Can Topic Modelling benefit from Word Sense Information?

Adriana Ferrugento[1]**, Hugo Gonçalo Oliveira**[1]**, Ana Oliveira Alves**[1,2]**, Filipe Rodrigues**[1]

[1]CISUC, Department of Informatics Engineering
University of Coimbra, Portugal
[2]IPC, Polytechnic Institute of Coimbra, Portugal
aferr@student.dei.uc.pt, {hroliv,ana,fmpr}@dei.uc.pt

Abstract

This paper proposes a new topic model that exploits word sense information in order to discover less redundant and more informative topics. Word sense information is obtained from WordNet and the discovered topics are groups of synsets, instead of mere surface words. A key feature is that all the known senses of a word are considered, with their probabilities. Alternative configurations of the model are described and compared to each other and to LDA, the most popular topic model. However, the obtained results suggest that there are no benefits of enriching LDA with word sense information.

Keywords: topic model, word senses, WordNet, semantics, SemLDA

1. Introduction

Topic models uncover the main subjects addressed in the documents of a corpus by inferring their themes or topics, which in turn correspond to probability distributions over words. Topic models are useful for improving traditional browsing and indexing, summarisation of large quantities of text, and text classification, among others. Classic topic modelling algorithms, such as the popular LDA (Blei et al., 2003), rely on the co-occurrences of surface words to capture their semantic proximity.

One limitation of existing topic models is that they do not consider semantic knowledge on the words, including their possible senses. Namely, they consider a surface word to be identical in different contexts and leverage on its co-occurrences with other words to differentiate topics. This may, for instance, result in topics with synonyms, both redundant and less informative.

Our goal was to develop SemLDA, a LDA-based topic model that would be sensitive to word senses. To minimise the aforementioned limitations, SemLDA considers not only the context where each word occurs, but also information on its possible senses, obtained, for example, from the lexical-semantic knowledge-base WordNet (Fellbaum, 1998). Similarly to other semantic topic models, instead of sets of surface words, the topics produced by SemLDA are based on word senses, inside their WordNet synsets. The main difference is that SemLDA considers all the possible senses of the words in a document, together with their probabilities. Moreover, it only requires a small intuitive change to the classic and popular LDA algorithm.

SemLDA was originally introduced in Ferrugento et al. (2015). This paper describes and performs an in-depth analysis of alternative configurations, involving different approaches for computing the probabilities of a word in a synset, which is the key step towards introducing word sense information in topic models and towards making SemLDA work in practice. Implemented configurations are compared among each other and with the classic LDA, first based on word association measures and then on a classification task. Despite resulting in more informative topic models, improvements over the classic LDA are not as clear as we would expect. In fact, the results obtained show that most versions of SemLDA are outperformed by the classic LDA, which hihglights some of the difficulties in enriching topic models with word sense information.

The remaining of this paper starts by briefly reviewing works on the automatic discovery of topics, with a focus on those that incorporate semantics. The proposed model is then introduced. After that, different configurations of SemLDA are described, in the form of experiments. The evaluation effort of SemLDA is finally presented, followed by a discussion of the main conclusions of this work.

2. Related Work

The first notable approach to reduce the dimensionality of documents was Latent Semantic Indexing (LSI) (Deerwester et al., 1990), which aimed at retaining the most of the variance present in the documents, thus leading to a significant compression of large datasets. Probabilistic Latent Semantic Indexing (pLSI) (Hofmann, 1999) later emerged as a variant of LSI, where different words in documents are modelled as samples from a simple mixture model where the mixture components are multinomial random variables that can be viewed as representations of “topics”. Nevertheless, pLSI was still not a proper generative model of documents, given that it provides no probabilistic model at the level of documents. With this limitation in mind, Blei et al. (2003) developed the Latent Dirichlet Allocation (LDA), a generalization of pLSI that is currently the most popular topic model. It allows documents to have a mixture of topics, given that it enables to capture significant intra-document statistical structure via the mixing distribution.

The main purpose of the previous models is to discover and assign different topics – represented by sets of surface words, each with a different probability – to the collection of documents provided. Those approaches have no concern with additional semantic knowledge about words, which can lead to some limitations in the generated topics. For instance, they might include synonyms, and thus be redundant and less informative. Alternative attempts address this problem using, for instance, WordNet (Miller, 1995), a lexical-semantic knowledge base of English. WordNet is structured in synsets, which are groups of synonymous words that may be seen as concept representations of a language. Synsets may be connected according to different semantic relations, such as hypernymy (generalization) or meronymy (part-of).

In an attempt to include semantics in topic modelling and, at the same time, perform word sense disambiguation (WSD), 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 word sense disambiguation (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.

Other LDA-based approaches incorporate semantics during the pre-processing phase (Guo and Diab, 2011), using WordNet as a sense repository, or during the generative process (Tang et al., 2014), this time inducing senses automatically from text. Similarly to the latter, there is additional work towards the discovery of concept-based topics, also not relying in WordNet. For instance, LDA was used as a ground model to generate topics based on the concepts of an ontology (Chemudugunta et al., 2008); and a commonsense knowledge-based algorithm was used to transform documents into commonsense concepts, which were then clustered to generate the topics (Rajagopal et al., 2013).

Despite some similarities, the model proposed in this paper differs from the previous in various ways. 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. But SemLDA considers all possible senses of a word, with a distribution over all the synsets that include it. Indeed, we do not benefit from similar words in the same topic to improve WSD, as in LDAWN. Rather, we try to avoid it.

3. Proposed Model

SemLDA extends Latent Dirichlet Allocation (Blei et al., 2003) by introducing a new set of parameters $\eta_{1:S}$, where S is the number of synsets where the word occurs (one for each of its senses). These parameters correspond to the probabilities of each word belonging to a synset (i.e. a concept). Hence, instead of assigning words (w_n) directly to topics, in SemLDA words are first assigned to concepts (c_n), which they are associated to, and each concept is then assigned to one or more topics by a mixture distribution (θ). Hence, contrarily to LDA, where topics correspond to distributions over words, in SemLDA the topics $\beta_{1:K}$ are probability distributions over concepts. The generative process of SemLDA can be summarized as follows:

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 $w_n|c_n, \eta_{1:S} \sim Mult(\eta_{c_n})$

The graphical model of SemLDA is depicted in Figure 1 where: D is the number of documents in the corpus, K is the number of topics, S is the number of available synsets and N is the number of words in a document. In this model, each word in a document, w_n, is drawn from a concept, c_n, and from a synset distribution, η. The concept c_n is determined by a discrete topic-assignment, z_n, picked from the document's distribution over topics θ and a topic distribution β.

In order to learn the proposed model, a variational Bayesian EM (VB-EM) algorithm was developed. The key difficulty is in estimating the parameters $\eta_{1:S}$, for which we explored different techniques, as we shall see in Section 4. The proposed model and inference algorithm are described in more detail in Ferrugento et al. (2015).

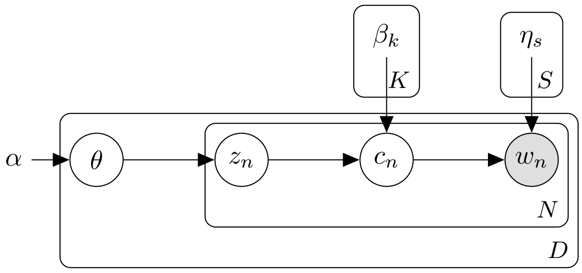

Figure 1: Graphical model representation of SemLDA.

4. Experiments

This section describes selected experiments towards possible implementations of SemLDA. After explaining their set up, the involved preprocessing is described together with the approach taken, a view on the results and an illustrative output. All experiments were performed in two English corpora, namely: 20 Newsgroups[1] and AP. The 20 Newsgroups is a popular dataset for experiments in text applications of machine learning techniques. It contains 20,000 documents, organized into 20 different newsgroups. AP is a large news corpus, from which we used only a part, more precisely, the sample data for the C implementation of LDA, available in David Blei's website[2].

In a preprocessing stage, the terms of all documents were lemmatized and stopwords in the Onix list[3] were removed. From the second experiment onward, part-of-speech (POS) tagging was applied to handle syntactical ambiguities and thus reduce the number of candidate synsets, as some surface words may have multiple POS (e.g. *plant* or *cover* can either nouns or verbs, *red* or *young* can either be nouns or adjectives).

The first experiment is the most basic, as it consists of just replacing each word in a document with its most frequent synset. The other four explore alternative configurations of SemLDA and were performed towards the validation of this model. Their main difference is the process

[1] http://qwone.com/~jason/20Newsgroups/
[2] http://www.cs.princeton.edu/~blei/lda-c/
[3] http://www.lextek.com/manuals/onix/stopwords1.html

of computing the probabilities of a word in a synset. More precisely, second experiment relies on the content of the SemCor corpus (Miller et al., 1994), version 3.0 – in SemCor, words are manually annotated according their WordNet senses and, in WordNet, both synsets and word senses are ordered according to their frequencies in SemCor. In the third, probabilities are either based on word sense disambiguation (WSD, Navigli (2009)) or SemCor. The fourth experiment only considers the probabilities from WSD, and does not require a sense-annotated corpus, which makes SemLDA more flexible and adaptable to other languages and/or wordnets. In this case, lexical-semantic knowledge was obtained from WordNet 3.0.

A few preliminary runs were performed to select preprocessing options and the parameters for the classic LDA algorithm, used in all the reported experiments. For instance, we empirically concluded that it did not make a big different to ignore some open-category words (e.g. adjectives or adverbs). LDA has an input parameter α, which may either be estimated through maximum likelihood or be assigned a fixed value. In this case, α was empirically set to 0.5 for all experiments. To discover the appropriate number of topics for each corpus, a Hierarchical Dirichlet process (HDP) (Teh et al., 2006) was used. Obtained results suggested that the 20 Newsgroups would have 15 topics and the AP 24 topics.

4.1. Most Frequent Sense

In this experiment, the occurrences of each word in a document were replaced by the ID of their first sense, according to WordNet 3.0, where word senses are ordered according to their frequency in SemCor. For example, the most frequent senses of the noun *student* and of the verb *to learn* are respectively in the following synsets:

- 10665698: {*student, pupil and educatee*}: a learner who is enrolled in an educational institution
- 597915: {*learn, larn, acquire*}: gain knowledge or skills

As several synonyms are used, this results in a lower number of words per document. For instance, the most frequent sense of the noun *pupil* is in the same synset as the most frequent sense of *student*.

After replacing words with synset IDs, the classic LDA was run on the resulting sequence of IDs. Table 1 shows one of the obtained topics.

4.2. Considering all possible senses

In this experiment, instead of choosing only the first sense, all the possible WordNet senses of a word were considered, though with different probabilities. This experiment was originally reported in Ferrugento et al. (2015), though with minor differences in the preprocessing.

Although many words have multiple senses, they often have distinct occurrence probabilities. To capture those, SemCor 3.0 was exploited. The probability of a word w given a synset s is obtained from equation 1, where $P(s|w)$ is the number of times w occurs in SemCor with its sense in s, and $P(s)$ is the number of times s occurs in SemCor.

$$P(w|s) = \frac{P(s|w)}{P(s)} \quad (1)$$

While this is straightforward for those WordNet synsets in SemCor, there are words and senses not covered by this corpus. To handle this, an extra preprocessing step was added: when a word in a document is not in SemCor, a new "dummy" synset is created, with a special negative ID, including just the uncovered word, with probability equals to 1. Table 2 displays one of the obtained topics.

This experiment can be seen as the first implementation of the SemLDA model, where the classic LDA is run with the computed probabilities as input.

4.3. Word Sense Disambiguation with Fallback

Most topics from the previous experiment include different senses of the same word, often unrelated. Apparently, the distribution of probabilities in SemCor is not enough to discriminate between each sense. Our next step was to perform WSD on each word to discover the most suitable sense for its context. For this purpose, we relied on the Adapted Lesk (Banerjee and Pedersen, 2002) algorithm to score each candidate WordNet synset for a word in context, provided by (Tan, 2014). $P(w|s)$ was computed from those scores, which are based on context overlaps. Still, when the context was not enough to rank the candidate synsets, the previous experiment was used as a fallback mechanism. Table 3 displays one of the obtained topics.

4.4. Just Word Sense Disambiguation

This experiment is similar to the previous but it does not use SemCor as a fallback, which makes SemLDA more flexible and adaptable to another language where such a corpus is not available. Instead, when the context is not enough to rank the candidates synset, probability is uniformly distributed among them. Table 4 displays on of the obtained topics.

4.5. SemCor-based Classifier

In this final experiment, the process of computing $P(w|s)$ is significantly different from the previous. First, LDA is applied to SemCor. Then, for every synset of each word in SemCor, a Logistic Regression classifier is trained with the probability distribution collected from LDA. Therefore, for each word in the corpus that is in one or more WordNet synsets, if there is a previously trained model, probabilities $P(w|s)$ are predicted from the distribution of LDA. When a word is not in WordNet or if there is not a trained model for a synset, a "dummy" synset is created. Table 5 displays one of the obtained topics.

5. Evaluation

Looking at the topics discovered with the last experiments, including the presented examples, results seem interesting. Yet, to have a more objective view and enable comparison between topics of different experiments, we conducted an evaluation that, at some level, enables the comparison of different approaches to topic modelling. In this case, topics obtained with the presented experiments were compared to those obtained with classic LDA. First, we relied on two

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 1: Topic from 20 Newsgroups, obtained when words are replaced with their most frequent synset.

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 2: Topic from 20 Newsgroups, obtained after considering all possible word senses.

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 3: Topic from 20 Newsgroups, obtained with the WSD with fallback experiment.

word association measures that have previously achieved high correlations with the human evaluation of topics. Second, we used the topic distributions to train a classifier that would predict the category of each document.

5.1. Word Association Measures

The measures of topic coherence (Mimno et al., 2011) and pointwise mutual information (PMI) (Newman et al., 2011) have previously shown to have high correlations with the human evaluation of topics. Both of them assess the coherence of the words in the topic, but topic coherence exploits the modelled documents, while PMI relies on an external independent corpus.

In order to apply the previous measures, in each SemLDA topic, only the first word of each synset was used. This enables a fair comparison with the topics produced by the

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: Topic from 20 Newsgroups, obtained with the WSD experiment.

Synset ID	POS	Words	Gloss
6431740	N	Bible, Christian Bible, Book, Good Book, Holy Scripture, Holy Writ, Scripture, Word of God, Word	The sacred writings of the christian religions.
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).
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.
9678009	N	Christian	A religious person who believes Jesus is the Christ and who is a member of a Christian denomination.
3560161	N	idol, graven image, god	A material effigy that is worshipped.
689344	V	think, believe, consider, conceive	Judge or regard; look upon; judge.
1260731	N	sin, hell	Violent and excited activity.
5916739	N	impression, feeling, belief, notion, opinion	A vague idea in which some confidence is placed.
10133307	N	god	A man of such superior qualities that he seems like a deity to other people.
8180190	N	multitude, masses, mass, hoi polloi, people, the great unwashed	The common people generally.

Table 5: Topic from 20 Newsgroups, obtained with the classifier experiment.

classic LDA, which are sets of surface words. We recall that, in WordNet, words are ordered in the synsets according to their probability to denote the synset meaning, computed in SemCor.

The PMI of a topic t is computed with equation 2, based on the co-occurrence probabilities of every pair of its top-10 words in an external corpus, more precisely, 45 pairs. As suggested by (Newman et al., 2010), the probability of each word was based on the number of Wikipedia articles where it occurred. So, in our case: $p(w)$, the probability of a word w, is the number of Wikipedia articles using this word; and $p(w_i, w_j)$, the probability of words w_i and w_j co-occurring, is the number of Wikipedia articles using both of these words. Wikipedia was used because it provides a large and wide-coverage source of text, completely independent from the datasets used and from WordNet.

$$PMI(t) = \frac{1}{45}\sum_{i<j}\log\frac{p(w_i, w_j)}{p(w_i)p(w_j)}, ij \in \{1...10\} \qquad (2)$$

For each topic, the measure dubbed topic coherence is computed with equation 3, where $D(v)$ is the number of documents with word v, $D(v, v')$ is the number of documents containing both words v and v', $V^{(t)} = (v_1^{(t)}, ..., v_M^{(t)})$ is a list of the M most probable words in topic t (in this case, $M = 10$), and 1 is a smoothing count to avoid the logarithm of zero.

$$C(t;V^{(t)}) = \sum_{m=2}^{M}\sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \tag{3}$$

This measure is very similar to PMI, but PMI is calculated on an independent corpus, which means that it evaluates topics as more generic instances and not just within the original collection.

Table 6 displays the average results of each measure on the topics obtained by the classic LDA as well as those obtained in each of the reported experiments, respectively for the 20 Newsgroups and for the AP corpora. The higher the scores of both measures, the better.

The only configuration of SemLDA that outperformed the classic LDA in both measures and both corpora is the one of experiment two, where all the senses of a word are considered, and their probabilities are based on SemCor. Yet, considering the standard deviation, there is still overlap with the results of LDA. For the remaining configurations, remarks should be given on: the fifth experiment (Classifier), which outperformed LDA in both measures for the 20 Newsgroups, but it performed worse in AP; and the fourth experiment (WSD), which got a better PMI for the 20 Newsgroups, but all the other measures are below LDA.

5.2. Classification

Besides the previous measures, the proposed model was also evaluated in a text classification task. The idea was to assess the predictive quality of the topic proportions inferred by SemLDA. For this purpose, the 20 Newsgroups corpus, also a popular benchmark dataset for text classification, was used. Its twenty thousand messages are divided among six super-classes, which are, in turn, partitioned in several sub-classes. For this experiment, only the six super-classes were used: “computers”, “science”, “politics”, “religion”, “recreative” and “sales”.

The proposed model was applied to the corpus in order to infer to topic proportions θ^d of each document d. The latter were then used to train a classifier. Using the Weka toolkit (Hall et al., 2009), several classifiers were tested. Table 7 reports the accuracy obtained by the different classifiers when trained on the topic proportions produced by SemLDA in comparison to those produced by classic LDA. Namely, the two versions of SemLDA from the second and fourth experiments were used (referred to in the table as “All senses/SemCor” and “WSD” respectively), since these were believed to be the most promising.

Classifier	LDA	All senses / SemCor	WSD
AdaBoostM1	37.1%	39.4%	33.8%
Bayes Network Classifier	71.8%	69.7%	57.8%
Decision Table	56.9%	58.7%	51.6%
KNN	**74.9%**	69.9%	59.1%
Logistic Regression	**76.1%**	74.4%	57.1%
MultiClassClassifier	**76.7%**	**75.1%**	58.4%
Naive Bayes Classifier	**45.0%**	44.6%	23.4%
NBTree	**71.2%**	69.6%	**61.6%**
SimpleLogistic	**76.0%**	74.3%	56.9%
SMO	**44.8%**	44.0%	34.9%

Table 7: Classification accuracy in the 20 Newsgroups.

The best results of this task were obtained with the classic LDA and a Multi Class classifier. None of the best accuracies obtained with a SemLDA-based classifier outperformed the best accuracies of the classic LDA. Comparing both configurations of the SemLDA, the topics obtained with the second experiment (All senses / SemCor) seem to be more suitable than those of the fourth experiment, where WSD was applied. While the latter is more flexible, it may also introduce new vocabulary, obtained from WordNet, especially when WSD does not select the most suitable synset. This, of course, results in lower performance in a classification task.

6. Concluding remarks

We have presented SemLDA, a topic model that considers word senses and their probabilities, and the implementation of different configurations of this model. Topics obtained were evaluated by automatic measures and results are not very positive, because none of the configurations of SemLDA clearly outperforms the classic LDA. In fact, the results of most configurations are below those for the classic LDA, which suggests that, similarly to other natural language processing tasks where vector models were enriched with the information of induced word senses (e.g. named entity recognition or sentiment analysis) (Li and Jurafsky, 2015), topic models do not benefit from word sense information.

However, we recall that, in order to enable a comparison with the classic LDA, SemLDA topics were oversimplified and only the first word of each synset was used. This fails to capture a key feature of SemLDA: instead of mere surface words, produced topics are based on synsets, and thus more informative than LDA’s. Therefore, one can always retrieve additional information from WordNet on the sense of each word in a SemLDA topic, through its gloss, synonyms and other relations. In the future, alternative evaluation approaches should be considered. For instance, topics can be presented to human judges with a random word / synset replaced by another (intruder) that does not belong to the topic. The rate of correctly identified intruders could be a sign of the quality of the topics. In this case, for SemLDA, the full synsets could be presented. The PMI measure could also be computed again, this time using the first n words of each synset in a SemLDA topic. The performance of SemLDA using different WSD algorithms should also be analysed. In order to assign different probabilities to the different senses of each word, WSD algorithms that give a numeric score to each candidate sense are more suitable.

Despite the previous issues, each performed experiment was helpful on their own. The second experiment was the only one that clearly outperformed LDA with both automatic measures and in both corpora. But this could be due to the occasional presence of the same word in different topics. This also confirmed that, despite difficulties on sorting them properly, our model considered all the senses of a word. Although it handled the previous issue, the impact of introducing WSD was below our expectations, most likely due to the process of computing probabilities, both with SemCor and WSD. The evaluation results of the fourth experiment, WSD without SemCor, were close to those of the classic LDA, but below. It is also the most flexible configu-

20 Newsgroups corpus

Measure	LDA	Top senses	All senses / SemCor	WSD w/ fallback	WSD	Classifier
PMI	1.175 ± 0.30	1.154 ± 0.35	1.302 ± 0.48	1.145 ± 0.33	1.215 ± 0.45	**1.321 ± 0.32**
Coherence	-35.186 ± 15.32	-34.928 ± 17.09	-32.491 ± 12.87	-42.118 ± 14.72	-40.235 ± 12.8	**-20.468 ± 3.55**

AP corpus

Measure	LDA	Top senses	All senses / SemCor	WSD w/ fallback	WSD	Classifier
PMI	1.286 ± 0.35	1.167 ± 0.36	**1.350 ± 0.36**	0.984 ± 0.17	1.175 ± 0.38	1.173 ± 0.34
Coherence	-21.184 ± 15.58	-26.492 ± 15.51	**-19.111 ± 16.07**	-29.622 ± 14.66	-28.806 ± 16.63	-33.066 ± 16.79

Table 6: Evolution of the automatic measures in the topics obtained from both corpora used.

ration, as it may be adapted to any language with a wordnet. To enable its utilization or improvement by others, the implementations of SemLDA are available from `https://github.com/aferrugento/SemLDA`, together with a list of the steps for performing the described experiments.

Acknowledgements

This work is partially supported by National Funds through the FCT –- Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) – within project REMINDS — UTAP-ICDT/EEI-CTP/0022/2014. Adriana Ferrugento was supported by the project InfoCrowds – FCT-PTDC/ECM-TRA/1898/2012.

Bibliographical References

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

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boyd-Graber, J. and Blei, D. (2007). A Topic Model for Word Sense Disambiguation. In *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, number June in EMNLP-CoNLL, pages 1024–1033.

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

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.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Ferrugento, A., Alves, A. O., Gonçalo Oliveira, H., and Rodrigues, F. (2015). Towards the improvement of a topic model with semantic knowledge. In *Proc. of the 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal*, volume 9273 of *LNCS*, pages 759–770, September.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations Newsletter*, 11(1):10–18, November.

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

Li, J. and Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, pages 1722–1732, Lisbon, Portugal, September. ACL Press.

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

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

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

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

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. ACL Press.

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.

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

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).