unimelb: Topic Modelling-based Word Sense Induction

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 shared task 13 “Word Sense Induction for Graded and Non-Graded Senses” of SemEval-2013. The task is on word sense induction (WSI), and builds on earlier SemEval WSI tasks in exploring the possibility of multiple senses being compatible to varying degrees with a single contextual instance: participants are asked to *grade* senses rather than selecting a single sense like most word sense disambiguation (WSD) settings. The evaluation measures are designed to assess how well a system perceives the different senses in a contextual instance. We adopt a previously-proposed WSI methodology for the task, which is based on a Hierarchical Dirichlet Process (HDP), a nonparametric topic model. Our system requires no parameter tuning, uses the English ukWaC as an external resource, and achieves encouraging results over the shared task.

1 Introduction

In our previous work (Lau et al., 2012) we developed a word-sense induction (WSI) system based on topic modelling, specifically a Hierarchical Dirichlet Process (Teh et al., 2006). In evaluations over the SemEval-2007 and SemEval-2010 WSI tasks we achieved performance on par with the current state-of-the art. The SemEval-2007 and SemEval-2010 WSI tasks assumed that each usage of a word has a single gold-standard sense. In this paper we apply this WSI method “off-the-shelf”, with no adaptation, to the novel SemEval-2013 task of “Word Sense Induction for Graded and Non-Graded Senses”. Given that the topic model allocates a multinomial distribution over topics to each word usage (“document”, in topic modelling terms), the SemEval-2013 WSI task is an ideal means for evaluating this aspect of the topic model.

2 System Description

Our system is based on the WSI methodology proposed by Lau et al. (2012), and also applied to SemEval-2013 Task 11 on WSI for web snippet clustering (Lau et al., to appear). 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, a distribution for each document. The generative process of LDA 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.

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 307–311, Atlanta, Georgia, June 14-15, 2013.

The number of topics, T, is a parameter in LDA. We relax this assumption by extending the model to be non-parametric, using a Hierarchical Dirichlet Process (HDP: (Teh et al., 2006)). HDP learns the number of topics based on data, with the concentration parameters γ and α_0 controlling the variability of topics in the documents (for details of HDP please refer to the original paper).

To apply HDP to WSI, the latent topics are interpreted as the word senses, and the documents are usages that contain the target word of interest. That is, given a target word (e.g. *paper*), a "document" in our application is a sentence context surrounding the target word. In addition to the bag of words surrounding the target word, we also include positional context word information, which was used in our earlier work (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 is given in Table 1.

2.1 Background Corpus and Preprocessing

The test dataset provides us with contextual instances for each target word, and these instances constitute the documents for the topic model. The text of the test data is tokenised and lemmatised using OpenNLP and Morpha (Minnen et al., 2001).

Note, however, that there are only 100 instances for most target words in the test dataset, and as such the dataset may be too small for the topic model to induce meaningful senses. To this end, we turn to the English ukWaC — a web corpus of approximately 1.9 billion tokens — to expand the data, by extracting context sentences that contain the target word. Each extracted usage is a three-sentence context containing the target word: the original sentence that contains the actual usage and its preceding and succeeding sentences. The extraction of usages from the ukWaC significantly increases the amount of information for the topic model to learn the senses for the target words from. However, HDP is computationally intensive, so we limit the number of extracted usages from the ukWaC using two sampling approaches:

UNIMELB (5P) Take a 5% random sample of usages;

UNIMELB (50K) Limit the maximum number of randomly-sampled usages to 50,000 instances.

The usages from the ukWaC are tokenised and lemmatised using TreeTagger (Schmid, 1994), as provided by the corpus.

To summarise, for each target word we apply the HDP model to the combined collection of the test instances (provided by the shared task) and the extracted usages from the English ukWaC (noting that each instance/usage corresponds to a topic model "document"). The topic model learns the senses/topics for all documents in the collection, but we only use the sense/topic distribution for the test instances as they are the ones evaluated in the shared 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.[1] For each target word, we apply HDP to induce the senses, and a distribution of senses is produced for each "document" in the model. To grade the senses for the instances in the test dataset, we apply the sense probabilities learnt by the topic model as the sense weights without any modification.

To illustrate the senses induced by our model, we present the top-10 words of the induced senses for the verb *strike* in Table 2. Although 13 senses in total are induced and some of them do not seem very coherent, only the first 8 senses — the more coherent ones — are observed (i.e., have non-zero probability for any usage) in the test dataset.

Two forms of evaluation are used in the task: WSD evaluation and clustering comparison. For WSD evaluation, three measures are used: (1) Jaccard Index (JI), which measures the degree of overlap between the induced senses and the gold senses; (2) positionally-weighted Kendall's tau (KT: (Kumar and Vassilvitskii, 2010)), which measures the correlation between the ranking of the induced senses and that of the gold senses; and (3) normalised discounted cumulative gain (NDCG), which

[1] These settings were considered "vague" priors in Teh et al. (2006). They were tested in Lau et al. (2012) and the model was shown to be robust under different parameter settings. As such we decided to keep the settings. The implementation of our WSI system can be accessed via GitHub: `https://github.com/jhlau/hdp-wsi`.

Target word	*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 the topic model features.

Sense Num	Top-10 Terms
1	strike @card@ worker union war iraq week pay government action
2	strike hand god head n't look face fall leave blow
3	strike @card@ balance court company case need balance_#1 order claim
4	strike ball @card@ minute game goal play player shot half
5	strike @card@ people fire disaster area road car ship lightning
6	@card@ strike new news post deal april home business week
7	strike n't people thing think way life book find new
8	@card@ strike coin die john church police age house william
9	div ukl syn color hunter text-decoration australian verb condom font-size
10	invent rocamadour cost mp3 terminal total wav honor omen node
11	training run rush kata performance marathon exercise technique workout interval
12	wrong qha september/2000 sayd — hawksmoor thyna pan salt common
13	zidane offering stone blow zidane_#-1 type type_#2 zidane_#1 blow_#3 materials

Table 2: The top-10 terms for each of the senses induced for the verb *strike* by the HDP model.

measures the correlation between the weights of the induced senses and that of the gold senses. For clustering comparison, fuzzy normalised mutual information (FNMI) and fuzzy b-cubed (FBC) are used. Note that the WSD systems participating in this shared task are not evaluated with clustering comparison metrics, as they do not induce senses/clusters in the same manner as WSI systems.

WSI systems produce senses that are different to the gold standard sense inventory (WordNet 3.1), and the induced senses are mapped to the gold standard senses using the 80/20 validation setting. Details of this mapping procedure are described in Jurgens (2012).

Results for all test instances are presented in Table 3. Note that many baselines are used, only some of which we present in this paper, namely: (1) Random — label instances with one of three random induced senses; (2) Semcor MFS — label instances with the most frequently occurring sense in Semcor; (3) Test MFS — label instances with the most frequently occurring sense in the test dataset. To benchmark our method, we present one or two of the best systems from each team.

Looking at Table 3, our system performs encouragingly well. Although not the best system, we achieve results close to the best system for each evaluation measure.

Most of the instances in the data were annotated with only one sense; only 11% were annotated with two senses, and 0.5% with three. As a result, the task organisers categorised the instances into single-sense instances and multi-sense instances to better analyse the performance of participating systems. Results for single-sense and multi-sense instances are presented in Table 4 and Table 5, respectively. Note that for single-sense instances, only precision is used for WSD evaluation as the Jaccard Index, positionally-weighted Kendall's tau and normalised discounted cumulative gain are not applicable. Our system performs relatively well, and trails marginally behind the best system in most cases.

4 Conclusion

We adopt a WSI methodology from Lau et al. (2012) for the task of grading senses in a WSD setting.

System	JI	KT	NDCG	FNMI	FBC
Random	0.244	0.633	0.287	0.018	0.382
SemCor MFS	0.455	0.465	0.339	—	—
Test MFS	0.552	0.560	0.412	—	—
AI-KU	0.197	0.620	**0.387**	**0.065**	0.390
AI-KU (remove5-ad1000)	**0.244**	**0.642**	0.332	0.039	0.451
La Sapienza (2)	0.149	0.510	0.383	—	—
UoS (top-3)	0.232	0.625	0.374	0.045	0.448
Unimelb (5p)	0.218	0.614	0.365	0.056	0.459
Unimelb (50k)	0.213	0.620	0.371	0.060	**0.483**

Table 3: Results for all instances. The best-performing system is indicated in boldface.

System	Precision	FNMI	FBC
Random	0.555	0.010	0.359
SemCor MFS	0.477	—	—
Test MFS	0.578	—	—
AI-KU	**0.641**	**0.045**	0.351
AI-KU (remove5-ad1000)	0.628	0.026	0.421
UoS (top-3)	0.600	0.028	0.414
Unimelb (5p)	0.596	0.035	0.421
Unimelb (50k)	0.605	0.039	**0.441**

Table 4: Results for single-sense instances. The best-performing system is indicated in boldface.

System	JI	KT	NDCG	FNMI	FBC
Random	0.429	0.548	0.236	0.006	0.113
SemCor MFS	0.283	0.373	0.197	—	—
Test MFS	0.354	0.426	0.248	—	—
AI-KU	0.394	0.617	0.317	0.029	0.078
AI-KU (remove5-ad1000)	**0.434**	0.586	0.291	0.004	0.116
La Sapienza (2)	0.263	0.531	**0.365**	—	—
UoS (#WN senses)	0.387	**0.628**	0.314	**0.036**	0.037
Unimelb (5p)	0.426	0.586	0.287	0.019	0.130
Unimelb (50k)	0.414	0.602	0.299	0.021	**0.134**

Table 5: Results for multi-sense instances. The best-performing system is indicated in boldface.

With no parameter tuning and using only the English ukWaC as an external resource, our system performs relatively well at the task.

References

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

David Jurgens. 2012. An evaluation of graded sense disambiguation using word sense induction. In *Proc. of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 189–198, Montréal, Canada.

Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized distances between rankings. In *Proc. of the 19th International Conference on the World Wide Web (WWW 2010)*, pages 571–580, Raleigh, USA.

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 for web snippet clustering. In *Proc. of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*.

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

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of the Conference on New Methods in Natural Language Processing*, Manchester, 1994.

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.