatures from a document is based on SLM, the language modeling technique introduced in this work that is also employed for both: (i) learning coherent topics and (ii) obtaining accurate topic descriptions. These two operations respectively correspond to the framework components *topic-definition* and *topic-description*.

Assessing topic meaningfulness (i.e., implementing the component *topic-meaningfulness* in the abstract framework) is carried out in this case by means of SDM, the new topic modeling approach proposed in this work to asses the meaningfulness of the topics underlying a set of lexical signatures that attempts to describe the main contents in a document.

Table 6.1 summarizes the notation followed in this chapter to specify the entire methodology.

6.3 SLM: learning coherent word distributions from lexically related words

SLM is a language modeling technique that proposes to learn a model of the language underlying the meaning of a set of lexically related words s in the context of a target collection $\mathcal{D} = \{d_1, .., d_{|\mathcal{D}|}\}$ with vocabulary $\mathcal{V} = \{w_1, .., w_{|\mathcal{V}|}\}$.

Table 6.1: Main notation used in the methodology to discover and describe coherent and meaningful topics by relying on statistical modeling of language and topics.

Variable	Description
$\mathcal{D} = \{d_1, \ldots, d_{\lvert\mathcal{D}\rvert}\}$	Set of documents defining the *target document collection*.
$\mathcal{V} = \{w_1, .., w_{\lvert\mathcal{V}\rvert}\}$	*Vocabulary* of target document collection.
s	An arbitrary lexical signature.
$s_{d,k}$	A lexical signature from document d.
$t_{s_{d,k}}$	Distribution of words representing the topic underlying signature $s_{d,k}$ in the context of $\mathcal{D}$.
$q_{s_{d,k}}$	Discrete distribution of words representing the description of topic $t_{s_{d,k}}$.
$\mathrm{s} = s_1, .., s_N$	A sample of lexical signatures.
z_i	Variable representing the topic assigned to the sample lexical signature s_i in SDM. The assigned topic is previously learned from a sample lexical signature that is not necessarily equal to s_i as SDM attempts to model a distribution of the most meaningful topics.
$\mathcal{U}(\alpha, \gamma, \mu_0, \ldots, \mu_K)$	A probability distribution of elements in $\mathbb{P}_K$ represented by an urn process. This distribution is responsible of modeling topic meaningfulness. Values $\alpha, \gamma, \mu_0, \ldots, \mu_K$ are the hyperparameters.
π	A sample from $\mathcal{U}(\alpha, \gamma, \mu_0, \ldots, \mu_K)$ that is learned in order to assess the meaningfulness of the topics underlying a sample of lexical signatures.

The aim of SLM is to obtain a probability distribution of words in which words likely to be accurately included in an explanation or description of s in the context of $\mathcal{D}$ are assigned with high probability values; whereas other words, including those ones that are very ambiguous in $\mathcal{D}$, receive marginalized values.

SLM is also aimed at performing in an unsupervised manner, without explicitly knowing which documents in $\mathcal{D}$ are relevant to the lexical signature s and which are not.

6.3.1 The SLM model

Given a set of lexically related words $s = \{w_{i_1}, .., w_{i_{|s|}}\}$ ($\forall j \in \{1, .., |s|\}$, $w_{i_j} \in \mathcal{V}$), SLM proposes to learn a probability distribution of words to represent the language model underlying s by relying on a statistical (stochastic) mapping $\tau = \{\mathrm{p}(w_i|w_j)\}_{w_i \in \mathcal{V}, w_j \in \mathcal{V}}$ deemed to reflect the lexical relationships between words from $\mathcal{D}$. Such a mapping can be directly estimated from word co-occurrences in $\mathcal{D}$.

Specifically, in SLM we consider to learn the distribution of words from s as a refined version of the posterior distribution $\{\mathrm{p}(w|s)\}_{w \in \mathcal{V}}$ defined as:

$$\mathrm{p}(w|s) \propto \prod_{j=1}^{|s|} \mathrm{p}(w_{i_j}|w)\, \mathrm{p}(w) \tag{6.1}$$

The aim of the refinement is mainly twofold: (i) to boost the likelihood of words that accurately describe the underlying meaning of s in the context of $\mathcal{D}$ and (ii) to decrease the likelihood of very common or ambiguous words than can be close to random contents from $\mathcal{D}$. Notice that some words co-occurring with words in s –or equivalently, words assigned with high probability values according to $\mathrm{p}(w|s)$– can be actually relevant to describe the meaning underlying s in $\mathcal{D}$, whereas some others cannot because they can be found co-occurring with other words likely to model other contents from the collection.

Thus, by representing the context of the target collection $\mathcal{D}$ with a probability distribution of words $\{\mathrm{p}(w)\}_{w \in \mathcal{D}}$, SLM learns the language underlying the meaning of s in $\mathcal{D}$ as the probability distribution $t_s = \{t_s(w)\}_{w \in \mathcal{V}}$ that minimize the cross entropy value:

$$H_s = -\sum_{w \in \mathcal{V}} \mathrm{p}(w|s) \log\left((1-\lambda) t_s(w) + \lambda\, \mathrm{p}(w)\right) \tag{6.2}$$

where the argument of the logarithm is a mixture in which λ is a mixture weight that accounts for the proportion of "context noise" in $\{\mathrm{p}(w|s)\}_{w \in \mathcal{V}}$, and $\mathrm{p}(w)$ is the probability of word w under the context model (i.e., the prior of w in $\mathcal{D}$).

This way of optimizing the language model underlying a lexical signature resembles the one employed in (Zhou et al., 2007) to learn the so called TSLM (see Equation 3.21). However, instead of relying on a mapping model between words and then considering cross-entropy to learn the model, TSLM learns a model from a signature s by relying on a set of document D_k ($D_k \subseteq \mathcal{D}$) deemed to be relevant for the contents behind the set of words under modeling. Specifically, TSLM aims to maximize the likelihood of word occurrences in D_k by regarding that each word is generated from a similar mixture to that in the argument of the logarithm above (specifically, TSLM aims to maximize $\prod_{w \in \mathcal{V}}((1-\lambda)t_s(w) + \lambda\, \mathrm{p}(w))^{c(w,D_k)}$, where $c(w, D_k)$ accounts for the number of times word w occurs in D_k).

Thus, the main concern with TSLM is that it is based on a knowledge that is as hard to model as that of modeling the language we are interested in. Indeed, in Zhou et al. (2007) the set of documents D_k is defined as the set of documents that simultaneously contain all words in the target signature; which –as we will show in our experiments– does not guarantee to learn coherent enough distributions.

Besides, the model proposed in (Zhou et al., 2007) is focused on providing an internal representation of documents that incorporate contextual information to be applied to ad-hoc document retrieval, instead of providing accurate definitions for specific contents included in a text collection.

6.3.2 Learning issues

From Equation 6.2, we base the learning of distribution t_s on an Expectation Maximization procedure that starting from initial values for $\{t_s(w)\}_{w \in \mathcal{V}}$, namely $\{t_s^{(0)}(w)\}_{w \in \mathcal{V}}$, it iteratively approximates the values in $\{t_s(w)\}_{w \in \mathcal{V}}$ until convergence by means of the following updates in the rth iteration:

From Equation 6.2, we base the learning of distribution t_s on an Expectation Maximization procedure that starting from initial values for $\{t_s(w)\}_{w \in \mathcal{V}}$, namely $\{t_s^{(0)}(w)\}_{w \in \mathcal{V}}$, it iteratively approximates the values in $\{t_s(w)\}_{w \in \mathcal{V}}$ until convergence by means of the following updates in the rth iteration:

$$t_s^{(r)}(w) = \frac{\mathrm{p}(w|s) Z_w}{\sum_{w' \in \mathcal{V}} \mathrm{p}(w'|s) Z_{w'}} \tag{6.3}$$

where Z_w is:

$$Z_w = \frac{(1-\lambda) t_s^{(r-1)}(w)}{(1-\lambda) t_s^{(r-1)}(w) + \lambda\, \mathrm{p}(w)} \tag{6.4}$$

In our work, we define both the mapping between words τ and the distribution of words representing the context model of the entire collection as

follows:

$$\mathrm{p}(w_i|w_j) = \frac{\mathrm{p}(w_i, w_j)}{\mathrm{p}(w_j)} \tag{6.5}$$

$$\mathrm{p}(w_j) = \sum_{w' \in \mathcal{V}} \mathrm{p}(w_j, w') \tag{6.6}$$

where $\mathrm{p}(w_i, w_j) \propto \sum_{d \in \mathcal{D}} \mathrm{p}(w_i|d)\,\mathrm{p}(w_j|d)\,\mathrm{p}(d)$. For all $w \in \mathcal{V}$ and all $d \in \mathcal{D}$, $\mathrm{p}(w|d)$ represents the MLE estimate of the probability of occurrence of word w in document d. Documents in the collection are assumed to be equally probable (i.e., $\mathrm{p}(d) = 1/|\mathcal{D}|$ for all $d \in \mathcal{D}$).

6.3.3 Summarizing a SLM model

Despite a set of lexically related words s and the model t_s learned from s can be used to describe some contents from the target collection of text documents $\mathcal{D}$, we propose an alternative way to describe such contents in terms of a discrete distribution of words q_s that summarizes t_s in a lower dimensional simplex than that determined by the entire vocabulary of the collection.

The aim is to obtain customizable descriptions of the contents underlying a set of lexically related words s that can be larger (possibly, more informative or richer) than s and shorter than t_s. This would allow to model documents relevant to s in a more efficient and accurate way.

Equation 6.2 allows to reduce the dimensionality of t_s according to two word features: one corresponding to the value $t_s(w)$ and another one regarding the posterior probability $\mathrm{p}(t_s|w) = (1-\lambda)t_s(w)/((1-\lambda)t_s(w) + \lambda\,\mathrm{p}(w))$; where the latter represents the probability of explaining an occurrence of w by means of the modeled content from s vs. explaining it by means of the context model of the collection.

Thus, given a likelihood cutoff θ_0 ($0 < \theta_0 < 1$) and a threshold β_0 ($0 < \beta_0 < 1$), we define a summary of t_s as the discrete distribution of words $q_s = \{q_s(w)\}_{w \in \mathcal{V}(q_s)}$, such that $q_s(w) \propto t_s(w)$ and:

$$\mathcal{V}(s) = \{w \in \mathcal{V} | t_s(w) \geq \theta_0, \mathrm{p}(t_s|w) \geq \beta_0\} \tag{6.7}$$

Table 6.2 shows some examples of distributions and their summaries learned from lexically related words found by analyzing one document in the context of TDT2 collection version 4.0. [1] As it can be seen, very frequent words in the collection (such as *world*, *today*, and *u.s.*), are removed from the top probable words after refinement.

Also, it can be noticed that the obtained summaries represent the aboutness of the contents behind each set of words as stressed in the collection documents. For example, in the case of the set {*nuclear*, *pakistan*}, that is closely

[1] `http://www.itl.nist.gov/iad/mig//tests/tdt/1998/`

Table 6.2: Examples of word distributions and their descriptions learned from lexically related words found by analyzing documents in TDT2. Distributions were refined by setting $\lambda = 0.75$. Stopwords were removed in a preprocessing step from the target document collection.

lexical signature $s_{d,k}$	top words regarding $p(w\|s_{d,k})$	top words regarding $t_{s_{d,k}}$ (Z_w) $\beta_0 = 0.75 \cdot \max_{w' \in \mathcal{V}, t_{s_{d,k}}(w') \geq \theta_0} \{Z_w(w')\}$	topic description $q_{s_{d,k}}$ $(\theta_0 = 0.1 \cdot \max_{w \in \mathcal{V}} \{t_{s_{d,k}}(w)\},$
{nuclear, pakistan}	test:0.0300 nuclear:0.0282 india:0.0269 pakistan:0.0216 minister:0.0097 weapon:0.0089 u.s.:0.0088 president:0.0085 country:0.0084 clinton:0.0068 prime:0.0066 conduct:0.0063 united:0.0058 pakistani:0.0057 world:0.0056 today:0.0055 official:0.0054 security:0.0050	test:0.0690 (0.92) nuclear:0.0638 (0.90) india:0.0616 (0.92) pakistan:0.0496 (0.93) minister:0.0161 (0.66) conduct:0.0139 (0.89) country:0.0135 (0.65) weapon:0.0132 (0.60) pakistani:0.0131 (0.92) prime:0.0118 (0.68) indian:0.0102 (0.88) device:0.0094 (0.93) u.s.:0.0093 (0.42) arm:0.0087 (0.83) sanction:0.0084 (0.80) testing:0.0080 (0.91) treaty:0.0080 (0.93) united:0.0074 (0.51)	$(\theta_0 = 0.10 \cdot 0.0690,$ $\beta_0 = 0.75 \cdot 0.93)$ test:0.2086 nuclear:0.1929 india:0.1862 pakistan:0.1499 conduct:0.0420 pakistani:0.0396 indian:0.0308 device:0.0284 arm:0.0263 sanction:0.0254 testing:0.0242 treaty:0.0242 indians:0.0215
{prime, minister}	minister:0.0186 prime:0.0136 president:0.0105 u.s.:0.0082 nuclear:0.0080 test:0.0066 iraq:0.0065 india:0.0063 israeli:0.0063 government:0.0053 clinton:0.0053 pakistan:0.0053 country:0.0051 netanyahu:0.0050 state:0.0049 weapon:0.0046	minister:0.0506 (0.83) prime:0.0386 (0.87) nuclear:0.0182 (0.69) israeli:0.0178 (0.86) test:0.0145 (0.67) india:0.0141 (0.68) netanyahu:0.0140 (0.86) pakistan:0.0121 (0.70) u.s.:0.0104 (0.39) pakistani:0.0102 (0.87) benjamin:0.0096 (0.88) meet:0.0088 (0.63) foreign:0.0087 (0.63) peace:0.0081 (0.77) palestinian:0.0080 (0.83) talk:0.0073 (0.51)	$(\theta_0 = 0.10 \cdot 0.0506,$ $\beta_0 = 0.75 \cdot 0.93)$ minister:0.2542 prime:0.1939 israeli:0.0889 netanyahu:0.0703 pakistan:0.0608 pakistani:0.0512 benjamin:0.0482 peace:0.0407 palestinian:0.0402 israel:0.0337 indian:0.0335 arafat:0.0301 nawaz:0.0281 sharif:0.0261
{indian}	nuclear:0.0229 test:0.0168 india:0.0142 minister:0.0136 pakistan:0.0120 indian:0.01129 prime:0.0094 president:0.0075 government:0.0072 country:0.0070 weapon:0.0058 pakistani:0.0054 foreign:0.0050 u.s.:0.0047 united:0.0047 indians:0.0045 clinton:0.0045 world:0.0045 party:0.0042 sanction:0.0039 make:0.0039 state:0.0037 states:0.0037 people:0.0037	nuclear:0.0514 (0.89) test:0.0370 (0.87) india:0.0307 (0.86) indian:0.0271 (0.95) pakistan:0.0265 (0.87) minister:0.0262 (0.76) prime:0.0193 (0.81) pakistani:0.0124 (0.91) country:0.0105 (0.59) government:0.0103 (0.57) indians:0.0103 (0.90) atal:0.0090 (0.97) foreign:0.0088 (0.69) party:0.0081 (0.76) sanction:0.0078 (0.79) vajpayee:0.0070 (0.93) kashmir:0.0067 (0.95) hindu:0.0065 (0.93) china:0.0058 (0.66) delhi:0.0057 (0.92) arm:0.0057 (0.76) weapon:0.0056 (0.39) election:0.0056 (0.76) treaty:0.0053 (0.89)	$(\theta_0 = 0.10 \cdot 0.0514,$ $\beta_0 = 0.75 \cdot 0.97)$ nuclear:0.1612 test:0.1159 india:0.0962 indian:0.0851 pakistan:0.0830 minister:0.0823 prime:0.0604 pakistani:0.0390 indians:0.0323 atal:0.0283 party:0.0256 sanction:0.0244 vajpayee:0.0219 kashmir:0.0211 hindu:0.0205 delhi:0.0180 arm:0.0178 election:0.0176 treaty:0.0167 conduct:0.0166

related to TDT2 topic 20070 "*India, a Nuclear Power?*", the learned summary includes words related to the "nuclear tests performed by India and Pakistan and theirs consequences", which was central on the news addressing the topic (i.e., the documents in the collection labeled with the topic). Besides, other semantically related words such as *weapon*, and *minister* that frequently appear in other topics related to war conflicts (e.g., topic 20015 "Current conflict with Iraq" and topic 20071 "Israeli-Palestinian Talks (London)") entail lower posterior probability values for the learned distribution.

On the other hand, the distributions underlying the sets {*prime*, *minister*} and {*indian*} merge descriptions of different subjects that are related to different (more meaningful) manually annotated topics. For example, the contents underlying {*prime*, *minister*} mainly refer to the prime ministers of Israel –related to topic 20071– and Pakistan –involved in topic 20070–. The contents behind {*indian*} include terminology related to topic 20070 and topic 20039 "India Parliamentary Elections" (mainly described by the underlined words in the table).

6.4 Assessing topic meaningfulness by means of SDM

This section introduces SDM, a probabilistic mixture model of word distributions built upon SLM that is aimed at modeling the meaningful contents addressed by a document d.

Our aim is to use SDM to asses the meaningfulness of the topics underlying a set of lexical signatures that is deemed to describe the different contents in a given document d. SDM takes as context the collection $\mathcal{D}$ represented by the word priors $\{\mathrm{p}(w)\}_{w\in\mathcal{V}}$ defined in the previous section.

6.4.1 The Signature Document Model

SDM assumes that for a given document $d \in \mathcal{D}$ there is a set of lexical signature $S_d = \{s_{d,1}, \ldots, s_{d,K}\}$ that describes the different contents in d (in the next section we will describe a method to obtain such a set). Then, from a sample of lexical signatures $\mathrm{s} = s_1, \ldots, s_N$ randomly drawn in a multinomial way from S_d, SDM proposes to model the contents in d in terms of the following distribution of lexical signatures from s:

$$\mathrm{p}_d(s) = \sum_{k=0}^{K} \mathrm{p}_d(z = t_{s_{d,k}})\, \mathrm{p}(s|z = t_{s_{d,k}}) \tag{6.8}$$

where $s \in \mathrm{s}$, $t_{s_{d,1}}, \ldots, t_{s_{d,K}}$ are the topics underlying lexical signatures $s_{d,1}, \ldots, s_{d,K}$ respectively (via SLM), $t_{s_{d,0}}$ is the context model of the target collection $\mathcal{D}$ (for all $d, d' \in \mathcal{D}$ $t_{s_{d,0}} = t_{s_{d',0}}$), and z is a random variable indicating the model (topic or context model) from which signature s is generated (z takes values

on the domain $\{t_{s_{d,0}}, \ldots, t_{s_{d,K}}\}$). The mixture coefficient $\mathrm{p}_d(z = t_{s_{d,k}})$ represents the probability of drawing model $t_{s_{d,k}}$ to generate a signature from d, and $\mathrm{p}(s|z = t_{s_{d,k}})$ represents the probability of generating s under model $t_{s_{d,k}}$.

In SDM, we assume that the vector $\left\langle \mathrm{p}_d(z = t_{s_{d,0}}), \ldots, \mathrm{p}_d(z = t_{s_{d,K}}) \right\rangle \in \mathbb{P}_K$ is distributed according to (the distribution of labels in) an urn process $\mathcal{U} = \mathcal{U}(\alpha, \gamma, \mu_0, \mu_1, \ldots, \mu_K)$ defined as follows.

Initially, the urn contains $\gamma + \alpha\mu_0 + \ldots + \alpha\mu_K$ balls, from which $\gamma + \alpha\mu_0$ balls are labeled with category c_0, and $\alpha\mu_k$ balls ($1 \leq k \leq K$) are labeled with category c_k (the parameters $\mu_0, \ldots, \mu_K$ can be seen as "priors" for categories $c_0, \ldots, c_K$ respectively, α plays the role of concentration parameter, and γ is a burstiness threshold for the context). Each time, one ball is drawn randomly from the urn and its label is inspected. If the label is in the set $\{c_1, .., c_K\}$, the ball is placed back in the urn together with an additional ball with the same label. Otherwise, the ball is placed back in the urn without adding a new ball.

Thus, we consider that the discrete distribution of models $\{\mathrm{p}_d(z = t_{s_{d,k}})\}_{0 \leq k \leq K}$ is actually conditioned on the parameters $\alpha, \gamma, \mu_0, .., \mu_K$. That is, $\mathrm{p}_d(z = t_{d,k}) = \mathrm{p}_d(z = t_{s_{d,k}}|\alpha, \gamma, \mu_0, \ldots, \mu_K)$.

Also, in accordance with $\mathcal{U}$, given an arbitrary collection of draws $\mathrm{z} = z_1, \ldots, z_N$, where $\forall i \in \{1, \ldots, N\}\, z_i \in \{t_{s_{d,0}}, \ldots, t_{s_{d,K}}\}$, the probability of drawing a new model $z = t_{s_{d,k}}$ is:

$$\mathrm{p}(z = t_{s_{d,k}}|\mathrm{z}, \alpha, \gamma, \mu_0, \ldots, \mu_K) = \begin{cases} \frac{N_k^* + \alpha\mu_k}{N^* + \gamma + \alpha \sum_{i=0}^{K} \mu_i}, & \text{if } k > 0 \\ \frac{\gamma + \alpha\mu_0}{N^* + \gamma + \alpha \sum_{i=0}^{K} \mu_i}, & \text{otherwise} \end{cases} \tag{6.9}$$

In this definition, N_k^* is the cardinal of $\{i : 1 \leq i \leq N, z_i = t_{s_{d,k}}\}$, which represents the number of balls that have been additionally added with label c_k to the urn, and $N^* = \sum_{k=1}^{K} N_k^*$.

Following our topic meaningfulness hypothesis H2, the idea underlying SDM is to consider the topic assignments of the sample signatures to the topics to asses topic meaningfulness in document d.

In SDM, we expect that lexical signatures representing random contents in the context of the collection be assigned to the context model $t_{s_{d,0}}$ since highly probable context words are less likely to be generated from (refined) SLM models. Besides topics close to the context model are expected to exhibit much lower burstiness (that is, less assignments to generate signatures) than true meaningful topics because both (a) the context model "subtracts" them burstiness and (b) words representing specific contents are less likely to be generated from these models. Finally, "supporting" topics are less likely to generate lexical signatures describing a more general content.

In this way, we propose to asses topic meaningfulness for a document d from the distribution of topics obtained from the assignment of topics to sig-

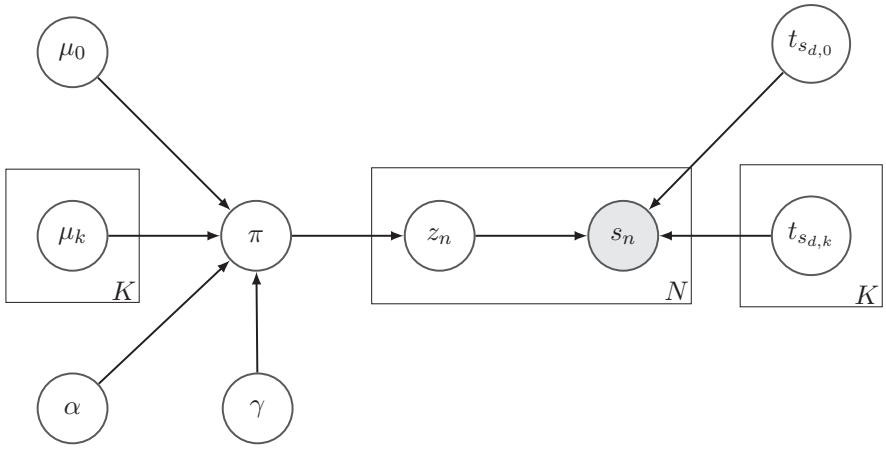


Figure 6.1: Generative model for SDM.

natures (i.e., the distribution of values in z) after inference; that is, the proportions $\{\frac{N_k^*}{N^*}\}$, where $1 \leq k \leq K$.

6.4.2 The generative process of SDM

Using plate notation, Figure 6.1 graphically represents the proposed generative model of lexical signatures with repeated sampling steps for producing a collection of signatures $\mathrm{s} = s_1, \ldots, s_N$. In the model, the multidimensional variable $\pi \in \mathbb{P}_K$ represents the discrete distribution $\{\mathrm{p}_d(z = t_{s_{d,k}})\}_{k \in \{0,\ldots,K\}}$ that models the topic proportions. The variables in this generative model are distributed as follows:

$$\pi \sim \mathcal{U}(\alpha, \gamma, \mu_0, \mu_1, \ldots, \mu_K) \tag{6.10}$$

$$z_n \sim \text{Discrete}(\pi) \tag{6.11}$$

$$s_n | z_n, t_{s_{d,0}}, \ldots, t_{s_{d,K}} \sim \text{Discrete}(t_{z_n}) \tag{6.12}$$

where z_n is the model that generates s_n.

6.4.3 Model inference

To infer the values in z, we consider a Gibbs sampling procedure that involves repeated sampling of the model utilized to generate each signature in s through a large number of iterations. The parameter π is collapsed from the sampling by considering the direct dependency between each variable in z and the hyperparameters $\alpha, \gamma, \mu_0, \ldots, \mu_K$.

Thus, the sampling of model (topic or context) assignment for z_n to generate lexical signature s_n is carried out from the following posterior:

$$\begin{aligned} \mathrm{p}(z_n = t_{s_{d,j}} | s_n, z_{-n}, \alpha, \gamma, \mu_0, \ldots, \mu_K) \propto \mathrm{p}(s_n | z_n = t_{s_{d,j}}) \times \\ \times \mathrm{p}(z_n = t_{s_{d,j}} | z_{-n}, \alpha, \gamma, \mu_0, \ldots, \mu_K) \end{aligned} \tag{6.13}$$

where z_{-n} represents the collection of models assigned to all lexical signatures in s except for s_n, $\mathrm{p}(s_n | z_n = t_{s_{d,j}})$ is the probability of lexical signature s_n under model $t_{s_{d,j}}$ (estimated as the normalized likelihood $\mathrm{p}(s_n | z_n = t_{s_{d,j}}) = \prod_{r=1}^{|s_n|} (t_{s_{d,j}}(w_{i_r}))^{1/|s_n|}$ (where $s_n = \{w_{i_1}, \ldots, w_{i_{|s_n|}}\}$), and $\mathrm{p}(z_n = t_{s_{d,j}} | z_{-n}, \alpha, \gamma, \mu_0, \ldots, \mu_K)$ is calculated from Equation 6.9 by replacing z with z_{-n}.

6.4.4 Parameter setting

We choose to define the context model $t_{s_{d,0}}$ using the same context model employed to learn the topics from lexical signatures (see Section 6.3.2).

Since parameters $\mu_0, \ldots, \mu_K$ can be seen as model priors for modeling document d, the values for $\mu_1, \ldots, \mu_K$ are estimated from the normalized likelihood:

$$\mu_i \propto \mathrm{p}(s_{d,k} | d) = \prod_{j=1}^{|s_{d,k}|} \mathrm{p}(w_{i_j} | d)^{(1/|s_{d,k}|)} \tag{6.14}$$

where the probability of word w_{i_j} under document d, $\mathrm{p}(w_{i_j} | d)$, is estimated using Jelineck-Mercer smoothing ($s_{d,k} = \{w_{i_1}, \ldots, w_{i_{|s_{d,k}|}}\}$). Similarly, the value of μ_0 is defined from the normalized likelihood of the top (20 percent) most probable words according to $t_{s_{d,0}}$. Despite $\mathcal{U}$ does not technically impose any normalization constraint to these parameters, they are taken in this work so that $\sum_{k=0}^{K} \mu_k = 1$.

The values of $\mu_1, \ldots, \mu_K$ are also employed to obtain the sample of N lexical signatures $s_1, \ldots, s_N$, which are drawn Multinomial$(\mu_1/\ell, \ldots, \mu_K/\ell)$, where $\ell = \sum_{k=1}^{K} \mu_k$. N was choosen as $10 \cdot K$; where K is the number of lexical signatures that describe the document.

So far, in the implementation of our approach we have empirically set $\alpha = 5 \cdot K$ and $\beta = \alpha \cdot K$ for all the documents modeled in each collection despite the collection specificities. We left the automatic learning of these hyperparameters for future work.

Thus, the meaningful values can be considered as en evolution of the priors μ_i according to the learned topical structure underlying the model's assignments.

6.4.5 Differences with respect to LDA

Comparing the generative model SDM (Figure 6.1) to the model that generates a document in LDA (Figure 2.1), the following differences arise:

(1) Each SDM model is learned from a collection of lexical signatures that describe the contents of a document instead of using the bag of words that corresponds to the document.

(2) In SDM, topics are not latent variables but parameters (i.e., SLM models) learned from lexically related words that provide a description for a document in the target collection. Thus, the modeled topics are deemed to be coherent, and the full inference problem becomes simpler than in LDA since it is only focused on estimating the joint posterior for the topic proportions (the value of variable π) and the topic assignments (the value of variable z).

(3) SDM additionally employs a context model of the entire target collection (i.e., $t_{s_{d,0}}$). The aim is to model random contents in the context to help assessing topic meaningfulness by decreasing the burstiness of random/abstract contents.

(4) The mixing proportions in SDM are distributed according to $\mathcal{U}$ instead of being distributed according to a Dirichlet distribution. The distribution $\mathcal{U}(\alpha, \gamma, \mu_0, \mu_1, \ldots, \mu_K)$ can be seen as a generalization of a Dirichlet distribution. A Dirichlet distribution can be obtained by setting $\gamma = \mu_0 = 0$.

Figure 6.4.5 summarizes the inference process of SDM for a document in TDT2 that has been labeled with topic 20070 "India, a Nuclear Power?". As it can be seen, the topic corresponding to the signature that most accurately describe the manually labeled topic is assigned with the largest topic proportion according to $\mathcal{U}$; whereas topics learned from more ambiguous signatures (describing more general and diverse contents) and, even, more specific signatures such as $\{pakistani\}$ (that is referred to a specific subject in the manually labeled topic) are assigned with smaller proportions.

6.5 Instantiating the abstract framework

Relying on SLM and SDM, in this framework instance we implement the framework components as follows:

- Lexical signatures (Component C1): Lexical signatures correspond to finite sets of word of arbitrary size that are calculated from individual documents in the target document collection. The computation of lexical signatures from a document $d \in \mathcal{D}$ is performed as follows.

  Firstly, an abstract summary for d is obtained by taking the following steps:

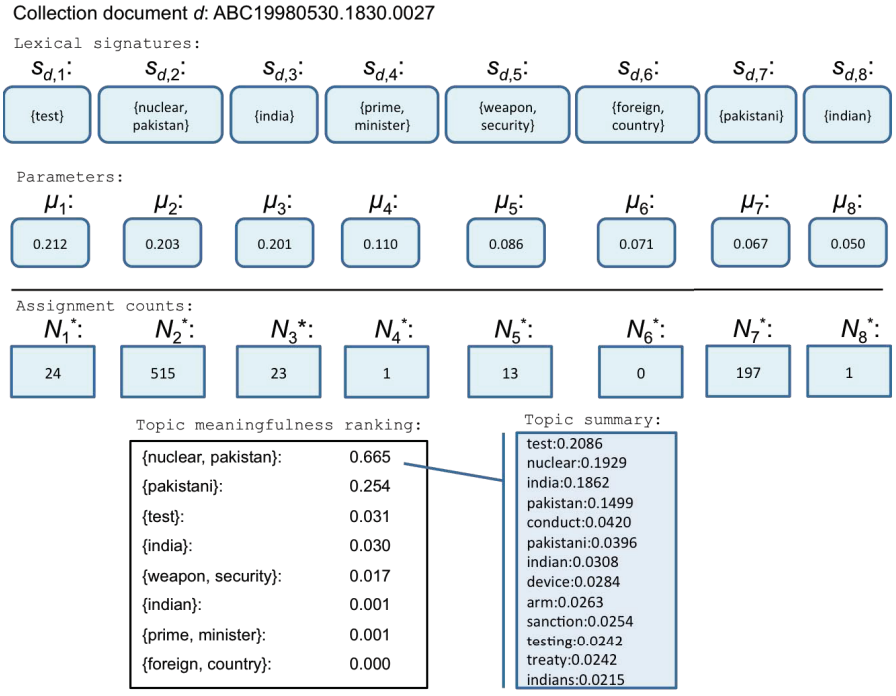


Figure 6.2: Summary of the SDM model inferred for a document in TDT2 collection labeled with topic 20070 "India, a Nuclear Power?".

(1) Obtain a clustering $G = \{g_1, .., g_{|G|}\}$ of the words in d (stop words are disregarded) by means of applying the *spectral clustering* criterion defined in (Shi et al., 2009) to a matrix of joint probabilities between words in d based on the following kernel function:

$$\mathrm{p}_T(w_i, w_j) = \exp\left\{-0.5\left(\frac{h(g(w_i), g(w_j))}{h_0}\right)^2\right\} \tag{6.15}$$

where h represents the geodesic distance between distributions, $g(w)$ is the posterior distribution of words in $\mathcal{V}$ conditioned on w according to the mapping τ, and h_0 is a distribution band width automatically calculated from the average of the distances between all words in d.[2]

(2) Calculate the mixture $\mathrm{p}_{lex}(w) = \frac{1}{|G|}\sum_{i=1}^{|G|}\mathrm{p}(w|g_i)$, where $\mathrm{p}(w|g_i)$ is a language model underlying the cluster of words g_i that is estimated using Equation 6.1.

(3) Following SLM, define the abstract summary of d as a summary of a refined version of p_{lex}.

Then, we define component *sample-lexical-signatures*($\bar{T}$) as the result of firstly choosing document d in the beginning of a topic search iteration as follows:

$$d = \underset{d' \in \mathcal{D}}{\operatorname{argmax}}\ \{\mathrm{p}(d') \cdot \prod_{(t_i, q_i) \in T}(1 - \mathrm{p}(q_i|d'))\} \tag{6.16}$$

and then defining *sample-lexical-signatures*($\bar{T}$) as the clusters of words obtained by applying the same clustering strategy employed in the above step 1 to the abstract summary of d.

- Topic definition/learning (Component C2) and topic description (Component C4): For each lexical signature s_i in *sample-lexical-signatures*(T), we define:

$$\textit{topic-definition}(s_i) = t_{s_i} \tag{6.17}$$

$$\textit{topic-description}(s_i) = q_{s_i} \tag{6.18}$$

where t_{s_i} is the SLM model learned from s_i, and q_i is a summary of t_{s_i}.

- Topic meaningfulness (Component C3): Based on the SDM model learned for document d and the set of signatures in *sample-lexical-signatures*(d),

[2] We have chosen the approach in (Shi et al., 2009) mainly because of such an approach does not require the number of cluster to be known in advance, and also because it has a statistical foundation that fits in with our proposal.

we define *topic-meaningfulness*(s_i) as:

$$\textit{topic-meaningfulness}(s_i) = \frac{N_i^*}{N^*} \tag{6.19}$$

where N_i^* is the number of samples assigned to the topic underlying s_i in SDM, and N^* is the number of samples assigned to a SLM model.

The framework parameter w_0 is chosen to select as meaningful topic that one learned from the lexical signature signature that maximizes the value N_i^*/N^*. Thus, without loss of generality, a single topic is selected as a meaningful one to describe document d.

- Stopping criterion of the search: We consider that a document $d' \in \mathcal{D}$ is covered by a topic t_r with description/summary q_r if $\mathrm{p}(q_r|d')$ is greater than a threshold $\epsilon = \epsilon(q_r)$. Then, the condition *stop-discovering-criterion* represents the condition that is satisfied when all documents in the collection are covered by at least one discovered topic in T.

  In our experiments, we define threshold ϵ as the geometric mean of the values $\mathrm{p}(q_r|d_1), \ldots, \mathrm{p}(q_r|d_{\mathcal{D}})$; that is, a Logarithmic Opinion Pool ensemble of $|\mathcal{D}|$ uniformly-weighted experts Hinton (1999).

From the above definitions, the topic discovery method proposed in this chapter can be described as an iterative search in which each iteration is focused on discovering a new coherent and meaningful topic from a collection document d. This document is chosen to find topic diversity regarding the previously discovered topics.

After choosing a document $d \in \mathcal{D}$ in the beginning of an iteration, the methodology continues to find a set of lexical signatures $\{s_{d,1}, .., s_{d,K}\}$ that describes the contents in d (the value of K is not prescribed beforehand, but automatically computed in the calculation of the signatures). Then, SLM is applied to learn the topic underlying each signature, and the meaningfulness of the topics is assessed through SDM.

At the end of the iteration, a single topic is chosen by means of maximum likelihood sampling from the inferred assignments of lexical signatures to topics. The selected topic is then stored together with its description. If all the documents in the collection are covered by the discovered topics, the iterative search finishes and the set of stored topics (and their descriptions) is returned. Otherwise, the topic discovery search proceeds with a new iteration in order to find new topics.

The general steps of this topic discovery methodology based on SLM and SDM are summarized in Algorithm 3. This algorithm is equivalent to Algorithm 1 under the substitution of the abstract components with the concrete ones as defined earlier in this section.

Algorithm 3 The proposed methodology based on SLM and SDM to discover and describe coherent and meaningful topics from a target collection of texts.

Entrada: A target collection of texts $\mathcal{D} = \{d_1, .., d_{|\mathcal{D}|}\}$.

Salida: The set of pairs $T = \{(t_i, q_i)\}_i$ that contains the topics and their descriptions.

1: $T \leftarrow \emptyset$

2: **repeat**

3: Let $d = \underset{d' \in \mathcal{D}}{\text{argmax}} \{\text{p}(d') \cdot \prod_{(t_i,q_i)\in T} (1 - \text{p}(q_i|d'))\}$.

4: Calculate the set of lexical signatures $\{s_{d,1}, .., s_{d,K}\}$ that describes d.

5: Relying on SLM, learn the topics $t_{s_{d,1}}, .., t_{s_{d,K}}$ that respectively correspond to $s_{d,1}, .., s_{d,K}$.

6: Apply SDM to infer the counts of topic assignments to lexical signatures; i.e., $N_1^*, .., N_K^*$.

7: Choose the pair $(t, q) = \underset{(t_{s_{d,k}}, q_{s_{d,k}})}{\text{argmax}} \{N_k^* / (\sum_{i=1}^{K} N_i^*)\}$ as a coherent and meaningful topic addressed by d together with its description.

8: $T \leftarrow T \cup \{(t, q)\}$

9: **until** *stop-discovering-criterion*

6.5.1 Computational complexity

The computational time complexity of this approach is determined by the time complexity of the iterative process. In the worst case, there will be n iterations; n being the number of documents in the target collection (i.e., $|\mathcal{D}| = n$). Thus, each sentence in an iteration (see steps 3 to 8 in Algorithm 3) will be executed at most n times.

For example, Sentence 3 in the algorithm will be executed n times in the worst case. Since the time complexity of this sentence is $O(n)$, the overall time complexity of executing this sentence during the whole search will be $O(n^2)$ in the worst case.

A similar analysis applies to Sentence 4 (i.e., the calculation of lexical signatures from a document). For a document d_i having l_i different words, the time complexity of calculating its lexical signatures is determined by the time complexity of building the matrix of joint probabilities (which is $O(|\mathcal{V}|l_i^2)$), and by the time complexity of analyzing the eigenvalues and eigenvectors of this matrix (which is $O(l_i^3)$). Thus, the time complexity of executing this sentence for all the documents in the target collection is $O(n(|\mathcal{V}|nL_2^2 + nL_3^3)) = O(n^2|\mathcal{V}|L_2^2 + n^2L_3^3)$; where L_2 and L_3 are the quadratic and cubic means of the number of different words in the documents of the target collection, respectively. By regarding that these two values and the size of the vocabulary are bounded by a constant, the time complexity of executing Sentence 4 n times is $O(n^2)$.

By applying a similar reasoning and also regarding that the number of

iterations to reach the convergence of SLM and SDM can be bounded by a constant, the complexity of executing n times sentences 5, 6 and 7 is $O(n^2)$. Hence, the overall time complexity of the proposed method is $O(n^2)$.

A simpler analysis can easily show that the space complexity of the method is $O(n^2)$.

6.6 Evaluation

In this case, our first experiment focuses on evaluating the coherence or degree of interpretability of the discovered topics. Specifically, we mainly compare the coherence of the topics obtained by our proposal to those ones produced by versions of the following three state-of-the-art approaches:

- LDA (Blei et al., 2003) (see Section 2.3.1), which is the simplest generative, PTM-based approach.
- HDP (Teh et al., 2006) (see Section 3.2.1), which is a bayesian PTM approach that does not require to know a priori the number of topics.
- LDA with asymmetric priors (Wallach et al., 2009) (see Section 3.2.2); namely Asym-LDA, which –like our method– relies on asymmetric priors though with a different definition from a Dirichlet distribution.
- Quad-Reg (Newman et al., 2011) (see Section 3.2.3), which aims to improve topic coherence by means of a MI-based regularization.

We also consider to compare our proposal to CTM Blei and Lafferty (2006) and Bg-LDA Chemudugunta et al. (2006). CTM aims at modeling documents regarding inter-topic relationships at the document level via the logistic normal distribution; whereas Bg-LDA uses a context model of the collection to model general aspects of the documents as well as a document-specific model to capture the specific aspects of each document. The aim of including these two approaches in the comparison is mainly to evaluate their impact in modeling topic meaningfulness since they rely on topic priors different from Dirichlet. The context model employed by Bg-LDA is similar to that one included in SDM. [3]

The versions of LDA, Quad-Reg, Asym-LDA, CTM and Bg-LDA were executed to produce the number of topics manually labeled in the target collection. In the case of HDP, hyperparameters were automatically estimated. To run Quad-Reg, we use the own target collection as corpus of reference from which to estimate the PMI values between words. This is to be fair with our strict no supervision assumptions.

[3] For example, CTM considers that if a document is about *scientific calculus* it is also likely to be about *maths*.

Table 6.3: Averaged values of coherence obtained for the discovered topics in the collections of news stories.

Method	TDT2		AFP	
	n=10	n=15	n=10	n=15
LDA	−789.85	−1887.0	−210.23	−421.56
HDP	−2623.3	−6415.3	−562.03	−2314.0
CTM	−504.70	−1306.8	−401.80	−956.47
Bg_LDA	−19 959.5	−46 847.8	−19 584.1	−45 807.7
Asym_LDA	−18 765.6	−43 983.9	−18 893.2	−43 593.2
Quad_Reg	**-125.96**	**-405.06**	**-77.68**	**-212.33**
Version-1	−748.58	−2422.9	−311.81	−1025.2
Our approach	−202.11	−744.94	−140.24	−377.26

6.6.1 Topic Coherence

We rely on the UMass measure of coherence as defined in Stevens et al. (2012), that regards word co-occurrence frequencies and a positive real value ϵ to define the coherence of a topic as in Equation 2.29.

We also consider a version of our method in order to evaluate the impact of using SLM to learn/define coherent topics from our lexical signatures instead of using TSLM (see Section 3.2.4). To define such a version (namely, Version-1), we replace Equation 6.2 in our approach by Equation 3.21.

The choice of values for n and ϵ in the formula of UMass can be critical to properly indicate a coherence score for the topics. For example, we employed the small values $n = 3$ and $n = 5$ to measure the coherence of the results obtained on the collections of news stories TDT2 and AFP; whereas in the case of the tweet collections we use the relatively large values $n = 10$ and $n = 15$.

This was because the manually labeled topics in the tweet collections shared a common vocabulary from their respective domains, so that each topic is actually described by using only a few words. The opposite situation occurs with the manual topics in TDT2 and AFP. Despite there is some overlapping between the vocabularies of a few topics, there is enough vocabulary to unambiguously describe each topic.

In the case of parameter ϵ, we have chosen $\epsilon = 10^{-100}$ in order to largely penalize off-topic words with high likelihood in the definition of a topic.

Tables 6.3 and 6.4 show the averaged values of UMass coherence obtained by each topic discovery approach on the collections of news stories and tweets respectively.

As can be seen, our approach clearly outperforms all other approaches except Quad-Reg in the four datasets. In particular, the lower values obtained by Version-1 corroborate the positive impact of learning/defining topics by means of the proposed SLM method instead of using TSLM.

Table 6.4: Averaged values of coherence obtained for the discovered topics in the collections of tweets.

Method	RL-M/A		RL-CARS	
	n=3	n=5	n=3	n=5
LDA	−220.28	−798.21	−388.82	−1330.79
HDP	−771.14	−3024.03	−130.02	−1093.84
CTM	−580.91	−1596.73	−690.74	−1716.92
Bg_LDA	−1381.15	−4606.22	−1380.10	−4002.65
Asym_LDA	−1382.23	−4601.91	−1379.32	−4600.86
Quad_Reg	−334.80	−1716.56	−235.32	−1341.86
Version-1	−44.46	−569.22	−35.23	−340.16
Our approach	**-33.72**	**-410.63**	**-22.82**	**-218.73**

Table 6.5: Comparison of the number of topics obtained by our approach to that obtained by HDP with respect to the number of manually labeled topics.

Method	TDT2	AFP	RL-M/A	RL-CARS
HDP	21	9	30	4
Our approach	100	23	1612	382
Manual topics	96	25	1018	1237

The coherence values obtained by Quad-Reg, which are higher than those ones obtained by our approach only in the case of news stories, are obviously due to regularization. To some extent, we expected that the topics obtained by means of Quad-Reg were the most coherent ones since our approach does not include any kind of topic regularization. However, it can be seen that in the case of the tweet collections, in which there is a high overlapping between the actual topics, Quad-Reg is unable to beat our approach in terms of coherence. This suggests that the use of MI-based scores is not useful enough to define topics when there is a lot of vocabulary overlapping between the actual topics.

In any case, from the obtained results we can claim that, overall, our approach discovers coherent topics, even when it is applied to collections of very short documents such as tweets with a lot of overlapping between the vocabularies of the manually labeled topics.

Another result from this experiment that shows the remarkable performance of our approach is that concerning the number of topics discovered. In Table 6.5, we compare the number of topics discovered by our approach in each test collection to that one produced by HDP. As it is shown, the number of topics obtained by our approach approximates much better to the number of manually labeled topics than that one obtained by HDP in each collection.

6.6.2 Topic Meaningfulness

In a second experiment, we empirically evaluate topic meaningfulness by measuring the correspondence between the word distributions that represent the automatically discovered topics and the word distributions that correspond to the manually labeled topics. MI was adopted to measure such a correspondence as in Equation 2.30.

We assume that the closer the obtained set of topics to the gold standard, the more coherent and meaningful the individual topics are (see Section 4.3.1). Thus, to show topic meaningfulness we only need to show a close correspondence between our topics and the gold standard according to MI.

We compare the peer topics to a series of refined versions of the gold-standard (i.e., to refined versions of the MLE distributions that correspond to the manually labeled topics). The refined versions were obtained by applying the refinement approach described in Section 6.3.1. To obtain the series, the parameter λ was varied in the interval $[0, 1)$ (the value $1 - \lambda = 0$, when $\lambda = 1$, corresponds to the unrefined MLE estimates of the topics).

Different refinements of the MLE estimate of a topic (i.e., different estimations that correspond to different values of the parameter λ) model different language granularities of the topic at the lexical level. For example, small values of $1 - \lambda$ (values of $1 - \lambda$ up to 0.4) mainly model the language of words in the background language or domain of the topic; whereas relatively large values of $1 - \lambda$ (values of $1 - \lambda$ about 0.7 and higher) focuses on modeling the language of very specific words in the topic. [4]

Our idea is then to validate if the peer topic definitions properly model the different language granularities of the gold-standard topics.

In addition to the state-of-the-art approaches, we consider in this experiment a version of our method (namely, Version-2) that is obtained by replacing the proposed distribution of topic priors described in Section 6.4.1 with a Dirichlet distribution. The aim is to validate the proposed model SDM.

Figures 6.3 to 6.6 show the results obtained in this experiment. As can be seen, our method clearly outperforms the state-of-the-art approaches for all of the values of λ except in the case of AFP for $1 - \lambda \leq 0.6$. In such a case, it is shown that HDP models the background language of the topics better than our approach. However, our approach is able to outperform HDP in the case of modeling the topics' specificities.

Regarding Version-2, it can be seen that in the case of RL-M/A (which can be seen as a collection of documents belonging to the same topic), Version-2 is clearly outperformed by our method. Whereas in the case of TDT2 and RL-CARS, the obtained values of MI by our method are significantly greater than those ones obtained by Version-2 (p-value < 0.01 according to Wilcoxon signed-rank test (Wilcoxon, 1992)). Whereas in AFP, the values obtained for

[4] In (Anaya-Sánchez et al., 2013), we have successfully employed SLM to individually model the backgorund language of four domains in order to perform entity disambiguation.

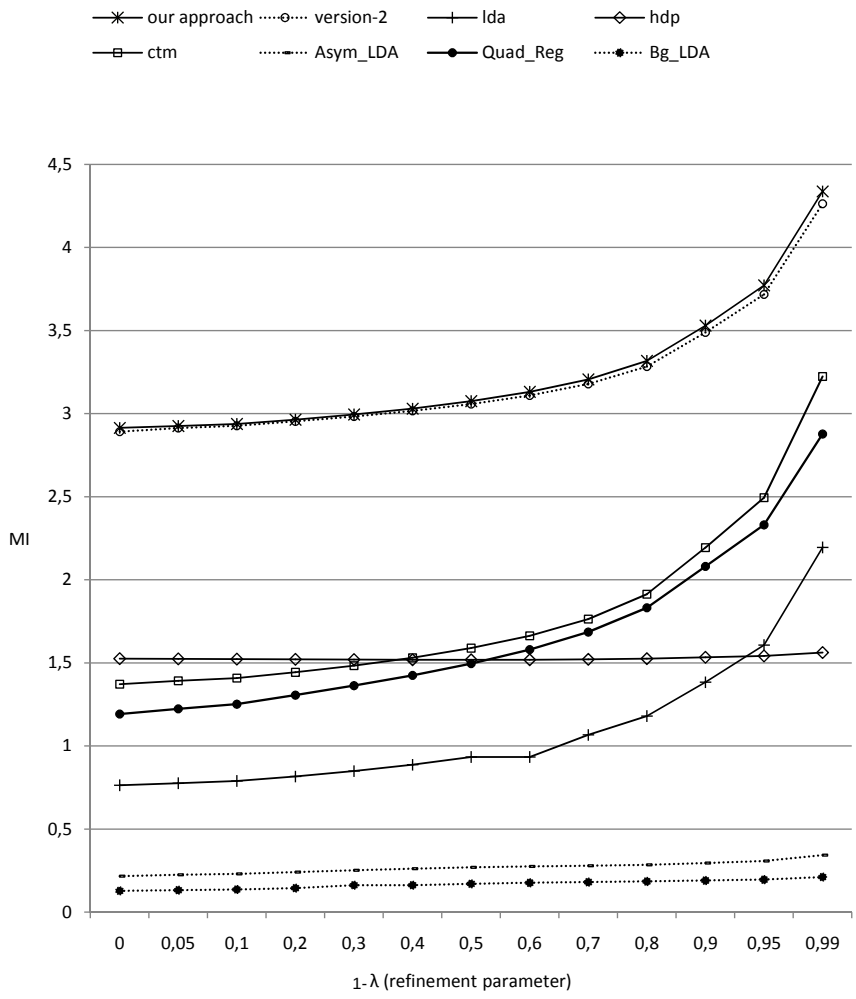


Figure 6.3: Values of MI obtained w.r.t. different refinements of the TDT2 topics in the gold standard estimated by using MLE.

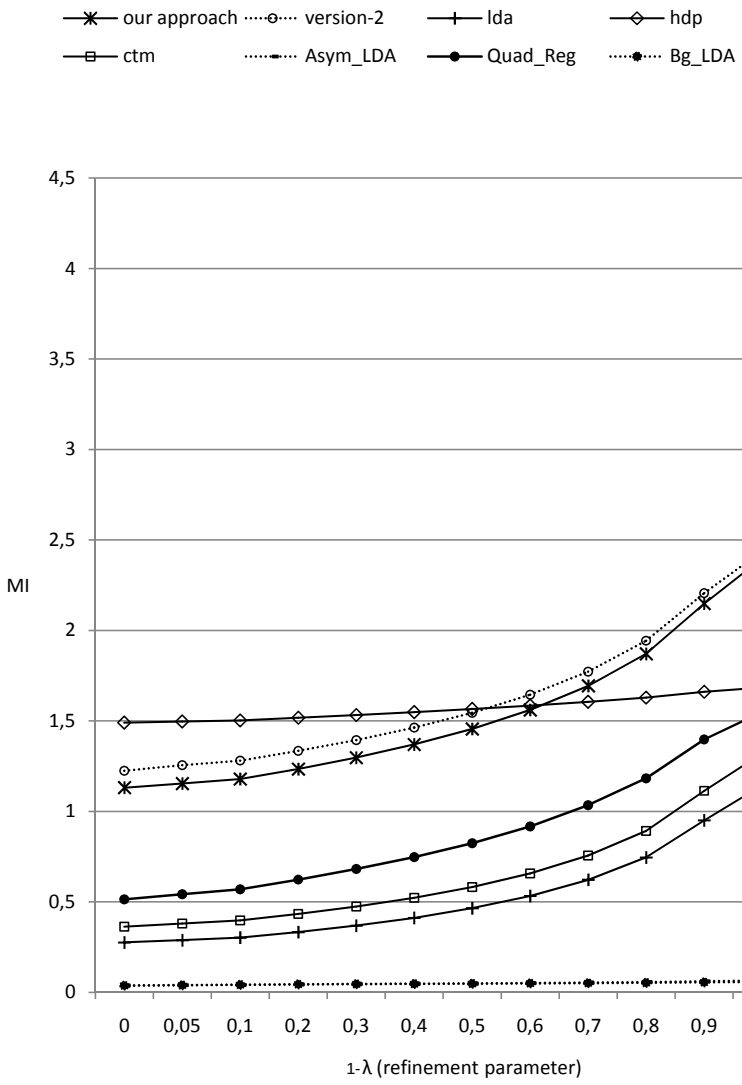


Figure 6.4: Values of MI obtained w.r.t. different refinements of the AFP topics in the gold standard estimated by using MLE.

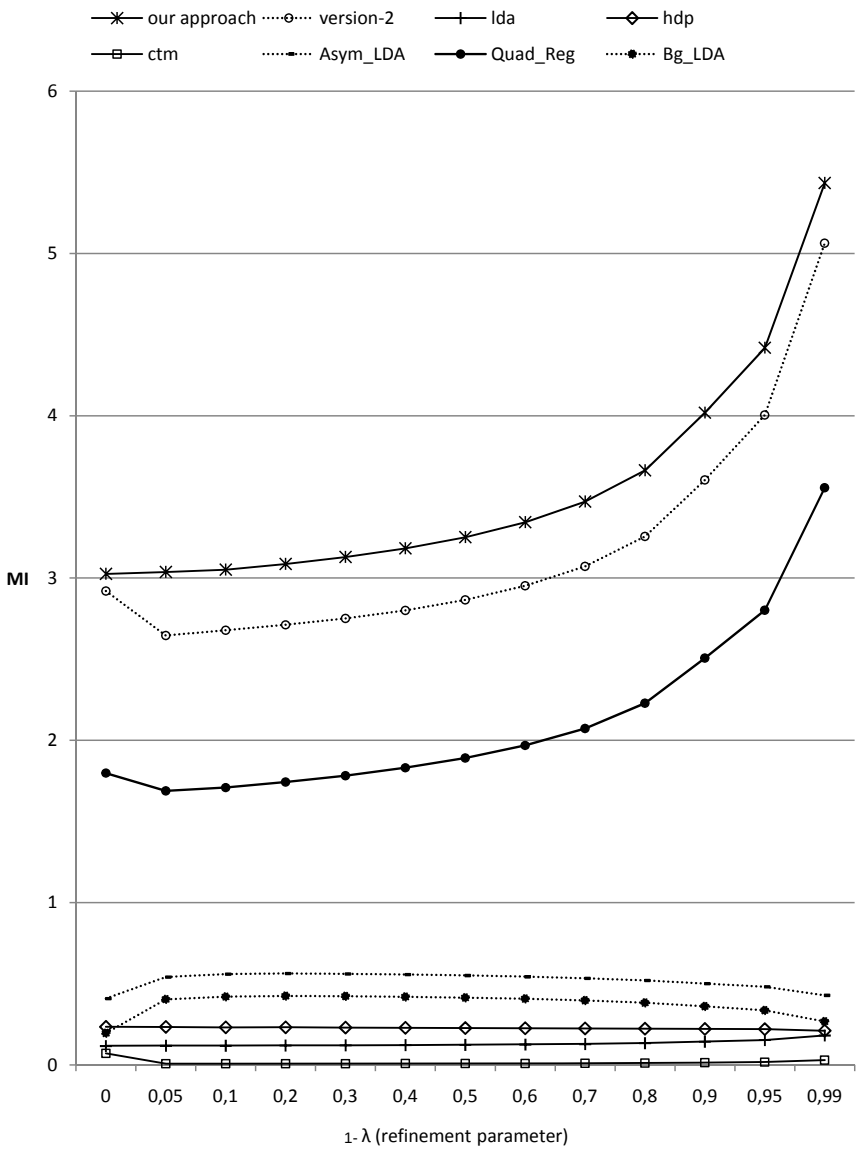


Figure 6.5: Values of MI obtained w.r.t. different refinements of the RL-M/A topics in the gold standard estimated by using MLE.

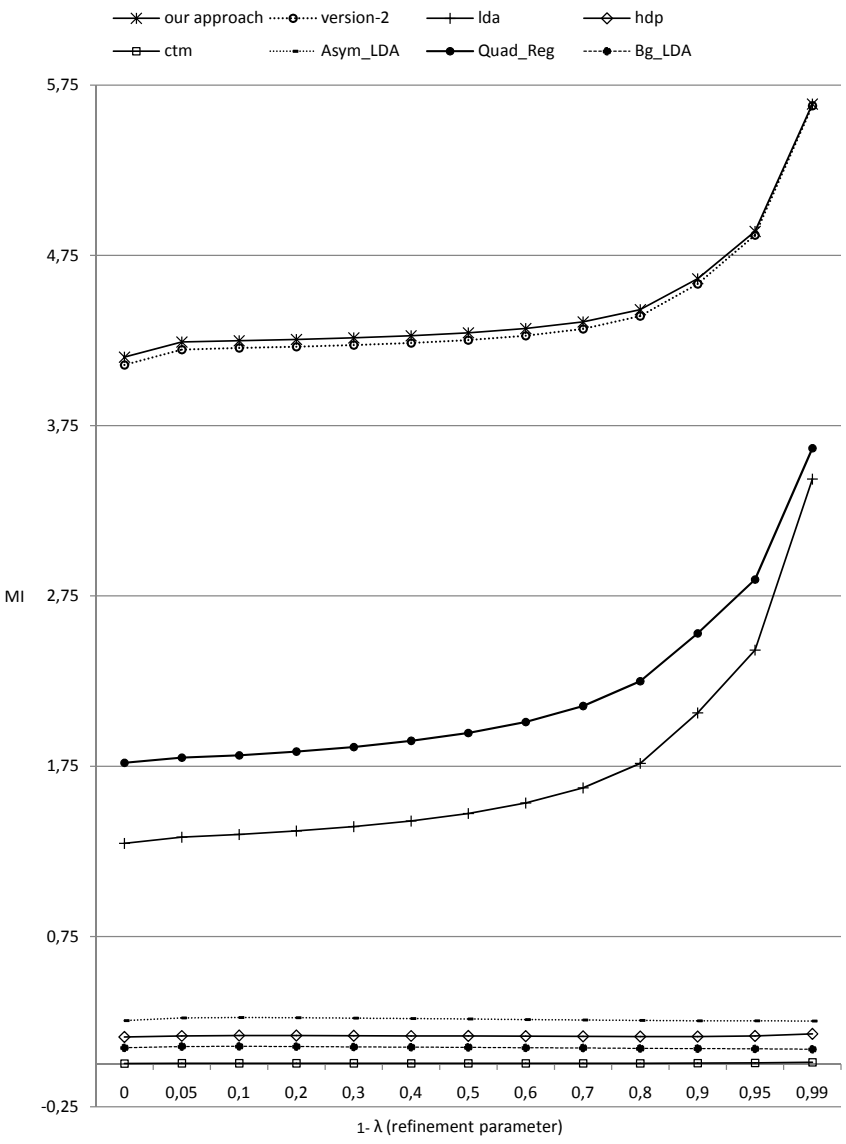


Figure 6.6: Values of MI obtained w.r.t. different refinements of the RL-CARS topics in the gold standard estimated by using MLE.

$1 - \lambda \geq 0.6$ are not significantly better than ours.

All these results statistically corroborate the positive impact of using SDM to model more meaningful topics than related approaches in PTM. It is also shown the overall capability of SDM to assess topic meaningfulness in the context of a target text collection.

Interestingly, CTM, Bg-LDA, HDP and Asym-LDA perform very bad in the collections of tweets; CTM being the worst. This shows that the basic assumption of CTM (i.e., that of modeling documents regarding inter-topic relationships) has a very negative impact to model the topics underlying RL-M/A and RL-CARS, in which each tweet has exactly one topic.

6.6.3 Descriptions

To illustrate how the generated descriptions are representative enough of the topics they describe –both in news stories and tweets–, we respectively show in tables 6.6 and 6.7 the descriptions obtained for some topics in TDT2 and RL-MA together with the topic agreement with respect to the manually labeled topics. For each manually labeled topic, the table includes the PMI value, the lexical signature and about the top 10 words in the topic description obtained for the best matched topic according to PMI. In the cases of LDA, HDP and Quad-Reg no lexical signatures are shown, and the topic description corresponds to about the top 10 more probable words under the topic.

It can be seen that the descriptions generated by our approach are in a close correspondence with the topic labels as provided by the human annotators. Also, unlike LDA, HDP and Quad-Reg, our approach systematically discovers semantically meaningful topics (according to human annotations) regardless the topic size. Both LDA and HDP do not always properly discover medium and small size topics; specially, if their vocabularies overlap (even marginally) with other topics (see for example the topics from RL-MA). Interestingly, in the case of TDT2 the approaches HDP and Quad-Reg seem to merge different manual topics through random topics/concepts in the context of the collection (e.g., "death or kill people by accidents or homicides", "sport competitions", "hidrographic subjects", etc.).

Despite the descriptions obtained by Quad-Reg in the table are similar to ours in many cases, the topics discovered by our approach are overall closer to the gold-standard as it is shown in the previous section. Indeed, it can be noticed in the tables that Quad-Reg includes some stop-words (e.g., *the*, *and*, etc.) in the top most probable words of each topic.

It is also worth mentioning that our generated descriptions are not only useful to describe the topics but also to properly discover the different topics comprised in the target text collection, since the topic search process directly relies on these descriptions to find topic diversity.

Table 6.6: Descriptions obtained for some topics in TDT2 together with their agreements with manually labeled topics.

Manual topic title / size	Method	Best PMI matching topic: PMI / lexical signature / description
Monica Lewinsky Case / 969	LDA	1.29 / - / lewinsky, monica, president, clinton, starr, lawyer, kenneth, tripp, linda, jones
	HDP	0.84 / - / president, clinton, lewinsky, monica, house, white, starr, lawyer, kenneth, jury
	Quad-Reg	2.05/ - / tripp, linda, tape, conversation
	Our approach	0.82 / {monica, lewinsky, white, house, starr} / lewinsky, monica, clinton, house, white, starr, grand, jury, intern, ken
Unabomber / 117	LDA	2.79 / - / kaczynski, theodore, lawyer, judge, trial, defence, unabomber, case,
	HDP	0.90 / - / people, woman, kaczynski, school, man, death, year, kill, lawyer, make
	Quad-Reg	3.63/ - / kaczynski, theodore, trial, judge, unabomber, lawyer, defence, defendant
	Our approach	2.33 / {ted, kaczynski, unabomber, trial} / kaczynski, unabomber, judge, ted, trial, defendant, guilty, theodore,
Superbowl '98 / 83	LDA	3.30 / - / super, bowl, denver, game, bronco, green, football, packer, team, play
	HDP	1.07 / - / olympic, win, game, team, medal, olympics, time, gold, nagano, world
	Quad-Reg	4.29/ - / denver, bowl, super, bronco, packer, green, bay, football, san, yard, fan
	Our approach	2.66 / {nfl} / denver, city, bronco, allen, nfl, downtown, celebrate, bowl, super, parade
Tornado in Florida / 53	LDA	1.61 / - / asia, asian, economy, growth, crisis, year, economic, rate, company, u.s., price, market, florida, federal, tornado, inflation
	HDP	0.34 / - / people, woman, kaczynski, school, man, death, year, kill, lawyer, make
	Quad-Reg	1.87/ - / people, kill, crash, train, car, accident, passenger, area, victim, northern, jet, scene
	Our approach	2.82 / {tornado, florida} / florida, tornado, central, victim, deadly, rip, storm, damage, loss, toll
Dr. Spock Dies / 15	LDA	3.15 / - / woman, child, man, parent, family, care, father, home, wife, spock
	HDP	0.89 / - / people, woman, kaczynski, school, man, death, year, kill, lawyer, make
	Quad-Reg	2.06/ - / brian, peterson, grossberg, amy, plead, guilty, newborn, baby, son, manslaughter
	Our approach	3.82 / {spock, baby} / child, baby, spock, die, dr, benjamin, care, book, parent
Great Lake Champlain?? / 5	LDA	2.85 / - / game, michael, jordan, bulls, chicago, jazz, karl, malone, final, utah
	HDP	2.34 / - / lake, river, great, lakes, water, donana, toxic, park, environmental, champlain
	Quad-Reg	1.75/ - / kwan, michelle, tara, lipinski, skating, skater, champion, figure, medallist
	Our approach	5.15 / {champlain} / great, lakes, lake, vermont, champlain, small, mile, square, research, york
LaSalle Boat FOUND! / 1	LDA	2.68 / - / interval, grishuk, corp, ice, platov, dance, time, year, olympic, microsoft
	HDP	0.72 / - / lake, river, great, lakes, water, donana, toxic, park, environmental, champlain
	Quad-Reg	2.95/ - / cohen, william, defense, secretary, gulf, aircraft, defence, carrier, persian, ship
	Our approach	6.65 / {artifact} / lasalle, ship, find, aimable, divers, explorer, historian, mouth, mississippi

Table 6.7: Descriptions obtained for some topics in RL-M/A together with their agreements with manually labeled topics.

Manual topic title / size	Method	Best PMI matching topic: PMI / lexical signature / description
Song-Video: Follow the leader / 30	LDA	2.01 / - / 14, 16, ac-dc, km, note, nc, x
	HDP	0.08 / - / the, be, i, to, a, you, and, of, in, on
	Quad-Reg	5.16 / - / leader, follow, wisin, yandel, ft., the
	Our approach	5.40 / {wisin} / yandel, leader, follow, jennifer, wisin, ft., lopez, *http://youtu.be/xmap94tcdns*, y
"Gangnam Style" spotlighted on CNN / 13	LDA	0.58 / - / the, and, in, of, a, with, love, she, you, be, britney, led, lady, it, houston
	HDP	0.05/ - / the, be, i, to, a, you, and, of, in, on
	Quad-Reg	5.55 / - / style, gangnam, psy, spotlight, cnn *http://www.allkpop.com*/2012/08/*psys-gangnam-style-spotlighted-on-cnn*
	Our approach	5.96/{spotlight, cnn, *http://www.allkpop.com*/2012/08/*psys-gangnam-style-spotlighted-on-cnn*} / cnn, gangnam, style, spotlight, *http://www.allkpop.com*/2012/08/*psys-gangnam-style-spotlighted-on-cnn*
WH performance in Cinderella movie / 11	LDA	2.33 / - / account, nw, fb, i, be, bb, in, the, mic, dimpledqueen, doktorbadass, my, you, u
	HDP	0.02 / - / the, be, i, to, a, you, and, of, in, on
	Quad-Reg	6.15 / - / brandy, cinderella, whitney, houston, apparently, and, watch
	Our approach	5.60 / {cinderella, impossible} / cinderella, houston, whitney, singing, impossible

6.6.4 Comparison to the clustering-based approach

Finally, we compare the results obtained by the approach presented in this chapter to those ones obtained by the method introduced in Chapter 5. To do this, we estimate a probability distribution of words for each cluster obtained by the clustering-based approach using MLE. Then, we rely on MI in order to perform a comparison with respect to two versions of these distributions: clust-0.0 (the MLE estimate) and clust-0.6 (a refined version of the MLE obtained by using $1 - \lambda = 0.6$). Results are shown in figures 6.7 and 6.8.

As it can be appreciated, the approach presented in this chapter clearly discovers topics of better quality, though, interestingly, the results obtained in the comparison to the language model of the topic specificities (i.e., when the gold-standard is refined using $1 - \lambda = 0.99$) are overall similar.

This indicates that the main aspects of each manual topic are captured in a similar way in both approaches. However, the overall topic definition is better captured with the approach introduced this chapter.

Besides, the differences between the two approaches are (significantly) more prominent in the case of the tweet collections; which reveals that this LM-based approach is more appropriate to discover topics from RL-M/A and RL-CARS than the clustering-based approach presented in Chapter 5.

It is worth mentioning that we carried out this comparison in terms of MI

because it is harder to produce a cluster-based partition from word distributions than estimating probability distributions from the clusters.

6.7 Conclusions

In this chapter, we have introduced a new fully-unsupervised approach to discover and describe topics that is derived from the abstract framework proposed in Chapter 4. Similar to PTM-based approaches, the method obtains a set of probability distributions each one representing a topic. To do so, it relies on the statistical modeling frameworks of LM and PTM to implement the framework components.

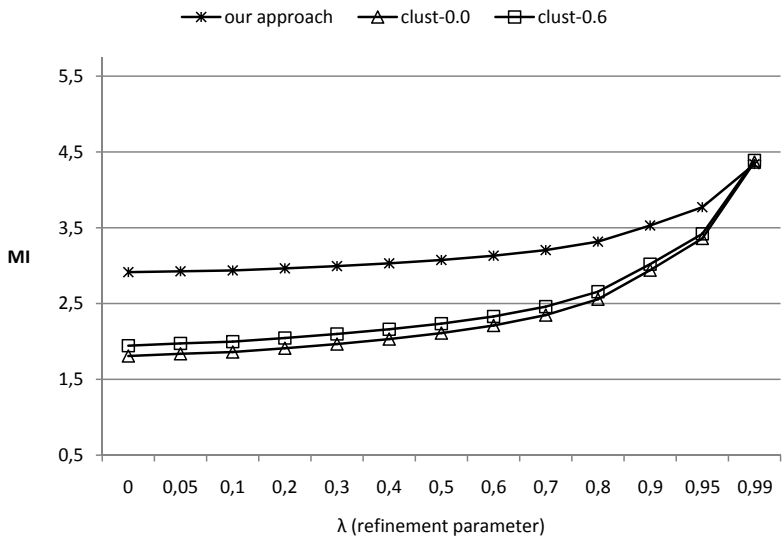
As part of the topic discovery method, two new modeling methods were proposed: SLM and SDM. The former relies on language modeling techniques and it is employed to set up the framework components related to both (i) the modeling of lexical signatures and (ii) learning and describing the coherent topics behind the lexical signatures. SDM is aimed to assess topic meaningfulness by means of a new PTM approach centered on modeling the main contents of individual documents from the target collection, which is implemented by means of a new urn model.

The experiments carried out over the AFP, TDT2, RL-M/A and RL-CARS collections show that the proposed approach is able to discover coherent enough (i.e., human-interpretable) topics, which are even more coherent than those ones produced by Quad-Reg (Newman et al., 2011) in the collections of tweets. Besides, the approach produces more meaningful topics than state-of-the-art approaches based on PTM (namely, LDA (Blei et al., 2003), HDP (Teh et al., 2006), CTM (Blei and Lafferty, 2006), Bg-LDA (Chemudugunta et al., 2006), Asym-LDA (Wallach et al., 2009) and Quad-Reg). It also outperforms the clustering-based approach proposed in Chapter 5 in terms of MI (mainly, in the case of the tweet collections).

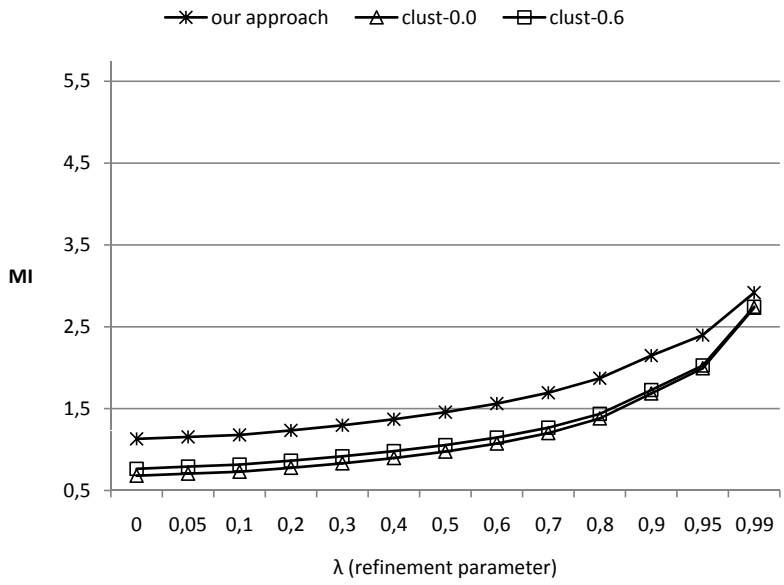
One strong point of this proposal is that it does not require to know the number of topics to be modeled beforehand; but even more, it approximates the best the number of manual topics (as labeled by human annotators) than the bayesian approach HDP.

The results also show that the proposed method can be successfully applied to collections of documents of any register (e.g., from tweets to news stories), and that in any case it is able to discover any kind of topic despite its size or broadness.

Similar to the framework instance proposed in Chapter 5, the time complexity of this method is $O(n^2)$; n being the number of documents in the target collection.

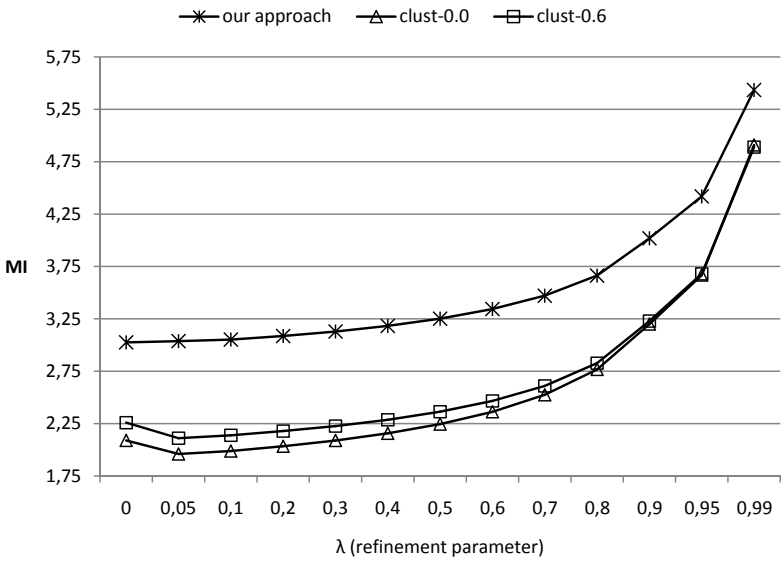


(a) TDT2

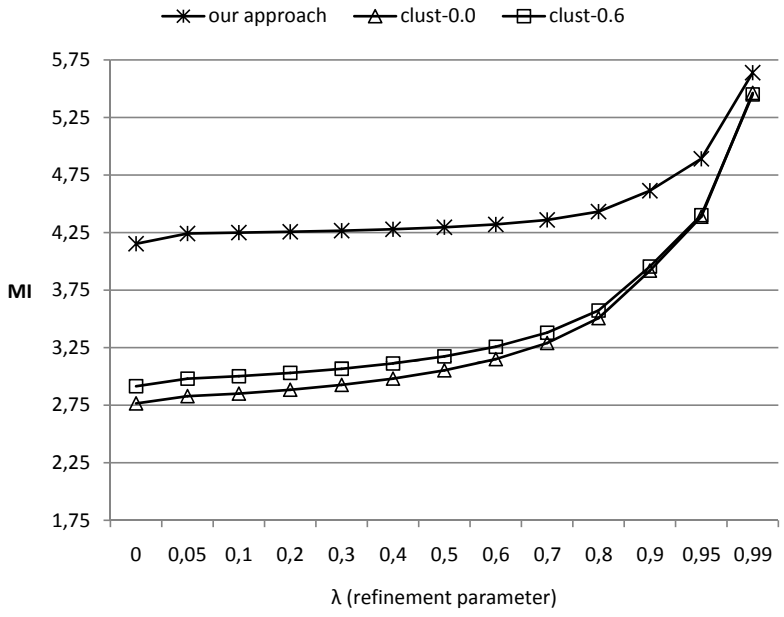


(b) AFP

Figure 6.7: Comparison to the clustering-based instance on news stories.



(a) RL-M/A



(b) RL-CARS

Figure 6.8: Comparison to the clustering-based instance on tweets.

Chapter 7

Conclusions

This thesis has been focused on addressing the unsupervised problem of discovering the coherent and meaningful topics underlying an unlabeled target collection of text documents, as well as simultaneously providing a human-interpretable description for each topic.

The main contributions of this thesis, the obtained results and future work are summarized in the next sections.

7.1 Contributions

The contribution of this thesis is manifold:

1. Firstly, it provides a review of methods aimed at simultaneously discovering and describing topics from a target collection of texts in a fully unsupervised manner. The review includes clustering methods based on frequent word sets as well as PTM approaches. From the review, the main limitations of the existing approaches are outlined.

2. Aligned with our two main hypotheses (see Section 1.3), an (unsupervised) abstract framework to discover and describe coherent and meaningful topics from a target collection is proposed. The framework implements the topic discovery as a search and mainly relies on the abstract concept of lexical signatures (i.e., lexically related words deemed to represent the topic aboutness) together with a set of abstract components aimed at supporting topic coherence and meaningfulness.

3. From the abstract framework, we firstly derive a new method to discover and describe topics with the form of document clusters. The concept of word pair is used to implement the lexical signatures that represent the topics' aboutness from which topics are discovered and pos-

teriorily described. Word pairs directly guide the topic search in this method.

4. As part of the method, a novel criterion of homogeneity of a support set is introduced to asses topic meaningfulness from a highly probable word pair. Such a criterion is based on a document similarity threshold that is automatically estimated from the target collection of documents.

5. From the above similarity threshold, a new procedure to define a coherent topic is introduced based on the concept of maximum β-similarity graph.

6. A new method based on the likelihood ratio test of word occurrences is employed to obtain enhanced descriptions for the document clusters.

7. There is also derived a new method to discover and describe topics with the form of probability distributions of words. The method is based on statistical language modeling and techniques of PTM. The method performs the search of coherent and meaningful topics based on the lexical signatures obtained from the documents in the collection.

8. As part of this framework instance, we propose SLM as method to learn and describe the topic underlying a lexical signature that is represented by a set of words. SLM combines both (i) a stochastic mapping between words and (ii) a language refinement procedure to perform the topic learning.

9. Following the general framework of PTM, the SDM model is introduced to assess topic meaningfulness based on a new urn model that models topic priors in the context of a generative model of individual documents.

10. We evaluate the proposed methods using collections of text documents of different registers (namely, news stories and tweets). These collections have been manually labeled with topics by human annotators. We compare the results obtained by each approach to those one obtained by their related state-of-the-art approaches. We also make a comparison of the two derived methodologies in terms of the MI measure.

7.2 Results

From the proposed framework and the two derived methods, we obtained the main result of this thesis, which is to corroborate the hypotheses of topic coherence (H1) and topic meaningfulness (H2) stated in Section 1.3. As part of this result, we also corroborate the following:

- the positive impact of combining the homogeneity criterion and the method based on the maximum β_0 similarity graph in the clustering-based approach introduced in Chapter 5 to obtain high quality topics (these topics outperform those one obtained by state-of-the-art approaches based on frequent word sets such as FIHC and relatives).

- the validity of SLM (see Section 6.3) to learn coherent (i.e., human-interpretable) topics and threir respective descriptions from sets of lexically related words (the topics learned by means of SLM are more coherent than those ones produced by LDA and similar methods such as HDP, and they are even more coherent than those ones obtained by Quad-Reg in the collections of tweets despite Quad-Reg focuses on enhancing topic coherence by means of regularization),

- the usefulness of SDM (see Section 6.4) to assess topic meaningfulness (by relying on SDM, the approach presented in Chapter 6 was able to produce more meaningful topics than state-of-the-art approaches based on PTM),

- the approach based on LM and PTM (Chapter 6) outperforms the clustering-based approach presented in Chapter 5 in terms of MI. It also approximates the best the number of manual topics as labeled by human annotators compared to HDP and the approach in Chapter 5, and

- the proposed framework can be used to derive methods to successfully discover and describe coherent and meaningful topics from a collection of text documents in a fully unsupervised manner, without regarding topic samples or prescribing the number of topics to be discovered.

7.2.1 Scientific publications

Related to Chapter 5:

- Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R., 2010. A document clustering algorithm for discovering and describing topics. Pattern Recognition Letters 31 (6), 502–510

- Berlanga-Llavori, R., Anaya-Sánchez, H., Pons-Porrata, A., Jiménez-Ruiz, E., 2008. Conceptual subtopic identification in the medical domain. In: Advances in Artificial Intelligence–IBERAMIA 2008. Springer, pp. 312–321

- Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R., 2008. A new document clustering algorithm for topic discovering and labeling. In: Proceedings of CIARP'08. Vol. 5197 of Lecture Notes in Computer Science. Springer, pp. 161–168

- Anaya-Sánchez, H., Berlanga-Llavori, R., Pons-Porrata, A., 2007. Retrieval of relevant concepts from a text collection. In: Current Topics in Artificial Intelligence. Springer, pp. 21–30

Related to Chapter 6:

- Anaya-Sánchez, H., Peñas, A., Cabaleiro, B., 2013. Uned-readers: Filtering relevant tweets using probabilistic signature models. In: CLEF 2013 Labs and Workshops Notebook Papers
- Anaya-Sánchez, H., Peñas, A., Berlanga-Llavori, R., 2015. Discovering Coherent and Meaningful Topics: a New Methodology Based on Signature Models. Submitted to IEEE Transactions on Knowledge and Data Engineering.

Other publications related to specific aspects of the proposed methods:

- Related to SDM to model tuple of classes for selectional preferences:

  Anaya-Sánchez, H., Peñas, A., 2015. Unsupervised induction of meaningful semantic classes through selectional preferences. In: Computational Linguistics and Intelligent Text Processing. Springer, pp. 361–371

- Related to the initial clustering employed to obtain the lexical signatures from a document:

  Anaya-Sánchez, H., Martínez-Sotoca, J., Martínez-Usó, A., 2011. Semi-supervised learning from a translation model between data distributions. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. p. 1165

7.3 Future work

As future work, we mainly consider to apply the proposed techniques for discovering coherent and/or meaningful topics in a variety of tasks that can benefit from the unsupervised categorization of texts. As an example, we can address the following tasks.

7.3.1 Subtopic discovery

We firstly propose to derive different methods from the abstract framework to address the special problem of *subtopic discovery*; which can be defined as the discovery of the different interpretations, aspects or facets underlying a user query from broad text collection.

The problem of discovering the different subtopics behind a query has been shown useful to address the task of *subtopic retrieval*, which has to do with

the retrieval of documents covering as many different subtopics of a query as possible.

So far, the subtopic retrieval approaches that rely on the discovery of subtopics have been mainly based on topics discovery approaches that often produce low quality topics such as LDA. Thus, it would be interesting to test whether or not discovering high quality subtopics (i.e., coherent and meaningful topics representing the different subtopics of a given query) can improve the performance of subtopic retrieval approaches.

7.3.2 Text generation for multi-document summarization

In second place, we might also address the task of generating multi-document summaries encoded in natural language for obtaining more interpretable topic descriptions. The motivation is mainly twofold:

(1) From an informational viewpoint, displaying a coherent piece of text might be more useful to an end-user than displaying only a simple set of terms.

(2) One of the methods proposed in this thesis is mainly based on language models to define the topics, which could provide direct support for generating natural language text from the topics' definition.

Rewriting the abstract framework into a fully generative (abstract) one should be also an important contribution to the framework of PTM in order to model more human-interpretable and meaningful topics in a generative fashion.

Bibliography

Anaya-Sánchez, H., Berlanga-Llavori, R., Pons-Porrata, A., 2007. Retrieval of relevant concepts from a text collection. In: Current Topics in Artificial Intelligence. Springer, pp. 21–30.

Anaya-Sánchez, H., Martínez-Sotoca, J., Martínez-Usó, A., 2011. Semi-supervised learning from a translation model between data distributions. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. Vol. 22. p. 1165.

Anaya-Sánchez, H., Peñas, A., 2015. Unsupervised induction of meaningful semantic classes through selectional preferences. In: Computational Linguistics and Intelligent Text Processing. Springer, pp. 361–371.

Anaya-Sánchez, H., Peñas, A., Cabaleiro, B., 2013. Uned-readers: Filtering relevant tweets using probabilistic signature models. In: CLEF 2013 Labs and Workshops Notebook Papers.

Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R., 2008. A new document clustering algorithm for topic discovering and labeling. In: Proceedings of CIARP'08. Vol. 5197 of Lecture Notes in Computer Science. Springer, pp. 161–168.

Anaya-Sánchez, H., Pons-Porrata, A., Berlanga-Llavori, R., 2010. A document clustering algorithm for discovering and describing topics. Pattern Recognition Letters 31 (6), 502–510.

Andrzejewski, D. M., 2010. Incorporating domain knowledge in latent topic models. Ph.D. thesis, UNIVERSITY OF WISCONSIN.

Antoniak, C. E., et al., 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The annals of statistics 2 (6), 1152–1174.

Arora, S., Raghavan, P., Rao, S., 1998. Approximation schemes for euclidean k-medians and related problems. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing. ACM, pp. 106–113.

Aslam, J. A., Pelekhov, E., Rus, D., 2004. The star clustering algorithm for static and dynamic information organization. J. Graph Algorithms Appl. 8, 95–129.

Asuncion, A., Welling, M., Smyth, P., Teh, Y., 2009. On smoothing and inference for topic models. In: In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence.

Beil, F., Martin, E., Xu, X., 2002. Frequent term-based text clustering. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 436–442.

Berlanga-Llavori, R., Anaya-Sánchez, H., Pons-Porrata, A., Jiménez-Ruiz, E., 2008. Conceptual subtopic identification in the medical domain. In: Advances in Artificial Intelligence–IBERAMIA 2008. Springer, pp. 312–321.

Blei, D., Carin, L., Dunson, D., 2010. Probabilistic topic models. Signal Processing Magazine, IEEE 27 (6), 55–65.

Blei, D., Lafferty, J., 2006. Correlated topic models. Advances in neural information processing systems 18, 147.

Blei, D. M., Apr. 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.
URL http://doi.acm.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., Jordan, M. I., Mar. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
URL http://dl.acm.org/citation.cfm?id=944919.944937

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., Blei, D., 2009. Reading tea leaves: How humans interpret topic models. In: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems.

Buckley, C., Salton, G., Allan, J., Singhal, A., 1995a. Automatic query expansion using smart: Trec-3. In: Proceedings of the Third TREC Conference. pp. 69–80.

Buckley, C., Salton, G., Allan, J., Singhal, A., 1995b. Automatic query expansion using smart: Trec-3. In: Proceedings of the Third TREC Conference. pp. 69–80.

Chemudugunta, C., Smyth, P., Steyvers, M., 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In: NIPS. Vol. 19. pp. 241–248.

Cutting, D. R., Pedersen, J. O., Karger, D. R., Tukey, J. W., 1992. Scatter/gather: a cluster-based approach to browsing large document collections. In: Belkin, N. J., Ingwersen, P., Pejtersen, A. M. (Eds.), Proceedings

of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, pp. 318–329.

Dillon, J., Mao, Y., Lebanon, G., Zhang, J., 2012. Statistical translation, heat kernels and expected distances. arXiv preprint arXiv:1206.5248.

Doddington, G., 1998. The topic detection and tracking phase 2 (tdt2) evaluation plan. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop. pp. 223–229.

Duda, R., Hart, P., Stork, D., 2001. Pattern classification. Wiley New York.

Dunning, T., 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19 (1), 61–74.

Feldman, R., Dagan, I., 1995. Knowledge discovery in textual databases (kdt). In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95). pp. 112–117.

Ferguson, T. S., 1973. A bayesian analysis of some nonparametric problems. The annals of statistics, 209–230.

Fung, B., Wang, K., Martin, E., 2003. Hierarchical document clustering using frequent itemsets. In: Barbar, D., Kamath, C. (Eds.), Proceedings of the Third SIAM International Conference on Data Mining. pp. 59–70.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (6), 721–741.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101 (Suppl 1), 5228.

Grossman, D., Frieder, O., 2004. Information Retrieval: Algorithms and Heuristics (2nd edition). Springer.

Harabagiu, S., Lacatusu, V., 2005. Topic themes for multi-document summarization. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 202–209.

Hartigan, J. A., Wong, M. A., 1979. Algorithm as 136: A k-means clustering algorithm. Applied statistics, 100–108.

Hinton, G. E., 1999. Products of experts. In: Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470). Vol. 1. pp. 1–6.

Jain, A., Dubes, R., 1988. Algorithms for clustering data. Prentice-Hall, Inc.

Lafferty, J., Lebanon, G., Jaakkola, T., 2005. Diffusion kernels on statistical manifolds. Journal of Machine Learning Research 6 (1).

Li, Y., Chung, S., Holt, J., 2008. Text document clustering based on frequent word meaning sequences. Data Knowledge and Engineering 64 (1), 381–404.

Lin, C.-Y., Hovy, E., 2000. The automated acquisition of topic signatures for text summarization. In: 18th International Conference on Computational Linguistics, COLING 2000. pp. 495–501.

Loftsgaarden, D., Quesenberry, C., 1965. A nonparametric estimate of a multivariate density function. Annals of Mathematical Statistics 36 (3), 1049–1051.

Malik, H., Kender, J., 2006. High quality, efficient hierarchical document clustering using closed interesting itemsets. In: Proceedings of the 6th International Conference on Data Mining. IEEE Computer Society Washington, DC, USA, pp. 991–996.

Mani, I., 2001. Automatic summarization. John Benjamin Publishing Co.

Manning, C. D., Schütze, H., 1999. Foundations of statistical natural language processing. MIT press.

Miller, G., 1995. Wordnet: A lexical database for english. Communications of the ACM 38 (11), 39–41.

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 262–272.

Newman, D., Bonilla, E., Buntine, W., 2011. Improving topic coherence with regularized topic models. In: Proceedings of 25th Annual Conference on Neural Information Processing Systems. pp. 496–504.

Newman, D., Karimi, S., Cavedon, L., 2009. External evaluation of topic models. In: Australasian Document Computing Symposium (ADCS). School of Information Technologies, University of Sydney, pp. 1–8.

Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J., 2002. On-line event and topic detection by using the compact sets clustering. Journal of Intelligent and Fuzzy Systems 12 (3–4), 185–194.

Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J., 2007a. Topic discovery based on text mining techniques. Journal of Information Processing and Management 43 (3), 752–768.

Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J., 2007b. Topic discovery based on text mining techniques. Information processing & management 43 (3), 752–768.

Santos, R., Macdonald, C., Ounis, I., 2010. Exploiting query reformulations for web search result diversification. In: Proceedings of the 19th International Conference on World Wide Web. pp. 881–890.

Shi, T., Belkin, M., Yu, B., 2009. Data spectroscopy: Eigenspaces of convolution operators and clustering. The Annals of Statistics, 3960–3984.

Sibson, R., 1973. Slink: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 16 (1), 30–34.

Steinbach, M., Karypis, G., Kumar, V., 2000. A comparison of document clustering techniques. In: ACM SIGKDD Workshop on Text Mining. Boston, pp. 109–110.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D., 2012. Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, pp. 952–961.

Steyvers, M., Griffiths, T., 2007. Probabilistic topic models. Handbook of latent semantic analysis 427 (7), 424–440.

Teh, Y., Jordan, M., Beal, M., Blei, D., 2006. Hierarchical dirichlet processes. Journal of the American Statistical Association 101 (476), 1566–1581.

Teh, Y., Newman, D., Welling, M., 2007. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Advances in Neural Information Processing Systems 19, 1353.

Wallach, H. M., Mimno, D. M., McCallum, A., 2009. Rethinking lda: Why priors matter. In: NIPS. Vol. 22. pp. 1973–1981.

Wilcoxon, F., 1992. Individual comparisons by ranking methods. In: Breakthroughs in Statistics. Springer, pp. 196–202.

Yu, H., Searsmith, D., Li, X., Han, J., 2004. Scalable construction of topic directory with nonparametric closed termset mining. In: Proceedings of the 4th IEEE International Conference on Data Mining (IDCM 2004). IEEE Computer Society, pp. 563–566.

Zamir, O., Etzioni, O., 1998. Web document clustering: a feasibility demonstration. In: Heckerman, D., Mannila, H., Pregibon, D. (Eds.), Proceedings of the 21st Annual International ACM SIGIR Conference (SIGIR'98). ACM, pp. 46–54.

Zhai, C., Cohen, W., Lafferty, J., 2003. Beyond independet relevance: Methods and evaluation metrics for subtopic retrieval. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. ACM Press New York, NY, USA, pp. 10–17.

Zhai, C., Lafferty, J., 2006. A risk minimization framework for information retrieval. Information Processing and Management 42 (1), 31–55.

Zhang, C., Zhang, X., Zhang, M., Li, Y., 2007. Neighbor number, valley seeking and clustering. Pattern Recognition Letters 28 (2), 173–180.

Zhou, X., Hu, X., Zhang, X., 2007. Topic signature language models for ad hoc retrieval. Knowledge and Data Engineering, IEEE Transactions on 19 (9), 1276–1287.