Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 15-22 recibido 26-03-2015 revisado 27-04-2015 aceptado 14-05-2015

# Topic Modeling and Word Sense Disambiguation on the Ancora corpus

## *Modelado de Categorías y Desambiguación del Sentido de las Palabras en el corpus Ancora*

**Rubén Izquierdo** **Marten Postma** **Piek Vossen**
VU University of Amsterdam, The Netherlands
{ruben.izquierdobevia, m.c.postma, piek.vossen}@vu.nl

**Resumen:** En este artículo se presenta una aproximación a la Desambiguación del Sentido de las Palabras basada en Modelado de Categorías (LDA). Nuestra aproximación consiste en dos pasos diferenciados, donde primero un clasificador binario se ejecuta para decidir si la heurística del sentido más frecuente se debe aplicar, y posteriormente otro clasificador se encarga del resto de sentidos donde esta heurística no corresponde. Se ha realizado una evaluación exhaustiva en el corpus en español Ancora, para analizar el funcionamiento de nuestro sistema de dos pasos y el impacto del contexto y de diferentes parámetros en dicho sistema. Nuestro mejor experimento alcanza un acierto de 74.53, lo cual es 6 puntos superior al *baseline* más alto. Todo el software desarrollado para estos experimentos se ha puesto disponible libremente para permitir la reprodubilidad de los experimentos y la reutilización del software
**Palabras clave:** Modelado de categorías, LDA, Sentido más frecuente, WSD, corpus Ancora

**Abstract:** In this paper we present an approach to Word Sense Disambiguation based on Topic Modeling (LDA). Our approach consists of two different steps, where first a binary classifier is applied to decide whether the most frequent sense applies or not, and then another classifier deals with the non most frequent sense cases. An exhaustive evaluation is performed on the Spanish corpus Ancora, to analyze the performance of our two–step system and the impact of the context and the different parameters in the system. Our best experiment reaches an accuracy of 74.53, which is 6 points over the highest baseline. All the software developed for these experiments has been made freely available, to enable reproducibility and allow the re–usage of the software.
**Keywords:** Topic Modeling, LDA, Most Frequent Sense, WSD, Ancora corpus

## *1 Introduction*

Word Sense Disambiguation (WSD) is a well–known task within the Natural Language Processing (NLP) field which consists of assigning the proper meaning to a word in a certain context. A very large number of works and approaches have addressed this task from different perspectives in the last decades. Despite all this effort, the task is considered to be still unsolved, and the performance achieved is not comparable to other tasks such as PoS–tagging (with an accuracy around 98%). This is especially problematic if we consider that sense information is used in almost all the high levels NLP tasks (event extraction, NER...). An extensive description of the WSD task and their approaches can be found in Agirre and Edmonds (2007).

Lately, more and more WSD unsupervised approaches have been exploiting, with a reasonable performance under some circumstances, the large resources that are becoming available. Nevertheless, the most widely applied techniques to WSD have been those based on supervised Machine Learning. These approaches tackle WSD as a classification problem, where the goal is to pick the best sense from a predefined list of possible values for a word in a given context, being WordNet (Fellbaum, 1998) the main sense repository selected.

Traditionally, one Machine Learning algorithm is selected (SVM, MaxEnt...), and local and topical features are used to represent the training examples and induce the models.

ISSN 1135-5948 

Nevertheless, the size of the context considered to model the problem is usually quite narrow (quite often not more than one sentence), and this may not be sufficient in some cases. Little attention has been paid to consider the role of broader contexts, such as the whole document, or even background information that could be found in external resources and it is not implicit in the document.

The most frequent sense (MFS) heuristic has been extensively used as a baseline for comparison and evaluation. This heuristic has turned to be very difficult to beat by any WSD system. Indeed, we think that in many cases the systems are too skewed towards assigning the MFS, and they do not address properly the problem, specially in the cases where the MFS does not apply. In this direction, we performed an error analysis on the previous SensEval/SemEval evaluations (Izquierdo, Postma, and Vossen, 2015). We found that the participant systems perform very well when the MFS is the correct sense (68% in average in SensEval2, 78% in SensEval3 or 80% in SemEval-2013), but the performance dramatically goes down when the correct label is not the MFS (20% for SensEval2, 18% for SensEval3 or 22% for SemEval2013). Besides to this, we found that, considering the SensEval–2 test dataset, when the correct sense is not the MFS (799 cases), in the 84% of the cases the systems still pick the MFS, which shows clearly the bias towards assigning the MFS that was mentioned before.

In this paper we propose to use topic modeling to perform WSD in the Ancora corpus (Taulé, Martí, and Recasens, 2008), which is a multilevel annotated corpus for Catalan and Spanish, and from which we will make use of the sense annotations for Spanish. Topic modeling is a statistical approach within the Machine Learning field, that tries to discover automatically what are the main topics for a given document or text. We will exploit this technique to create a supervised WSD system that automatically learns the topics related with different senses of a target word and uses these topics to select the proper sense for a new unknown word. The impact of the context and the number of topics on the performance of the WSD system will be also explored. Besides to this, the phenomenon of the most frequent sense will be analyzed and considered as an individual step in the entire WSD problem. To our knowledge there is no other work presenting such an analysis based on topic modeling for Spanish. With the experimentation presented in this paper, an improvement around 6 points in accuracy is obtained over the most frequent sense baseline. All the software developed and the data used for these experiment has been made freely available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015` enabling the reproducibility of these experiments as well as the reuse of the code and data created by the NLP community.

Section 2 will introduce some works applying topic modeling to perform WSD. Then section 3 will present our system architecture. The evaluation framework will be introduced in section 4. Finally the results will be presented in section 5 and some conclusions and future work will be drawn in section 6.

## 2 Related work

Latent Dirichlet Analysis (LDA) and Topic Modeling in general have been largely applied in NLP tasks, mainly in document classification, topic classification and information retrieval. In these areas, the strong relation between the definition and objective of the task and the application and relevance of topics is quite obvious. In addition, Topic modeling has been also applied in some works to perform Word Sense Disambiguation, which is the main focus of our paper. For instance in Cai, Lee, and Teh (2007), LDA is applied to extract topic features from a large unlabeled corpus. These features are fed into a Naïve Bayes classifier, together with traditional features (part-of-speech, bag-of-words, local collocations...). They perform the evaluation on the SensEval-3 corpora, showing a significant improvement with the use of the topic features. Also in Boyd-Graber and Blei (2007) the authors extend the predominant sense algorithm presented in McCarthy et al. (2004) to create an unsupervised approach for WSD. The topics obtained via LDA are used to calculate similarity measures and pre-dictions for each word in the document, also considering frequencies and features from the surrounding words.

In Li, Roth, and Sporleder (2010) the task of WSD is approached by selecting the best sense based on the conditional probability of sense paraphrases given a context. Two mod-

els are proposed for WSD. One requires prior knowledge of the conditional probability of senses; the second one uses the cosine similarity of two topic-document vectors (sense and context). They prove to get good results (comparable to state-of-the-art) when evaluating at different granularity levels on several SemEval and SenseEval datasets.

The structure of WordNet is exploited in another unsupervised approach presented by Boyd-Graber, Blei, and Zhu (2007). WordNet senses are incorporated as additional latent variables. Each topic is associated not just with simple words, but with a random walk through the WordNet hierarchy. Top-ics and synsets are generated together. An improvement is obtained in some cases, but in some other cases the structure of WordNet affects the accuracy of the system.

Topics and topic modeling have been extensively applied to word sense induction. For instance in (Brody and Lapata, 2009) sense induction is performed as a Bayesian problem by modeling the contexts of the ambiguous word as samples from a multinomial distribution over senses which are in turn characterized as distributions over words. Other works facing word sense induction from a topic modeling point of view are (Wang et al., 2015) or (Knopp, Völker, and Ponzetto, 2013).

## 3 Our WSD approach

Our WSD system[1] is a supervised machine learning framework based on topic modeling, in particular Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003). LDA is one of the algorithms for topic modeling that has shown a higher performance and some advantages compared to others such as Latent Semantic Indexing (LSI) or Random Indexing (RI). In our case we have used the LDA implementation available in the Gensim python library[2]. The main idea is to induce a topic model for every sense of every polysemous word based on token features within a certain context. Giving a new word (on the tagging or evaluation phase), we will pick the sense that maximizes the similarity of the feature document created for the new word with each of the models induced for every sense of the same lemma (in the training phase). The features used for representing one target example are the bag-of-words (token based) within a certain number of sentences around the target word. These classifiers assign the proper sense for a target word, and we will refer to them as *sense–LDA* classifiers.

In order to analyze what is the effect of the most frequent sense phenomenon (MFS) in our WSD system, we isolate the problem by considering two different steps in the classification task for a specific case:

1. Decide if the MFS applies in this case
2. If it applies, the MFS is selected
3. Otherwise, the sense returned by the *sense–LDA* is selected[3]

This means that basically a binary classifier is applied first (the MFS classifier) to decide if the most frequent sense applies in this case or not, and the second classifier (sense–LDA) is only queried in the cases where the MFS classifier does not apply. In fact the two tasks could be quite different in nature. On the one hand, deciding if the MFS applies can depend on clues found in larger contexts, related to the topics of the document or even derived from external knowledge sources. On the other hand, learning the topics for less frequent senses could rely on different types of information, linked to more specific and small contexts. Tackling both tasks in one step would not allow to specialize and exploit the proper information for each task. The classifiers derived for deciding on the MFS are based also in LDA and will be named *mfs–LDA* classifiers. The features in this case are the same bag-of-words on larger contexts.

## 4 Evaluation framework

For the evaluation of our WSD system we performed a folded–cross validation (3-FCV) on the Ancora corpus[4]. We first converted the Ancora corpus to NAF[5] format, as it is the format used by all the tools and linguistic processors developed in our group. Then the folds for training and evaluation were created for every lemma in the Ancora corpus,

[1] As stated previously, all the software and data used for these experiments can be found at `http://kyoto.let.vu.nl/lda_wsd_sepln2015`

[2] `http://radimrehurek.com/gensim/`

[3] The MFS can not be selected anymore in this step

[4] The folds created for our evaluation are available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015/data/`

[5] `http://www.newsreader-project.eu/files/2013/01/techreport.pdf`

making sure to keep the sense distribution in every fold for a fair evaluation.

Our evaluation has been focused only to the polysemous lemmas, and, from these, just in those with at least three manually annotated instances on all the corpus (otherwise the 3-FCV is not possible). There are a total of 7119 unique lemmas annotated in the Ancora corpus. Out of these, 4907 (almost 69%) are monosemous (or annotated just with one sense). From the remaining 31% polysemous[6], 589 lemmas fulfill the requirement of having at least three annotated instances per sense. This set of 589 lemmas compose our evaluation set.

For obtaining the evaluation figures, we use the traditional precision, recall and F-score, micro-averaging across the three folds to get the figures per lemma, and micro–averaging over all the lemmas to get the overall performance of the system. As the total coverage is 100% (the system always provides an answer for every test instance), precision, recall and F–score have the same value and we will refer to this value as *accuracy*. All the lemma output files for the the different experiments presented in next section can be found at `http://kyoto.let.vu.nl/lda_wsd_sepln2015/data/`.

## 5 Results

In this section the figures obtained by our WSD system for different configurations are shown. As we explained previously, we focus on the polysemous lemmas annotated with at least three instances for each sense (589 lemmas in total). All the results shown in this section refer to that set of 589 lemmas. In order to establish a reference for comparison, three baselines on the Ancora corpus have been derived following different heuristics:

- *Random*: selecting a random sense in each case
- *MFS–overall*: the well–known most frequent sense baseline considering the whole corpus to obtain the sense distribution
- *MFS–folded*: the most frequent sense heuristic using the evaluation folds to calculate the MFS

The *MFS–folded* baseline establishes a better comparison for our WSD system, as the information available for both is exactly the same. In table 1 we can see the figures for these baselines.

| Exp | Accuracy |
|---|---|
| *Random* | 40.10 |
| *MFS–overall* | 67.68 |
| *MFS–folded* | 68.63 |

Table 1: Baselines on the Ancora corpus

Both MFS baselines are quite high, as expected *a priori* and similarly to the same heuristics calculated for other languages and other sense annotated corpora.

Our first experiment evaluates the behavior of our WSD system when the disambiguation process is done just in one step by the *sense–LDA* classifier (so no MFS classifier is involved). As mentioned previously, one topic model is induced for every sense of each lemma (regardless the MFS or non MFS cases), and the classifier picks the sense that maximizes the similarity of a test instance against the possible sense models. In all our experiments involving LDA classifiers, there are two main parameters that can play a crucial role and that will be analyzed:

- Sentence size: number of sentences considered around a target word to extract the bag-of-word features. The possible values for this parameter are 0, 3 or 50 (where 0 means only the same sentence where the target word is contained). With these values we aim to examine what is the impact of three different context sizes (small, medium and large) on the topic induction task.
- Number of topics: the number of topics set to build the LDA models. In this case this parameter can take the values 3, 10 or 100, which represent three different levels of abstraction.

Combining these three values for the sentence window with the three values for the number of topics we obtain nine possible experiments. The results of these parameter combinations in our first experiment (just *sense–LDA* classifiers) can be seen in Table 2

[6] There 1318 with 2 senses, 449 with 3, 227 with 4, 110 with 5, 41 with 6, 38 with 7, 11 with 8, 10 with 9, 5 with 10 senses, 2 lemmas with 11 senses and one lemma with 12 senses

(*MFSfolded* baseline is included for easy comparison).

| Sentences | Topics | Accuracy |
|---|---|---|
| *MFSfolded* | - | 68.63 |
| 0 | 3 | 67.54 |
| | 10 | 65.56 |
| | 100 | 58.34 |
| 3 | 3 | 66.30 |
| | 10 | 64.62 |
| | 100 | 60.07 |
| 50 | 3 | 66.04 |
| | 10 | 63.42 |
| | 100 | 59.06 |

Table 2: Results the sense– classifiers (no MFS–classifier)

As can be seen, the *sense–LDA* classifier is not able to reach the *MFSfolded* baseline in any case. This could mean that indeed considering the task in just one step (with no MFS specialization) makes it very difficult for the LDA models to induce the correct topics. The best results are obtained by considering only the same sentence of the target word to get the features and three as the number of topics for LDA (67.54%). It seems that in this task, the most informative clues are to be found in near contexts. Besides to this, apparently there is certain relation between the two parameters. For instance, the result for $\{sentences = 0; topics = 100\}$ is 58.34 while the result for the same number of topics with $\{sentences = 10\}$ is 59.06, which could imply than for modeling broader contexts (with a larger number of tokens and features), a higher number of topics is required in order to get good results.

The second experiment consists in evaluating our two–steps approach, chaining together the *mfs–LDA* and the *sense–LDA* classifiers. Before doing this, we will evaluate which would be the performance of the whole WSD system if we could use a perfect *mfs–LDA* classifier. In order to simulate this, all the test instances where the correct label is the MFS are considered to be classified correctly, and the rest of instances are classified automatically by the *sense–LDA* classifier (this classifier does not assign the MFS in any case). In other words, this evaluation will examine the performance of the *sense–LDA* classifier on just the non–MFS instances. The results are shown in Table 3.

| Sentences | Topics | Accuracy |
|---|---|---|
| *MFSfolded* | - | 68.63 |
| 0 | 3 | 92.48 |
| | 10 | 92.12 |
| | 100 | 90.5 |
| 3 | 3 | 92.45 |
| | 10 | 92.11 |
| | 100 | 91.60 |
| 50 | 3 | 92.41 |
| | 10 | 92.12 |
| | 100 | 91.43 |

Table 3: Results of the WSD system with 2 steps: perfect *mfs–LDA* and automatic *sense–LDA*

The figures in this case are extremely high. This indicates that the *sense–LDA* is able to classify the non MFS cases with a high accuracy, which reinforces our idea of separating both tasks. The conclusions drawn about the combinations of number of sentences and topics are the same as in the previous experiment with only the *sense–LDA* classifier.

The next two experiments will show the evaluation of the two–steps WSD framework with both *mfs–LDA* and *sense–LDA* classifiers induced automatically. Two tables will be shown, the first one for a context of 5 sentences to build the *mfs–LDA* classifier, and the second one using 50 sentences instead.

The first table is Table 4. Each row presents the result for a certain combination of the sentence window and number of topics parameters for the *sense–LDA* classifiers. The last two columns represent the accuracy for different settings of the *mfs–LDA* classifier. In this table the number of sentences for the *mfs–LDA* classifier is set to 5, and there are two experiments for different values of the number of topics: 100 (column *"MFS s5 t100"*) and 1000 (column "*MFS s5 t1000*").

As derived from the table, in the all the cases using the *mfs–LDA* with options $\{sentences = 5; topics = 100\}$, the results are higher than the baseline. In concrete, the best experiment correspond to the *sense–LDA* with option $\{sentences = 0; topics = 3\}$ (74.53), with an improvement around 6 points over the Apparently, using a context of 5 sentences, 100 topics are more informative than 1000 to represent the main features that characterize the most frequent sense. Analyzing the different *sense–LDA* experiments

| Sentences | Topics | MFS s5 t100 | MFS s5 t1000 |
|---|---|---|---|
| *MFSfolded* | - | 68.63 | 68.63 |
| 0 | 3 | 74.53 | 66.73 |
| | 10 | 74.00 | 66.41 |
| | 100 | 72.61 | 64.91 |
| 3 | 3 | 74.30 | 66.61 |
| | 10 | 73.87 | 66.36 |
| | 100 | 73.39 | 65.76 |
| 50 | 3 | 74.26 | 66.48 |
| | 10 | 73.90 | 66.24 |
| | 100 | 73.53 | 65.75 |

Table 4: Results for different *sense–LDA* classifier with *mfs–LDA* (S=5 T=100) and *mfs–LDA* (S=5 T=1000)

| Sentences | Topics | MFS s50 t100 | MFS s50 t1000 |
|---|---|---|---|
| *MFSfolded* | - | 68.63 | 68.63 |
| 0 | 3 | 73.34 | 67.15 |
| | 10 | 72.92 | 66.76 |
| | 100 | 71.43 | 65.13 |
| 3 | 3 | 73.21 | 67.02 |
| | 10 | 72.88 | 66.60 |
| | 100 | 72.40 | 66.24 |
| 50 | 3 | 73.21 | 66.95 |
| | 10 | 72.83 | 66.58 |
| | 100 | 72.15 | 66.20 |

Table 5: Results for different *sense–LDA* classifier with *mfs–LDA* (S=50 T=100) and with *mfs–LDA* (S=50 T=1000)

in all the cases (the same sentence, 3 or 50 sentences) the results are quite similar. This would indicate that the most rich information to disambiguate the no MFS cases is to be found in local contexts (as the contexts of 3 and 50 sentences already include the smaller context of just the same sentence where the target words are found). Finally about what is the best number of topics to build the *sense–LDA* classifiers, the best performance is reached by using just 3 topics, indicating that larger number of topics may just introduce noise and no relevant information to the disambiguation process.

Following with the experimentation, Table 5 shows the same evaluation as in the previous table, but in this case the context used to build the *mfs–LDA* classifiers is 50 sentences. Similarly, there are two columns with the accuracy of the whole system when 100 or 1000 topics are selected to build the *mfs–LDA* classifiers.

The analysis of this table is similar to the previous one (with only 5 sentences used to generate the *mfs–LDA* classifiers). Comparing both tables, in this case the performance is a bit lower. This might point out that the clues for learning when the MFS applies or not are found in medium sizes contexts (at least for the simple bag–of–words feature model that is being used). Regarding the number of sentences or topics used to build the *sense–LDA* in this experiment, the behavior is the same as in the previous table with a context of 5 sentences.

Finally, we have evaluated individually the *mfs–LDA* classifiers on the task of deciding when the MFS applies or not. In next table, Table 6, the performance of the *mfs–LDA* classifiers for different settings of the parameters (*Sents* for the number of sentences considered as context and *Tops* as the number of Topics) is shown. Specifically, we show the accuracy on predicting the MFS cases, as these cases are those that can affect the overall performance of our system.

| **Sents. \\Tops.** | **100** | **1000** |
|---|---|---|
| **5** | 74.41 | 62.82 |
| **50** | 72.17 | 62.93 |

Table 6: Evaluation of the *mfs–LDA* classifiers on detecting the MFS cases

The results endorse the conclusions drawn from the previous experiments. The *mfs–LDA* classifier obtains a better performance by considering 100 topics to induce the models. Furthermore, and as expected, a *mfs–LDA* classifier with a better performance leads to a better overall accuracy when integrated in the two steps (*mfs–LDA* + *sense–LDA*) WSD system.

## 5.1 Lemma comparison

In this section we will compare the best of our experiments[7], with an accuracy of 74.53 with the baseline (68.63) at lemma level. Out of

[7] *mfs–LDA* with $\{sentences = 5; topics = 100\}$ and *sense–LDA* with $\{sentences = 0; topics = 3\}$

the 589 lemmas evaluated (those lemmas that are polysemous and at least with 3 senses annotated for each sense), a total of 399 (67.7%) lemmas were improved by our best run over the baseline, 126 were under the baseline (21.4%) and for 64 (10.9%) the accuracy of our system and the baseline was equal. In Table 7 we can see the top 5 lemmas with the highest improvement over the baseline. The columns *MFS* and *LDA* represent the accuracy for the MFS baseline and for our LDA system, the column *Var.* shows the variation of our system with respect to the baseline and the last column (*#S*) contains the number of senses of the lemma.

| lemma | MFS | LDA | Var. | #S |
|---|---|---|---|---|
| castigo | 50 | 100 | +50 | 2 |
| ética | 50 | 100 | +50 | 2 |
| veto | 50 | 100 | +50 | 2 |
| mediación | 50 | 100 | +50 | 2 |
| rebeldía | 50 | 100 | +50 | 2 |

Table 7: Lemmas with highest improvement

We can see that in all the five cases, the number of senses of these lemmas is two. This makes sense with our two–step approach, so if the *mfs–LDA* detects correctly the MFS cases, the rest of the cases become monosemous for the *sense–LDA* classifier. In next table, Table 8, we show the variation in accuracy of our system compared to the MFS baseline for the top 10 lemmas with the highest number of annotations in Ancora (the number of annotations for the lemma is presented in the column *#A.*).

| lemma | MFS | LDA | Var. | #A. |
|---|---|---|---|---|
| año | 89.15 | 91.19 | 2.04 | 1275 |
| país | 72.29 | 83.55 | 11.26 | 695 |
| presidente | 70.31 | 73.94 | 3.63 | 690 |
| partido | 55.87 | 64.48 | 8.61 | 641 |
| equipo | 98.32 | 98.88 | 0.56 | 539 |
| mes | 54.29 | 80 | 25.71 | 315 |
| hora | 61.39 | 56.11 | -5.28 | 305 |
| caso | 61.05 | 91.58 | 30.53 | 286 |
| mundo | 47.31 | 40.14 | -7.17 | 279 |
| semana | 85.06 | 92.34 | 7.28 | 263 |

Table 8: Improvement on the most frequent lemmas

In this case we can see a general positive effect, mainly with improvement over the baseline. These lemmas with a large number of annotations are those that can be most affected by the MFS bias. The improvement in these cases could show the robustness of our two–step WSD system. Finally we include in Table 9 those lemmas where our LDA systems presents the largest decrease with respect to the MFS baseline.

| lemma | MFS | LDA | Var. |
|---|---|---|---|
| colisión | 66.67 | 33.33 | -33.34 |
| filosofía | 66.67 | 33.33 | -33.34 |
| garantía | 60 | 26.67 | -33.33 |
| prestigio | 50 | 16.67 | -33.33 |
| congreso | 56.25 | 27.08 | -29.17 |

Table 9: Lemmas with highest reduction of accuracy

In the majority of these cases, the number of senses is around 2 or 3. This would indicate that either the *mfs–LDA* is not been modeling the context properly in these cases, or that the non most frequent senses for these lemmas are problematic and difficult to disambiguate (which is pointed out too by the discrete results of the MFS baseline in these cases).

## *6 Conclusions*

In this paper we have presented an approach for WSD based on Topic Modeling (LDA), and it has been evaluated on the Ancora Spanish corpus. The whole WSD task is split into two tasks: when the most frequent sense heuristic applies and when it does not. These subtasks have different nature and they might need to be approached in different steps. Our WSD system implements a two–step approach, where first a classifier is applied to decide whether or not the most frequent sense heuristic should be applied. In the cases where this heuristic does not correspond, a traditional sense classifier is employed to return the proper sense.

An exhaustive evaluation of our system has been performed following fold–cross validation on the Ancora corpus, in order to analyze all the different parameters that can play a role in our system. We have found that the best run reaches an accuracy of 74.53 by using the two step system, which is 6 points better than the most frequent sense baseline (68.63). In general, it seems that the best clues for deciding on the most frequent sense

are to be found on contexts around 50 sentences (medium sized) and using 100 topics to induce the models. For the traditional sense classifier, the best models are induced using few topics (3) within small sentence windows (just the sentences where the training instances occur).

All the code and software developed for these experiments, as well as the evaluation data and experiment outputs, can be found freely available at `http://kyoto.let.vu.nl/lda_wsd_sepln2015`. This will enable the reproduction of our experiments as well as the reuse of our programs for further research within the NLP community. Also the data used in these experiments and the output files produced are available.

As future work, we plan to incorporate external knowledge through the detection of named entities and their links to DBpedia in the whole process to enrich the classifiers. Some experiments have been already conducted in this direction with promising results, but some analysis are further experiments are still required. Furthermore, we will carry on a similar evaluation for other languages, starting with English, to reproduce our experiments and analyze our approach in other resources.

## *References*


Agirre, E. and P. Edmonds. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Blei, D. M., A.Y. Ng, and M. I. Jordan. 2003 Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Boyd-Graber, J. and D. Blei. 2007. Putop: Turning predominant senses into a topic model for word sense disambigua-tion. In *Proceedings of the Fourth International Workshop SemEval-2007*, pages 277–281. Association for Computational Linguistics.

Boyd-Graber, J. L., D. M. Blei, and X. Zhu. 2007. A topic model for word sense disambiguation. In *EMNLP*, pages 1024–1033. ACL.

Brody, S. and M. Lapata. 2009. Bayesian word sense induction. In *Pro-ceedings of the 12th EACL Conference*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cai, J., W. Sun Lee, and Y. Whye Teh. 2007. Improving word sense disambiguation using topic features. In *Proceedings of the 2007 Joint EMNLP and CoNLL conferences)*, pages 1015–1023.

Fellbaum, C., editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Izquierdo, R., M. Postma, and P. Vossen. 2015. Error analysis of word sense disambiguation. In *Proceedings of the Computational Linguistics in The Netherlands (CLIN)*, volume 25.

Knopp, J., J. Völker, and S. P. Ponzetto. 2013. Topic modeling for word sense induction. In Iryna Gurevych, Chris Biemann, and Torsten Zesch, editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 97–103. Springer Berlin Heidelberg.

Li, L., B. Roth, and C. Sporleder. 2010. Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th ACL conference*, ACL '10, pages 1138–1147, Stroudsburg, PA, USA. Association for Computational Linguistics.

McCarthy, D., R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42nd ACL conference*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taulé, M., M. A. Martí, and M. Recasens. 2008. Ancora: Multi-level annotated corpora for catalan and spanish. In Nicoletta Calzolari et al., edi-tor, *Proceedings of the Sixth International LREC*. European Language Resources As-sociation (ELRA).

Wang, J., M. Bansal, K. Gimpel, B. Ziebart, and C. Yu. 2015. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71.