ISSN 2007-9737

A Behavior Analysis of the Impact of Semantic Relationships on Topic Discovery

Ana Laura Lezama Sánchez[1], Mireya Tovar Vidal[1], José Alejandro Reyes Ortiz[2]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Mexico

[2] Universidad Autónoma Metropolitana, Departamento de Sistemas, Azcapotzalco, Mexico

ana.lezama@alumno.buap.mx, mireya.tovar@correo.buap.mx, jaro@azc.uam.mx

Abstract. Information Technologies have generated large amounts of documents available for analysis and use. Information systems can provide the user with the necessary data for a specific purpose without human intervention, saving time in providing the response expected by the user. Some traditional models of topic discovery provide essential information in the literature, but it is still necessary to incorporate the knowledge that a person can use when reading a document. In this work, an analysis of the behavior of the techniques of Latent Dirichlet Analysis, Latent Semantic Analysis, and Probabilistic Latent Semantic Analysis is carried out incorporating the semantic relationships of the type hypernym, hyponym, synonymy, holonymy, and meronymy extracted from an external source of knowledge as WordNet. In order to improve the results obtained by applying the three mentioned techniques in a set of documents without adding external knowledge. Compared to the initial results, our experimental results improved when incorporating semantic relationships, such as hypernyms and synonyms. The best result was obtained when using the Lesk algorithm for word sense disambiguation and subsequently applying Latent Dirichlet Analysis.

Keywords. Topic discovery, Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), WordNet, Latent Dirichlet Analysis (LDA).

1 Introduction

The increase in available textual information is increasing, becoming the basis for automatic information systems, searching, organizing, and understanding large amounts of text in a short time. Some information systems need to incorporate modules that extract or discover the topics presented in the documents to be analyzed. *Topic discovery* is a Natural Language Processing (NLP) technique designed to analyze large volumes of documents automatically. NLP is a branch of artificial intelligence capable of providing machines with the knowledge needed to process human language into any language.

One of the branches of the NLP study is topic discovery, which aims to obtain the central idea of a document by analyzing the texts to determine the words grouped for a set of documents. In the literature, there are techniques for topic discovery, such as Latent Semantic Analysis (LSA), Dirichlet Latent Analysis (LDA), and Probabilistic Latent Semantic Analysis (PLSA).

The existence of computational techniques for topic discovery made this task easier to solve, providing adequate information and in a short time to those who use them.

Some information systems can quickly extract the topics presented in large volumes of text compared to the time used by a person when carrying out the same activity. The traditional techniques that can discover topics present in the texts cannot incorporate external knowledge with the certainty that it is related to the work. From a linguistic point of view, semantics study the word's meaning and the sentence [15]. Hence, semantic

relationships refer to relations between meanings in the sentences.

Semantic relationships are external knowledge that can discern ambiguity in natural language, since the latter is often difficult to identify, even for a human. The words that form a document can have more than one meaning, improving the results of a topic discovery module. This type of knowledge is easy to apply when the topic discovery is manually made because when a person reads a text, they can quickly apply the knowledge that they already have stored in their memory, that is, semantic memory. From the linguistic point of view, it is the present semantic relations that are easy and fast to recognize, knowing that for a computer, if it is not provided, it is not possible to use it.

There are several algorithms for topic discovery. The most used are described below. Latent Dirichlet Analysis (LDA) is a probabilistic generative model for discrete data collections formed by text collections [4].

Latent Semantic Analysis (LSA) is an NLP technique for semantic analysis; the meaning of a corpus is extracted using statistical methods as a tool. This technique does not use input text dictionaries and later analyze the words defined as unique character strings or significant samples such as sentences [17].

Probabilistic Latent Semantic Analysis (PLSA), based on LSA, is a NLP technique; its purpose is to find a probabilistic model with latent topics to obtain the observed data in a document-term matrix. PLSA uses the decomposition of a mixed matrix that belongs to a generative model [16].

An important tool in NLP is WordNet. WordNet is a lexical database, developed at Princeton University in 1986, providing semantic knowledge that can be computer-readable, grouping words into synsets providing meanings and semantic relationships between such synsets [9]. Authors expose LDA, LSA, and PLSA for the topic discovery. We exposed six different experiments for topic discovery existing in nine datasets extracting semantic relationships from the WordNet lexical database as hypernyms, hyponyms, synonyms, holonyms, and meronyms of each word that forms the corpus the Lesk disambiguation algorithm. The objective is to provide additional knowledge of each word in the corpora but its semantic features.

This research is structured as follows. In section 2, some proposals are presented by various authors for the topic discovery. In section 3, the proposed method and the dataset are displayed. In section 4.2, the experiments are presented, and finally, in Section 5, the conclusions and future work are provided.

2 Related Work

This section summarizes works related to topic discovery that use traditional literature techniques or develop new methods with the same purpose.

In [20], the authors propose a method called lda2vec; learn word vectors with mixes of thematic vectors by producing mixes of documents through a non-negative simplex constraint. The model proposed by the authors is based on modifying the negative-sampling skip-gram (SGNS) to work with feature vectors throughout the document and topic vectors. The method employs pairs of pivots and targets words previously extracted when they co-occur in a moving window that scans the entire corpus. The lda2vec model embeds words and document vectors in the same space and carries out training. The corpora used for the experiments were 20newsgroups and Hacker News comments. The model proposed by the authors extended SGNS for the construction of unsupervised document representations to obtain consistent topics by training word vectors, topics and documents jointly and integrated into a space of representation that preserved semantic regularities between the vectors of learned words.

Perera et al. [21] propose an extension for the algorithm KeyGraph by adding hypernyms extracted from the lexical database WordNet and terms of the FIBO ontology. The results showed that the presence of hypernyms in the KeyGraph algorithm provides groups of improved topics. The experiments made use of the corpus called EMMA composed of 1,424 documents; the first consisted of finding terms of FIBO with their respective frequencies; the second with the terms found in the first experiment filter the empty words and their respective frequencies; and the third for each term

obtained in experiment 2 extracts from WordNet the Hypernyms for each term.

In [25], it is presented a method to incorporate the knowledge of correlation of external words to improve the coherence of the topic modeling. The authors mention that the topic discovery assumes that words are generated independently and lack mechanisms for the use of similarity relationships between words to obtain more coherent topics, so they built an LDA assignment model Markov Random Field regularization (MRF) so that the words labeled as similar share the same topical tag.

In [1], the authors show a method that incorporates LDA, Twitter, WordNet, and hashtags to improve the labels of the most representative words of each topic. The authors stress the importance of keywords for different topics based on semantic relationships and co-occurrences of keywords present in hashtags. In addition, they proposed another method to find the ideal number of topics for representing collections of text documents. The experimental results showed the authors that the proposed method works better than the original LDA in complexity and topic coherence. The datasets used in the experiments were snippet and BaiduQA.

Bougteb et al. [5] propose the use of an autoencoder for topic discovery present in a set of tweets about news and companies such as Apple, Google, and Microsoft. The authors made tests with a window model, obtaining significant results of 5 and 10, subsequently occupying the k-means++ algorithm to obtain the number of groups needed to classify topics. The second dataset was tested with CluStream, DenStream, and Dstream clustering algorithms, improving the results by working with autoencoders, representing its dataset with word2vec, and using cosine similarity. The model was evaluated with precision, accuracy, and F_1-measure metrics.

He et al. [14] propose a model called the Representative Latent Dirichlet thematic assignment model (RLDA). The model exposes the close and complex relationships between diabetes, obesity, and other diseases. The authors tested a corpus of over 337,000 publications on diabetes and obesity. An analysis of his model's results revealed significant relationships between diabetes mellitus, obesity, and other diseases. The results indicated that the diseases closely related to obesity in the last ten years were asthma, gastric disease, and heart disease.

The proposed method achieved results for discovering the relationship analysis of relevant points on diabetes and obesity. The model extracts significant relationships between them and other diseases. The RLDA model is based on LDA, word2vec, and affinity propagation grouping.

In [27], it is proposed a model for novel topic discovery for a short text corpus using word embedding. The word embeddings were incorporated into the model to provide additional semantic. Therefore, it modeled each short document as a Gaussian topic on word embedding in vector space. Furthermore, considering that the background words in a short text are generally not semantically related. They introduced a discrete background mode on the types of words to complement Gauss' continuous themes. The model was evaluated with a corpus of the news titles from data sources such as *abcnews*, which shows that the model can extract topics with greater coherence compared to reference methods and learn a better representation of topics for each document.

A new statistical learning approach to combine topical modeling and document grouping is proposed in [8]. In particular, they developed a Bayesian generative model of collections of texts, in which the above two tasks are incorporated as latent coupled factors governing the drafting of the document. The latter consisted of embedding word2vec words to capture semantic and syntactic regularities between words. The approach used collapsed Gibbs sampling and parameter estimation to perform topic modeling and document grouping through Bayesian reasoning. Comparative evidence from the real-world reference corpus reveals the approach's effectiveness devised to group collections of text documents and coherently retrieve their semantic. The approach was tested with two English corpora: financial and news domains. Authors evaluated employing accuracy and standardized mutual information; they also incorporated word embedding to capture semantic

and syntactic knowledge. The datasets used were Reuters-21578 and 20-Newsgroup.

A model for learning a modular taxonomy is shown in [26]. The model collects terms with information about a topic, and at the same time, it is possible to discover hypernyms and relationships between the terms collected. Using LDA, the terms of each subdomain are divided and related subdomains in ontology modules to automatically assign topical characteristics to terms. They employed an information insertion technique (concept substitution) during the process, obtaining a significant improvement in the learning of modular taxonomy.

In [7], it proposed a model to work with topic model. The model uses lexical-semantic relations as synonyms, antonyms, and adjective attributes to obtain more topic coherence. Given that a word can have more than one meaning, but not be of use, propose a model GK-LDA to exploit the knowledge of the lexical relationships used. The experiments performed were done with datasets on online purchases showing that the model proposed by the authors performs better than conventional methods. The evaluation was carried out using the topic coherence metric.

In [2], it is proposed an entity-based model for the correct integration of the DBpedia ontology for topic modeling with the topic discovery. The proposed model presented an increase in the coherence of the topic discovery. The authors aimed to include entity information contained in each document and at the same time to integrate concepts of ontology and the relationships between them in the topic discovery, exploring knowledge to discover automatically topic coherence.

Baldwin et al. [3] show a model for disambiguation consisting of three modules designed to discover referential ambiguity divided into three modules. First, examine n-grams of the corpus of contemporary American English dataset as the authors are based on the understanding that terms appearing in contexts of named entities may not be referential and terms that may be non-referential references can be used to detect ambiguity. Subsequently, the second module was designed to detect ambiguity between domains, and implemented a module that employs ontologies, where terms with multiple senses were labeled as ambiguous. Finally, they used Latent Dirichlet Analysis to detect non-referential ambiguity between domains, grouping the contexts in which the words appear. A term is labeled as ambiguous if some of the three modules predict that it is ambiguous, but it is only labeled as unambiguous if all three modules make that prediction.

Fortuna et al. [11] develop a system capable of incorporating latent and k-means semantic indexing techniques and integrating them for the semi-automatic construction of topical ontologies, which provides support to its users for the construction by providing suggestions for topics present in the documents that make up a dataset. The ontology will be composed of a set of topics related to different types of relationships. The user can select each topic discovered for the ontology to edit which documents are assigned to a specific topic. The cosine similarity is used to calculate the similarity between each corpus and the topic centroid, the selected topic, and all existing topics in an ontology.

An approach to finding latent topics from large corpora is presented in [19]. The corpora used were NIPS 14-23 and NIPS 00-13. Since LDA has the deficiency that documents lack semantic knowledge, the authors proposed to add synonyms extracted from WordNet. They performed two different experiments; the first consists of adding the extracted synonyms to the original corpus and then applying LDA, contrary to the second experiment that first apply LDA and then add the synonyms extracted from WordNet carrying out a morphological reduction. Later explore if two words are synonyms eliminating the word with less probability.

In [6], it is proposed the development of a model relies on correlating mixed words to find hidden topics in the corpus. The model extracts semantic knowledge by constructing a Markov mixed random field and incorporating links through word embedding. The authors carried out experiments with two datasets, 20-newsgroup divided into 20 categories and NIPS with 1,500 papers. The proposed model was able to extract

more abundant knowledge than that extracted by the reference models. The metric used for the evaluation of the model was topic coherence.

In [22], the authors propose a tool that implements text classification and clustering techniques and topic discovery. They use the LDA technique, which is one of the most used for topic discovery, and the topic coherence metric to calculate the coherence of topics using an additional dataset using Wikipedia as a meta-document to qualify word pairs co-occurrence of the term.

In [12], it is proposed a method based on min-hashing for the discovery of topics. The datasets used are 20-Newsgroup, Reuters, and Wikipedia in Spanish and English. The proposed method does not require a predefined number of topics and is capable of highly recurring word sets that are then grouped to obtain latent topics. The authors evaluated the proposed method using the normalized topic coherence metric.

A method for topic discovery in short texts that includes the use of the word embedding to capture semantic and syntactic information about words is exposed [13]. The model also incorporates global and local semantic correlations through the use of conditional random fields. The datasets used for the experiments were StackOverflow and News. The evaluation metric used was topic coherence and accuracy. The experimental results reflected the effectiveness of their model compared to traditional models.

Tovar et al. [24] present an approach for the automatic identification of relations in ontologies of the restricted domain, specifically taxonomic and nontaxonomic relationships. The proposed approach is based on Formal Concept Analysis (FCA) for the evaluation of ontological relationships. The authors aimed to search for evidence of ontological relationships to evaluate in a reference corpus. The authors incorporated two types of variants in selecting properties or attributes to construct an incidence matrix that the FCA needs to obtain the formal concepts. The main difference between these two variants is the dependency parser applied in the preprocessing phase, Stanford, and Minitorque. From the results obtained, the authors found evidence that Stanford obtained the best results for nontaxonomic relationships. The evaluation metric used was precision. The precision obtained was 96% for taxonomic relationships and 100% for nontaxonomic relationships. The authors point out that the results obtained show the quality of the ontology.

Sánchez et al. [18] propose an approach to enrich the LDA model by extracting the Hypernyms of each word that make up a sentence supported by WordNet. The model considered it essential to include a hypernym for each word that makes up the dataset of documents. The results suggest that the behavior of LDA with hypernyms provides higher results compared to when only datasets are provided with traditional preprocessing. The datasets used in the experiments were nine domains such as news, biomedicine, technology, social networks, among others. The evaluation metric used was topic coherence.

In this work, we propose to add the hyponyms, hypernyms, synonyms, holonyms, and meronyms of each word that make up the corpus. Each semantic relation was extracted from WordNet. In addition, a disambiguation method is incorporated. The aim is to improve the levels of coherence of each technique to be used. The evaluation process obtains the coherence of topics supported by the Wikipedia corpus with 3M documents.

3 Proposed Method

This paper proposes a method for extracting semantic relations such as hypernyms, synonyms, holonyms, hyponyms, and meronyms existing in WordNet. In addition, we have an extra module of disambiguation through the Lesk algorithm. The purpose is to incorporate semantic knowledge into the corpus for topic discovery. The techniques used are LDA, PLSA, and LSA, which are used to discover the latent topics from documents. Below are the stages of the proposed method:

1. Corpus preprocessing: This stage includes the following actions:
   (a) Removal of special symbols, numbers, and punctuation marks.
   (b) Split of the corpus into unigrams.

2. Semantic relations extraction. Several relations are extracted from WordNet for each unigram of the corpus to evaluate the topic discovery. It is presented as an algorithm as follows.

   Begin (Relation extraction)

   According to the option do:

   2.1: Extract from WordNet, the hypernym, synonym, holonym, hyponym or corresponding meronym of the input word, from the first sense and add it to the corpus.

   2.2: Extract from WordNet the hypernym and synonym of the input word, from the first sense and add it to the corpus.

   2.3: Extract from WordNet the hypernym and the hierarchy of hypernyms of the first sense and add them to the corpus.

   2.4: Extract from WordNet the hypernym and root of the first sense hypernym hierarchy and add them to the corpus.

   else: a subcorpus is formed with pairs of the type ”word, sentence”.
   Lesk algorithm is applied in pairs ”word,sentence”.
   The corpus is updated by storing the original sentence and words provided by Lesk.

   End According

   End (Relation extraction)

3. Preprocessing of enriched corpus removed special symbols, punctuation signs, and stopwords.

4. Apply the topical discovery method: LDA, LSA, or PLSA as appropriate.

5. Evaluation: The evaluation of the obtained results is carried out with the topic coherence metric that uses Equation 1 to measure how coherent the recovered topics are and, in our case, also evaluate how coherent it was to add semantic relationships to the corpus. It should be noted that this metric uses an external corpus for its operation. In our case, we used a corpus from Wikipedia with 3M documents [23]:

$$PMI(w_i, w_j) = \frac{2}{T(T-1)} \sum_{1 \leq i < j \leq T} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{1}$$

where T is the main words $p(w_i)$ (resp. $p(w_j)$) is the probability that the word w_i (resp. $p(w_j)$) appears in a text window of a given size, while $p(w_i, w_j)$ denotes the probability that w_i y w_j co-occur in the same window.

4 Results

This section exposes the datasets used in this work and the results obtained with the method proposed. In this work, we proposed six experiments with relations extracted from WordNet and the Lesk algorithm. The characteristics of each are presented below and described in more detail in the following sections:

1. Apply and evaluate the three algorithms LDA, LSA, and PLSA without relationship enrichment.

2. Extract from WordNet the hypernym (LDA+HE, LSA+HE, PLSA+HE), synonym (LDA+S, LSA+S, PLSA+S), holonym (LDA+HO, LSA+HO, PLSA+HO), hyponym (LDA+HP, LSA+HP, PLSA+HP) or corresponding meronym (LDA+ME, LSA+ME, PLSA+ME) of the input word, from the first sense and they are add to the corpus, and finally apply LDA, PLSA y LSA.

3. Extract from WordNet, the hypernym, and synonym (LDAhys) from the first sense and add them to the corpus, and apply LDA.

4. Extract from WordNet the hypernym (LDAhN) and the hierarchy of hypernyms of the first sense and add them to the corpus, and apply LDA.

5. Extract from WordNet, the hypernym (LDAhR), the root of the first sense hypernym hierarchy, add them to the corpus, and apply LDA.

6. A subcorpus is formed with pairs of the type ”word, sentence”. Lesk algorithm is applied on the pairs ”word, sentence” and apply LDA (LDA+ Lesk).

4.1 Datasets

In this section we present the datasets used (see Table 1) and the experimental results with the proposed method. The effects and differences of applying the LDA, LSA, and PLSA models with hypernyms, hyponyms, synonyms, holonyms and meronyms, and without them are analyzed in this Section. We selected 9 datasets [22]. First, preprocessing was performed which included removing orthographic signs. The information of the 9 datasets is shown in the Table 1, where D represents the number of documents in each dataset and V is the total vocabulary including stopwords[1].

Table 1. Dataset

Dataset	D	V
Technology	12,295	11,167
SearchSnippets	12,295	895,710
StackOverflow	16,407	426,594
Biomedical	19,448	604,676
Tweet	2,472	115,552
GoogleNews	11,109	396,349
PascalFlickr	4,834	202,097
20-newsgroup	20,000	14,537,363
Reuters	18,456	7,976,744

4.2 Experimental Results

The experimental results obtained with the implementation of the proposed method are presented below. Six experiments were developed based on the proposed method, which are described below.

The first experiment involves the testing the three methods of topic discovery, without incorporating semantic relationships to corpora. Table 2 shows the second experiment, which includes the experimental results obtained with the topic coherence evaluation metric are shown and the best results by corpus are highlighted. The

[1] https://github.com/qiang2100/STTM/tree/master/dataset

Table 2. Results obtained with LDA, LSA and PLSA without enriching the dataset with semantic relations

Dataset	LDA	LSA	PLSA
Technology	0.78	1.24	**1.67**
Biomedical	1.39	2.07	**2.36**
Tweets	**1.23**	1.17	0.92
StackOverflow	2.13	**2.33**	1.52
SearchSnippets	0.78	**0.84**	0.79
PascalFlickr	0.80	0.87	**1.02**
GoogleNews	1.26	**1.43**	1.04
20-newsgroup	**1.52**	1.30	0.52
Reuters	**1.31**	0.89	1.04

the application of the method using cases 1 to 5 (semantic relations) and the call to the LDA. Table 3 presents the results obtained with topic coherence with LDA without semantic relationships (LDA), with LDA adding hypernyms (LDA+HE), hyponyms (LDA+HP), synonyms (LDA+S), holonyms (LDA+HO), and meronyms (LDA+ME) respectively from WordNet. The third experiment involves the application of the method using the cases of 1 to 5 (semantic relationships) and the call to the technique of LSA. In Table 4, we present the results obtained with the metric topic coherence, when applying the technique of LSA without semantic relationships (LSA) with LSA adding hypernyms (LSA+HE), hyponyms (LSA+HP), synonyms (LSA+S), holonyms (LSA+HO), and meronyms (LSA+ME) respectively from WordNet.

The fourth experiment comprises the method using cases 1 to 5 (semantic relationships) and the call to the technique of PLSA. The results are shown in Table 5, an increase in the results obtained with the topic coherence metric is observed in the corpora Tweets, SearchSnippets, PascalFlickr, 20-newsgroup, and Reuters compared to the results obtained applying PLSA without some type of semantic relationship.

Therefore, a fifth experiment is proposed using hypernyms and synonymy relations incorporating both relationships at the same time, and subsequently the results were obtained with the LDA model.

These results are shown in Table 6, where the original results are presented without any

Table 3. Results obtained with LDA and the proposed semantic relationships

Dataset	LDA	LDA+HE	LDA+HP	LDA+S	LDA+HO	LDA+ME
Technology	0.78	**1.34**	1.01	1.30	1.01	1.04
Biomedical	1.39	1.84	1.84	**2.70**	2.00	2.34
Tweets	1.23	**1.51**	1.07	0.99	1.00	1.13
StackOverflow	2.13	**2.40**	1.27	1.37	1.37	1.11
SearchSnippets	0.78	1.29	1.48	**2.45**	2.29	2.23
PascalFlickr	0.80	1.23	1.30	**2.10**	1.84	1.94
GoogleNews	1.26	**1.87**	1.02	1.02	1.01	0.70
20-newsgroup	1.52	1.59	1.17	**1.66**	1.25	1.33
Reuters	1.31	1.79	1.46	**1.80**	1.27	1.34

Table 4. Results obtained with LSA and the proposed semantic relationships

Dataset	LSA	LSA+HE	LSA+HP	LSA+S	LSA+HO	LSA+ME
Tecnology	1.24	1.14	**1.37**	1.28	1.03	1.19
Biomedical	**2.07**	2.03	1.82	2.02	1.43	1.97
Tweets	1.17	1.15	**1.18**	1.08	1.12	1.09
StackOverflow	**2.33**	1.50	1.51	1.39	1.60	1.51
SearchSnippets	**0.84**	0.74	0.78	0.77	0.74	0.77
PascalFlickr	**0.87**	0.75	0.77	0.79	0.66	0.79
GoogleNews	**1.43**	1.14	1.09	1.18	1.04	1.16
20-newsgroup	**1.30**	1.05	0.88	0.91	1.02	1.11
Reuters	0.89	0.89	0.93	0.88	**0.94**	0.82

semantic relation (LDA), LDA with hypernyms and synonyms (LDAhys), LDA with one hypernym per level (LDAhN), and LDA with the root hypernym (LDAhR).

According to the results, it is observed that in general the best results obtained are those where the hierarchy of hypernyms was incorporated into the corpus (see column LDAhN) and therefore, an increase in the levels of coherence is obtained.

Taking the best results from Table 3, the last experiment was defined in order to incorporate the else case of step 2 of the proposed method, that is, a disambiguation algorithm is used that determines the word of the most appropriate sense given the context. Table 7 exhibits the results obtained by incorporating the Lesk disambiguation algorithm to disambiguate the meaning of the words, then LDA is applied. This last experiment obtained the best results compared to the original results with LDA and those obtained with the added semantic relationships because it was possible to add new meanings of the words according to the context of this and therefore, LDA was able to generate the words that they form each topic with greater semantic knowledge.

Table 8, shows the best results obtained for each corpus with some of the proposed experiments. The topic coherence columns show the coherence levels obtained with any of the techniques used without any type of relation or algorithm used (T), with aggregate relations (T + R) or with the Lesk algorithm (T + L). As can be seen in most cases, the topic coherence results of the Lesk method are higher than the results obtained with a traditional technique and with semantic relationships.

The Lesk algorithm selects WordNet as a source of knowledge, which allowed the comparison of glosses between words that are connected by various semantic relationships with words that are being disambiguated, in this experiment the algorithm uses the hyponymic semantic relation to carry out this process [10].

Based on the results obtained, by the LDA technique with semantic relationships, as shown in

Table 5. Results obtained with PLSA and the proposed semantic relationships

Dataset	PLSA	PLSA+HE	PLSA+HP	PLSA+S	PLSA+HO	PLSA+ME
Technology	**1.67**	1.61	1.61	1.55	1.32	1.37
Biomedical	**2.36**	1.96	2.03	2.08	1.78	1.88
Tweets	0.92	0.87	**1.03**	0.84	0.94	0.88
StackOverflow	**1.52**	1.18	1.25	1.30	1.23	1.43
SearchSnippets	0.79	0.74	**0.92**	0.73	0.81	0.89
PascalFlickr	1.02	0.97	0.90	**1.15**	0.91	0.86
GoogleNews	**1.04**	1.00	1.01	0.87	0.87	0.96
20-newsgroup	0.52	0.87	**0.96**	0.76	0.93	0.91
Reuters	1.04	1.11	1.04	1.12	**1.23**	1.07

Table 6. Results obtained with LDA and variants with synonyms and hypernyms

Dataset	LDA	LDAhys	LDAhN	LDAhR
Technology	0.78	1.24	**1.43**	1.15
Biomedical	1.39	**1.87**	1.40	1.46
Tweets	1.23	1.13	**1.43**	1.10
StackOverflow	**2.13**	1.27	1.49	1.20
SearchSnippets	0.78	1.32	**1.44**	1.00
PascalFlickr	0.80	1.19	**1.41**	1.21
GoogleNews	1.26	1.01	**1.45**	1.10
20-newsgroup	**1.52**	1.01	1.40	1.16
Reuters	1.31	1.16	**1.41**	1.20

Table 7. Results obtained with LDA and the Lesk algorithm

Dataset	LDA	LDA+Lesk
Technology	0.78	**1.40**
Biomedical	1.39	**2.79**
Tweets	1.23	**1.55**
StackOverflow	2.13	**2.43**
SearchSnippets	0.78	**2.46**
PascalFlickr	0.80	**2.12**
GoogleNews	1.26	**1.87**
20-newsgroup	1.52	**1.85**
Reuters	1.31	**1.85**

Table 3, that the results of the columns LDA+HE and LDA+S increase the result of the level of topic coherence in comparison with that obtained without adding some type of semantic relationship (see column LDA). The increase is due to the fact that a hypernym contains the semantic features of a word, and the synonyms contain the similarity between the words of the same grammatical category, which provides additional knowledge that is useful in topic discovery.

In addition, given that LDA considers a topic as a distribution on a fixed vocabulary, and having embedded within that vocabulary new words, but not alien to the original corpus, it generates, as a result, topics with greater semantic wealth. The increase in the level of coherence was obtained mainly in the corpora: Technology, Tweets, StackOverflow, and GoogleNews when adding hypernyms, where the corpus with the highest level of coherence obtained was StackOverflow obtaining an increase of 0.27 when adding hypernyms, that is, a coherence of 2.13 without hypernyms and 2.40 with them was obtained.

For the synonymy type relationship, the corpus with the greatest increase in the level of coherence was Biomedical, with an increase of 1.31. The corpora SearchSnippets, PascalFlickr, 20-newsgroup, and Reuters also obtained an increase in the levels of coherence compared to those obtained without adding any type of relationship.

Based on the obtained results, varied results are observed, that is in comparison with LSA without some type of relationship when adding hyponyms; an increase is observed with corpora technology and tweets. For the rest of the corpora, no improvement is reflected compared to the original results. The increase registered with hyponyms is due to the fact that this type of word has all the semantic features of a more general one, and since it is a domain over Technology we find a broad

Table 8. Summary of the best results

Corpus	Topic coherence	T	Topic coherence	T+R	Topic coherence	T+L
Technology	**1.670**	PLSA	1.610	PLSA+HE	1.400	LDA+Lesk
Biomedical	2.360	PLSA	2.70	LDA+S	**2.790**	LDA+Lesk
Tweets	1.230	LDA	1.510	LDA+HE	**1.550**	LDA+Lesk
StackOverflow	2.330	LSA	2.40	LDA+HE	**2.430**	LDA+Lesk
SearchSnippets	0.840	LSA	2.450	LDA+S	**2.460**	LDA+Lesk
PascalFlickr	1.020	PLSA	2.100	LDA+S	**2.120**	LDA+Lesk
GoogleNews	1.430	LSA	1.870	LDA+HE	**1.875**	LDA+Lesk
20-newsgroup	1.520	LDA	1.660	LDA+S	**1.850**	LDA+Lesk
Reuters	1.310	LDA	1.80	LDA+S	**1.850**	LDA+Lesk

vocabulary with many semantic characteristics that can be differentiated.

In addition, LSA represents semantic concepts of each document, representing the text as a term-document matrix; therefore, when adding new concepts, that semantic field is enriched since new intersections are generated in each row and column.

In the case of the corpus Reuters, according to Tables 4, and 5, an increase in coherence levels was obtained by adding holonyms because their meaning includes that of others with the same semantic traits. The corpus Reuters has a varied domain, that is, ranging from topics about organizations, people, places, and technology.

Furthermore, despite the fact that LSA and PLSA are very similar, the latter represents the documents as a bag of words and is based on an aspect model, and the documents are represented by a co-occurrence matrix and by means of a probability distribution, a word is generated the hidden topic in the observed document. According to the experimental results presented in Table 3, the best results obtained when incorporating semantic relations were those that included hyperonymy and synonymy relationships.

On the other hand, some related works that use these traditional techniques were also reviewed with some corpus used in the experiments and their comparison with the results obtained by the proposed method.

Table 9 presents the results obtained by some authors of state of the art are for the corpus 20-newsgroup, most proposing a variant of the LDA technique as in [12], and [20] proposing new methods of topic discovery.

In addition, the results obtained with the method proposed in this paper are shown, which allowed obtaining a significant improvement compared to the authors mentioned. It should be noted that the results presented were obtained by applying the Lesk algorithm for disambiguation that is based on the use of a dictionary, that is, WordNet.

Also, Table 10, presents the results obtained by testing the corpus Reuters with the methods proposed by the authors state of the art and those proposed in this work.

An improvement is observed as in the previous corpus when applying the disambiguation algorithm and later LDA. Likewise, Table 11 shows the results obtained by the two authors when applying a variant of LDA and the method proposed in this work on the corpus StackOverflow.

Table 9. Comparative results obtained for corpus 20-newsgroup

Authors	Topic coherence	Techique
[20]	0.5670	proposed by authors
[8]	0.2983	LDA+HAC
[8]	0.3572	LDA+k-means
[8]	0.3572	LDA+k-means
[12]	0.3000	min-hashing
This work	**1.6800**	LDA+Lesk

This work presents six different experiments with semantic relationships, and the last experiment uses the Lesk disambiguation algorithm. The best results obtained were when applying the Lesk

Table 10. Comparative results obtained for corpus Reuters

Authors	Topic coherence	Technique
[8]	0.3647	LDA+k-means
[8]	0.3498	LDA+HAC
[12]	0.6000	min-hashing
This work	**1.8500**	LDA+Lesk

Table 11. Comparative results obtained for corpus StackOverflow

Authors	Topic coherence	Technique
[13]	0.3647	LDA+k-means
[22]	1.1500	LDA
This work	**2.4300**	LDA+Lesk

algorithm because the new words incorporated are in strict adherence to the context of the sentence where the word is found.

This experiment allowed us to visualize that when applying a disambiguation algorithm, words (new information) with more excellent semantic knowledge are added between them.

On the other hand, the incorporation of words, through semantic relationships, to the corpus also obtained good results of topic coherence with the LDA technique. On the other hand, it also highlighted that this method using Lesk significantly improves the experimental results of the related works for some corpus reported in the literature.

5 Conclusion and Future Work

This work presents an analysis of six different experiments applying the LDA, LSA, and PLSA techniques in 9 datasets in the English language. For each word that forms the corpus, semantic relationships were extracted from WordNet and incorporated a disambiguation algorithm to obtain new words related to the corpus, but based on the context of the sentence.

On the basis of experimental results, it is observed that the addition of semantic relations improves the results obtained in certain cases, but when integrating a disambiguation algorithm, the results obtained have improved significantly.

In general, we concluded that the LDA technique generates better results with the proposed method than the other two techniques because it is a three-level model that considers each topic as a distribution over the corpus being used.

As future work, we propose incorporating other topic discovery techniques with the same procedure presented here, in addition to incorporating, to the preprocessing step, a morphological analysis to determine the form, class, or grammatical category of each word that forms the corpus.

We plan to incorporate an artificial neural network to previously obtain a classification of the different categories present in the corpora and obtain the topics by categories to implement a disambiguation algorithm of the sense that, given the context, information is obtained about which sense is the most appropriate and to be able to include more than one synonym.

Acknowledgments

The authors would like to thank Universidad Autónoma Metropolitana, Azcapotzalco. The present work has been funded by the research project SI001-18 at UAM Azcapotzalco, partly supported by project VIEP 2021 at BUAP and by the Consejo Nacional de Ciencia y Tecnología (CONACYT) with the scholarship number 788155.

References

1. **Alkhodair, S. A., Fung, B. C., Rahman, O., Hung, P. C. (2018).** Improving interpretations of topic modeling in microblogs. Journal of the Association for Information Science and Technology, Vol. 69, No. 4, pp. 528–540.
2. **Allahyari, M., Kochut, K. (2016).** Discovering coherent topics with entity topic models. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, pp. 26–33.
3. **Baldwin, T., Li, Y., Alexe, B., Stanoi, I. R. (2013).** Automatic term ambiguity detection. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 804–809.

4. **Blei, D. M., Ng, A. Y., Jordan, M. I. (2003).** Latent dirichlet allocation. Journal of machine learning research, Vol. 3, pp. 993–1022.
5. **Bougteb, Y., Ouhbi, B., Frikh, B., others (2019).** Deep learning based topics detection. 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), IEEE, pp. 1–7.
6. **Chen, J., Zhang, K., Zhou, Y., Chen, Z., Liu, Y., Tang, Z., Yin, L. (2019).** A novel topic model for documents by incorporating semantic relations between words. Soft Computing, pp. 1–17.
7. **Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. (2013).** Discovering coherent topics using general knowledge. Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pp. 209–218.
8. **Costa, G., Ortale, R. (2020).** Document clustering meets topic modeling with word embeddings. Proceedings of the 2020 SIAM International Conference on Data Mining, SIAM, pp. 244–252.
9. **Fellbaum, C. (2010).** Wordnet. In Theory and applications of ontology: computer applications. Springer, pp. 231–243.
10. **Fernández Reyes, F. D. L. C., Leyva Pérez, E. C., Fernández, R. L. (2011).** Consideraciones de diseõ para una herramienta de análisis semántico. RLA. Revista de lingüística teórica y aplicada, Vol. 49, No. 1, pp. 51–68.
11. **Fortuna, B., Mladenič, D., Grobelnik, M. (2005).** Semi-automatic construction of topic ontologies. In Semantics, Web and Mining. Springer, pp. 121–131.
12. **Fuentes-Pineda, G., Meza-Ruiz, I. V. (2019).** Topic discovery in massive text corpora based on min-hashing. Expert Systems with Applications, Vol. 136, pp. 62–72.
13. **Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G. (2019).** Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems, Vol. 61, No. 2, pp. 1123–1145.
14. **He, G., Liang, Y., Chen, Y., Yang, W., Liu, J. S., Yang, M. Q., Guan, R. (2018).** A hotspots analysis-relation discovery representation model for revealing diabetes mellitus and obesity. BMC systems biology, Vol. 12, No. 7, pp. 116.
15. **Hjørland, B. (2007).** Semantics and knowledge organization. Vol. 41, No. 1, pp. 367–405.
16. **Hofmann, T. (2013).** Probabilistic latent semantic analysis. arXiv preprint arXiv:1301.6705.
17. **Landauer, T. K., Foltz, P. W., Laham, D. (1998).** An introduction to latent semantic analysis. Discourse processes, Vol. 25, No. 2-3, pp. 259–284.
18. **Lezama Sánchez, A. L., Tovar Vidal, M., Reyes-Ortiz, J. A. (2021).** Hypernyms-based topic discovery using LDA. **Batyrshin, I., Gelbukh, A., Sidorov, G.**, editors, Advances in Soft Computing, Springer International Publishing, Cham, pp. 70–80.
19. **Li, C., Feng, S., Zeng, Q., Ni, W., Zhao, H., Duan, H. (2018).** Mining dynamics of research topics based on the combined LDA and WordNet. IEEE Access, Vol. 7, pp. 6386–6399.
20. **Moody, C. E. (2016).** Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019.
21. **Perera, K., Karunarathne, D. (2015).** Keygraph and wordnet hypernyms for topic detection. 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE, pp. 303–308.
22. **Qiang, J., Qian, Z., Li, Y., Yuan, Y., Wu, X. (2020).** Short text topic modeling techniques, applications, and performance: a survey. IEEE Transactions on Knowledge and Data Engineering.
23. **Saorín, T. (2012).** Wikipedia de la A a la W, volume 8. Editorial UOC.
24. **Tovar, M., Pinto, D., Montes, A., González, G., Vilarino, D. (2015).** Identification of ontological relations in domain corpus using formal concept analysis. Engineering Letters, Vol. 23, No. 2.
25. **Xie, P., Yang, D., Xing, E. (2015).** Incorporating word correlation knowledge into topic modeling. Proceedings of the 2015 conference of the north American chapter of the association for computational linguistics: human language technologies, pp. 725–734.
26. **Xu, Z., Harzallah, M., Guillet, F., Ichise, R. (2019).** Modular ontology learning with topic modelling over core ontology. Procedia Computer Science, Vol. 159, pp. 562–571.
27. **Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J., Zhang, A. (2016).** Topic discovery for short texts using word embeddings. 2016 IEEE 16th international conference on data mining (ICDM), IEEE, pp. 1299–1304.

Article received on 20/07/2021; accepted on 30/09/2021.
Corresponding author is José Alejandro Reyes Ortiz.