# unimelb: Topic Modelling-based Word Sense Induction for Web Snippet Clustering

**Jey Han Lau, Paul Cook** and **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
jhlau@csse.unimelb.edu.au, paulcook@unimelb.edu.au,
tb@ldwin.net

## Abstract

This paper describes our system for Task 11 of SemEval-2013. In the task, participants are provided with a set of ambiguous search queries and the snippets returned by a search engine, and are asked to associate senses with the snippets. The snippets are then clustered using the sense assignments and systems are evaluated based on the quality of the snippet clusters. Our system adopts a pre-existing Word Sense Induction (WSI) methodology based on Hierarchical Dirichlet Process (HDP), a non-parametric topic model. Our system is trained over extracts from the full text of English Wikipedia, and is shown to perform well in the shared task.

## 1 Introduction

The basic premise behind research on word sense disambiguation (WSD) is that there exists a static, discrete set of word senses that can be used to label distinct usages of a given word (Agirre and Edmonds, 2006; Navigli, 2009). There are various pitfalls underlying this premise, including: (1) what sense inventory is appropriate for a particular task (given that sense inventories can vary considerably in their granularity and partitioning of word usages)? (2) given that word senses tend to take the form of prototypes, is discrete labelling a felicitous representation of word usages, especially for non-standard word usages? (3) how should novel word usages be captured under this model? and (4) given the rapid pace of language evolution on real-time social media such as Twitter and Facebook, is it reasonable to assume a static sense inventory? Given this backdrop, there has been a recent growth of interest in the task of word sense induction (WSI), where the word sense representation for a given word is automatically inferred from a given data source, and word usages are labelled (often probabilistically) according to that data source. While WSI has considerable appeal as a task, intrinsic cross-comparison of WSI systems is fraught with many of the same issues as WSD (Agirre and Soroa, 2007; Manandhar et al., 2010), leading to a move towards task-based WSI evaluation, such as in Task 11 of SemEval-2013, titled "Evaluating Word Sense Induction & Disambiguation within an End-User Application".

This paper presents the UNIMELB system entry to SemEval-2013 Task 11. Our method is based heavily on the WSI methodology proposed by Lau et al. (2012) for novel word sense detection. Largely the same methodology was also applied to SemEval-2013 Task 13 on WSI (Lau et al., to appear).

## 2 System Description

Our system is based on the WSI methodology proposed by Lau et al. (2012) for the task of novel word sense detection. The core machinery of our system is driven by a Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003). In LDA, the model learns latent topics for a collection of documents, and associates these latent topics with every document in the collection. A topic is represented by a multinomial distribution of words, and the association of topics with documents is represented by a multinomial distribution of topics, with one distribution per document. The generative process of LDA



*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 217–221, Atlanta, Georgia, June 14-15, 2013. 

for drawing word $w$ in document $d$ is as follows:

1. draw latent topic $z$ from document $d$;

2. draw word $w$ from the chosen latent topic $z$.

The probability of selecting word $w$ given a document $d$ is thus given by:

$$P(w|d) = \sum_{z=1}^{T} P(w|t = z)P(t = z|d).$$

where $t$ is the topic variable, and $T$ is the number of topics.

The number of topics, $T$, is a parameter in LDA, and the model tends to be highly sensitive to this setting. To remove the need for parameter tuning over development data, we make use of a non-parametric variant of LDA, in the form of a Hierarchical Dirichlet Process (HDP: Teh et al. (2006)). HDP learns the number of topics based on data, and the concentration parameters $\gamma$ and $\alpha_0$ control the variability of topics in the documents (for details of HDP please refer to the original paper, Teh et al. (2006)).

To apply HDP in the context of WSI, the latent topics are interpreted as the word senses, and the documents are usages that contain the target word of interest (or search query in the case of Task 11). That is, given a search query (e.g. *Prince of Persia*), a "document" in our application is a sentence/snippet containing the target word. In addition to the bag of words surrounding the target word, we also include positional context word information, as used in the original methodology of Lau et al. (2012). That is, we introduce an additional word feature for each of the three words to the left and right of the target word. An example of the topic model features for a context sentence is given in Table 1.

### 2.1 Background Corpus and Preprocessing

As part of the task setup, we were provided with snippets for each search query, constituting the documents for the topic model for that query (each search query is topic-modelled separately). Our system uses only the text of the snippets as features, and ignores the URL information. The text of the snippets is tokenised and lemmatised using OpenNLP and Morpha (Minnen et al., 2001).

As there are only 64 snippets for each query in the test dataset, which is very small by topic modelling standards, we turn to English Wikipedia to expand the data, by extracting all context sentences that contain the search query in the full collection of Wikipedia articles.[1] Each extracted usage is a three-sentence context containing the search query: the original sentence that contains the actual usage and its preceding and succeeding sentences. The extraction of usages from Wikipedia significantly increases the amount of information for the topic model to learn the senses for the search queries. To give an estimate: for very ambiguous queries such as *queen* we extracted almost 150,000 usages from Wikipedia; for most queries, however, this number tends to be a few thousand usages.

To summarise, for each search query we apply the HDP model to the combined collection of the 64 snippets and the extracted usages from Wikipedia. The topic model learns the senses/topics for all documents in the collection, but we only use the sense/topic distribution for the 64 snippets as they are the documents that are evaluated in the shared task.

Our English Wikipedia collection is tokenised and lemmatised using OpenNLP and Morpha (Minnen et al., 2001). The search queries provided in the task, however, are not lemmatised. Two approaches are used to extract the usages of search queries from Wikipedia:

**HDP-CLUSTERS-LEMMA** Search queries are lemmatised using Morpha (Minnen et al., 2001), and both the original and lemmatised forms are used for extraction;[2]

**HDP-CLUSTERS-NOLEMMA** Search queries are not lemmatised and only their original forms are used for extraction.

[1] The Wikipedia dump was retrieved on November 28th 2009.

[2] Morpha requires the part-of-speech (POS) of a given word, which is determined by the majority POS aggregated over all of that word's occurrences in Wikipedia.

| | |
|---|---|
| **Search query** | *dogs* |
| **Context sentence** | Most breeds of *dogs* are at most a few hundred years old |
| **Bag-of-word features** | most, breeds, of, are, at, most, a, few, hundred, years, old |
| **Positional word features** | most_#-3, breeds_#-2, of_#-1, are_#1, at_#2, most_#3 |

Table 1: An example of topic model features.

| System | F1 | ARI | RI | JI | Avg. No. of Clusters | Avg. Cluster Size |
|---|---|---|---|---|---|---|
| HDP-CLUSTERS-LEMMA | **0.6830** | 0.2131 | **0.6522** | 0.3302 | 6.6300 | 11.0756 |
| HDP-CLUSTERS-NOLEMMA | 0.6803 | **0.2149** | 0.6486 | 0.3375 | 6.5400 | 11.6803 |
| TASK11.DULUTH.SYS1.PK2 | 0.5683 | 0.0574 | 0.5218 | 0.3179 | 2.5300 | 26.4533 |
| TASK11.DULUTH.SYS7.PK2 | 0.5878 | 0.0678 | 0.5204 | 0.3103 | 3.0100 | 25.1596 |
| TASK11.DULUTH.SYS9.PK2 | 0.5702 | 0.0259 | 0.5463 | 0.2224 | 3.3200 | 19.8400 |
| TASK11-SATTY-APPROACH1 | 0.6709 | 0.0719 | 0.5955 | 0.1505 | 9.9000 | 6.4631 |
| TASK11-UKP-WSI-WACKY-LLR | 0.5826 | 0.0253 | 0.5002 | **0.3394** | 3.6400 | 32.3434 |
| TASK11-UKP-WSI-WP-LLR2 | 0.5864 | 0.0377 | 0.5109 | 0.3177 | 4.1700 | 21.8702 |
| TASK11-UKP-WSI-WP-PMI | 0.6048 | 0.0364 | 0.5050 | 0.2932 | 5.8600 | 30.3098 |
| RAKESH | 0.3949 | 0.0811 | 0.5876 | 0.3052 | 9.0700 | 2.9441 |
| SINGLETON | 1.0000 | 0.0000 | 0.6009 | 0.0000 | 64.0000 | 1.0000 |
| ALLINONE | 0.5442 | 0.0000 | 0.3990 | 0.3990 | 1.0000 | 64.0000 |
| GOLD | 1.0000 | 0.9900 | 1.0000 | 1.0000 | 7.6900 | 11.5630 |

Table 2: Cluster quality results for all systems. The best result for each column is presented in boldface. SINGLETON and ALLINONE are baseline systems and GOLD is the theoretical upper-bound for the task.

## 3 Experiments and Results

Following Lau et al. (2012), we use the default parameters ($\gamma = 0.1$ and $\alpha_0 = 1.0$) for HDP.[3] For each search query, we apply HDP to induce the senses, and a distribution of senses is produced for each "document" in the model. As the snippets in the test dataset correspond to the documents in the model and evaluation is based on "hard" clusters of snippets, we assign a sense to each snippet based on the sense (= topic) which has the highest probability for that snippet.

The task requires participants to produce a ranked list of snippets for each induced sense, based on the relative fit between the snippet and the sense. We induce the ranking based on the sense probabilities assigned to the senses, such that snippets that have the highest probability of the induced sense are ranked highest, and snippets with lower sense probabilities are ranked lower.

Two classes of evaluation are used in the shared task:

1. cluster quality measures: Jaccard Index (JI), RandIndex (RI), Adjusted RandIndex (ARI) and F1;

2. diversification of search results: Subtopic Recall@K and Subtopic Precision@r.

Details of the evaluation measures are described in Navigli and Vannella (2013).

The idea behind the second form of evaluation (i.e. diversification of search results) is that search engine results should cluster the results based on senses (of the query term in the documents) given an ambiguous query. For example, if a user searches for *apple*, the search engine may return results related to both the computer brand sense and the fruit sense of *apple*. Given this assumption, the best WSI/WSD system is the one that can correctly identify the diversity of senses in the snippets.

[3] Our implementation can be accessed via `https://github.com/jhlau/hdp-wsi`.

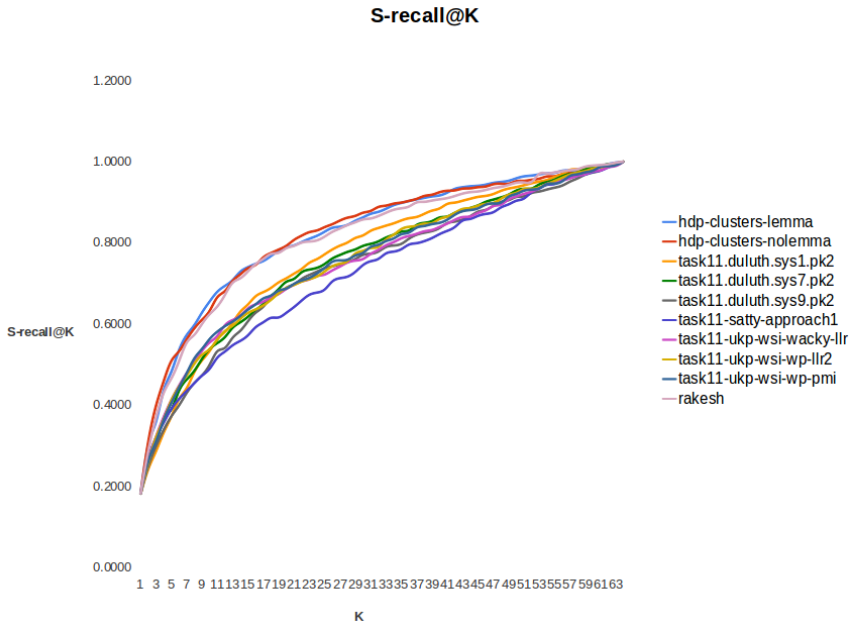


Figure 1: Subtopic Recall@K for all participating systems.

Cluster quality, subtopic recall@K and subtopic precision@r results for all systems entered in the task are presented in Table 2, Figure 1 and Figure 2, respectively.

In terms of cluster quality, our systems (HDP-CLUSTERS-LEMMA and HDP-CLUSTERS-NOLEMMA) consistently outperform the other teams for all measures except for the Jaccard Index (where we rank second and third, by a narrow margin). The average number of induced clusters and the average cluster size of our systems are similar to those of the gold standard system (GOLD), indicating that our systems are learning an appropriate sense granularity.

In terms of diversification of search results, our systems perform markedly better than most teams, other than RAKESH which trails closely behind our systems (despite a relatively low ranking in terms of the cluster quality evaluation). Overall, the results are encouraging and our system performs very well over the task.

## 4 Discussion and Conclusion

Our system adopts the WSI system proposed in Lau et al. (2012) with no parameters tuned for this task, and performs very well over it. Parameter tuning and exploiting URL information in the snippets could potentially boost the system performance further. Other background corpora (such as news articles) could also be used to increase the size of the training data. We leave these ideas for future work.

Inspecting the difference between the HDP-CLUSTERS-LEMMA and HDP-CLUSTERS-NOLEMMA approaches, only 6 out of the 100 lemmas have a lemmatised form which differs from the original query composition: *pods* (*pod*), *ten commandments* (*ten commandment*), *guild wars* (*guild war*), *stand by me* (*stand by i*), *sisters of mercy* (*sister of mercy*) and *lord of the flies* (*lord of the fly*). In most cases, including the lemmatised query results in the extraction of additional useful usages, e.g. using only the original form *lord of the flies* would extract no usages from Wikipedia (because this corpus has itself been lemmatised). In other cases, however, including the lemmatised forms results in many common noun usages, e.g. the number of usages of the lemmatised *pod* is significantly greater than that of the original form *pods* (which corresponds to proper noun usages in the lemmatised corpus), resulting in senses being induced only for common noun usages of *pods*. The

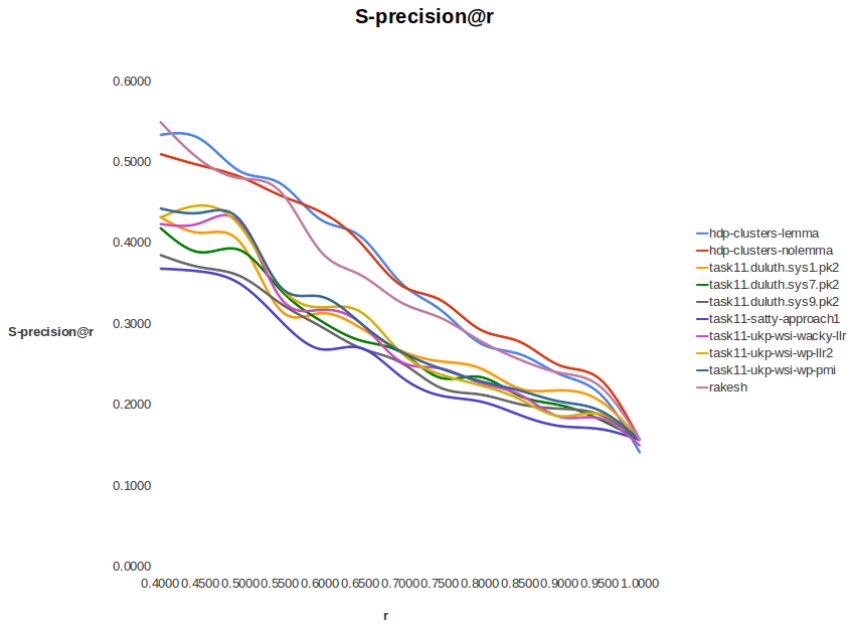


Figure 2: Subtopic Precision@r for all participating systems.

advantages and disadvantages of both approaches are reflected in the results: performance is mixed and no one method clearly outperforms the other.

To conclude, we apply a topic model-based WSI methodology to the task of web result clustering, using English Wikipedia as an external resource for extracting additional usages. Our system is completely unsupervised and requires no annotated resources, and appears to perform very well on the task.

## References


Eneko Agirre and Philip Edmonds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Springer, Dordrecht, Netherlands.

Eneko Agirre and Aitor Soroa. 2007. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proc. of the 4th International Workshop on Semantic Evaluations*, pages 7–12, Prague, Czech Republic.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proc. of the 13th Conference of the EACL (EACL 2012)*, pages 591–601, Avignon, France.

Jey Han Lau, Paul Cook, and Timothy Baldwin. to appear. unimelb: Topic modelling-based word sense induction. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

Roberto Navigli and Daniele Vannella. 2013. SemEval-2013 task 11: Evaluating word sense induction & disambiguation within an end-user application. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013)*, Atlanta, USA.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.