Mestrado em Engenharia Informática
Dissertação
Relatório Final

Semantic Topic Modelling

Adriana Figueiredo Ferrugento
aferr@student.dei.uc.pt

Orientadores:
Hugo Gonçalo Oliveira
Ana Oliveira Alves
Data: 6 de Julho de 2015

FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Departamento de Engenharia Informática Faculdade de Ciências e Tecnologia, Universidade de Coimbra

Mestrado em Engenharia Informática
Dissertação - Relatório Final

Estágio: [1865] - Semantic Topic Modelling
Autor: Adriana Figueiredo Ferrugento (aferr@student.dei.uc.pt)
Orientador DEI: Hugo Gonçalo Oliveira
Co-Orientador: Ana Oliveira Alves
Juri Arguente: Alexandre Miguel Pinto
Juri Vogal: António Jorge Silva Cardoso

Acknowledgements

This work was supported by FCT, in the scope of the InfoCrowds project - FCT-PTDC/ECM-TRA/1898/2012FCT.

First, I would like to thank my advisors, Hugo Oliveira and Ana Alves, for your guidance and support throughout the past year. I could not have accomplish this without you. I also want to give a big thanks to Filipe Rodrigues. He acted like an advisor and was crucial in this thesis's development.

I am very grateful to my whole family, always showing encouragement and support.

A special thanks to my boyfriend, Mário Balça, who was always there when I needed him.

And, finally, to all my friends, you know who, I thank you for being there and making my days better. And especially to Mariana Lourenço, without whom I could not have come as far as I did.

Abstract

Topic models came to improve the way search, browse and summarization of large sets of texts is performed. These models are used for uncovering the main theme of the documents in a corpus, where topics are probability distributions over a collection of words that is representative of a document. The most widely used topic model is called Latent Dirichlet Allocation (LDA) and it enables for documents to be characterized by more than one topic. This allows for a more accurate representation of what happens with real documents, where a text may have more than one underlying theme. However, this popular model is still far from producing excellent topics, given that it does not account for the semantic relations between words. It may thus result in redundant topics that contain different words, but with the same meaning.

This thesis offers a way to improve the LDA algorithm and, hence, solve the problem of not considering the semantics of words. The model proposed here uses the LDA algorithm as a starting point, however some changes are made, since it is our interest to introduce semantic relations in this model. A main component of the proposed model is the use of a lexical database for English, WordNet, which enables the integration of semantics by accessing its content.

Resumo

A existência de topic models veio melhorar a maneira como se pesquisa, navega e resume grandes quantidades de textos. Estes modelos são utilizados para descobrir qual o principal tema dos documentos de um corpus, onde tópicos são distribuições de probabilidade sobre um conjunto de palavras que é representativa do documento. O topic model mais vulgarmente utilizado é Latent Dirichlet Allocation (LDA) e permite que os documentos sejam caracterizados por mais do que um tópico. Isto permite uma representação mais precisa do que acontece com documentos reais, onde um texto pode ter mais do que um tema subjacente. No entanto, este modelo popular ainda está longe de produzir tópicos excelentes, uma vez que não tem em conta as relações semânticas entre as palavras. Isto pode resultar em tópicos redundantes, que contêm palavras diferentes, mas com o mesmo significado.

Esta tese propõe uma maneira de melhorar o algoritmo LDA e, portanto, resolver o problema de não considerar a semântica das palavras. O modelo aqui apresentado utiliza o algoritmo LDA como ponto de partida, no entanto, algumas alterações serão realizadas, uma vez que é do nosso interesse introduzir as relações semânticas neste modelo. Uma componente principal deste modelo é o uso de uma base de dados lexical, para inglês, WordNet, que permite a integração de semântica no novo modelo, ao se explorar o seu conteúdo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Topic models are algorithms used for uncovering the main theme of documents in a corpus. Topics are probability distributions over a collection of words which should inform what the documents are about. These algorithms are appropriate for evolving traditional searches and browsing, summarizing large quantities of texts, classification, novelty detection, similarity and relevance judgments. They are not limited to text mining applications, as they can be used in fields with different types of data, such as computer vision (Wang et al. (2009)), collaborative filtering, content-based image retrieval, or bioinformatics (Flaherty et al. (2005)). For example, in computer vision they are applied to natural images in order to perform image retrieval, classification, and indexing, where instead of documents they deal with images. In bioinformatics their goal is to classify correctly a specific type of genes. This thesis proposes an alternative to the traditional topic model, applied only to text mining applications, given that the result is based on the semantics of the content of the documents (words).

The first proposal of a topic model was called Latent Semantic Indexing (LSI) (Deerwester et al. (1990)) and its purpose was to retain the most of the variance present in the documents, which can lead to significant compression in large datasets. Then came the Probabilistic Latent Semantic Indexing (PLSI) (Hofmann (1999)), a variant of LSI where different words could generate different topics. However this model was still incomplete given that it provides no probabilistic model at the level of documents. Later, Blei et al. (2003) developed Latent Dirichlet Allocation (LDA), which is currently the most commonly used topic model and is a generalization of PLSI. It allows documents to have a mixture of topics, given that it enables the capture of significant intra-document statistical structure via the mixing distribution. This model is explained with further detail in chapter 2.

1.1 Motivation

This thesis subject arose due to the existence of topics, created by models like those described above, where the included words are sometimes semantically related to other words in the same topic. For example, if the words *automobile*, *car* and *auto* appeared in the same topic, it would be redundant, because each of them does not provide additional information to the other two, given that they have the same meaning. This means that the result of

the traditional Topic Modelling would not be very informative, given that there are words semantically related. So, our goal is to have topics that do not have these similar words. Further ahead, in chapter 3, there will be a more detailed explanation on what motivated the choice of this theme, alongside with real examples to support it.

The problem at hand is how to implement such a model, with a semantic background. The solution chosen to explore it is to use the LDA model as a starting point and then adapt it until it satisfies the purpose intended. The input for this model will be a collection of documents, but they will not be just words, like in the traditional way. Instead, every (open class) word will be represented with all its possible synonyms, each with its own probability. For example, for the word *dog*, there will be (`2084071:prob1, 10114209:prob2, ...`), where `2084071` is a sense identifier (`dog, domestic dog, Canis familiaris`), `10114209` of (`frump, dog`) and `prob1` and `prob2` will be values calculated based on other assumptions. A word sense is one of the possible meanings of a word. In a lexical database as WordNet (Miller, 1995), it corresponds to the presence of the word in a synset (set of synonyms), which may be seen as the representation of concept by its possible lexicalizations. So, in a similar fashion to the task of word sense disambiguation (Navigli, 2012), this model decides which is the most suitable synset for a word, based on its context and probabilities. Having decided on an input based on the identifiers of each synset, then the output of the algorithm will consist of sets of topics with these IDs instead of words, as in the classic LDA. However, for usability purposes, the most representative word of the synset (the most frequently used with that sense, according to WordNet) can be chosen to represent the concept.

The proposed model is expected to be of great value to all the areas that currently use topic models. Even though it is currently more oriented to text mining applications, it does not mean that it cannot be adapted to the other areas, especially when there is ambiguity in the data used.

1.2 Objectives

The goal is to implement a topic model based on LDA, but sensitive to the semantics of words. In order to avoid redundant and less informative topics, it must consider not only the context in which the words occur, but also information about the words of a language and its meanings, obtained, for example, from knowledge bases as wordnets.

As a starting point, the datasets chosen have their content in English, due to the existence of a wider range of tools and resources to select from. To broaden the language domain, it is thought of adapting the model to accept documents in Portuguese. This will allow the analysis of the model's behavior with other languages, as well as a better validation of it. However, in order for this to happen, it is necessary to replace some tools and resources that are only designed for the English language. For example, WordNet can no longer be used, but a valid replacement can be selected from the available Portuguese wordnets (Gonçalo Oliveira et al., 2015). The model itself will not need to be changed, only its input will have to be generated by other tools and resources.

1.3 Contributions

The main contribution of this thesis is a fully functional topic model, with a semantic strand. Towards the current version of this model, several versions were developed, with variations, especially on the way semantics was considered. This thesis describes each of those versions, and experiments performed to validate them, which constitute important steps to reach the current model. One of the preliminary versions was described in a paper that will be presented in the Portuguese Conference on Artificial Intelligence (EPIA 2015[1]) (Ferrugento et al., 2015). A second paper is being prepared about the current version of the model.

The work described in this thesis was developed in the scope of the project InfoCrowds – Social Web Information Retrieval for crowds mobility management[2], where we are also currently working on. For InfoCrowds, it is necessary to perform information retrieval of events on the Web, where they will afterwards be clustered into different groups with topic modelling. This project mainly focuses on events that occurred in Lisbon, which means that the information is in Portuguese. In this context, the integration of the model is a plus, and InfoCrowds will surely benefit from having more informative topics.

1.4 Thesis outline

After this introductory chapter, in chapter 2 of this thesis, there is a section to introduce pertinent background knowledge, including an explanation on probabilistic graphical models, approximate inference and natural language processing. The second section of chapter 2 is intended to enumerate related work on this topic. The proposed model is presented in chapter 3, alongside with additional motivation behind the choice of this thesis subject and the technological choices for the development of the model. Chapter 4 presents all the experiments performed, as well as their outcomes and evaluation. Chapter 5 has some conclusions, a discussion about the work done so far, and an enumeration of future work.

[1] http://epia2015.dei.uc.pt/

[2] https://www.cisuc.uc.pt/projects/show/176

Chapter 2

Background and Related work

This chapter is dedicated to the background knowledge for the subject of this thesis, and it presents as well related work, which is why it is divided in two sections.

The first section explains what probabilistic graphical models are, how to perform approximate inference, and a few methods that can do that, and also a short introduction to natural language processing. This section is fundamental to understand the model that is proposed in our thesis.

The second section presents the state of the art of the thesis. It has different categories, allowing the content to be organized by theme. The first one presents several models, such as the one whose results inspired this thesis, the classic LDA, and also some of the current models that were also based in this algorithm. In this category there are also models that can be used to perform WSD or WSI. The next presents some models that were not based on LDA, but have the same purpose as our thesis, because they are still topic models.

In the third section, a small comparison on the techniques and resources that some models used is presented and the final one presents the different measures used to evaluate topics models.

2.1 Background

In this section some basic concepts and methods that are used in topic models are presented. This is important in the way that it helps the understanding of existing models and the one proposed in this thesis. First, there is an introduction to Probabilistic Graphical Models, followed by an explanation of Approximate Inference and finally an overview on the topic of Natural Language Processing.

2.1.1 Probabilistic Graphical Models

For a better understanding of a real-world phenomenon, it is sometimes helpful to represent it visually and for that there are probabilistic graphical models, which use a graph-based representation as basis for encoding a complex probability distribution in a compact way.

This is considered the best approach for representing the system because it provides a simple and transparent way to visualize and understand the structure of a probabilistic

model. Moreover, it gives insight on the properties of the model, like conditional independence, and it allows the inference algorithms to work much faster than they would if using the joint distribution explicitly. Furthermore, this graphical model offers the possibility of representing the distribution in a tractable way, even if the representation of the joint distribution is very complex.

A probabilistic graphical model consists of a set of nodes, which are connected by edges. These graphical models are divided in two major groups: they can either be a directed graphical model, like a Bayesian Network (Jensen, 1996), that has links defined by arrows which imply directionality, or an undirected graphical model, like Markov random fields, where the link is not represented by an arrow and therefore, there is no explicit direction. A directed graph is more appropriate for expressing relationships between random variables, which are represented by nodes, whilst undirected graphs are more convenient for expressing soft constraints between random variables.

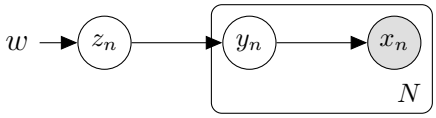

Figure 2.1: Example of a Probabilistic Graphical model.

Figure 2.1 shows an example of a Bayesian network model representing a factorization of the joint distribution $p(x, z, y)$. This model consists of random variables and parameters, where w is a parameter and z, y_n and x_n are variables, represented with circles. To illustrate several nodes (variables or parameters), but in a compact way, a plate is added, which surrounds the affected nodes with a box, with N being the number of times that the nodes will be repeated within it, as it is visible in the example.

There are different types of variables, namely: latent (hidden) or observed. Variables can also be either discrete or continuous. The observed variable, x_n, is represented in figure 2.1 by shading the corresponding node, whereas the latent variables are just open circles, z_n and y_n. Finally, there are parameters, w, which in other notations are represented with an open circle, but in Bishop et al. (2006) they are smaller solid circles. We chose to adapt the latter notation. These variables are considered parameters because we estimate or identify the most likely values of that variable, instead of determining their posterior distribution as we do for latent variables. This was the notation used throughout the thesis for expressing graphical models.

By translating dependencies expressed in the probabilistic graphical model of figure 2.1, the factorization of the joint distribution of the model, given the parameters w, is in equation 2.1.

$$p(x_n, y_n, z|w) = p(z|w) \prod_{n=1}^{N} p(y_n|z)p(x_n|y_n) \tag{2.1}$$

A joint distribution, with the variables x, y and z, can be factorized in various ways, by connecting these variables, the probabilistic graphical model specifies how a joint distribution

factorizes. This means that this distribution will be represented by a product of smaller factors, which correspond to conditional probability distributions.

However it is also important to comprehend how each factor is created. A good way to do this is with the generative process of the model. A generative process of the model in figure 2.1 can be something like the following:

1. Choose $z|w \sim Dir(z|w)$.
2. For each of n:
   (a) Choose $y_n|z \sim Binomial(y_n|z)$
   (b) Choose $x_n|y_n \sim Binomial(x_n|y_n)$

Generative processes explain how observable data, such as x_n in our example, is randomly generated, typically given some latent variables, such as z and y_n. Another approach is called a discriminative approach, where it provides a model only for the latent variables conditioned on the observed variables, which contrasts with generative models. In this generative process, the variable z is sampled from a Dirichlet distribution with parameter w. Similarly, the variable y_n is sampled according to a Binomial distribution with parameter z. Both these distributions will be discussed in the next section.

The use of generative processes to complement a model provides additional details on the latter. So, we are going to make use of generative processes to present our model.

2.1.2 Approximate Inference

In this section, some methods to perform approximate inference will be described. If we want to perfom exact inference, one can follow a Bayesian inference, which uses the Bayes theorem to compute the posterior distribution of variables, this theorem says that:

$$posterior = \frac{prior \times likelihood}{evidence} \tag{2.2}$$

When the posterior is intractable, too difficult to compute, we can use an approximate inference approach. An example on how the posterior can be intractable will be given when referring the LDA algorithm (Blei et al. (2003)) in the next section.

Two different methods for approximate inference will be presented: variational inference and Gibbs sampling. The first one is used in the LDA algorithm, which is the starting point for the proposed model and the second one is sometimes used in models in the related work.

Variational inference

One of many possibilities of variational inference is to perform the variational Bayesian Inference. The reason why variational Bayes is chosen is because it provides a locally-optimal, exact analytical solution to an approximation of the posterior. The results obtained have a similar accuracy as in sampling methods, such as Gibbs sampling, however variational Bayes is considerably faster, even though it usually requires a larger amount of work. A particular

sub-class of variational inference algorithms are mean-field methods, which assume that the approximate distribution is fully factorized.

Given an intractable posterior distribution, an approximate distribution will be selected from a tractable family, $q(Z)$, with variational parameters that need to be optimized. The main goal is to make this variational distribution as close as possible to the true posterior $p(Z|X)$, which is why we use the Kullback-Leibler (KL) divergence (MacKay, 2003) to measure the distance between the two distributions. Where X are the observed variables and Z the model parameters.

So, the goal is to minimize the KL divergence, however this cannot be minimized directly, which is why we find a function that we can minimize, the one that is in equation 2.3.

$$\begin{aligned}
\mathbb{KL}(q(Z)\|p(Z|X)) &= \int q(Z)\log\frac{q(Z)}{p(Z|X)} \\
&= \mathbb{E}_q\left[\log\frac{q(Z)}{p(Z|X)}\right] \\
&= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z|X)] \\
&= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q\left[\log\frac{p(Z,X)}{p(X)}\right] \\
&= -\underbrace{(\mathbb{E}_q[\log p(Z,X)] - \mathbb{E}_q[\log q(Z)])}_{\mathcal{L}(q)} + \underbrace{\log p(X)}_{const.}
\end{aligned} \tag{2.3}$$

Minimizing the KL divergence between the variational distribution q(Z) and the true posterior distribution $p(Z|X)$ is then equivalent to maximizing $\mathcal{L}(q)$, which is called the evidence lower bound or log marginal likelihood. The fact that $\mathcal{L}(q)$ is a lower bound on the log marginal likelihood, $\log p(X)$, can be verified by making use of Jensen's inequality (Kuczma, 2009), which established that $\log\mathbb{E}[p(X)] \geq \mathbb{E}[\log p(X)]$, thus resulting in equation 2.4. The integration value is $[-\infty, +\infty]$, which is why we chose not to represent it in the equation.

$$\begin{aligned}
\log p(X) &= \log\int p(Z,X) \\
&= \log\int\frac{q(Z)}{q(Z)}p(Z,X) \\
&= \log\mathbb{E}_q\left[\frac{p(Z,X)}{q(Z)}\right] \\
&\geq \underbrace{\mathbb{E}_q[\log p(Z,X)] - \mathbb{E}_q[\log q(Z)]}_{\mathcal{L}(q)}
\end{aligned} \tag{2.4}$$

The inference problem then becomes an optimization problem, where if the lower bound is maximized and the KL divergence minimized, then it will be produced a very tight lower bound, with results close to the ones obtained with exact inference. To optimize the lower bound, the coordinate ascent algorithm will be used, that iteratively optimizes the each

variational parameter of the approximate posterior distribution, while the others are fixed, until the convergence to a fixed value. This algorithm was chosen due to its simplicity and efficiency to solve the problem at hand.

The next step is to expand this lower bound in terms of the model parameters and the variational parameters, so that it is possible to maximize it with respect to the latter. In the description of the LDA model, (Blei et al., 2003), there will be a further explanation on this step.

Expectation-maximization, EM, is a method used to estimate the parameters and hidden variables, when there is missing data and/or latent variables. If the variables and the parameters were known it would be easier to use a different method for the estimation. The EM algorithm consists in alternating between inferring the missing values given the parameters (which is the E-step), and then optimizing the parameters given the values obtained from the previous step (this is the M-step). For the E-step Bayesian inference could be used to infer the variables, and for the M-step, for example, the Maximum Likelihood to estimate the parameters.

Gibbs Sampling

The Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm and another popular method used for approximate inference. This method supports that it is easier to sample from a conditional distribution than to marginalize by integrating over a joint distribution. The basic idea is the sampling of each variable in different turns, where it is conditioned by the values of the rest of the variables of the distribution. The expected value of a variable is then given by averaging over all the samples. It is a random method given that the values of the variables are determined randomly. Furthermore, this method does not consider all samples, given that it ignores a number of samples in the beginning for being too far from the desired distribution. This is called the burn-in period. Only when the Markov chain has converged (or mixed), we can start collecting samples.

The downside of this method is that it might need infinite samples so that it converges to the correct result. With this method it is possible to obtain the most accurate results, however its computation time can be too high, which makes it less suitable for larger-scale models such as the one proposed in this thesis.

A variant of this method is the collapsed Gibbs sampling (Porteous et al., 2008), which is much more efficient, due to sampling in a lower dimensional space.

The variational inference method was still chosen over the collapsed Gibbs sampling to be applied in the proposed model, given the simplicity of already being implemented in the classic LDA Blei et al. (2003).

2.1.3 Natural Language Processing

Natural Language Processing (NLP) (Chowdhury, 2003; Jurafsky and Martin, 2009) is a subfield of artificial intelligence whose goal is to allow computers to understand and manipulate human language, in the form of text or speech. Some of its applications include summarization (Rau et al., 1989), information retrieval (Voorhees, 1999), speech recognition

(Martin and Jurafsky, 2000), human computer interaction (Allen et al., 2001), among others. Researchers in this area aim to study on how human beings understand and use language.

The main challenge of NLP is that, in opposition to programming languages, natural languages are ambiguous. In fact, ambiguity can occur at different levels (e.g. morphological, syntactical, semantic, speech).

The morphological level deals with the identification, analysis and description of the structure of words. It might cover the task of lemmatization, which consists in the normalization of words into their dictionary form. This means that it converts nouns and adjectives to their masculine and singular forms (e.g. cars becomes car, or feet becomes foot) and verbs to their infinite form (e.g. walks to walk or knew to know). This level might identify and remove stopwords. A word is considered to be a stop-word if it is one of the most common words in a language. Typical stop-words are articles, prepositions and, depending on the purpose, very common verbs or adjectives may also be considered as such.

The syntactical level studies the structural relationships between words in a sentence. It may include the task of part-of-speech (POS) tagging, which is the process of assigning to a word a particular part-of-speech, depending on its definition and context it is in. Parts-of-speech include, for instance, the categories of nouns, verbs, adjectives or adverbs, typically identified by a predefined tag. For instance, the sentence *I deposit my money in the bank*, can be POS-tagged as follows: *I_PRP deposit_VBP my PRP money_NN in_IN the_DT bank_NN*, which clearly identifies, for example, *bank* as a noun.

The semantic level deals more with the meaning of words and sentences. Word Sense Disambiguation (WSD) (Navigli, 2009) and Word Sense Induction (WSI) (Navigli, 2012) both aim to identify the most suitable meaning of a word in context. The main difference between them is that WSD resorts to a sense inventory (list of senses of a given word, available in traditional dictionaries, but ignore named entities), whereas WSI automatically discovers word senses from a text. For example, in the sentence *I deposit my money in the bank*, the word *bank* should refer to a financial institution.

In WSD, polysemous senses can be created at any level of granularity, thus leading to possibly very fine-grained sense distinctions. For the sense distinctions to be more coarse one can follow two approaches: a manual one (creating sense distinctions by iteratively submitting new partitions of senses for the given word to sense annotators) or an automatic one (clustering, by automatically mapping WordNet senses to the Oxford Dictionary of English semantically similar senses using WSD techniques). However, it is a very challenging task, mostly due to the representation of senses, but the higher the high-quality knowledge the higher the performance.

WSD is a key step for performing other natural language processing tasks, such as machine translation (Chan et al., 2007) or information retrieval (Zhong and Ng, 2012). On the latter, queries can be expanded with synonyms and other related words, thus improving recall, and search results can be narrowed towards the desired senses, thus improving precision.

WSI uses unsupervised techniques to automatically identify the set of senses denoted by a word. Instead of assigning words to an existing sense inventory, it induces word senses from raw text by clustering word occurrences on the basis of the distributional hypothesis (a given word tends to co-occur with the same neighbouring words). WSI has the potential to harvest even more senses than those available in a traditional predefined sense inventory.

In the proposed model we apply WSD, more specifically, we use the algorithm Adapted Lesk for WordNet, which is an improved version of Lesk's (Lesk, 1986), developed by (Banerjee and Pedersen, 2002). It identifies the best sense for a specific word, given the words of its document, by computing the overlap between synset glosses and related words. Adapted Lesk was chosen because it is the most widely used WSD algorithm, it is very easy to understand, and it was available out-of-the-box. However, there are many algorithms developed, that use WordNet as a standard sense inventory, among which supervised (Pedersen and Bruce, 1997), unsupervised (Yarowsky, 1995) and knowledge-based approaches (Resnik, 1997; Navigli and Velardi, 2005; Agirre and Soroa, 2009). The latter have been more successful (Agirre et al., 2009; Ponzetto and Navigli, 2010), especially in the disambiguation of all words and not just a lexical sample. They typically exploit the structure of WordNet itself, among other sources. There are knowledge-based WSD algorithms that consider semantic classes and their implications (Resnik, 1997), or measures that exploit WordNet graph structure (Navigli and Velardi, 2005; Agirre and Soroa, 2009).

WordNet (Miller, 1995; Fellbaum, 1998) is commonly used as a standard sense inventory in English WSD, which turns the goal into the assignment of a word, in context, to one of its possible WordNet senses. WordNet is also a large lexical database of English, somewhat like a dictionary or a thesaurus. It is structured in the so-called synsets — groups of synonymous words that may be seen as the representation of concepts by their possible lexicalizations. Apart from its words, a synset has a part-of-speech and a gloss, which is like a dictionary definition. In fact, WordNet can be seen as a dictionary, because the presence of each word in a synset denotes a different word sense. In the scope of this work, synsets are linked by conceptual-semantic and lexical relations, like part-of, hyperonymy and hyponymy. The latter are super-subordinate relations. Synsets and relations may be exploited to identify which words are similar to others in the same topic. Moreover, for a specific word, synsets including it are ordered according to the most frequent senses and, at the same time, words in the same synset are ordered by their frequency to denote the synset meaning. These frequencies were computed based on the annotations of SemCor[1],where each word is connected to the most suitable WordNet synset. Figure 2.2 is an example of what a WordNet response would look like. In that figure, a query for the word *dice* was performed in the web search engine. We can see that there is more than one synset with this word, each with different synonym and relations.

SemCor (Miller et al., 1994) is an annotated corpus created by the same researchers that developed WordNet. This corpus is a subset of the English Brown Corpus and all the content words are annotated with POS, lemma, and WordNet synset. However, as expected, it is not large enough to cover all the information in WordNet.

Nowadays, there are many toolkits available, in different programming languages, whose purpose is to perform natural language processing (e.g. NLTK[2], Standford NLP[3], OpenNLP[4], among others). We now enumerate some of those toolkits that were used in this work:

[1] `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`
[2] `http://www.nltk.org/`
[3] `http://nlp.stanford.edu/`
[4] `https://opennlp.apache.org/`

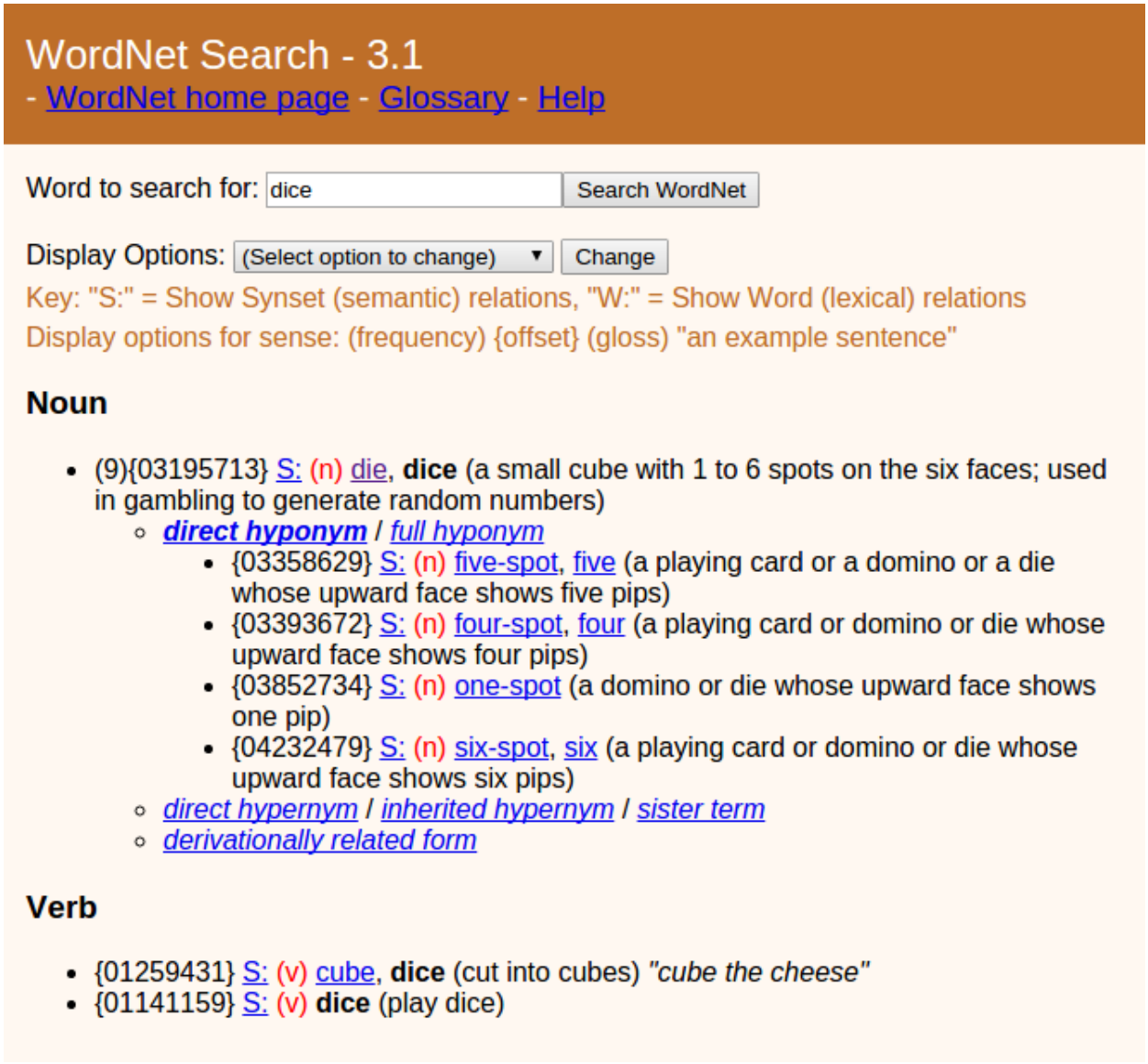

Figure 2.2: Example of a query in WordNet search engine.

- NLTK is implemented with the Python programming language. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, lemmatization, POS tagging, parsing, and semantic reasoning. Given all its tools, this toolkit was used to tag, lemmatize and to access the WordNet corpus. It is easy to use and provides fast access to all the features that WordNet has. It also contains the implementation of a few WSD algorithms out-of-the-box.

- OpenNLP is used for a wide range of languages. It allows, amongst others, part-of-speech tagging (assigns parts of speech to each word) and chunking (splits sentences in syntactically correlated parts of words), which were the ones used in SemEval. It is mainly oriented to be used with Java programming language.

- CoreNLP[5], toolkit from Stanford, has basically the same tools as OpenNLP, however, it can perform lemmatization.

It is common to organize events aimed at evaluating specific NLP tasks, like for example SemEval. SemEval 2015[6] is a shared task on semantic evaluation, where participants may create different semantic analysis systems, for a wide range of tasks available. Our group participated in task 2 (Semantic Textual Similarity), where the main goal was to score two sentences based on their similarity. By entering this contest, it was necessary to deal with several tools. For example, regarding the preprocessing phase a method to identify named entities was used, OpenNLP POS tagging, stop-words removal, OpenNLP Chunking and lemmatization (with NLTK).

However, given that NLP is a challenging research topic, where there has been continuous improvement, none of them is perfect.

2.2 Related work

This section presents the related work on this subject, divided into different categories: LDA based topic models, which also includes models used to perform WSI and WSD, and other topic models which are not based in LDA. Whenever there is a topic model based on LDA and we present its graphical model or generative process, the differences between the classic algorithm are highlighted. Also, there is a section explaining some methods on how to evaluate topic models, both the model and the results produced by it.

2.2.1 LDA based Topic Models

The Latent Dirichlet Allocation (Blei et al., 2003) is currently the most popular topic model and it is the starting point for the model proposed in this thesis. It should thus be described in detail, so that its process becomes clearer.

By using this algorithm it is possible to describe a document as a mixture over latent topics, given that a topic is a distribution over words. The main goal is to automatically

[5] `http://nlp.stanford.edu/software/corenlp.shtml`

[6] `http://alt.qcri.org/semeval2015/`

assign these documents with topic distributions, where a document may contain several topics that were learned with the help of statistical inference. LDA adopts the bag-of-words assumption, since it does not take in consideration the order of the words in a document.

This is the generative process of LDA, for a document $\mathbf{w}$ of the corpus $\mathcal{D}$:

1. Choose $\theta \sim Dir(\alpha)$.
2. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Mult(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability distribution conditioned on the topic assignment z_n.

There is a need to understand what each variable represents, so that it is possible to understand the generative process. A word is the item in a vocabulary from 1 to V, a document consists of N words, $\mathbf{w} = (w_1,\ w_2,\ .\ .\ .\ ,\ w_N\ )$ and a corpus is a set of M documents, $\mathcal{D} = (\mathbf{w}_1,\ \mathbf{w}_2\ ,\ .\ .\ .\ ,\ \mathbf{w}_M)$. Also, θ represents the distribution over topics of a document, α is a parameter, β are the topics distributions over words and it is represented in a matrix of $K \times V$, where K is a fixed value that indicates the number of topics that will exist and automatically defines the dimensionality of the topic assignment vector, z.

Analyzing the generative process more thoroughly, in step 1 there will be a draw from a Dirichlet distribution, with a parameter α. The results of this step represent the topics distribution over a document, θ. Then, in step (a), a topic z_n, will be selected given the multinomial distribution with parameters θ, so that, in step (b), a word is chosen depending on z_n and the multinomial distribution distribution over words of the selected topic z_n. A multinomial distribution is the generalization of the binomial distribution. In the binomial only two events are considered and the probability distribution is the successes given these events. Whereas, in the multinomial there are N events, each with its own probability, and the distribution will highlight the success of one of the events.

It was referred previously that θ is drawn from a Dirichlet distribution. The values that the variable θ can take are confined to a $(K-1)$ simplex, where K is the dimensionality of the distribution.

The Dirichlet distribution was chosen because, since it belongs to the exponential family, it has finite dimensional sufficient statistics, and because it is the conjugate prior to the multinomial distribution. All of these characteristics make the Dirichlet a convenient choice as a prior given that it simplifies the algorithm of statistical inference. The plus of belonging to the exponential family is that it allows the distribution's probability density function to be represented solely by its sufficient statistics, which are enough to describe the distribution without presenting all the data. For example, in a coin toss, which is a Bernoulli distribution that belongs to the exponential family, the sufficient statistics are the number of times that heads occurred and this is enough to specify its distribution. The probability density of the Dirichlet is shown in equation 2.5.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{(\alpha_i - 1)} \tag{2.5}$$

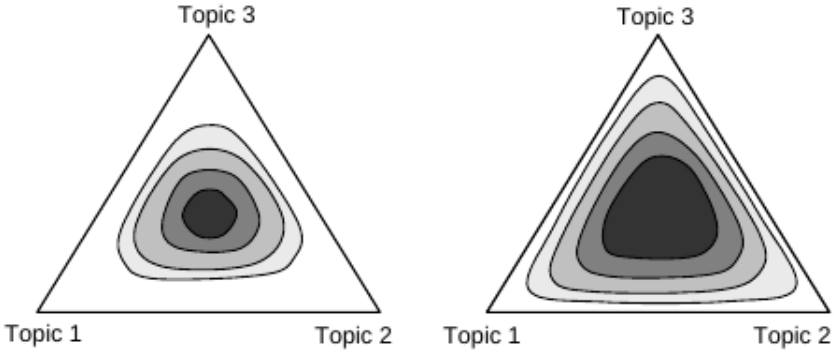

Figure 2.3: Symmetric Dirichlet distribution with k = 3, where dark colors represent a higher probability. On the left: $\alpha = 4$ and on the right: $\alpha = 2$.

The α parameter is called a hyperparameter, to distinguish it from the model's parameters, and each α_i represents the number of times that topic i was sampled in a specific document, without taking into account existing knowledge from that document. Usually, a symmetric Dirichlet distribution, where all the α_i values are all the same, is used.

In figure 2.3, there is a 2D simplex, which means it involves 3 topics[7], where each topic corresponds to one vertex. Given a value for α it is feasible to visualize the probability of drawing the topics. The greater the value for α the higher is the concentration of the probability of the topics is in the center, which means that all the topics collected will have very similar probabilities. If $\alpha = 1$, then the simplex will have a uniform color, meaning that it has the same probability in every position. It also leads to topics with the same distribution, which is sometimes not the desired given that the documents will not be characterized by a specific topic. When $\alpha < 1$, the highest probability will become located in the corners of the simplex and the lowest in the center, which is the reverse of what happens in figure 2.3. This value for α is the most frequent choice by the users, since it generates topic distributions where a topic has a more distinguished probability from the others. This means that the hyperparameter α is responsible for determining the sparsity of the topics.

Figure 2.4 shows the probabilistic graphical model of LDA in a three-level representation. The first level is the sampling of the corpus-level parameters, which are α and β, chosen only once for a corpus. Then there is the document-level variable, θ, which is going to be sampled once for each document. Finally, there are word-level variables, z_n and w_n, that will be drawn for every word present in a document. The fact that the topic node is sampled more than once for a document is what allows for a document to be characterized as a mixture of topics, i.e. each word in a document can be assigned to a different topic.

The joint distribution of the model is obtained by splitting the $p(\theta, z_{1:N}, w_{1:N}|\Theta)$ into separate factors according to the graphical model, where there is a factor for each variable

[7]This dimensionality was chosen for demonstration purposes only, given that it is visually easier to understand.

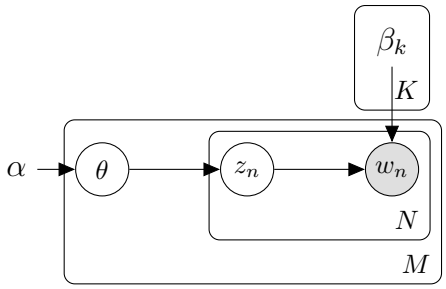

Figure 2.4: Graphical model representation of LDA.

and they are conditioned by their parents, which results in equation 2.6.

$$p(\theta, z_{1:N}, w_{1:N}|\Theta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta_{1:K}) \tag{2.6}$$

The next step is inference and parameter estimation.

According to Bayes theorem, the formula for the posterior distribution is given by equation 2.7.

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{2.7}$$

Which in the LDA model translates to equation 2.8.

$$p(\theta, z_{1:N}|w_{1:N}, \Theta) = \frac{p(\theta, z_{1:N}, w_{1:N}|\Theta)}{p(w_{1:N}|\Theta)} \tag{2.8}$$

The numerator is the joint distribution and the denominator is obtained by marginalizing over θ and z in the joint distribution by using the sum rule of probability (Murphy, 2012). The sum rule is different when dealing with discrete or continuous variables. If a variable is continuous, then it is necessary to integrate through all of its values, whereas if it is discrete then we have to sum all its values. Equation 2.9 shows the sum rule being applied.

By analyzing the denominator of equation 2.8 in equation 2.10, it is clear that due to the coupling of θ and β, when it comes to calculate the log of this probability it is not possible to separate θ and β, so this computation becomes intractable for exact inference. However, there are several available methods to solve this problem by using approximate inference, like the Laplace approximation, variational inference, and Markov chain Monte Carlo. The one chosen for this model was the variational Bayesian inference.

$$p(w_{1:N}|\Theta) = \int_\theta p(\theta|\alpha)\sum_{z_{1:N}}\left(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta_{1:K})\right) \tag{2.9}$$

$$= \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\int_\theta\left(\prod_{i=1}^{K}\theta_i^{(\alpha_i-1)}\right)\left(\prod_{n=1}^{N}\sum_{i=1}^{K}\prod_{j=1}^{V}(\theta_i\beta_{i,j})^{w_n^j}\right)d\theta \tag{2.10}$$

To obtain a tractable family of approximate distributions q for the true posterior distribution of the latent variables, it was necessary to remove the problematic edges and nodes. In this model the edges between θ, z_n and w_n and the node w_n, were the troublesome so they were removed, as figure 2.5 shows.

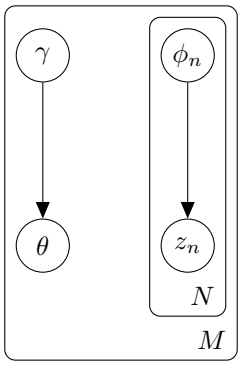

Figure 2.5: Graphical model of the variational approximation of LDA.

The distribution q is referred as the variational distribution and it is represented in equation 2.11, where γ is the Dirichlet parameter, ϕ the multinomial parameter and these are called the variational parameters.

$$q(\theta, z_{1:N}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{K} q(z_n|\phi_n) \tag{2.11}$$

The main goal of variational inference is to obtain the optimal values for the variational parameters, which consists in maximizing the lower bound on the log marginal likelihood and, at the same time, minimizing the KL divergence between the variational distribution q and the true posterior distribution(Bishop et al., 2006; Jordan et al., 1999). As it was explained previously, it is necessary to obtain the lower bound the log marginal likelihood of the data.

Making use of Jensen inequality, the lower bound is given by equation 2.12.

$$\log(w_{1:N}|\Theta) = \mathbb{E}_q[\log p(\theta, z_n, w_n|\Theta)] - \mathbb{E}_q[\log q(\theta, z_n)] \tag{2.12}$$

The next step is to expand the lower bound by using the factorizations of $p(\theta, z, w|\theta)$ and $q(\theta, z)$, which translates in equation 2.13.

$$\begin{aligned}\mathcal{L}(w_{1:N}|\Theta) = \mathbb{E}_q[\log p(\theta|\alpha)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(w_n|z_n, \beta)] \\ - \mathbb{E}_q[\log q(\theta|\gamma) - \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_n|\phi_n)]\end{aligned} \tag{2.13}$$

The probabilities of each term are known and are given by the distributions, accordingly to their definition. So, in equation 2.13, each term is computed according to the distributions defined.

To infer the variational parameters γ and ϕ, the lower bound is maximized in turn with respect to each. Taking the derivative of the lower bound, and setting this equation to zero gives the update equations that allow the estimation of the variational parameters. However, when the parameter follows a multinomial distribution this maximization is a constrained optimization problem. So the appropriate Lagrange multipliers have to be added to the lower bound with respect to the variational parameter.

For the estimation of the model parameters, the variational parameters have to be known, so it is used the variational EM procedure. In the E-step the optimal values for the variational parameters are found and the M-step maximizes the lower bound on the log marginal likelihood with respect to the parameters α and β. This step assumes that the variational parameters are fixed, given the values obtained from the E-step. This algorithm alternates between the E-step and the M-step until it reaches a specific conversion criteria.

Many authors design their semantic approaches based on the LDA model, which was described above. The remainder of this section is dedicated to introduce some of those models.

The model described in Chemudugunta et al. (2008) uses the LDA algorithm as ground model. In this model, the authors modified the LDA model, instead of having a variable that represents the distribution of words over a topics, it is now the distribution of concepts. A concept consists in a set of words, that only belong to a small subset of the vocabulary and these are defined a priori. The concepts are part of an ontology, which is a collection of human-defined concepts, with a hierarchical structure. The authors only consider ontological concepts and associated vocabulary. They view concepts as topics but with restrictions, for example words that are not mentioned a priori will have 0 probability. The model can be viewed in figure 2.6, where Ψ is the word-concept distribution and β_Ψ is a Dirichlet prior on Ψ. The rest of the variables are the same as in the LDA model. In orange is represented the difference of this model to the classic LDA.

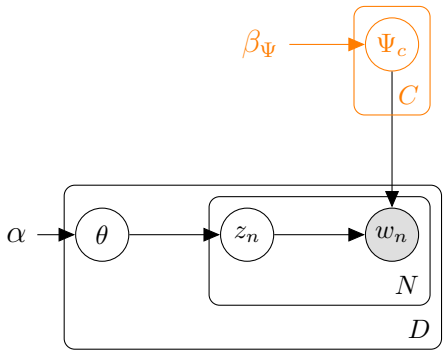

Figure 2.6: Graphical model representation of the Concept model.

For the inference process, the authors make use of Gibbs sampling, which assigns concepts to words in documents. However, this method has the restriction that a word can only be linked to a concept if it is assigned to that concept in the ontology. The result that this model provides is the same as in the traditional LDA, in which is each word of the corpus will be assigned to a concept from the ontology.

To evaluate the quality of the concept models they use perplexity, which is a quantitative measure used for comparing language models. It is the distance between the word distribution learned and the distribution of words in test documents. If this measure produces low scores, it indicates that the distribution produced by the model is close to the text. In their experiments they used the following corpora: TASA, CIDE[8], and ODP[9]. They also evaluated it in a visual way, by comparing the different models when tagging web pages or texts.

Another model that makes use of LDA is presented in Griffiths et al. (2004), which proposed a generative model with two different components that considers both short-range syntactic dependencies, which do not go beyond a sentence, and long-range semantic dependencies, from different sentences throughout the document. The main goal is to discover syntactic classes and semantic classes of words without having prior knowledge on syntax or semantics, besides the statistical dependencies. To achieve this, the authors use two different methods, Hidden Markov Models (HMM) and topic models, like LDA (Blei et al., 2003).

The HMM will split the sentences into function words and group them together given their syntactic role, which can vary in different contexts. The topic model's function is to assign determiners to topics, although they lack in semantic content. With the HMM they obtain syntactic classes and with the topic model, semantic topics.

To combine the two components they use a composite model, where it is replace one of the probability distributions over words from the syntactic model with the semantic model. This merging of methods allows the syntactic model, HMM, to determine when a content word should be emitted and the semantic model, a topic model, can choose which word to emit.
The generative process is as follows, where in orange is the difference of this process to the classic LDA:

1. Sample θ from a Dirichlet(α) prior
2. For each word w_i
   (a) Draw z_i from θ
   (b) Draw c_i from $\pi^{(c_i-1)}$
   (c) If $c_i = 1$, then draw w_i from $\phi^{(z_i)}$, else draw w_i from $\phi^{(c_i)}$

In the generative process above, ϕ is a distribution over words for both topics and classes, c_i is a sequence of classes, where the first is designated the semantic class, and π is the distribution that the classes follow. So, to generate sentences is this model one has to follow a path through the model, where first a word will be selected from the distribution associated with each syntactic class, and then a word from a topic, which follows a distribution.

The authors make use of Gibbs sampling to perform the inference in this model, which will, in every sample, assign to a word a class and a topic. For the evaluation phase, they

[8] www.cambridge.org/elt/cide
[9] available at http://www.dmoz.org

tested the composite model and the LDA model on the Brown corpus[10], a concatenation of the Brown and TASA corpora and also, NIPS papers[11]. After this, the results were evaluated based on their accuracy.

Word Sense Induction In Tang et al. (2014), a similar approach to the one presented previously was found. This model incorporates word sense as a latent variable in the LDA model, which replaces the latent variable word. It is unsupervised and does not depend on external resources, like WordNet or Wikipedia, for the word senses, because it induces them automatically from the corpora. It was also the first to perform WSI with the help of a topic model.

They use Hierarchical Dirichlet Processes (HDP) (Teh et al., 2006), as a prior to induce the word senses, in an unsupervised way, given that it avoids linking the number of word senses to each word.

Two models are specified in this article, the Standalone SLDA (SA-SLDA) and the Collaborative SLDA (CO-SLDA). In SA-SLDA, WSI and the document representation are independent, which means that the output from the WSI, word senses, is the input for the document representation that consists in assigning topics to a word.

Whilst in CO-SLDA, WSI and the document representation occur simultaneously. The main purpose of this model is to have constant feedback on WSI whether the topics assigned to a word are appropriate or not. For that, this model uses the output of the SLDA (sense based model), which are topics of senses, as a feedback for WSI, so that it is possible from then on to infer, simultaneously topics and senses.

For the inference the authors used the collapsed Gibbs sampling and to evaluate the topics the following datasets, Reuters[12] and TDT4[13], in a clustering task F measure.

Word Sense Disambiguation The model proposed by Guo and Diab (2011) integrates words semantics explicitly in the topic modelling framework. As a lexical resource the authors use WordNet, so all the senses come from there, but also their definitions will be used and will be treated as documents. These definitions will be useful in a way that provide a better understanding on the semantics of a word.

A sense node was added to the LDA model, between the topic node and the word node, in order that it is possible to disambiguate the meaning of a word. When it comes to choosing a word, it depends on the relatedness of the sense and its fit to the document context.

They adopted the WSD local window strategy, where a fixed window size is chosen so that only K neighboor words are considered semantically related to that word.

This model has two components. For the window size. In first part a word is generated from a specific sense and the words it can emit. This sense is drawn from a specific topic and a distribution of senses over topics. The second component is the definition one, where it is possible to draw a sense based on its definition.

[10] http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[11] http://papers.nips.cc/

[12] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[13] https://catalog.ldc.upenn.edu/LDC2005S11

Collapsed Gibbs sampling is also used here and, during the inference, the choice of the topic/sense for each word, requires that the topic is supported by the context and that the sense definition matches that topic. For the experiments several datasets were used, such as, Brown Corpus, New York Times (NYT)[14] and WordNet definitions. To validate the model the authors used two tasks, a text categorization task for evaluating the quality of the values of topic nodes and a WSD task for evaluating the quality of the values of the sense nodes.

In an attempt to include semantics in topic modelling and, at the same time, perform WSD, Boyd-Graber and Blei (Boyd-graber and Blei, 2007) presented LDAWN, a modified LDA algorithm that includes a hidden variable for representing the sense of a word, according to WordNet. Each topic consists of a random walk through the WordNet hypernymy hierarchy, which is used to infer topics and their synsets, based on the words from documents. LDAWN was also applied to WSD, although its authors accept the worse performance when compared with state-of-the-art WSD algorithms. One of the proposed solutions is to acquire local context to improve WSD, in the future.

Instead of words, the produced topics are also distributions over concepts (synsets) and, similarly to LDAWN, it exploits WordNet and modifies the basic LDA by adding a sense variable. They benefit from similar words in the same topic to improve WSD.

They use Gibbs sampling for the inference. When it comes to evaluation they used two datasets, SemCor[15] and British National Corpus[16].

The evaluation of the system was performed on real-world WSD data, by comparing the accuracy results obtained with WSD.

2.2.2 Other Topic Models

The model presented in Li et al. (2010) chooses the best sense based on the conditional probability of sense paraphrases given a context. Its purpose is to automatically determine the correct sense for a target word given the context, with the use of topic models. It is resource poor and only requires a large unlabeled corpus, to estimate topic distributions, and paraphrases, which can be user supplied or extracted from existing resources, for possible target senses. The senses described in the article are a collection of paraphrases that capture its meaning.

They created three different models. The first model needs to have prior knowledge of the distribution over the senses and will maximize the conditional probability of a sense. The second model does not need to know the prior distribution over the senses and will indirectly maximize the sense-context probability. Model III calculates probability of a sense given a context according to the component words of the sense paraphrase. By maximizing the conditional probability of senses given a context, it is possible to assign the most appropriate sense to a word.

In Model I and II sense paraphrases and context are treated as documents. The context is taken into account because it is where the word occurs. However in Model III contexts are treated as documents but sense paraphrases are treated as sequences of independent words,

[14] `https://catalog.ldc.upenn.edu/LDC2008T19`

[15] `http://www.gabormelli.com/RKB/SemCor_Corpus`

[16] `http://corpus.byu.edu/bnc/`

these independent words will capture the meaning of the idioms. For the inference phase they make use of the Gibbs sampling.

To evaluate the model they use three different tasks, coarse-grained WSD, fine-grained WSD and task idiom detection, this involves distinguishing literal from non-literal sense. The first two tasks are where the first two models are tested, since the third model was specifically created to perform idiom detection. For this, they used two datasets, Semeval 2007 task-07 dataset and Semeval 2007 task-1.

Rajagopal et al. (2013b) proposed a commonsense knowledge based algorithm for document topic modeling and, unlike probabilistic topic models, does not involve training nor depend on word co-occurrence and particular word distributions. Commonsense knowledge is the way humans understand the world, which is acquired in the daily life. To improve the most common topic models, it obtains knowledge of word meanings that are in a commonsense knowledge database, structured in the INTELNET (Olsher, 2013) formalism. Instead of a bag-of-words model, they propose a bag-of-concepts, where a lexical item is an index of a set of semantic atoms, which contain a piece of knowledge regarding a specific concept. A concept is defined by being either a single lexical item or a multi-word expression. The first step of their model is to extract the commonsense concepts from a natural language text and this is done by using a graph-based approach, which is further explained in Rajagopal et al. (2013a). The knowledge base concepts can be extracted from different sources, one of those is WordNet. To generate the topics, they used a clustering technique, Group Average Agglomerative Clustering (GAAC), which proved to have the highest accuracy. For the evaluation part they chose standard measures, like precision, recall and F-measure and for that they used the Brown corpus.

2.2.3 Comparison of different models

After gathering so much information on all these different models, they should be compared to each other, in order to highlight the major differences between them. The models were analyzed based on two aspects: the methods and tools used and the datasets chosen to validate them.

The first table, table 2.1, is organized the following way: each line corresponds to a method/tool used in, at least, one of the models presented in chapter 2, each column represents the authors of a model. The table was filled in with ✓that show which methods where used by the different authors. However, in this table not all the authors of models where mentioned, such as Blei et al. (2003), given that it was the starting point for many models..

By analyzing the content of this first table, it is feasible to understand that not as many authors use external resources, such as WordNet, to retrieve semantic knowledge. The majority uses the LDA model as its foundation, which indicates that the authors have this model in high consideration. To perform the approximate inference, the most common choice is the Gibbs sampling or its variant, Collapsed Gibbs sampling.

On table 2.2, the data is organized in a similar way to table 2.1, except that now each line corresponds to a dataset used with the models proposed by the authors in the columns. All the authors are presented here.

		Chemudugunta et al. (2008)	Tang et al. (2014)	Guo and Diab (2011)	Boyd-graber and Blei (2007)	Li et al. (2010)	Griffiths et al. (2004)	Rajagopal et al. (2013b)
WordNet	Senses			✓		✓		
	Definitions			✓				
	WORDNET-WALKs				✓			
WSD				✓	✓	✓		
WSI			✓					
HDP			✓					
LDA model		✓	✓	✓	✓		✓	
Gibbs sampling		✓			✓	✓		
Collapsed Gibbs sampling			✓	✓				
Common-sense knowledge								✓
Group Average Agglomerative Clustering (GAAC)								✓

Table 2.1: Methods/Tools used by the models

Given the data presented on table 2.2, it was possible to retrieve a large list of datasets, which can be used in the proposed model. Overall, there was not a dataset that stood out, each model was tested different corpus.

Since most of the datasets on table 2.2 were not available for download, we did not have many corpus to choose from, so we could use with our model. Overall, we used SemCor, Reuters, 20 Newsgroups and the Associated Press (AP), all in English.

The Reuters corpus contains several news text divided into the 10 most populated categories, however given that its vocabulary is very specific to a determined domain and abbreviations are very common, the results obtained with it were not informative enough.

The 20 Newsgroups[17] is a popular dataset for experiments in text applications of machine learning techniques (Baker and McCallum, 1998). It contains 20,000 documents, organized into 20 different newsgroups. This corpus replaced Reuters, because its outcome of the same tests was significantly better, in terms of interpretation.

Finally, AP is a large news corpus, from which we use only a part. More precisely, the sample data for the C implementation of LDA, available in David Blei's website[18], which includes 2,246 documents.

Both the AP corpus and the 20 Newsgroups were used in all the experiments performed in chapter 4, given that it is important to test our model with different data sources to verify if it behaves always the same way.

[17] `http://qwone.com/~jason/20Newsgroups/`
[18] `http://www.cs.princeton.edu/~blei/lda-c/`

	Blei et al. (2003)	Chemudugunta et al. (2008)	Tang et al. (2014)	Guo and Diab (2011)	Boyd-graber and Blei (2007)	Li et al. (2010)	Griffiths et al. (2004)	Rajagopal et al. (2013b)
Associated Press	✓							
TASA		✓						
CIDE		✓						
ODP		✓						
Reuters	✓		✓					
TDT4			✓					
Brown Corpus				✓			✓	✓
Brown Corpus + TASA							✓	
NIPS papers							✓	
New York Times (NYT)				✓				
SemCor					✓			
British National Corpus					✓			
Semeval 2007 task-07 dataset						✓		
Semeval 2007 task-1						✓		

Table 2.2: Datasets used by the models

2.2.4 Evaluation of Topic Models

As it happens in most research domains, different proposed solutions for the same purpose have to be evaluated according to certain guidelines, in order to select the best ones or to assess progress. The evaluation of topic models is not a trivial task, because it involves large collections of documents, sometimes on a specific domain, and sometimes subjective criteria around words and their relevance. There are however approximations that have been applied to evaluate, manually and also automatically, the quality of topic models, some of them presented in this section. Here are only referred the metrics that were used later to validate the proposed model in this thesis.

A standard measure for estimating the performance of a probabilistic model is the perplexity. It evaluates how well a probability model predicts the classification of a sample, by measure the log-likelihood of a held-out test set. A lower perplexity score indicates a better generalization performance.

In Blei et al. (2003), the authors use this measure to evaluate LDA, following to equation 2.14, where the numerator is the likelihood obtained for a sample.

$$perplexity(\mathcal{D}_{test}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d} \right\} \tag{2.14}$$

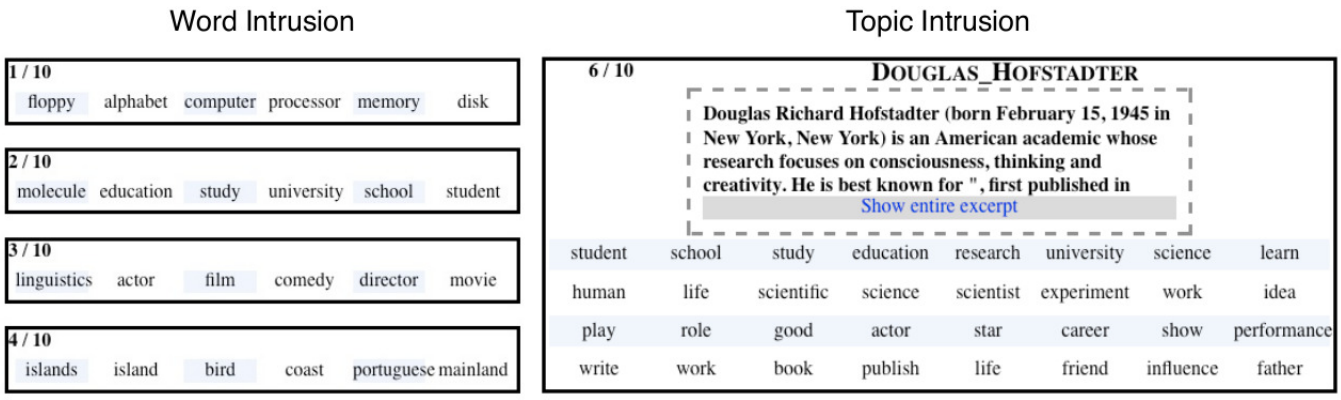

Figure 2.7: Example of a word intrusion task (on the left) and a topic intrusion task (on the right), from Chang et al. (2009).

Whereas in Chang et al. (2009) are presented quantitative methods to measure the semantic meaning in the inferred topics. For this, the authors designed two tasks, based on human evaluation, that evaluate the quality of the topics, as well as the capability to assign topics to documents. These tasks are called *word intrusion* and *topic intrusion*. Figure 2.7 gives an example of both tasks, where in the one on the left a word intruder must be identified, and in the one on the right a topic intruder must be recognized, based on the first sentences of a document. The first measures if the topics inferred by the model were semantically cohesive and if these topics are grouped in a natural way, for humans to understand. The second one measures whether or not the association of a document and a topic is coherent, but it requires reading a complete document to assess each evaluation, which makes it impractical in the scope of this thesis. The authors suggest evaluating by also using predictive metrics. With these, they can analyze how well the model can predict a test set, without seeing these documents after having learned its parameters from a training set. One of the predictive metrics used was the predictive log likelihood, also known as perplexity.

The tasks based on human evaluation were offered on Amazon Mechanical Turk, where other people, through the Internet and for a fee, performed these evaluations. Amazon Mechanical Turk has been used successfully, for instance, in the creation of gold-standard data for natural language processing.

The concept of mutual information was initially introduced by Fano (1961), where it states that if two words, x and y, have probabilities $P(x)$ and $P(y)$, then their mutual information, I(x,y), is defined by equation 2.15.

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2.15}$$

It compares the probability of observing words x and y together (the joint probability) with the probabilities of observing words x and y independently (chance). If the two words are associated, then the value of mutual information will be high.

According to Newman et al. (2011), word association is highly correlated with human-judged topic coherence. This was proven in Newman et al. (2010), where the authors asked

users to annotate topics on how coherent these were. Then, they compared the Spearman rank correlation values (Zar, 1998) obtained with all the measures, which included Pointwise Mutual Information (PMI), a measure that calculates the mutual information between words based on Fano (1961), and the annotators judgement. The conclusion reached was that PMI with Wikipedia was the most consistent performer, achieving the best results and approaching the manual evaluation.

PMI is calculated for each topic, based on the co-occurrence probabilities of every pair of its words in an external corpus. Equation 2.16 is the PMI's formula, where $p(w)$ is the probability of a word w (in our case, the number of Wikipedia articles using this word), and $p(w_i, w_j)$ is the probability of words w_i and w_j co-occurring (in our case, the number of Wikipedia articles using both words). It also uses the 10 most probable words of each topic, which results in 45 different pairs. The higher the results, the more coherent the topics are.

$$PMI - Score(\mathbf{w}) = \frac{1}{45} \sum_{i<j} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, ij \in \{1...10 \tag{2.16}$$

Mimno et al. (2011) presented an equation that also measures the co-occurrence, within the modelled documents, of pairs of words in the same topic. This measure is very similar to PMI but, in some situations, it achieved higher correlation with human judges. However, PMI is calculated based on an independent corpus of the used documents, which means that it evaluates topics as more generic instances and not just within the collection. So, even though the coherence measure, sometimes outperformed PMI, they both contribute with relevant information when evaluating the topics.

The higher the coherence value, the more coherent the topics are. Equation 2.17 shows the formula of this measure, where $D(v_m^{(t)}, v_l^{(t)})$ is the co-document frequency of two words, 1 is a smoothing count to avoid the logarithm of zero and $D(v_l^{(t)})$ is the document frequency of a word.

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{2.17}$$

Chapter 3

Semantic Topic Modelling

In this chapter, a more extensive explanation is presented on the reason that lead us to choose the subject, alongside with some examples that support it.

The first section, where the first experiment performed with the classic LDA (Blei et al., 2003) is described, motivates this work. The results obtained were thoroughly analyzed, in order to support our thesis. The proposed model is disclosed here, where it is explained in detail. This section is mostly to explain the mathematical component of our model. The goal of all these deductions is to obtain the formulas for the variational parameters, so that they can be used in the programming phase.

The final section presents the pseudo-code of the algorithm for our model. The technological choices made throughout the implementation are also numerated.

3.1 Motivation and examples

The main subject of this thesis arose due to the discovery of topics where some words were too similar and thus redundant, which came to prove that these were not satisfying enough. This means that in the words *car*, *auto* and *automobile* could co-occur in a topic. However, the presence of all these words would be redundant, because they are synonyms and denote the same information. In order to confirm that it actually happens and, consequently, to demonstrate that the the thesis subject made sense, some examples were generated. For their creation, two English text corpora were used, namely the 20 Newsgroups[1] dataset and the Associated Press (AP)[2], already presented in chapter 2. The content of those corpora went through a preprocessing phase, where stop-words and numbers were removed and lemmatization was applied.

When applying the LDA algorithm implemented in C, (Blei et al., 2003), 15 topics with 10 words each were created for the 20 Newsgroups and 24 topics with 10 words each for the AP corpus, but we chose to only present 3 of those topics. In table 3.1, the topics selected are presented. They help to validate the purpose of this work.

Just by looking at these topics, we can understand that there are some words which are in some way related to others. Nevertheless, an algorithm was created by us, alongside with

[1] `http://qwone.com/~jason/20Newsgroups/`

[2] `http://www.cs.princeton.edu/~blei/lda-c/`

AP			**20 Newsgroups**		
Topic 1	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
police	military	time	game	window	image
kill	force	film	hit	font	file
people	troop	world	player	file	graphic
arrest	u.s	book	article	run	format
shot	army	woman	write	application	jpeg
attack	president	write	team	server	color
officer	soldier	life	time	program	program
fire	american	game	pitch	include	bit
death	war	movie	play	use	available
wound	defense	play	baseball	sun	software

Table 3.1: Topics extracted with LDA from AP and 20 Newsgroups.

the lexical database WordNet. It was used to retrieve the connections between the terms, such as relations of specialization (hyponyms), generalization (hypernyms) and synonyms. A relation of specialization indicates that a word's semantic field is included in that of another word. Whereas a relation of generalization is the exact opposite of the latter. For example, *pidgeon* is a hyponym for *bird* and *bird* is a hypernym for *pidgeon*. However, the hypernymy/hyponymy relations are still not being taken into account in the current version of the model, only synonyms are. For the AP corpus, in topic 1, it was identified that *death* is a specialization of *kill*. In topic 2, *force* is a generalization of *military*. Finally, in topic 3, *movie* is a synonym of *film*. In table 3.2 more of these relations are presented for both datasets, AP and 20 Newsgroups.

	Relations	
	AP	**20 Newsgroups**
Topic 1	*death* hyponym of *kill* *wound* hypernym of *shot* *fire* hyponym of *attack*	*play* hypernym of *game* *pitch, play* hyponyms of *hit*
Topic 2	*force* hypernym of *military*	*application* hyponym of *use*
Topic 3	*movie* synonym of *film* *play* hypernym of *game*	*image* hypernym of *graphic* *program* hyponym of *software*

Table 3.2: Relationships between some words of each topic.

It was already possible to identify some relations inside each topic. However, this algorithm can still be improved since some very explicit connections are not being accounted for, such as *jpeg* being a format of an *image*, in Topic 2 of the 20 Newsgroups, because they are not covered by WordNet. These relations will be analyzed to the point of upgrading the current algorithm, so that they are included, which may be possible by using a different lexical database.

With these results, it is obvious that the topics can be improved by removing the similar words and representing them by the most suitable one and thus creating a better set of topics, where redundant information is minimized.

3.2 Model

This section is where the proposed model is presented. It is divided in different sections, first the generative process is described, then we reveal the respective graphical model and, finally, we explain the necessary calculations to perform the inference and estimation of, respectively, our variables and parameters.

3.2.1 Generative process

This thesis proposes a semantic topic model that acesses an external lexical-semantic database, such as WordNet. It is based on Latent Dirichlet Allocation model (Blei et al., 2003) and it has a new parameter, $\eta_{1:S}$, which represents the probabilities of each word in a synset. These are represented in a matrix $S \times V$ and are fixed, provided by WordNet, where S is the number of synsets and V the size vocabulary of the corpus.

Given a corpus $\mathcal{D} = \{\boldsymbol{w}\}_{d=1}^{D}$ of size D and the probabilities of each word in a synset, η_n, this model is going to estimate the topics. It is possible to achieve this by considering that each document is represented by a distribution of topics, θ, and a topic consists of a distribution over the synsets in the vocabulary, β.

This is a major difference from the traditional LDA, where the vocabulary consists of words. Here every word has associated all the possible synsets, with the respective probabilities. Since a concept has many possible words, there is a need to choose the best one. So, the word chosen to illustrate the concept, w, which is related with the topic-synset distribution β and the document-topic distribution θ, is linked on the other hand with the concept c itself and the synset-word distribution η. These synset-word distributions will be calculated with the help of SemCor, in an initial phase, and in a final experiment, word sense disambiguation was used for this purpose.

So, in the end, this model is going to be a more enriched version of LDA, where the semantics of the words are considered.

The generative process of a document d under the semantic LDA (semLDA) model is the following:

1. Choose topic proportions $\theta|\alpha \sim Dir(\alpha)$
2. For each concept, c_n
   (a) Choose topic assignment $z_n|\theta \sim Mult(\theta)$.
   (b) Choose concept c_n $|z_n, \beta_{1:K} \sim Mult(\beta_{z_n})$.
   (c) Choose word to represent concept $w_n|c_n, \eta_{1:S} \sim Mult(\eta_{c_n})$.

According to their definitions, the remainder distributions are given by equations 3.1, 3.2, 3.3 and 3.4.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{(\alpha_i-1)} \tag{3.1}$$

$$p(z_n|\theta) = \prod_{i=1}^{K}\theta_i^{z_{n,i}} \tag{3.2}$$

$$p(c_n|z_n,\beta_{1:K}) = \prod_{j=1}^{S}(\beta_{z_n,j})^{c_{n,j}} \tag{3.3}$$

$$p(w_n|c_n,\eta_{1:S}) = \prod_{i=1}^{V_{c_n}}(\eta_{c_n,i})^{w_{n,i}} \tag{3.4}$$

Where V_{c_n} is the number of possible words to express the concept c_n.

3.2.2 Graphical model

The graphical model of the proposed model, semLDA, can be viewed in Figure 3.1, where D is the number of documents in the corpus, K is the number of topic and, N is the number of words in a document. In this model each word of a document, w_n, is going to be drawn from a concept, c_n to represent it, and from a synset distribution, $\boldsymbol{\eta}$, which is going to be fixed. The concept c_n is determined by a discrete topic-assignment z_n, which is picked from the document's distribution over topics θ and a topic distribution $\boldsymbol{\beta}$. It follows the same reasoning as the LDA model, however it is slightly different.

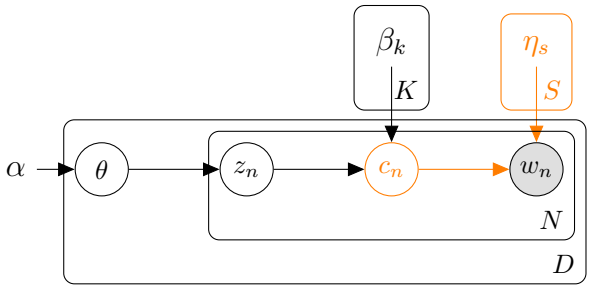

Figure 3.1: Graphical model representation of SemLDA.

3.2.3 Approximate inference

The goal of the generative probabilistic model of semLDA is to estimate latent variables and the model's parameters from the observed data. Given a document, the joint distribution, as represented by the graphical model in figure 3.1, is given by equation 3.5, where we defined the model parameters $\Theta = \{\alpha, \beta_{1:K}\}$.

$$p(\theta, z_{1:N}, c_{1:N}, w_{1:N}|\Theta) = p(\theta|\alpha)\left(\prod_{n=1}^{N} p(z_n|\theta)p(c_n|z_n, \beta_{1:K})p(w_n|c_n)\right) \tag{3.5}$$

The posterior distribution over the latent variables $\theta, z_{1:N}, c_{1:N}$ is given by equation 3.6, according to the Bayes theorem. This is used to compute exact inference of the variables.

$$\begin{aligned} p(\theta, z_{1:N}, c_{1:N}|w_{1:N}, \Theta) &= \frac{p(\theta, z_{1:N}, c_{1:N}, w_{1:N}|\Theta)}{p(w_{1:N}|\Theta)} \\ &= \frac{p(\theta|\alpha)\Big(\prod_{n=1}^{N} p(z_n|\theta)p(c_n|z_n, \beta_{1:K})p(w_n|c_n)\Big)}{\int_\theta p(\theta|\alpha)\sum_{z_{1:N}}\Big(\prod_{n=1}^{N} p(z_n|\theta)\sum_{c_{1:N}} p(c_n|z_n, \beta_{1:K})p(w_n|c_n)\Big)} \end{aligned} \tag{3.6}$$

However, the computation on this equation is intractable because of its denominator, so a different approach will be used, variational inference.

Variational inference setup

Let $q(\theta, z_{1:N}, c_{1:N})$ denote a variational distribution of the latent variables. Since we are using a fully-factorized (mean-field) approximation, we have equation 3.7, where $\gamma, \phi_{1:N}, \lambda_{1:N}$ are the variational parameters that correspond to, respectively, $\theta, z_{1:N}$ and $c_{1:N}$.

$$q(\theta, z_{1:N}, c_{1:N}) = q(\theta|\gamma)\left(\prod_{n=1}^{N} q(z_n|\phi_n)q(c_n|\lambda_n)\right) \tag{3.7}$$

The variational objective function (or the evidence lower bound or ELBO) is given by equation 3.8 and the entropy $H(q)$ of the variational distribution is given by equation 3.9.

$$\begin{aligned} \log p(w_{1:N}|\alpha, \beta_{1:K}, \eta_{1:S}) &= \log \int_\theta \sum_{z_{1:N}} \sum_{c_{1:N}} \frac{p(\theta, z_{1:N}, c_{1:N}, w_{1:N}|\Theta)q(\theta, z_{1:N}, c_{1:N})}{q(\theta, z_{1:N}, c_{1:N})} \\ &\geqslant \mathcal{L}(w_{1:N}|\Theta) \\ &= \mathbb{E}_q[\log p(\theta, z_{1:N}, c_{1:N}, w_{1:N})] \underbrace{-\mathbb{E}_q[\log q(\theta, z_{1:N}, c_{1:N})]}_{H(q)} \\ &= \mathbb{E}_q[\log p(\theta|\alpha)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(c_n|z_n, \beta_{1:K})] \\ &+ \sum_{n=1}^{N} \mathbb{E}_q[\log p(w_n|c_n, \eta_{1:S})] + H(q) \end{aligned} \tag{3.8}$$

$$H(q) = -\mathbb{E}_q[\log q(\theta|\gamma)] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_n|\phi_n)] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(c_n|\lambda_n)] \tag{3.9}$$

Terms needed for the lower bound

The next step is to split equation 3.8 into terms and expand each one with the distributions defined in the beginning. After expanding and simplifying each element, the outcomes are in equations 3.10, 3.11, 3.12 and 3.13.

$$\begin{aligned}\mathbb{E}_q[\log p(\theta|\alpha)] &= \mathbb{E}_q\left[\log \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_i^{(\alpha_i-1)}\right] \\ &= \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K}\log\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\mathbb{E}_q[\log\theta_i]\end{aligned} \tag{3.10}$$

$$\mathbb{E}_q[\log p(z_n|\theta)] = \mathbb{E}_q\left[\log\prod_{i=1}^{K}\theta_i^{z_{n,i}}\right] = \sum_{i=1}^{K}\mathbb{E}_q[z_{n,i}]\mathbb{E}_q[\log\theta_i] = \sum_{i=1}^{K}\phi_{n,i}\mathbb{E}_q[\log\theta_i] \tag{3.11}$$

$$\begin{aligned}\mathbb{E}_q[\log p(c_n|z_n,\beta_{1:K})] &= \mathbb{E}_q\left[\log\prod_{j=1}^{S}(\beta_{z_n,j})^{c_{n,j}}\right] = \sum_{j=1}^{S}\mathbb{E}_q[c_{n,j}]\mathbb{E}_q[\log\beta_{z_n,j}] \\ &= \sum_{j=1}^{S}\lambda_{n,j}\mathbb{E}_q[\log\beta_{z_n,j}] = \sum_{j=1}^{S}\lambda_{n,j}\mathbb{E}_q\left[\sum_{i=1}^{K}z_{n,i}\log\beta_{i,j}\right] \\ &= \sum_{j=1}^{S}\lambda_{n,j}\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,j} = \sum_{j=1}^{S}\sum_{i=1}^{K}\lambda_{n,j}\phi_{n,i}\log\beta_{i,j}\end{aligned} \tag{3.12}$$

$$\begin{aligned}\mathbb{E}_q[\log p(w_n|c_n,\eta_{1:S})] &= \mathbb{E}_q\left[\log\prod_{i=1}^{V_{c_n}}(\eta_{c_n,i})^{w_{n,i}}\right] = \mathbb{E}_q\left[\sum_{i=1}^{V_{c_n}}w_{n,i}\log\eta_{c_n,i}\right] \\ &= \sum_{j=1}^{S}\lambda_{n,j}\sum_{i=1}^{V_j}w_n\log\eta_{j,i} = \sum_{j=1}^{S}\sum_{i=1}^{V_j}\lambda_{n,j}w_{n,i}\log\eta_{j,i}\end{aligned} \tag{3.13}$$

Similarly, for the corresponding terms of the variational distribution, the ones in equation 3.9, we have the equations 3.14, 3.15 and 3.16.

$$\begin{aligned}\mathbb{E}_q[\log q(\theta|\gamma)] &= \mathbb{E}_q\left[\log\frac{\Gamma(\sum_{i=1}^{K}\gamma_i)}{\prod_{i=1}^{K}\Gamma(\gamma_i)}\prod_{i=1}^{K}\theta_i^{(\gamma_i-1)}\right] \\ &= \log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right) - \sum_{i=1}^{K}\log\Gamma(\gamma_i) + \sum_{i=1}^{K}(\gamma_i-1)\mathbb{E}_q[\log\theta_i]\end{aligned} \tag{3.14}$$

$$\mathbb{E}_q[\log q(z_n|\phi_n)] = \mathbb{E}_q\left[\log \prod_{i=1}^{K} \phi_{n,i}^{z_{n,i}}\right] = \sum_{i=1}^{K} \mathbb{E}_q[z_{n,i}]\mathbb{E}_q[\log \phi_{n,i}] = \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \tag{3.15}$$

$$\mathbb{E}_q[\log q(c_n|\lambda_n)] = \mathbb{E}_q\left[\log \prod_{j=1}^{S} (\lambda_{n,j})^{c_{n,j}}\right] = \sum_{j=1}^{S} \mathbb{E}_q[c_{n,j}]\mathbb{E}_q[\log \lambda_{n,j}] = \sum_{l=1}^{S} \lambda_{n,j} \log \lambda_{n,j} \tag{3.16}$$

The expectation of the log of the Dirichlet that appears in various of the equations above is given by equation 3.17, where $\Psi(\cdot)$ is the digamma function. This function is the first derivative of the $\log \Gamma$ function, which is computable via Taylor approximations (Abramowitz et al., 1966). See appendix A.1 in Blei et al. (2003) for the derivation of this standard result.

$$\mathbb{E}_q[\log \theta_i] = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K} \gamma_j\right) \tag{3.17}$$

Lower bound

With the results obtained from the previous equations, the lower bound is now given by equation 3.18, where all the terms where replaced by its respective simplification.

$$
\begin{aligned}
\mathcal{L}(w_{1:N}|\Theta) &= \mathbb{E}_q[\log p(\theta|\alpha)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(z_n|\theta)] + \sum_{n=1}^{N} \mathbb{E}_q[\log p(c_n|z_n, \beta_{1:K})] \\
&+ \sum_{n=1}^{N} \mathbb{E}_q[\log p(w_n|c_n, \eta_{1:S})] - \mathbb{E}_q[\log q(\theta|\gamma)] \\
&- \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_n|\phi_n)] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(c_n|\lambda_n)] \\
&= \log \Gamma\Big(\sum_{i=1}^{K} \alpha_i \Big) - \sum_{i=1}^{K} \log \Gamma(\alpha_i) + \sum_{i=1}^{K} (\alpha_i - 1) \Big(\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{K} \gamma_j \Big) \Big) \\
&+ \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \Big(\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{K} \gamma_j \Big) \Big) \\
&+ \sum_{n=1}^{N} \sum_{j=1}^{S} \sum_{i=1}^{K} \lambda_{n,j} \phi_{n,i} \log \beta_{i,j} \\
&+ \sum_{n=1}^{N} \sum_{j=1}^{S} \sum_{i=1}^{V_j} \lambda_{n,j} w_{n,i} \log \eta_{j,i} \\
&- \log \Gamma\Big(\sum_{i=1}^{K} \gamma_i \Big) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) - \sum_{i=1}^{K} (\gamma_i - 1) \Big(\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{K} \gamma_j \Big) \Big) \\
&- \sum_{n=1}^{N} \sum_{i=1}^{K} \phi_{n,i} \log \phi_{n,i} \\
&- \sum_{n=1}^{N} \sum_{j=1}^{S} \lambda_{n,j} \log \lambda_{n,j}
\end{aligned} \tag{3.18}
$$

Optimizing the lower bound

Now the goal is to optimize the lower bound with respect to the variational parameters γ, ϕ and λ, using the algorithm coordinate ascent, which allows to make these results as close as possible to the true posterior, equation 3.6. By optimizing w.r.t every variational parameter we will obtain expressions that allows us to compute their value.

Optimizing w.r.t. γ_i

To optimize w.r.t. the parameter γ_i, it is necessary to collect only the terms in the lower bound, equation 3.18, that contain this parameter. This translates into equation 3.19.

$$\begin{aligned}
\mathcal{L}_{[\gamma]} &= \sum_{i=1}^{K}(\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) \\
&\quad + \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{n,i}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) \\
&\quad - \log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right) + \log\sum_{i=1}^{K}\Gamma(\gamma_i) - \sum_{i=1}^{K}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) \\
&= \sum_{i=1}^{K}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right)\left(\alpha_i + \sum_{n=1}^{N}\phi_{n,i} - \gamma_i\right) - \log\Gamma\left(\sum_{i=1}^{K}\gamma_i\right) + \log\sum_{i=1}^{K}\Gamma(\gamma_i)
\end{aligned} \tag{3.19}$$

The next step is to take derivatives w.r.t. γ_i, which, in the end, gives equation 3.20.

$$\begin{aligned}
\frac{\partial\mathcal{L}_{[\gamma]}}{\partial\gamma_i} &= \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^{N}\phi_{n,i} - \gamma_i\right) - \Psi(\gamma_i) - \Psi'\left(\sum_{j=1}^{K}\gamma_j\right)\sum_{j=1}^{K}\left(\alpha_j + \sum_{n=1}^{N}\phi_{n,j} - \gamma_j\right) \\
&\quad + \Psi\left(\sum_{j=1}^{K}\gamma_j\right) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \Psi(\gamma_i) \\
&= \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^{N}\phi_{n,i} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^{K}\gamma_j\right)\sum_{j=1}^{K}\left(\alpha_j + \sum_{n=1}^{N}\phi_{n,j} - \gamma_j\right)
\end{aligned} \tag{3.20}$$

Setting this derivative to zero in order to get a maximum (notice that the solutions for the different γ_i are coupled, hence they have to be solved as a system of linear equations), we get the solution in equation 3.21. This can be easily verified by submitting the value for γ_i above in the expression for the partial derivatives. This update equation is the same as in standard LDA Blei et al. (2003).

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{n,i} \tag{3.21}$$

Optimizing w.r.t. $\phi_{n,i}$

Now regarding the parameter $\phi_{n,i}$, we start once again by collecting only the terms in the bound that contain this parameter. Since this parameter is from a multinomial distribution, this is constrained maximization problem, and $\sum_{k=1}^{K} \phi_{n,k} = 1$, which is necessary for it to be a valid probability distribution. Hence, we need to also add the necessary Lagrange multipliers. Equation 3.22, Lagrangian, contains the respective terms from the lower bound, as well as the Lagrangian multipliers.

$$\begin{aligned}\mathcal{L}_{[\phi_{n,i}]} = \phi_{n,i}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right)\right) + \sum_{j=1}^{S}\lambda_{n,j}\phi_{n,i}\log\beta_{i,j} - \phi_{n,i}\log\phi_{n,i} \\ + \mu\left(\sum_{k=1}^{K}\phi_{n,k} - 1\right)\end{aligned} \tag{3.22}$$

By taking the derivatives w.r.t. $\phi_{n,i}$, the result is equation 3.23.

$$\frac{\partial\mathcal{L}_{[\phi_{n,i}]}}{\partial\phi_{n,i}} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j} - \log\phi_{n,i} - 1 + \mu \tag{3.23}$$

Setting this derivative to zero and solving for $\phi_{n,i}$ translates into equation 3.24.

$$\begin{aligned}&\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j} - \log\phi_{n,i} - 1 + \mu = 0 \\ &\Leftrightarrow \log\phi_{n,i} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j} - 1 + \mu \\ &\Leftrightarrow \phi_{n,i} = \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j} - 1 + \mu\right)\end{aligned} \tag{3.24}$$

Since the Lagrangian multipliers where added, there is an extra step that consists in plugging this expression in the constraint and solving for μ (or $\exp(\mu)$), which results in equation 3.26.

$$\sum_{k=1}^{K}\phi_{n,k} = 1 \tag{3.25}$$

$$\begin{aligned}&\Leftrightarrow \sum_{k=1}^{K}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{k,j} - 1 + \mu\right) = 1 \\ &\Leftrightarrow \sum_{k=1}^{K}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{k,j} - 1\right)\exp(\mu) = 1 \\ &\Leftrightarrow \exp(\mu) = \frac{1}{\sum_{k=1}^{K}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{j=1}^{K}\gamma_j\right) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{k,j} - 1\right)}\end{aligned} \tag{3.26}$$

To obtain the expression for $\phi_{n,i}$, we need to plug in the previous equation back in the equation 3.24, so that it results in equation 3.27.

$$\begin{aligned}\phi_{n,i} &= \frac{\exp\Big(\Psi(\gamma_i) - \Psi\big(\sum_{j=1}^{K}\gamma_j\big) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j} - 1\Big)}{\sum_{k=1}^{K}\exp\Big(\Psi(\gamma_k) - \Psi\big(\sum_{j=1}^{K}\gamma_j\big) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{k,j} - 1\Big)} \\ &\propto \exp\left(\Psi(\gamma_i) - \Psi\Big(\sum_{j=1}^{K}\gamma_j\Big) + \sum_{j=1}^{S}\lambda_{n,j}\log\beta_{i,j}\right) \end{aligned} \tag{3.27}$$

Optimizing w.r.t. $\lambda_{n,j}$

Finally, for the parameter $\lambda_{n,j}$, we again collect only the terms in the bound that contain it. Since this parameter is from a multinomial distribution, this is constrained maximization problem, and $\sum_{k=1}^{S}\lambda_{n,k} = 1$, which is necessary for it to be a valid probability distribution. Hence, we need to also add the necessary Lagrange multipliers. The Lagrangian is then given by equation 3.28.

$$\mathcal{L}_{[\lambda_{n,j}]} = \sum_{i=1}^{K}\lambda_{n,j}\phi_{n,i}\log\beta_{i,j} + \sum_{i=1}^{V_j}\lambda_{n,j}w_{n,i}\log\eta_{j,i} - \lambda_{n,j}\log\lambda_{n,j} + \mu\Big(\sum_{k=1}^{S}\lambda_{n,k} - 1\Big) \tag{3.28}$$

By taking the derivatives w.r.t. $\lambda_{n,j}$, we obtain equation 3.29.

$$\frac{\partial\mathcal{L}_{[\lambda_{n,j}]}}{\partial\lambda_{n,j}} = \sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,j} + \sum_{i=1}^{V_j}w_{n,i}\log\eta_{j,i} - \log\lambda_{n,j} - 1 + \mu \tag{3.29}$$

We need to set this derivative to zero and solve for $\lambda_{n,j}$, so that it results in equation 3.30.

$$\begin{aligned}&\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,j} + \sum_{i=1}^{V_j}w_{n,i}\log\eta_{j,i} - \log\lambda_{n,j} - 1 + \mu = 0 \\ &\Leftrightarrow \log\lambda_{n,j} = \sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,j} + \sum_{i=1}^{V_j}w_{n,i}\log\eta_{j,i} - 1 + \mu \\ &\Leftrightarrow \lambda_{n,j} = \exp\left(\sum_{i=1}^{K}\phi_{n,i}\log\beta_{i,j} + \sum_{i=1}^{V_j}w_{n,i}\log\eta_{j,i} - 1 + \mu\right)\end{aligned} \tag{3.30}$$

Plugging this expression in the constraint and solving for μ (or $\exp(\mu)$) gives equation 3.32.

$$\sum_{k=1}^{S} \lambda_{n,k} = 1 \tag{3.31}$$

$$\Leftrightarrow \sum_{k=1}^{S} \exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,k} + \sum_{i=1}^{V_k} w_{n,i} \log \eta_{k,i} - 1 + \mu \right) = 1$$

$$\Leftrightarrow \sum_{k=1}^{S} \exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,k} + \sum_{i=1}^{V_k} w_{n,i} \log \eta_{k,i} - 1 \right) \exp(\mu) = 1$$

$$\Leftrightarrow \exp(\mu) = \frac{1}{\sum_{k=1}^{S} \exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,k} + \sum_{i=1}^{V_k} w_{n,i} \log \eta_{k,i} - 1 \right)} \tag{3.32}$$

Finally by plugging this previous equation back in the equation 3.30, gives the solution 3.33 for $\lambda_{n,j}$.

$$\begin{aligned} \lambda_{n,j} &= \frac{\exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,j} + \sum_{i=1}^{V_j} w_{n,i} \log \eta_{j,i} - 1 \right)}{\sum_{k=1}^{S} \exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,k} + \sum_{i=1}^{V_k} w_{n,i} \log \eta_{k,i} - 1 \right)} \\ &\propto \exp\left(\sum_{i=1}^{K} \phi_{n,i} \log \beta_{i,j} + \sum_{i=1}^{V_j} w_{n,i} \log \eta_{j,i} \right) \end{aligned} \tag{3.33}$$

3.2.4 Parameter estimation

Given a corpus of D documents, $\mathcal{D} = \{w_{1:N}^d\}_{d=1}^D$, we find maximum likelihood estimates for the text topics $\beta_{1:K}$. In order to do this, we will use variational Bayesian EM, which replaces the E-step of the expectation-maximization algorithm with variational inference to find an approximate posterior for each document. In the M-step, as in exact EM, we find approximate maximum likelihood estimates of the parameters using the expected sufficient statistics computed in the E-step.

The corpus-level log-likelihood is given by equation 3.34.

$$\mathcal{L}(\mathcal{D}) = \sum_{d=1}^{D} \log p(w_{1:N_d} | \alpha, \beta_{1:K}, \eta_{1:S}) \tag{3.34}$$

where $\log p(w_{1:N_d} | \alpha, \beta_{1:K}, \eta_{1:S})$ is given by equation 3.8, i.e. the lower bound.

Estimating $\beta_{i,j}$

We start by collecting only the terms in the log-likelihood (equation 3.34) that contain $\beta_{i,j}$. Notice that this is constrained maximization problem, since $\sum_{k=1}^{V} \beta_{i,k} = 1$, which is necessary for it to be a valid probability distribution. Hence, we need to also add the necessary Lagrange multipliers. The Lagrangian is then given by equation 3.35, where N

denotes the number of words in document D, and where we made use of the "long" form for $\mathbb{E}_q[\log p(c_n|z_n, \beta_{1:K})]$ from equation 3.12.

$$\mathcal{L}_{[\beta_{i,j}]} = \sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{i=1}^{K}\sum_{j=1}^{S}\lambda_{n,j}^{d}\phi_{n,i}^{d}\log\beta_{i,j} + \sum_{i=1}^{K}\mu_i\left(\sum_{k=1}^{S}\beta_{i,k} - 1\right) \tag{3.35}$$

Taking derivatives w.r.t. $\beta_{i,j}$ gives equation 3.36.

$$\frac{\partial\mathcal{L}_{[\beta_{i,j}]}}{\partial\beta_{i,j}} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,j}^{d}\phi_{n,i}^{d}\frac{1}{\beta_{i,j}} + \mu_i \tag{3.36}$$

By setting this derivative to zero and solving for $\beta_{i,j}$ we obtain equation 3.37.

$$\beta_{i,j} = -\frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,j}^{d}\phi_{n,i}^{d}}{\mu_i} \tag{3.37}$$

Plugging this expression in the constraint and solving for μ gives equation 3.38.

$$\begin{aligned}
&\sum_{k=1}^{V}\beta_{i,k} = 1 \\
&\Leftrightarrow -\frac{\sum_{k=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,k}^{d}\phi_{n,i}^{d}}{\mu_i} = 1 \\
&\Leftrightarrow \mu_i = -\sum_{k=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,k}^{d}\phi_{n,i}^{d}
\end{aligned} \tag{3.38}$$

By plugging this previous equation back in equation 3.37, allows us to obtain equation 3.39 for $\beta_{i,j}$, which is very similar to the update in standard LDA (Blei et al., 2003).

$$\begin{aligned}
\beta_{i,j} &= \frac{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,j}^{d}\phi_{n,i}^{d}}{\sum_{k=1}^{V}\sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,k}^{d}\phi_{n,i}^{d}} \\
&\propto \sum_{d=1}^{D}\sum_{n=1}^{N_d}\lambda_{n,j}^{d}\phi_{n,i}^{d}
\end{aligned} \tag{3.39}$$

3.3 Algorithm implementation

Our model was implemented with two main stages:

- Preprocessing of the input, where the raw text is transformed in a suitable input for the algorithm, considering, among others, the possible senses of the words, and their probabilities.

- The algorithm itself (detailed in section 3.3).

The preprocessing phase is a critical stage in this work that may have a huge impact on the results of the algorithm. It is explained in further detail in the next chapter, so what matters now are the technologies used for this purpose. The programming language Java was used to read the data from the different files and to remove special characters and stop-words. The processing of the text was all performed in the programming language Python.

Both the classic LDA and SemLDA need a special input file, which can be created with the help of Gensim. Gensim (Řehůřek and Sojka, 2010) is a software that, aside from other features, can perform topic modelling of large corpora. It has a different implementation of the LDA algorithm, in Python programming language. It also allows the creation of the input file for the classic LDA.

Given these three documents: D_1 = *(student private baptist allegedly)*, D_2 = *(oil israel discount promise)* and D_3 = *(woman hostage attempt steal jewelry)*, which were initially sentences, but after a preprocessing phase the result is sets of words. For LDA, the entry file is something like:

- D_1: `4 0:1 1:1 2:1 3:1`
- D_2: `4 4:1 5:1 6:1 7:1`
- D_3: `5 8:1 9:1 10:1 11:1 12:1`

Where the first number of every document is the number of different words in it, for example in document 1 there are 4 different words. For every word there is format like `0:1`, where the left side is the identifier of the word, in this case *student* and the right side is the number of times that word occurs in that document.

Whereas for the SemLDA entry file, there is a slight change, it becomes:

- D_1: `4 0:1:2[6:0.891891891892 1:0.666666666667] 1:1:4[12:1.0 3:1.0 4:1.0 5:1.0] 2:1:1[0:1] 3:1:[7:1.0]`
- D_2: `4 5:1:4[8:0.0819672131148 9:0.333333333333 18:1.0 11:1.0] 6:1:2[2:1.0 13:0.875] 7:1:4[10:1.0 17:0.0666666666667 29:0.985714285714 19:1.0] 8:1:2[15:0.5 14:1.0]`
- D_3: `5 10:1:4[20:1.0 16:1.0 22:0.5 23:1.0] 11:1:1[24:1] 12:1:2[25:1.0 26:0.761904761905] 13:1:3[27:1.0 28:1.0 18:0.727272727273] 14:1:1[30:1]`

The difference here that there is an addition to the format of every word, we now have `0:1:2[6:0.891891891892 1:0.666666666667]`, where *2* indicates the number of senses of that word. Inside the square brackets are the identifiers of the synsets and their respective probabilities, which are obtained by accessing WordNet.

The pseudo-code of the Variational Expectation-Maximization method for the proposed model is presented in algorithm 1, where the differences towards the classic LDA algorithm are highlighted. When it came to the implementation of this algorithm, the programming language chosen was C, given that there was an implementation of the classical model, provided by the author (Blei et al., 2003), in this language. So we used this available code to

adapt it to SemLDA. The complexity of the classic algorithm in the E-step is $O(D \times N \times k)$. However, there is a cycle in the algorithm until it converges to a value, so the number of these iterations, according to (Blei et al., 2003), is on the order of the number of words in the document, so the complexity is now $O(D \times N^2 \times K)$. In the M-step the complexity is $O(D \times K \times V)$. The difference of the number of iterations between the two models, is not very signficant. In our model, the complexity in the E-step is $O(D \times N^2 \times (K + S))$, following the same reasoning as before, and in the M-step is $O(D \times K \times S)$. If the number of synsets S is very high, it might increase the computation time of our algorithm.

Algorithm 1: Variational Expectation-Maximization Semantic LDA

Input : Number of Topics K
Number of Synsets S
Corpus with M documents and N_d words in document d
Output: Model parameters: β, θ, z

initialize $\phi^0_{ni} := 1/k$ for all i in k and n in N_d
initialize $\lambda^0_{nj} := 1/s$ for all j in s and n in N_d
initialize $\gamma_i := \alpha_i + N/k$ for all i in k
initialize $\alpha := 50/k$
initialize $\beta_{ij} := 0$ for all i in k and j in V

//E-Step (determine ϕ, γ and λ and compute expected likelihood)
loglikelihood := 0
for $d = 1$ **to** D **do**
  repeat
    for $n = 1$ **to** N_d **do**
      for $i = 1$ **to** K **do**
        $\phi^{t+1}_{dni} := \exp(\Psi(\gamma^t_{di}) + \sum_{j=1}^{S} \lambda^t_{nj} \log \beta_{ij})$
      end
      normalize ϕ^{t+1}_{dni} to sum to 1
      for $j = 1$ **to** S **do**
        $\lambda^{t+1}_{dnj} := \exp(\sum_{k=1}^{K} \phi^t_{n,k} \log \beta_{k,j} \sum_{m=1}^{V_j} w_{n,m} \log \eta_{j,m})$
      end
      normalize λ^{t+1}_{dnj} to sum to 1
    end
    $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi^{t+1}_{dn}$
  until *convergence of* ϕ_d, γ_d *and* λ_d;
  loglikelihood := loglikelihood + $L(\gamma, \phi, \lambda; \alpha, \beta)$ // See equation 3.18
end

//M-Step (maximize the log likelihood of the variational distribution)
for $d = 1$ **to** D **do**
  for $i = 1$ **to** K **do**
    for $j = 1$ **to** S **do**
      $\beta_{ij} := \lambda_{dnj}\ \phi_{dni}$
    end
    normalize β_i to sum to 1
  end
end
estimate α via Eq. 8 (tutorial)
if *loglikelihood converged* **then**
  return parameters
else
  go back to E-Step
end

Chapter 4

Experiments

This chapter describes the experiments performed throughout the year, and presents the respective results. Overall, four main experiments were performed and these are explained with further detail in the following sections, including the preprocessing involved, approach taken and respective results. Those are described after enumerating the set up performed for the experiments and evaluation of the results. Three measures were selected, Pointwise Mutual Information (PMI), topic coherence and perplexity, each already introduced in section 2.2. The first experiment, Experiment 1, is basically where we started to introduce semantics in the LDA algorithm. It consists in replacing the words of each document by their most probable synset. The other 3 experiments have the purpose of validating the proposed model, SemLDA. They differ amongst each other on how the probabilities of a word in a synset are calculated. The second experiment, Experiment 2, relies on the content of the SemCor corpus to perform the calculations. On the third, Experiment 3, probabilities are either based on WSD or SemCor, because in some cases it was necessary to resort to this annotated corpora. The final experiment, Experiment 4, only considered the probabilities obtained from WSD, and does not require an external annotated corpus, which makes SemLDA more flexible and thus adaptable to other languages/wordnets. In the last section, some observations are presented, regarding the results obtained with all the experiments.

4.1 Set up for experiments and evaluation

Set up for experiments A few initial experiments were performed in order to select the best parameters and preprocessing options for the classic LDA algorithm. The selected parameters were used in all the SemLDA experiments.

Regarding the preprocessing, topics were discovered using only nouns (typically the most informative words), or nouns and verbs, or nouns, verbs and adjectives, or all the parts-of-speech. By applying the different measures to the results, we concluded that there was no improvement when ignoring specific grammatical classes. Therefore, all the latter experiments take into account all open class words, namely nouns, verbs, adjectives and adverbs.

Both the LDA model and SemLDA have an input parameter α, with a specific value, which can either be estimated throughout the algorithm or it can be a fixed value. So that a random value was not chosen, we experimented different values with α fixed or not.

The conclusion we arrived at was that, it made no difference if the argument was fixed or not. Regarding the value, we noticed that, in some occasions, the value 0.1 produced better results and in others the value 0.5 surpassed the previous. So, for the presented experiments α was always set as fixed with the value 0.5.

In SemEval 2015, it was necessary to explore different implementations in different programming languages of the LDA algorithm, such as Gensim, a Java version and the original implementation in C, to analyze the different outputs obtained. Given the results obtained from the work done for the SemEval, only the C implementation of LDA was used thenceforth.

Hierarchical Dirichlet process (HDP) (Teh et al., 2006) is a nonparametric Bayesian model for clustering problems involving multiple groups of data. Its goal is to cluster the information into different groups that have something in common. This method was used to discover the appropriate number of topics for each dataset, instead of trial and error with different numbers of topics. The results obtained suggested that the 20 Newsgroups dataset contains 15 topics and the AP corpus 24.

Set up for evaluation Although, at a first glance, some results seem promising, to have a more objective view, they were validated automatically, using metrics previously applied to the context of topic modelling, referred in the related work: pointwise mutual information (PMI), topic coherence and perplexity.

We recall that, topics discovered by SemLDA are sets of synsets and not of surface words. Therefore, to enable a fair comparison with the classic LDA, before computing the PMI scores and the topic coherence, we converted our topics to a plain word representation. For this purpose, instead of full synsets, we used only their first word. We recall that, to WordNet, this is the word most frequently used to denote the synset concept, in the SemCor corpus.

Regarding the PMI measure, for both datasets, co-occurrence is computed from Wikipedia, which provides a large and wide-coverage source of text, completely independent from the datasets used and from WordNet. After computing PMI for all topics, we computed the average score for the full topic set.

For the coherence measure, the average is also computed for the full topic set. Assuming that, in every document, there is an explicit theme, by calculating this, we can analyze if the grouping of words is coherent, given their co-occurrence.

The perplexity was computed after splitting each dataset into two subsets of randomly selected documents: one for training (70%) and another for testing (30%).

It was necessary to have baseline results obtained from the classic LDA (Blei et al., 2003), so that we could compare with our model. Given that, we have table 3.1 with the topics produced and table 4.2 that presents the results obtained with the evaluation metrics. In all experiments, we compare the results from SemLDA with the ones from table 4.2 to verify if they show an improvement.

AP			20 Newsgroups		
Topic 1	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
drug	party	bush	game	god	image
school	government	dukakis	fan	christian	file
student	president	campaign	team	believe	graphic
charge	rebel	vote	hockey	people	format
attorney	gorbachev	republican	win	write	jpeg
federal	political	president	play	article	color
teacher	leader	jackson	goal	sin	program
cocaine	communist	democratic	pit	mean	bit
department	panama	candidate	article	belief	available
prison	republic	election	player	homosexual	software

Table 4.1: Topics extracted with LDA from AP and 20 Newsgroups.

	Associated Press (AP)	**20 Newsgroups**
PMI	1.286 ± 0.35	1.175 ± 0.30
Coherence	-21.184 ± 15.58	-35.186 ± 15.32
Perplexity	17426.799	9961.437

Table 4.2: LDA base results with the Onix stop-words.

4.2 LDA with WordNet

This experiment is the first attempt for introducing WordNet in the preprocessing of the documents used as input for the classic LDA.

Preprocessing Some of the techniques used in SemEval were initially applied in the preprocessing of the datasets. However, after the first semester it was clear that some of them were not producing the expected output, so alternatives were found. For instance, the lemmatizer used for the initial experiments, from CoreNLP, had a poor performance. For example, in the same topic, the words *gun* and *guns* co-occurred, when the second one should have been transformed to *gun*. In order to improve these results, a different lemmatizer was used, from NLTK. With this new method it was possible to correct most of the problems found.

The stop-words were initially removed based on a list called the *Snowball stop-words*[1], however we began to realize that some words, that occurred many times in news, could also be considered stop-words. So a new list of stop-words was discovered from the *Onix Text Retrieval Toolkit*[2]. It is said to be the list most widely used, covering a large amount of words without being too agressive. This new list has 429 words, in comparison to the 175 words of the other, so it is expected to lead to an improvement in the topics, in terms of the appearance of too frequent and not very informative words. To verify this, both lists of stop-words were used in the first and second approach, and the result was compared.

[1] `http://snowball.tartarus.org/algorithms/english/stop.txt`
[2] `http://www.lextek.com/manuals/onix/stopwords1.html`

Approach The first experiments performed were already described in the motivation section and at that point the proposed model still had not been thought of, which means that those results made it clear that there were problems to be solved. After designing a possible solution, the first experiments with semantics took place.

In those experiments, after preprocessing, the occurrence of each word was assigned to its first sense, in WordNet. We recall that this is the sense that this word most frequently denotes, in the SemCor corpus. After having a colletion of synsets, represented by their ID, with the respective words that belong to it, we altered the content of our corpus to the IDs of the synsets instead of words. An example is given for a document containing the following words *student, school, teacher*. By accessing WordNet, we retrieved all the possible senses for each word and selected the most probable one, which is the first presented. So, for *student*, the most probable sense has the **ID 10665698**, refers to a learner who is enrolled in an educational institution and has the words *student, pupil and educatee*. *School* has the **ID 8276720**, refers to an educational institution and has the word *school*. Finally, *teacher* has the **ID 10694258**, refers to a person whose occupation is teaching and the words *teacher and instructor*. Given this information, the document is altered, so instead of having *student, school, teacher* it will have *10665698, 8276720, 10694258*. This way, the number of different words per document was reduced, given that several words may belong to the same synset.

Outcome Afterwards, by running the LDA algorithm using the preprocessed documents, the topics presented in tables 4.3 and 4.4 were obtained, for each dataset, using the list of stop-words from the *Onix Text Retrieval Toolkit*. For the sake of simplicity, we only show the top 10 synsets for each LDA topic, with their Synset ID, POS-tag, words and gloss. The underlined words are those that occur in the corpus.

These topics are based on synsets and WordNet can be used to retrieve additional information on the concept they denote, including their definition (gloss), POS and other words with the same meaning. With both models, the top words of each topic are consistently nouns, which should transmit more content. It is already evident that several words from the corpus were grouped in the same synset, consequently reducing the similarity between the content of topics. However, these topics are still evaluated with the evaluation metrics presented.

We obtained the results presented in table 4.5. It was helpful to apply these metrics to the topics obtained here, so that we could compare the different lists of stop-words. The Snowball stop-word list outperformed the Onix stop-word list, in the Coherence and Perplexity measure. However, we found that the topics obtained with the latter were more satisfying.

The underlying problem with this experiment is that we only considered the most probable sense of each word, and there may be events of a word in a sense that is not the most frequent. Given that LDA is a probabilistic model, we can try and consider the different senses of a word, which is what inspired the next experiment.

LDA with WordNet			
Synset ID	**POS**	**Words**	**Gloss**
3247620	N	drug	A substance that is used as a medicine or narcotic.
10020890	N	doctor, doc, physician, MD, Dr., medico	A licensed medical practitioner.
644503	N	survey, study	A detailed critical inspection.
1698271	V	write, compose, pen, indite	Produce a literary work.
2760116	ADJ	medical	Relating to the study or practice of medicine.
14447908	N	health, wellness	A healthy state of wellbeing free from disease.
2547586	V	help, assist, aid	Give help or assistance; be of service.
6268096	N	article	Nonfictional prose forming an independent part of a publication.
10182913	N	homosexual, homophile, homo, gay	Someone who practices homosexuality; having a sexual attraction to persons of the same sex.
10405694	N	patient	A person who requires medical care.

Table 4.3: Illustrative topics from 20 Newsgroups, obtained with Experiment 1.

LDA with WordNet			
Synset ID	**POS**	**Words**	**Gloss**
13817526	N	percentage, percent, per centum, pct	A proportion in relation to a whole (which is usually the amount per hundred).
5145118	N	monetary value, price, cost	The property of having material worth (often indicated by the amount of money something would bring if sold).
1097292	N	market, marketplace, market place	The world of commercial activity where goods and services are bought and sold.
13333833	N	stock	The capital raised by a corporation through the issue of shares entitling holders to an ownership interest (equity).
1968569	V	rise, lift, arise, move up, go up, come up, uprise	Move upward.
13664521	N	cent	A fractional monetary unit of several countries.
15286249	N	rate	A magnitude or frequency relative to a time unit.
965035	V	report, describe, account	To give an account or representation of in words.
1212469	ADJ	low	Less than normal in degree or intensity or amount.
156601	V	increase	Become bigger or greater in amount.

Table 4.4: Illustrative topics from AP, obtained with Experiment 1.

		Associated Press (AP)	**20 Newsgroups**
Onixstopwords	*PMI*	**1.167 ± 0.36**	**1.154 ± 0.35**
	Coherence	-26.492 ± 15.51	-34.928 ± 17.09
	Perplexity	13827.146	7970.342
Snowballstopwords	*PMI*	1.011 ± 0.28	0.991 ± 0.21
	Coherence	**-17.319 ± 12.97**	**-34.221 ± 15.62**
	Perplexity	**9475.427**	**7541.267**

Table 4.5: Results obtained with Experiment 1.

4.3 Semantic LDA with SemCor

The difference from the first experience to this one is that we are not only interested in the most probable synset. Here, all the senses of a word are considered.

Preprocessing The preprocessing was essentially the same. The words were lemmatized using NLTK and stop-words were removed given both lists presented in the previous chapter. The only difference is that here words have a POS tag associated with them. POS-tagging the words solves syntactical ambiguities and thus reduces the number of candidate synsets, given that there are some words that can belong to multiple grammatical classes. For example, *plant* can either be the noun plant, as in a living organism or the verb plant, as in putting seeds into the ground.

Approach First, given the SemCor 3.0 annotations, we counted the number of times a word occurred with a specific sense. Then we counted how many times each synset occurred, and, with this, it was possible to calculate the probabilities of a word given a synset. This is a straightforward task for those WordNet synsets that are in SemCor. But SemCor is a limited corpus and does not cover all words and senses in WordNet. To handle this issue, an extra preprocessing step was added, where all documents were reviewed and, when a word did not occur in SemCor, a new "dummy" synset was created with a special negative id, and probability equal to 1. This value was chosen given that the "dummy" synset would only have one word, and that word had a probability of 1, given that synset.

Outcome The next step was to run the proposed model already implemented, the SemLDA algorithm, with the new preprocessing. The topics obtained, with the Onix list, are presented in tables 4.6 and 4.7, where we only show the top 10 synsets for each topic. The format of those tables is exactly the same, however we tried to find an analogous topic by the classic LDA, to prove they share similar domains. The presented examples share many words and the other are closely related to each other (eg. *team* and *fan*, or *student* and *school*). Both topics have, sometimes, the same word in different synsets. While this might sometimes be undesirable, and a possible sign of incoherence, it also shows that the algorithm is correctly handling different senses of the same word. These situations are minimized in the next experiment, by acquiring sense probabilities from a WSD algorithm, instead of relying blindly in SemCor for this purpose. This will also minimize the number of dummy synsets.

This experiment was the first to use the algorithm proposed in this thesis and the results are in table 4.8. As such, these results were compared with the baseline. First we only compared the values of the different list of stop-words. Once again, the Snowball list surpassed the other list in almost every measure. However, since the topics produced with the Onix list are still more visually appealing, from this experiment on, this was the only list of stop-words used. Now, comparing the results obtained with the chosen stop-word list with the base results, it is noticeable that SemLDA outperformed LDA in the evaluation of PMI and topic coherence. The perplexity, on the other hand, has much worse results with our model. We think that the PMI and coherence have such high values, maybe due to the

LDA
game, fan, team, hockey, win, play, goal, pit, article, player

SemLDA with SemCor			
Synset ID	**POS**	**Words**	**Gloss**
7985384	N	team	Two or more draft animals that work together to pull something.
456199	N	game	A single play of a sport or other contest.
2152991	N	game	Animal hunted for food or sport.
430606	N	game	An amusement or pastime.
1100145	V	win	Be the winner in a contest or competition; be victorious.
2799071	N	baseball	A ball used in playing baseball.
6268096	N	article	Nonfictional prose forming an independent part of a publication.
10639925	N	sports fan, fan, rooter	An enthusiastic devotee of sports.
-1596	N	hockey	
9843956	N	batter, hitter, slugger, batsman	(baseball) a ballplayer who is batting.

Table 4.6: Illustrative (analogous) topics from 20 Newsgroups, obtained with the classic LDA (top) and with Experiment 2 (bottom).

existence of repeated words in the same topic. This was later improved by introducing WSD to the model.

LDA
drug, school, student, charge, attorney, federal, teacher, cocaine, department, prison

SemLDA with SemCor			
Synset ID	**POS**	**Words**	**Gloss**
10665698	N	student, pupil, educatee	A learner who is enrolled in an educational institution.
8276720	N	school	An educational institution.
5757536	N	school, schooling	The process of being formally educated at a school.
10694258	N	teacher, instructor	A person whose occupation is teaching.
15203229	N	school, schooltime, school day	The period of instruction in a school; the time period when school is in session.
10399491	N	parent	A father or mother; one who begets or one who gives birth to or nurtures and raises a child; a relative who plays the role of guardian.
8275185	N	school	A body of creative artists or writers or thinkers linked by a similar style or by similar teachers.
8286163	N	university	The body of faculty and students at a university.
9917593	N	child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling	A young person of either sex.
8278169	N	college	The body of faculty and students of a college.

Table 4.7: Illustrative (analogous) topics from AP, obtained with the classic LDA (top) and with Experiment 2 (bottom).

		Associated Press (AP)	**20 Newsgroups**
Onixstopwords	*PMI*	1.350 ± 0.36	**1.302 ± 0.48**
	Coherence	-19.111 ± 16.07	**-32.491 ± 12.87**
	Perplexity	23801.624	18505.091
Snowballstopwords	*PMI*	**1.429 ± 0.25**	1.168 ± 0.39
	Coherence	**-8.196 ± 10.07**	-34.090 ± 13.19
	Perplexity	**15897.313**	**10289.625**

Table 4.8: Results obtained with Experiment 2.

4.4 Semantic LDA with WSD and SemCor

Given the topics obtained from the previous experiment, it was clear that there was an issue that needed fixing: the same word could appear in the same topic with different meanings. A solution to this problem was to perform WSD on each word, to discover the most appropriate sense, based on the context where it is inserted.

Preprocessing For this experiment the preprocessing phase did not change at all and was exactly the same as in the experiment with SemCor. Given the outcome obtained in the

previous experiments, we established that there was no need to continue using the Snowball stop-words list for the next experiments, given that with the other list the results improved substantially.

Approach To perform WSD, an implementation of Adapted Lesk (Banerjee and Pedersen, 2002), provided by Tan (2014), was used. Given the context and the word in question, this algorithm returns a set of scores for each candidate synset for the word. Based on those scores, we calculated the word probability given a synset.

However, when the algorithm did not rank the senses, because there were not overlaps with any sense context and the document, we resorted to the probabilities obtained from SemCor. When neither Lesk nor SemCor returned results, a 'dummy' synset was created, just like before.

By performing WSD, computation time was increased in, approximately, two days, as compared to the previous experiment.

Outcome The topics obtained with this version of SemLDA are presented in tables 4.6 and 4.7, where we only show the top 10 synsets for each topic. The format of these tables is exactly the same as before, with the analogous topic from the classic LDA.

With this experiment we have both good and not so good topics. Some have a clear theme and others were confusing, since the synsets did not express the same theme. The topics presented in the tables are an example of the good topics.

It is not clear why this happens and, so, we were not able to find a possible solution to this problem. However, even though the topics could be better, there is no occurrence, in the same topic, of the same word in different synsets. Given these results, calculating the probabilities with WSD is a good choice.

Table 4.11, shows the results obtained with this experiment. Overall, we can say that these results are not great, especially when comparing with the previous experiment and the baseline. This was already foreseeable, given the topics obtained. The perplexity has even worse values than before and the PMI and coherence are also inferior to the baseline results.

LDA
god, christian, believe, people, write, article, sin, mean, belief, homosexual

SemLDA with SemCor and WSD			
Synset ID	**POS**	**Words**	**Gloss**
9505418	N	deity, divinity, god, immortal	Any supernatural being worshipped as controlling some part of the world or some aspect of life or who is the personification of a force.
5916739	N	impression, feeling, belief, notion, opinion	A vague idea in which some confidence is placed.
11083656	N	Jesus, Jesus of Nazareth, the Nazarene, Jesus Christ, Christ, Savior, Saviour, Good Shepherd, Redeemer, Deliverer	A teacher and prophet born in bethlehem and active in nazareth; his life and sermons form the basis for christianity (circa 4 bc - ad 29).
5946687	N	religion, faith, religious belief	A strong belief in a supernatural power or powers that control human destiny.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
9820044	N	atheist	Someone who denies the existence of god.
1260731	N	sin, hell	Violent and excited activity.
14526182	N	spirit, tone, feel, feeling, flavor, flavour, look, smell	The general atmosphere of a place or situation and the effect that it has on people.
689344	V	think, believe, consider, conceive	Judge or regard; look upon; judge.
8082602	N	church, Christian church	One of the groups of christians who have their own beliefs and forms of worship.

Table 4.9: Illustrative (analogous) topics from 20 Newsgroups, obtained with the classic LDA (top) and with Experiment 3 (bottom).

LDA
party, government, president, rebel, gorbachev, political, leader, communist, panama, republic

SemLDA with SemCor and WSD			
Synset ID	**POS**	**Words**	**Gloss**
798245	N	campaign, cause, crusade, drive, movement, effort	A series of actions advancing a principle or tending toward a particular end.
13112664	N	shrub, bush	A low woody perennial plant usually having several major stems.
11076566	N	Jackson, Jesse Jackson, Jesse Louis Jackson	United states civil rights leader who led a national campaign against racial discrimination and ran for presidential nomination (born in 1941).
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
8078020	N	family, household, house, home, menage	A social unit living together.
746718	V	order, tell, enjoin, say	Give instructions to or direct somebody to do something with authority.
9623038	N	leader	A person who rules or guides or inspires others.
15224692	N	prison term, sentence, time	The period of time a prisoner is imprisoned.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
13817526	N	percentage, percent, per centum, pct	A proportion in relation to a whole (which is usually the amount per hundred).

Table 4.10: Illustrative (analogous) topics from AP, obtained with the classic LDA (top) and with Experiment 3 (bottom).

	Associated Press (AP)	**20 Newsgroups**
PMI	0.984 ± 0.17	1.145 ± 0.33
Coherence	-29.622 ± 14.66	-42.118 ± 14.72
Perplexity	3.42E+09	9.62E+09

Table 4.11: Results obtained with Experiment 3.

4.5 Semantic LDA with Word Sense Disambiguation

This experiment makes the model more flexible, and opens the way for its adaptations to other languages that have an available WordNet, even if they do not have a sense-annotated corpus, like SemCor. Portuguese is amongst many languages that have a WordNet available or something equivalent (Gonçalo Oliveira et al., 2015) , which is why it would be interesting to perform this experiment.

Preprocessing In this final experiment the preprocessing was the same as in the experiment described in section 4.4.

Approach This final experiment had the purpose of changing the model so that it is not dependent of an annotated corpus. So, to achieve this, SemCor was put aside and whenever the probabilities from this corpora were once used, now there is the creation of a 'dummy' synset, just like it was explained before.

Outcome The final step was to run the SemLDA algorithm, with the probabilities only obtained from WSD. The topics obtained are presented in table 4.6 and 4.7, where we only show the top 10 synsets for each topic. The format of these tables is exactly the same as before, with the analogous topic from the classic LDA.

These topics were satisfying, given that, once again, there was no occurrence, in the same topic, of the same word in different synsets. And, also, because the words in all these topics do express a common theme, which is a major improvement from the previous experiment.

We found it pertinent that all of the topics, from both datasets, obtained from this experiment were presented in appendix A, given that they were the best topics of the proposed model, so far.

The results obtained are in table 4.14. These were surprisingly better than the previous experiment. The perplexity values remain above the expected value, however they were a little better than the previous experiment. We were able to surpass the PMI base values for the 20 Newsgroups dataset, which shows an improvement.

LDA
image, file, graphic, format, jpeg, color, program, bit, available, software

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3931044	N	picture, image, icon, ikon	A visual representation (of an object or scene or person or abstraction) produced on a surface.
3336839	N	file	A steel hand tool with small sharp teeth on some or all of its surfaces; used for smoothing wood or metal.
6566077	N	software, software program, computer software, software system, software package, package	(computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory.
3453696	N	graphic, computer graphic	An image that is generated by a computer.
4956594	N	color, colour, coloring, colouring	A visual attribute of things that results from the light they emit or transmit or reflect.
4677385	N	format	The general appearance of a publication.
6264398	N	mail, mail service, postal service, post	The system whereby messages are transmitted via the post office.
183053	ADJ	available	Obtainable or accessible and ready for use or service.
10741590	N	user	A person who makes use of a thing; someone who uses or employs something.
6634376	N	information, info	A message received and understood.

Table 4.12: Illustrative (analogous) topics from 20 Newsgroups, obtained with the classic LDA (top) and with Experiment 4 (bottom).

LDA
bush, dukakis, campaign, vote, republican, president, jackson, democratic, candidate, election

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13112664	N	shrub, bush	A low woody perennial plant usually having several major stems.
798245	N	campaign, cause, crusade, drive, movement, effort	A series of actions advancing a principle or tending toward a particular end.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
13421462	N	budget	A summary of intended expenditures along with proposals for how to meet them.
10522495	N	republican	An advocate of a republic (usually in opposition to a monarchy).
8161477	N	senate	Assembly possessing high legislative powers.
715140	ADJ	democratic	Characterized by or advocating or based upon the principles of democracy or social equality.
11076566	N	Jackson, Jesse Jackson, Jesse Louis Jackson	United states civil rights leader who led a national campaign against racial discrimination and ran for presidential nomination (born in 1941).
8324514	N	committee, commission	A special group delegated to consider some matter.
10002031	N	democrat, populist	An advocate of democratic principles.

Table 4.13: Illustrative (analogous) topics from AP, obtained with the classic LDA (top) and with Experiment 4 (bottom).

	Associated Press (AP)	**20 Newsgroups**
PMI	1.175 ± 0.38	1.215 ± 0.45
Coherence	-28.806 ± 16.63	-40.235 ± 12.80
Perplexity	865253609.4	1.68E+12

Table 4.14: Results obtained with Experiment 4.

4.6 Observations

After analyzing all the obtained results, we can draw some observations. Even though we cannot evidently state that our model is better than Blei's algorithm, we are moving towards it. With our model, the obtained topics are more informative, in a way that it is possible to understand the sense of each word in it. This is very helpful when it is necessary to understand clearly the theme of a specific topic.

There are still several improvements to be made, given that we want to increase the results obtained with just WSD. For this, we can explore different ways to calculate the probabilities and tools to use.

Overall, each of these experiments was helpful in their own. With the first two experiments we observed the benefits of both stop-words lists. With one, we had better results in the measures and with the other the topics were visually more attractive.

In Experiment 2 we obtained topics where several synsets had the same word repeated. Which proved that our model was taking into account the different senses of a word.

The results in Experiment 3 were so below what was expected, most likely due to the way the probabilities were being calculated, both with SemCor and WSD. However, it showed some improvements in a way that it solved the problem that occurred in the previous experiment, due to the addition of WSD. Even though this task increases the computation time significantly, we think that it is worth, given the obtained results.

Tables 4.15 and 4.16 show how the values of each metric evolved throughout the experiments, respectively, for the AP corpus and the 20 Newsgroups. In both datasets, the experiment that clearly outperformed LDA in almost every measures was the second, SemLDA with SemCor. However, since there is not a SemCor-like corpus for every language, we believe that the results of the last experiment are the most promising, as this approach is more flexible and also show improvements over the classic LDA, by introducing semantics.

	Base results	**Exp 1**	**Exp 2**	**Exp 3**	**Exp 4**
PMI	1.286 ± 0.35	1.167 ± 0.36	1.350 ± 0.36	0.984 ± 0.17	1.175 ± 0.38
Coherence	-21.184 ± 15.58	-26.492 ± 15.51	-19.111 ± 16.07	-29.622 ± 14.66	-28.806 ± 16.63
Perplexity	17426.799	13827.146	23801.624	3.42E+09	865253609.4

Table 4.15: Evolution of the results with the AP corpus.

	Base results	**Exp. 1**	**Exp. 2**	**Exp. 3**	**Exp 4**
PMI	1.175 ± 0.30	1.154 ± 0.35	1.302 ± 0.48	1.145 ± 0.33	1.215 ± 0.45
Coherence	-35.186 ± 15.32	-34.928 ± 17.09	-32.491 ± 12.87	-42.118 ± 14.72	-40.235 ± 12.8
Perplexity	9961.437	7970.342	18505.091	9.62E+09	1.68E+12

Table 4.16: Evolution of the results with the 20 Newsgroups corpus.

Perplexity was the only measure that continuously had lower results than LDA. We should examine closely our model and figure out a solution to this problem.

Chapter 5

Concluding remarks

Throughout the years, a lot of work has been done in the field of machine learning, natural language processing and artificial intelligence, more specifically in the area of topic modelling. Latent Dirichlet Allocation (LDA) is, perhaps, the most cited algorithm that performs topic modelling. But, even though the LDA algorithm is such a popular topic model, it does not mean that it does not have limitations and generates always perfect topics, which we confirmed after a thorough analysis on the topics that were produced. As explained in this thesis, LDA's limitations were identifiable, not only with the naked eye, but also with the contribution of a simple computer program developed by us, specifically for this purpose. These limitations led to the formulation of a model that could solve this problem, which meant taking into account the possible meanings of the words when creating the topics.

An appropriate amount of time of the first semester was dedicated to examining existing models and techniques that already attempted to solve this problem. The conclusions reached were that there is not a model created exactly for what is proposed in this thesis, which was a relevant discovery since it inspired us to succeed in designing our model.

So far, we have successfully implemented a fully functional model. It is based on LDA, with some modifications, but it still uses the same method to perform inference, which is variational inference. The changes to the model itself are not that innovative. However, we have a specific method to process the data and with different techniques working alongside the algorithm, that is how our model stands out. The participation in the SemEval 2015 task was very beneficial, since it improved the preprocessing method and introduced some different techniques. Also, by participating in SemEval 2015 we managed to have a paper accepted on the approach we submitted.

We developed different variations of our model, which are explained in chapter 4. In all of them, the model accesses external resources, such as WordNet, which was an idea already explored in the related work. These versions all helped in the process if validating our model.

One of the experiments consisted of using probabilities that were obtained from SemCor, a semantically-annotated corpus. With this experiment we noticed that semantics had been successfully introduced in the model, given that we had topics with repeated words but with different senses. At this phase of our work, we were able to get a paper published (Ferrugento et al., 2015) in the Portuguese Conference on Artificial Intelligence (EPIA 2015), to be held in Coimbra. This paper introduces our approach, SemLDA, and presented the results obtained

with the previous experiment.

This meant that we were in the right direction. However, those repeated words should not co-occur. So, we applied WSD to correct it. First, we experimented a model which used both probabilities from SemCor and WSD and then we changed it to just use WSD. We acknowledged that if we wanted to adapt our model to different languages, such as Portuguese, we needed to put SemCor aside. SemCor is specific for the English WordNet and has not yet been replicated for most languages, unfortunately.

In the experiment where only WSD was used, we obtained the results nearest to the baseline. After performing all of the experiments, we considered the previous to be the best approach, since it will allow us to adapt to other languages more easily.

We still have not outperformed the state of the art models in all the evaluation metrics, but we are close to it. The biggest advantage that our model has is that it produces topics that are more informative, given that WordNet allows us to retrieve all type of information, like a gloss, part-of-speech and synonyms. When we surpass the classic LDA, our model can enhance, even more, search engines, browsing the Internet and the summarization of documents.

Since it might be of interest for the rest of the community, we decided to share our algorithm, given that it solves a serious problem in LDA and in an innovative way. So, our algorithm, SemLDA, is avaliable in `https://github.com/aferrugento`.

In order to be recognized as valuable work presented by the scientific community in the field, at least one more research paper will be written and submitted to a relevant conference, describing the word sense disambiguation approach and the adaptation to other languages.

As future work there is still a lot we plan on doing. Our main focus is to adapt our model to the Portuguese language, by using one or more of the available wordnets for this language (Gonçalo Oliveira et al., 2015). After this, it will be possible to apply our algorithm to events synopsis, a part of the work in Infocrowds, given that the content is in Portuguese. Another possible application of our work is the next edition of SemEval.

We intend to use datasets that do not have journalistic content, so we can analyze how our algorithm behaves. For instance, SemLDA should be tested in other kinds of text, such as encyclopedia articles or social network text, just to mention a few kinds of text available in large quantities. We can also broaden the semantic relations being considered, and start taking into account hypernymy and hyponymy of the words.

Another important task we have is to improve our approach that uses word sense disambiguation. For this, we will experiment other algorithms, such as those mentioned in section 2.1.

When we are able to gather a large number of people, we also aim at assessing the quality of the topics manually. A way to do this is with the topic intrusion test, referred in the state of the art, so that we can verify if our topics make that big of a difference, given that they are more informative.

To conclude, we believe that the work developed throughout this year will be of great value to the community. We obtained some interesting results in our experiments and in a near future we have high expectations that the results will be even better. At the first opportunity, the current outcome of this thesis is going to be applied in the Infocrowds project.

References

Abramowitz, M., Stegun, I. A., et al. (1966). Handbook of mathematical functions. *Applied Mathematics Series*, 55:62.

Agirre, E., Lacalle, O. L. D., and Soroa, A. (2009). Knowledge-based WSD on specific domains: performing better than generic supervised WSD. In *Proceedings of 21st International Joint Conference on Artificial Intelligence*, IJCAI 2009, pages 1501–1506, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Agirre, E. and Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, pages 33–41. ACL Press.

Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., and Stent, A. (2001). Toward conversational human-computer interaction. *AI magazine*, 22(4):27.

Baker, L. D. and McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–103. ACM.

Banerjee, S. and Pedersen, T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, LNCS, pages 136–145, London, UK. Springer.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*, volume 1. springer New York.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Boyd-graber, J. and Blei, D. (2007). A Topic Model for Word Sense Disambiguation. (June):1024–1033.

Chan, Y. S., Ng, H. T., and Chiang, D. (2007). Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic. ACL Press.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Chemudugunta, C., Holloway, A., Smyth, P., and Steyvers, M. (2008). *Modeling documents by combining semantic concepts with unsupervised statistical learning.* Springer.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

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

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, MA.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Ferrugento, A., Alves, A. O., Oliveira, H. G., and Rodrigues, F. (2015). Towards the improvement of a topic model with semantic knowledge. In *17th Portuguese Conference on Artificial Intelligence (EPIA 2015)*, LNCS, page in press. Springer.

Flaherty, P., Giaever, G., Kumm, J., Jordan, M. I., and Arkin, A. P. (2005). A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293.

Gonçalo Oliveira, H., de Paiva, V., Freitas, C., Rademaker, A., Real, L., and Simões, A. (2015). As wordnets do português. In Simões, A., Barreiro, A., Santos, D., Sousa-Silva, R., and Tagnin, S. E. O., editors, *Linguística, Informática e Tradução: Mundos que se Cruzam*, volume 7(1) of *OSLa: Oslo Studies in Language*, pages 397–424. University of Oslo.

Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544.

Guo, W. and Diab, M. (2011). Semantic topic models: combining word distributional statistics and dictionary definitions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 552–561. Association for Computational Linguistics.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, Englewood Cliffs, NJ, 2nd edition.

Kuczma, M. (2009). *An introduction to the theory of functional equations and inequalities: Cauchy's equation and Jensen's inequality*. Springer Science & Business Media.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

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

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Martin, J. H. and Jurafsky, D. (2000). Speech and language processing. *International Edition*.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Proceedings of ARPA Human Language Technology Workshop*, Plainsboro, NJ, USA.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272. ACL Press.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. In *SOFSEM 2012: Theory and practice of computer science*, pages 115–129. Springer.

Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(7):1075–1086.

Newman, D., Bonilla, E. V., and Buntine, W. (2011). Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems*, pages 496–504.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108. ACL Press.

Olsher, D. J. (2013). Cogview & intelnet: Nuanced energy-based knowledge representation and integrated cognitive-conceptual framework for realistic culture, values, and concept-affected systems simulation. In *Computational Intelligence for Human-like Intelligence (CIHLI), 2013 IEEE Symposium on*, pages 82–91. IEEE.

Pedersen, T. and Bruce, R. (1997). A new supervised learning algorithm for word sense disambiguation. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference*, AAAI/IAAI, pages 604–609.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, ACL 2012, pages 1522–1531, Uppsala, Sweden. ACL Press.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM.

Rajagopal, D., Cambria, E., Olsher, D., and Kwok, K. (2013a). A graph-based approach to commonsense concept extraction and semantic similarity detection. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 565–570. International World Wide Web Conferences Steering Committee.

Rajagopal, D., Olsher, D., Cambria, E., and Kwok, K. (2013b). Commonsense-based topic modeling. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 6. ACM.

Rau, L. F., Jacobs, P. S., and Zernik, U. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419–428.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57, Washington, D.C.

Tan, L. (2014). Pywsd: Python implementations of word sense disambiguation (wsd) technologies [software]. https://github.com/alvations/pywsd.

Tang, G., Xia, Y., Sun, J., Zhang, M., and Zheng, T. F. (2014). Topic models incorporating statistical word senses. In *Computational Linguistics and Intelligent Text Processing*, pages 151–162. Springer.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Voorhees, E. M. (1999). Natural language processing and information retrieval. In *Information*

Extraction, pages 32–48. Springer.

Wang, C., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. ACL Press.

Zar, J. H. (1998). Spearman rank correlation. *Encyclopedia of Biostatistics*.

Zhong, Z. and Ng, H. T. (2012). Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 273–282, Stroudsburg, PA, USA. ACL Press.

Appendix A

Topics obtained from Experiment 4

A.1 Topics with 20 Newsgroups

This appendix lists all the topics obtained in Experiment 4 for the corpus 20 Newsgroups. The synset-based topics are shown in tables A.1 to A.15.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3931044	N	picture, image, icon, ikon	A visual representation (of an object or scene or person or abstraction) produced on a surface.
3336839	N	file	A steel hand tool with small sharp teeth on some or all of its surfaces; used for smoothing wood or metal.
6566077	N	software, software program, computer software, software system, software package, package	(computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory.
3453696	N	graphic, computer graphic	An image that is generated by a computer.
4956594	N	color, colour, coloring, colouring	A visual attribute of things that results from the light they emit or transmit or reflect.
4677385	N	format	The general appearance of a publication.
6264398	N	mail, mail service, postal service, post	The system whereby messages are transmitted via the post office.
183053	ADJ	available	Obtainable or accessible and ready for use or service.
10741590	N	user	A person who makes use of a thing; someone who uses or employs something.
6634376	N	information, info	A message received and understood.

Table A.1: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9682291	N	Muslim, Moslem	A believer in or follower of islam.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
6352117	N	Armenian, Armenian alphabet	A writing system having an alphabet of 38 letters in which the armenian language is written.
9681351	N	Jew, Hebrew, Israelite	A person belonging to the worldwide group claiming descent from jacob (or converted to it) and connected by cultural or religious ties.
8792295	N	Israel	An ancient kingdom of the hebrew tribes at the southeastern end of the mediterranean sea; founded by saul around 1025 bc and destroyed by the assyrians in 721 bc.
3247620	N	drug	A substance that is used as a medicine or narcotic.
9729530	N	Arab, Arabian	A member of a semitic people originally from the arabian peninsula and surrounding territories who speaks arabic and who inhabits much of the middle east and northern africa.
973077	N	war, warfare	The waging of armed conflict against an enemy.
11410625	N	consequence, effect, outcome, result, event, issue, upshot	A phenomenon that follows and is caused by some previous phenomenon.
1323958	V	kill	Cause to die; put to death, usually intentionally or knowingly.

Table A.2: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9394007	N	planet, major planet	(astronomy) any of the nine large celestial bodies in the solar system that revolve around the sun and shine by reflected light; mercury, venus, earth, mars, jupiter, saturn, uranus, neptune, and pluto in order of their proximity to the sun; viewed from the constellation hercules, all the planets rotate around the sun in a counterclockwise direction.
6839190	N	space, blank	A blank character used to separate successive words in writing or printing.
9358550	N	moon	Any object resembling a moon.
998886	V	write, save	Record data on a computer.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
8403225	N	mission, missionary post, missionary station, foreign mission	An organization of missionaries in a foreign land sent to carry on religious work.
4264914	N	spacecraft, ballistic capsule, space vehicle	A craft capable of traveling in outer space; technically, a satellite around the sun.
9270894	N	Earth, earth, world, globe	The 3rd planet from the sun; the planet we live on.
14514039	N	sphere, domain, area, orbit, field, arena	A particular environment or walk of life.
4137444	N	satellite, artificial satellite, orbiter	Man-made equipment that orbits around the earth or the moon.

Table A.3: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
5174653	N	right	An abstract idea of that which is due to a person or governmental body by law or tradition or nature; it is something that nobody can take away".
4565375	N	weapon, arm, weapon system	Any instrument or instrumentality used in fighting or hunting.
8050678	N	government, authorities, regime	The organization that is the governing authority of a political unit.
9682291	N	Muslim, Moslem	A believer in or follower of islam.
6268096	N	article	Nonfictional prose forming an independent part of a publication.
10405694	N	patient	A person who requires medical care.
1699896	V	spell, write	Write or name the letters that comprise the conventionally accepted form of (a word or part of a word).
7192129	N	call, claim	A demand especially in the phrase "the call of duty".
5814291	N	topic, subject, issue, matter	Some situation or event that is thought about.

Table A.4: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3180969	N	detector, sensor, sensing element	Any device that receives a signal or stimulus (as heat or pressure or light or motion etc.) and responds to it in a distinctive manner.
4004767	N	printer	(computer science) an output device that prints the results of data processing.
812526	N	clasp, clench, clutch, clutches, grasp, grip, hold	The act of grasping.
6798750	N	mark, print	A visible indication made on a surface.
2958343	N	car, auto, automobile, machine, motorcar	A motor vehicle with four wheels; usually propelled by an internal combustion engine.
6264398	N	mail, mail service, postal service, post	The system whereby messages are transmitted via the post office.
-10088	V	windows	
998886	V	write, save	Record data on a computer.
6566077	N	software, software program, computer software, software system, software package, package	(computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory.
15122231	N	time	An indefinite period (usually marked by specific attributes or activities).

Table A.5: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
6825399	N	font, fount, typeface, face, case	A specific size and style of type within a type family.
3336839	N	file	A steel hand tool with small sharp teeth on some or all of its surfaces; used for smoothing wood or metal.
5154676	N	resource	A source of aid or support that may be drawn upon when needed.
3793489	N	mouse, computer mouse	A hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad.
1031256	V	mail, post, send	Cause to be directed or transmitted to another place.
1158872	V	use, utilize, utilise, apply, employ	Put into service; make work or employ for a particular purpose or for its inherent or natural purpose.
13774404	N	batch, deal, flock, good deal, great deal, hatful, heap, lot, mass, mess, mickle, mint, mountain, muckle, passel, peck, pile, plenty, pot, quite a little, raft, sight, slew, spate, stack, tidy sum, wad	(often followed by 'of') a large number or amount or extent.
15122231	N	time	An indefinite period (usually marked by specific attributes or activities).
14526182	N	spirit, tone, feel, feeling, flavor, flavour, look, smell	The general atmosphere of a place or situation and the effect that it has on people.
3104594	N	copy	A thing made to be similar or identical to another thing.

Table A.6: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9222051	N	bit, chip, flake, fleck, scrap	A small fragment of something broken off from the whole.
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
8590909	N	key, paint	(basketball) a space (including the foul line) in front of the basket at each end of a basketball court; usually painted a different color from the rest of the court.
3045228	N	clipper, clipper ship	A fast sailing ship used in former times.
615887	N	encoding, encryption	The activity of converting data or information into code.
1277097	ADJ	cardinal, central, fundamental, key, primal	Serving as an essential component.
7111047	N	phone, speech sound, sound	(phonetics) an individual sound unit of speech without concern as to whether or not it is a phoneme of some language.
8462320	N	data, information	A collection of facts from which conclusions may be drawn.
5847438	N	algorithm, algorithmic rule, algorithmic program	A precise rule (or set of rules) specifying how to solve some problem.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.

Table A.7: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
5916739	N	impression, feeling, belief, notion, opinion	A vague idea in which some confidence is placed.
9466280	N	universe, existence, creation, world, cosmos, macrocosm	Everything that exists anywhere.
5888929	N	hypothesis, possibility, theory	A tentative insight into the natural world; a concept that is not yet verified but that if true would explain certain facts or phenomena.
11410625	N	consequence, effect, outcome, result, event, issue, upshot	A phenomenon that follows and is caused by some previous phenomenon.
6648724	N	argument, statement	A fact or assertion offered as evidence that something is true.
11452218	N	energy, free energy	(physics) a thermodynamic quantity equivalent to the capacity of a physical system to do work; the units of energy are joules or ergs.
6223669	N	theism	The doctrine or belief in the existence of a god or gods.
11428023	N	beam, beam of light, light beam, ray, ray of light, shaft, shaft of light, irradiation	A column of light (as from a beacon).
5981230	N	aim, object, objective, target	The goal intended to be attained (and which is believed to be attainable).
1548193	ADJ	moral	Concerned with principles of right and wrong or conforming to standards of behavior and character based on those principles.

Table A.8: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
2958343	N	car, auto, automobile, machine, motorcar	A motor vehicle with four wheels; usually propelled by an internal combustion engine.
998886	V	write, save	Record data on a computer.
2834778	N	bicycle, bike, wheel, cycle	A wheeled vehicle that has two wheels and is moved by foot pedals.
6392935	N	article, clause	A separate section of a legal document (as a statute or contract or will).
1699896	V	spell, write	Write or name the letters that comprise the conventionally accepted form of (a word or part of a word).
8131530	N	Department of Defense, Defense Department, United States Department of Defense, Defense, DoD	The federal department responsible for safeguarding national security of the united states; created in 1947.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
2670683	N	accelerator, accelerator pedal, gas pedal, gas, throttle, gun	A pedal that controls the throttle valve.
6268096	N	article	Nonfictional prose forming an independent part of a publication.
3684823	N	locomotive, engine, locomotive engine, railway locomotive	A wheeled vehicle consisting of a self-propelled engine that is used to draw trains along railway tracks.

Table A.9: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9505418	N	deity, divinity, god, immortal	Any supernatural being worshipped as controlling some part of the world or some aspect of life or who is the personification of a force.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
9678009	N	Christian	A religious person who believes jesus is the christ and who is a member of a christian denomination.
10182913	N	homosexual, homophile, homo, gay	Someone who practices homosexuality; having a sexual attraction to persons of the same sex.
6431740	N	Bible, Christian Bible, Book, Good Book, Holy Scripture, Holy Writ, Scripture, Word of God, Word	The sacred writings of the christian religions.
5946687	N	religion, faith, religious belief	A strong belief in a supernatural power or powers that control human destiny.
8082602	N	church, Christian church	One of the groups of christians who have their own beliefs and forms of worship.
5916739	N	impression, feeling, belief, notion, opinion	A vague idea in which some confidence is placed.
1260731	N	sin, hell	Violent and excited activity.
856847	N	homosexuality, homosexualism, homoeroticism, queerness, gayness	A sexual attraction to (or sexual relations with) persons of the same sex.

Table A.10: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
13480848	N	fire, flame, flaming	The process of combustion of inflammable materials producing heat and light and (often) smoke.
8136260	N	Federal Bureau of Investigation, FBI	A federal law enforcement agency that is the principal investigative arm of the department of justice.
9146813	N	Waco	A city in east central texas.
9917593	N	child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling	A young person of either sex.
8050678	N	government, authorities, regime	The organization that is the governing authority of a political unit.
8140219	N	Bureau of Alcohol Tobacco and Firearms, ATF	The law enforcement and tax collection agency of the treasury department that enforces federal laws concerning alcohol and tobacco products and firearms and explosives and arson.
6268096	N	article	Nonfictional prose forming an independent part of a publication.
345761	V	get down, begin, get, start out, start, set about, set out, commence	Take the first step or steps in carrying out an action.
10902409	N	Clinton, DeWitt Clinton	United states politician who as governor of new york supported the project to build the erie canal (1769-1828).

Table A.11: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3033986	N	circuit board, circuit card, board, card, plug-in, add-in	A printed circuit that can be inserted into expansion slots in a computer to increase the computer's capabilities.
9222051	N	bit, chip, flake, fleck, scrap	A small fragment of something broken off from the whole.
3924069	N	phonograph record, phonograph recording, record, disk, disc, platter	Sound recording consisting of a disk with a continuous groove; used to reproduce music by rotating while a phonograph needle tracks in the groove.
3702719	N	macintosh, mackintosh, mac, mack	A waterproof raincoat made of rubberized fabric.
1208797	N	thanks	With the help of or owing to.
4245218	N	small computer system interface, SCSI	Interface consisting of a standard port between a computer and its peripherals that is used in some computers.
6566077	N	software, software program, computer software, software system, software package, package	(computer science) written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory.
2995345	N	central processing unit, CPU, C.P.U., central processor, processor, mainframe	(computer science) the part of a computer (a microprocessor chip) that does most of the data processing.
3578656	N	interface, port	(computer science) computer circuit consisting of the hardware and associated circuitry that links one device with another (especially a computer and a hard disk drive or other peripherals).
3777754	N	modem	(from a combination of modulate and demodulate) electronic equipment consisting of a device used to connect computers by a telephone line.

Table A.12: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
430606	N	game	An amusement or pastime.
10439851	N	player, participant	A person who participates in or is skilled at some game.
7985384	N	team	Two or more draft animals that work together to pull something.
1111816	V	score, hit, tally, rack up	Gain points in a game.
471613	N	baseball, baseball game	A ball game played with a bat and ball between two teams of nine players; teams take turns at bat trying to score runs.
8208560	N	team, squad	A cooperative unit (especially in sports).
10077593	N	fan, buff, devotee, lover	An ardent follower and admirer.
8231184	N	league, conference	An association of sports teams that organizes matches for its members.
920336	V	determine, check, find out, see, ascertain, watch, learn	Find out, learn, or determine with certainty, usually by making an inquiry or other effort.
15122231	N	time	An indefinite period (usually marked by specific attributes or activities).

Table A.13: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9505418	N	deity, divinity, god, immortal	Any supernatural being worshipped as controlling some part of the world or some aspect of life or who is the personification of a force.
11083656	N	Jesus, Jesus of Nazareth, the Nazarene, Jesus Christ, Christ, Savior, Saviour, Good Shepherd, Redeemer, Deliverer	A teacher and prophet born in bethlehem and active in nazareth; his life and sermons form the basis for christianity (circa 4 bc - ad 29).
10388440	N	overlord, master, lord	A person who has general authority over others.
9536363	N	Godhead, Lord, Creator, Maker, Divine, God Almighty, Almighty, Jehovah	Terms referring to the judeo-christian god.
8082602	N	church, Christian church	One of the groups of christians who have their own beliefs and forms of worship.
14526182	N	spirit, tone, feel, feeling, flavor, flavour, look, smell	The general atmosphere of a place or situation and the effect that it has on people.
1009240	V	state, say, tell	Express in words.
4827957	N	sin, sinfulness, wickedness	Estrangement from god.
11161412	N	Mary, Virgin Mary, The Virgin, Blessed Virgin, Madonna	The mother of jesus; christians refer to her as the virgin mary; she is especially honored by roman catholics.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.

Table A.14: Illustrative topic from 20 Newsgroups obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8078020	N	family, household, house, home, menage	A social unit living together.
913065	V	shout, shout out, cry, call, yell, scream, holler, hollo, squall	Utter a sudden loud cry.
6352117	N	Armenian, Armenian alphabet	A writing system having an alphabet of 38 letters in which the armenian language is written.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
9041785	N	Istanbul, Stambul, Stamboul, Constantinople	The largest city and former capital of turkey; rebuilt on the site of ancient byzantium by constantine i in the fourth century; renamed constantinople by constantine who made it the capital of the byzantine empire; now the seat of the eastern orthodox church.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
9917593	N	child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling	A young person of either sex.
15122231	N	time	An indefinite period (usually marked by specific attributes or activities).
345761	V	get down, begin, get, start out, start, set about, set out, commence	Take the first step or steps in carrying out an action.
10029729	N	dragon, tartar	A fiercely vigilant and unpleasant woman.

Table A.15: Illustrative topic from 20 Newsgroups obtained with SemLDA.

A.2 Topics with Associated Press (AP)

This appendix lists all the topics obtained in Experiment 4 for the corpus AP. The synset-based topics are shown in tables A.16 to A.39.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8208016	N	force, personnel	Group of people willing to obey orders.
10210137	N	insurgent, insurrectionist, freedom fighter, rebel	A person who takes part in an armed rebellion against the constituted authority (especially in the hope of improving conditions).
8274354	N	troop	A group of soldiers.
9612447	N	Contra	A member of the guerrilla force that opposed a left-wing government in nicaragua.
1517081	ADJ	military	Associated with or performed by members of the armed services as contrasted with civilians.
1207609	N	aid, assist, assistance, help	The activity of contributing to the fulfillment of a need or furtherance of an effort or purpose.
1518386	ADJ	military	Characteristic of or associated with soldiers or the military.
2859184	N	boater, leghorn, Panama, Panama hat, sailor, skimmer, straw hat	A stiff hat made of straw with a flat crown.
1124794	N	government, governing, governance, government activity, administration	The act of governing; exercising authority.
2207647	N	soldier	A wingless sterile ant or termite having a large head and powerful jaws adapted for defending the colony.

Table A.16: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9887850	N	caller, company	A social or business visitor.
15224692	N	prison term, sentence, time	The period of time a prisoner is imprisoned.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
8264897	N	party, company	A band of people associated temporarily in some activity.
8059412	N	corporation, corp	A business firm whose articles of incorporation have been approved in some state.
10372373	N	official, functionary	A worker who holds or is invested with an office.
9976728	N	creditor	A person to whom money is owed by a debtor; someone to whom an obligation exists.
10225219	N	judge, justice, jurist	A public official authorized to decide questions brought before a court of justice.
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
434374	V	fail, go bad, give way, die, give out, conk out, go, break, break down	Stop operating or functioning.

Table A.17: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3247620	N	drug	A substance that is used as a medicine or narcotic.
8264897	N	party, company	A band of people associated temporarily in some activity.
13308999	N	tax, taxation, revenue enhancement	Charge against a citizen's person or property or activity for the support of government.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
582388	N	occupation, business, job, line of work, line	The principal activity in your life that you do to earn money.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
8143321	N	Internal Revenue Service, IRS	The bureau of the treasury department responsible for tax collections.
6548671	N	tax return, income tax return, return	Document giving the tax collector information about the taxpayer's tax liability.
5726596	N	arrangement, organization, organisation, system	An organized structure for arranging or classifying.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.

Table A.18: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13112664	N	shrub, bush	A low woody perennial plant usually having several major stems.
798245	N	campaign, cause, crusade, drive, movement, effort	A series of actions advancing a principle or tending toward a particular end.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
13421462	N	budget	A summary of intended expenditures along with proposals for how to meet them.
10522495	N	republican	An advocate of a republic (usually in opposition to a monarchy).
8161477	N	senate	Assembly possessing high legislative powers.
715140	ADJ	democratic	Characterized by or advocating or based upon the principles of democracy or social equality.
11076566	N	Jackson, Jesse Jackson, Jesse Louis Jackson	United states civil rights leader who led a national campaign against racial discrimination and ran for presidential nomination (born in 1941).
8324514	N	committee, commission	A special group delegated to consider some matter.
10002031	N	democrat, populist	An advocate of democratic principles.

Table A.19: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
8078020	N	family, household, house, home, menage	A social unit living together.
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
8523483	N	center, centre, middle, heart, eye	An area that is approximately central within some larger region.
10187557	N	hostage, surety	A prisoner who is held by one party to insure that another party will meet specified terms.
746718	V	order, tell, enjoin, say	Give instructions to or direct somebody to do something with authority.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
7319652	N	miscarriage, abortion	Failure of a plan.
10162991	N	head, chief, top dog	A person who is in charge.
1861205	ADJ	public	Not private; open to or concerning the people as a whole.

Table A.20: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8082899	N	church	The body of people who attend or belong to a particular local church.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
6613686	N	movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick	A form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement.
2622234	V	connect, link, link up, join, unite	Be or become joined or united or linked.
4446276	N	toilet, lavatory, lav, can, john, privy, bathroom	A room or building equipped with one or more toilets.
6998748	N	artwork, art, graphics, nontextual matter	Photographs or other visual representations in a printed publication.
10453533	N	pope, Catholic Pope, Roman Catholic Pope, pontiff, Holy Father, Vicar of Christ, Bishop of Rome	The head of the roman catholic church.
9857200	N	bishop	A senior member of the christian clergy having spiritual and administrative authority; appointed in christian churches to oversee priests or ministers; considered in some churches to be successors of the twelve apostles of christ.
6267145	N	newspaper, paper	A daily or weekly publication on folded sheets; contains news and articles and advertisements.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.

Table A.21: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
2691156	N	airplane, aeroplane, plane	An aircraft that has a fixed wing and is powered by propellers or jets.
10433164	N	pilot, airplane pilot	Someone who is licensed to operate an aircraft in flight.
2690081	N	airline, airline business, airway	A commercial enterprise that provides scheduled flights for passengers.
300441	N	air travel, aviation, air	Travel via aircraft.
2692232	N	airport, airdrome, aerodrome, drome	An airfield equipped with control tower and hangars as well as accommodations for passengers and cargo.
301192	N	flight	A scheduled trip by plane between designated airports.
9917593	N	child, kid, youngster, minor, shaver, nipper, small fry, tiddler, tike, tyke, fry, nestling	A young person of either sex.
2686568	N	aircraft	A vehicle that can fly.
2725829	ADJ	federal	Of or relating to the central government of a federation.
10403876	N	passenger, rider	A traveler riding in a vehicle (a boat or bus or car or plane or train etc) who is not operating it.

Table A.22: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
1323958	V	kill	Cause to die; put to death, usually intentionally or knowingly.
6711159	N	fire, attack, flak, flack, blast	Intense adverse criticism.
8078020	N	family, household, house, home, menage	A social unit living together.
10372373	N	official, functionary	A worker who holds or is invested with an office.
1793177	V	hurt, wound, injure, bruise, offend, spite	Hurt the feelings of.
13480848	N	fire, flame, flaming	The process of combustion of inflammable materials producing heat and light and (often) smoke.
7469325	N	mile	A footrace extending one mile.
2958343	N	car, auto, automobile, machine, motorcar	A motor vehicle with four wheels; usually propelled by an internal combustion engine.

Table A.23: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8913434	N	Iraq, Republic of Iraq, Al-Iraq, Irak	A republic in the middle east in western asia; the ancient civilization of mesopotamia was in the area now known as iraq.
8929243	N	Kuwait, State of Kuwait, Koweit	An arab kingdom in asia on the northwestern coast of the persian gulf; a major source of petroleum.
9714694	N	Iraqi, Iraki	A native or inhabitant of iraq.
11068401	N	Hussein, Husain, Husayn, Saddam Hussein, Saddam, Saddam bin Hussein at-Takriti	Iraqi leader who waged war against iran; his invasion of kuwait led to the gulf war (born in 1937).
9296121	N	gulf	An arm of a sea or ocean partly enclosed by land; larger than a bay.
3075191	ADJ	Iranian, Persian	Of or relating to iran or its people or language or culture.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
10372373	N	official, functionary	A worker who holds or is invested with an office.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
8910668	N	Iran, Islamic Republic of Iran, Persia	A theocratic islamic republic in the middle east in western asia; iran was the core of the ancient empire that was known as persia until 1935; rich in oil.

Table A.24: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
3649459	N	court, lawcourt, court of law, court of justice	A tribunal that is presided over by a magistrate or by one or more judges who administer justice according to the laws.
10249950	N	lawyer, attorney	A professional person authorized to practice law; conducts lawsuits or gives legal advice.
10225219	N	judge, justice, jurist	A public official authorized to decide questions brought before a court of justice.
4005630	N	prison, prison house	A correctional institution where persons are confined while on trial or for punishment.
220522	N	murder, slaying, execution	Unlawful premeditated killing of a human being by a human being.
15224692	N	prison term, sentence, time	The period of time a prisoner is imprisoned.
6561942	N	charge, complaint	(criminal law) a pleading describing some wrong or offense.
791078	N	test, trial, run	The act of testing something.
8078020	N	family, household, house, home, menage	A social unit living together.

Table A.25: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13664521	N	cent	A fractional monetary unit of several countries.
14980579	N	petroleum, crude oil, crude, rock oil, fossil oil, oil	A dark oil consisting mainly of hydrocarbons.
7164546	N	offer, offering	Something offered (as a proposal or bid).
3408721	N	future	Bulk commodities bought or sold at an agreed price for delivery at a specified future date.
1212469	ADJ	low	Less than normal in degree or intensity or amount.
5145118	N	monetary value, price, cost	The property of having material worth (often indicated by the amount of money something would bring if sold).
8403907	N	Iraqi National Congress, INC	A heterogeneous collection of groups united in their opposition to saddam hussein's government of iraq; formed in 1992 it is comprised of sunni and shiite arabs and kurds who hope to build a new government.
8264897	N	party, company	A band of people associated temporarily in some activity.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
9791530	N	analyst	An expert who studies financial data (on credit or securities or sales or financial patterns etc.) and recommends appropriate business actions.

Table A.26: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13395897	N	dollar, dollar bill, one dollar bill, buck, clam	A piece of paper money worth one dollar.
5145118	N	monetary value, price, cost	The property of having material worth (often indicated by the amount of money something would bring if sold).
13333833	N	stock	The capital raised by a corporation through the issue of shares entitling holders to an ownership interest (equity).
8420278	N	depository financial institution, bank, banking concern, banking company	A financial institution that accepts deposits and channels the money into lending activities.
79398	N	trading	Buying or selling securities or commodities.
8072837	N	market, securities industry	The securities markets in the aggregate.
13342135	N	share	Any of the equal portions into which the capital stock of a corporation is divided and ownership of which is evidenced by a stock certificate.
13367070	N	store, stock, fund	A supply of something available for future use.
816481	ADJ	late	Being or occurring at an advanced period of time or after a usual or expected time.
10720453	N	trader, bargainer, dealer, monger	Someone who purchases and maintains an inventory of goods to be sold.

Table A.27: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8320201	N	soviet	An elected governmental council in a communist country (especially one that is a member of the union of soviet socialist republics).
8500433	N	outer space, space	Any location outside the earth's atmosphere.
4211970	N	shuttle	Public transport that consists of a bus or train or airplane that plies back and forth between two points.
8126290	N	National Aeronautics and Space Administration, NASA	An independent agency of the united states government responsible for aviation and spaceflight.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
2959406	ADJ	Soviet	Of or relating to or characteristic of the former soviet union or its people.
9270894	N	Earth, earth, world, globe	The 3rd planet from the sun; the planet we live on.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
6636524	N	record, record book, book	A compilation of the known facts regarding something or someone.
103140	N	launching, launch	The act of propelling with force.

Table A.28: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
10557854	N	scholar, scholarly person, bookman, student	A learned person (especially in the humanities); someone who by long study has gained mastery in one or more disciplines.
183505	N	vote, ballot, voting, balloting	A choice that is made by counting the number of people in favor of each alternative.
181781	N	election	A vote to select the winner of a position or political office.
8277393	N	school	An educational institution's faculty and students.
8256968	N	party, political party	An organization to gain political power.
798245	N	campaign, cause, crusade, drive, movement, effort	A series of actions advancing a principle or tending toward a particular end.
9889691	N	campaigner, candidate, nominee	A politician who is running for public office.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
15203229	N	school, schooltime, school day	The period of instruction in a school; the time period when school is in session.
13817526	N	percentage, percent, per centum, pct	A proportion in relation to a whole (which is usually the amount per hundred).

Table A.29: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
6613686	N	movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick	A form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement.
8078020	N	family, household, house, home, menage	A social unit living together.
2603056	V	unite, unify	Bring together for a common purpose or action or ideology or in a shared situation.
10467395	N	President of the United States, United States President, President, Chief Executive	The person who holds the office of head of state of the united states government.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
6947479	N	American English, American language, American	The english language as used in the united states.
6636524	N	record, record book, book	A compilation of the known facts regarding something or someone.
8159924	N	York, House of York	The english royal house (a branch of the plantagenet line) that reigned from 1461 to 1485; its emblem was a white rose.
6642138	N	news, intelligence, tidings, word	Information about recent and important events.
15224692	N	prison term, sentence, time	The period of time a prisoner is imprisoned.

Table A.30: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
3956922	N	plant, works, industrial plant	Buildings for carrying on industrial labor.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
6737394	N	contract, declaration	(contract bridge) the highest bid becomes the contract setting the number of tricks that the bidder must make.
8114861	N	department, section	A specialized division of a large organization.
5726596	N	arrangement, organization, organisation, system	An organized structure for arranging or classifying.
582388	N	occupation, business, job, line of work, line	The principal activity in your life that you do to earn money.
9632518	N	worker	A person who works at a specific occupation.
7200813	N	refutation, defense, defence	The speech act of answering an attack on your assertions.
10372373	N	official, functionary	A worker who holds or is invested with an office.
9961012	N	contractor, declarer	The bridge player in contract bridge who wins the bidding and can declare which suit is to be trumps.

Table A.31: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13817526	N	percentage, percent, per centum, pct	A proportion in relation to a whole (which is usually the amount per hundred).
5145118	N	monetary value, price, cost	The property of having material worth (often indicated by the amount of money something would bring if sold).
15206296	N	month	A time unit of approximately 30 days.
15286249	N	rate	A magnitude or frequency relative to a time unit.
13754293	N	addition, increase, gain	A quantity that is added.
582388	N	occupation, business, job, line of work, line	The principal activity in your life that you do to earn money.
13308999	N	tax, taxation, revenue enhancement	Charge against a citizen's person or property or activity for the support of government.
13279262	N	wage, pay, earnings, remuneration, salary	Something that remunerates.
7218470	N	report, study, written report	A written document describing the findings of some individual or group.
192613	N	economy, saving	An act of economizing; reduction in cost.

Table A.32: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8256968	N	party, political party	An organization to gain political power.
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
8954611	N	Korea, Korean Peninsula, Dae-Han-Min-Gook, Han-Gook	An asian peninsula (off manchuria) separating the yellow sea and the sea of japan; the korean name is dae-han-min-gook or han-gook.
8078020	N	family, household, house, home, menage	A social unit living together.
5872742	N	rudiment, first rudiment, first principle, alphabet, ABC, ABC's, ABCs	The elementary stages of any subject (usually plural).
181781	N	election	A vote to select the winner of a position or political office.
8050678	N	government, authorities, regime	The organization that is the governing authority of a political unit.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
1105840	ADJ	national	Concerned with or applicable to or belonging to an entire nation or country.
6536853	N	bill, measure	A statute in draft before it becomes law.

Table A.33: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
8209687	N	police, police force, constabulary, law	The force of policemen and officers.
9225146	N	body of water, water	The part of the earth's surface covered with water (such as a river or lake or ocean).
8420278	N	depository financial institution, bank, banking concern, banking company	A financial institution that accepts deposits and channels the money into lending activities.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
8078020	N	family, household, house, home, menage	A social unit living together.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
13750844	N	thousand, one thousand, 1000, M, K, chiliad, G, grand, thou, yard	The cardinal number that is the product of 10 and 100.
965035	V	report, describe, account	To give an account or representation of in words.
15163979	N	Monday, Mon	The second day of the week; the first working day.

Table A.34: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
654885	N	care, attention, aid, tending	The work of providing treatment for or attending to someone or something.
3247620	N	drug	A substance that is used as a medicine or narcotic.
14447908	N	health, wellness	A healthy state of wellbeing free from disease.
7218470	N	report, study, written report	A written document describing the findings of some individual or group.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
14070360	N	disease	An impairment of health or a condition of abnormal functioning.
13817526	N	percentage, percent, per centum, pct	A proportion in relation to a whole (which is usually the amount per hundred).
1207609	N	aid, assist, assistance, help	The activity of contributing to the fulfillment of a need or furtherance of an effort or purpose.
5898568	N	plan, program, programme	A series of steps to be carried out or goals to be accomplished.
10405694	N	patient	A person who requires medical care.

Table A.35: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
9767197	N	actor, doer, worker	A person who acts and gets things done.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
582388	N	occupation, business, job, line of work, line	The principal activity in your life that you do to earn money.
6613686	N	movie, film, picture, moving picture, moving-picture show, motion picture, motion-picture show, picture show, pic, flick	A form of entertainment that enacts a story by sound and a sequence of images giving the illusion of continuous movement.
8078020	N	family, household, house, home, menage	A social unit living together.
15164957	N	day, daytime, daylight	The time after sunrise and before sunset while it is light outside.
10372373	N	official, functionary	A worker who holds or is invested with an office.
1124794	N	government, governing, governance, government activity, administration	The act of governing; exercising authority.
8540903	N	city	An incorporated administrative district established by state charter.
9767700	N	actress	A female actor.

Table A.36: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8320201	N	soviet	An elected governmental council in a communist country (especially one that is a member of the union of soviet socialist republics).
11007750	N	Gorbachev, Mikhail Gorbachev, Mikhail Sergeyevich Gorbachev	Soviet statesman whose foreign policy brought an end to the cold war and whose domestic policy introduced major reforms (born in 1931).
2959406	ADJ	Soviet	Of or relating to or characteristic of the former soviet union or its people.
8166552	N	nation, land, country	The people who live in a nation or country.
8256968	N	party, political party	An organization to gain political power.
9623038	N	leader	A person who rules or guides or inspires others.
1110274	N	deal, trade, business deal	A particular instance of buying or selling.
10372373	N	official, functionary	A worker who holds or is invested with an office.
10468962	N	president, chairman, chairwoman, chair, chairperson	The officer who presides at the meetings of an organization.
2716739	ADJ	economic, economical	Of or relating to an economy, the system of production and management of material wealth.

Table A.37: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
13649791	N	inch, in	A unit of length equal to one twelfth of a foot.
11501381	N	rain, rainfall	Water falling in drops from vapor condensed in the atmosphere.
329831	ADJ	central	In or near a center or constituting a center; the inner area.
9141526	N	Texas, Lone-Star State, TX	The second largest state; located in southwestern united states on the gulf of mexico.
9911226	N	charwoman, char, cleaning woman, cleaning lady, woman	A human female employed to do housework.
5011790	N	temperature	The degree of hotness or coldness of a body or environment (corresponding to its molecular activity).
3540595	N	hospital, infirmary	A health facility where patients receive treatment.
8159924	N	York, House of York	The english royal house (a branch of the plantagenet line) that reigned from 1461 to 1485; its emblem was a white rose.
8078020	N	family, household, house, home, menage	A social unit living together.
1105840	ADJ	national	Concerned with or applicable to or belonging to an entire nation or country.

Table A.38: Illustrative topic from AP obtained with SemLDA.

SemLDA with WSD			
Synset ID	**POS**	**Words**	**Gloss**
8766988	N	Germany, Federal Republic of Germany, Deutschland, FRG	A republic in central europe; split into east germany and west germany after world war ii and reunited in 1990.
9747722	N	German	A person of german nationality.
9189411	N	Africa	The second largest continent; located to the south of europe and bordered to the west by the south atlantic and to the east by the indian ocean.
8561835	N	west	A location in the western part of a country, region, or city.
9715833	N	Israeli	A native or inhabitant of israel.
8166552	N	nation, land, country	The people who live in a nation or country.
7942152	N	people	(plural) any group of human beings (men or women or children) collectively.
9050730	N	South	The region of the united states lying to the south of the mason-dixon line.
8563180	N	East, eastern United States	The region of the united states lying to the north of the ohio river and to the east of the mississippi river.
9634494	N	African	A native or inhabitant of africa.

Table A.39: Illustrative topic from AP obtained with SemLDA.