Mixing in Some Knowledge: Enriched Context Patterns for Bayesian Word Sense Induction

Rachel Chasin
MIT CSAIL
Cambridge, MA
rchasin@mit.edu

Anna Rumshisky
Department of Computer Science
University of Massachusetts, Lowell, MA
arum@cs.uml.edu

Abstract

Bayesian topic models have recently been shown to perform well in word sense induction (WSI) tasks. Such models have almost exclusively used bag-of-words features, and failed to attain improvement by including other feature types. In this paper, we investigate the impact of integrating syntactic and knowledge-based features and show that both parametric and non-parametric models consistently benefit from additional feature types. We perform evaluation on the SemEval2010 WSI verb data and show statistically significant improvement in accuracy ($p < 0.001$) both over the bag-of-words baselines and over the best system that competed in the SemEval2010 WSI task.

1 Introduction

The resolution of lexical ambiguity in language is essential to true language understanding. It has been shown to improve the performance of such applications as statistical machine translation (Chan et al., 2007; Carpuat and Wu, 2007), and cross-language information retrieval and question answering (Resnik, 2006). Word sense induction (WSI) is the task of automatically grouping the target word's contexts of occurrence into clusters corresponding to different senses. Unlike word sense disambiguation (WSD), it does not rely on a pre-existing set of senses.

Much of the classic bottom-up WSI and thesaurus construction work – as well as many successful systems from the recent SemEval competitions – have explicitly avoided the use of existing knowledge sources, instead representing the disambiguating context using bag-of-words (BOW) or syntactic features (Schütze, 1998; Pantel and Lin, 2002; Dorow and Widdows, 2003; Pedersen, 2010; Kern et al., 2010).

This particularly concerns the attempts to integrate the information about semantic classes of words present in the sense-selecting contexts. Semantic roles (such as those found in PropBank (Palmer et al., 2005) or FrameNet (Ruppenhofer et al., 2006)) tend to generalize poorly across the vocabulary. Lexical ontologies (and WordNet (Fellbaum, 2010) in particular) are not always empirically grounded in language use and often do not represent the relevant semantic distinctions. Very often, some parts of the ontology are better suited for a particular disambiguation task than others. In this work, we assume that features based on such ontology segments would correlate well with other context features.

Consider, for example, the expression "to deny the visa". When choosing between two senses of 'deny' ('refuse to grant' vs. 'declare untrue'), we would like our lexical ontology to place 'visa' in the same subtree as approval, request, recognition, commendation, endorsement, etc. And indeed, WordNet places all of these, including 'visa', under the same node. However, their least common subsumer is 'message, content, subject matter, substance', which also subsumes 'statement', 'significance', etc., which would activate the other sense of 'deny'. In other words, the distinctions made at this level in the nominal hierarchy in WordNet would not

be useful in disambiguating the verb 'deny', unless our model can select the appropriate nodes of the subtree rooted at the synset 'message, content, subject matter, substance'. Our model should also infer the associations between such nodes and other context relevant features that select the sense 'refuse to grant' (such as the presence of ditransitive constructions, etc.)

In this paper, we use the topic modeling approach to identify ontology-derived features that can prove useful for sense induction. Bayesian approaches to sense induction have recently been shown to perform well in the WSI task. In particular, Brody and Lapata (2009) have adapted the Latent Dirichlet Allocation (LDA) generative topic model to WSI by treating each occurrence context of an ambiguous word as a document, and the derived topics as sense-selecting context patterns represented as collections of features. They applied their model to the SemEval2007 set of ambiguous nouns, beating the best-performing system in its WSI task. Yao and Van Durme (2011) used a non-parametric Bayesian model, the Hierarchical Dirichlet Process (HDP), for the same task and showed that following the same basic assumptions, it performs comparably, with the advantage of avoiding the extra tuning for the number of senses.

We investigate the question of how well such models would perform when some knowledge of syntactic structure and semantics is added into the system, in particular, when bag-of-words features are supplemented by the knowledge-enriched syntactic features. We use the SemEval2010 WSI task data for the verbs for evaluation (Manandhar et al., 2010). This data set choice is motivated by the fact that (1) for verbs, sense-selecting context patterns often most directly depend on the nouns that occur in syntactic dependencies with them, and (2) the nominal parts of WordNet tend to have much cleaner ontological distinctions and property inheritance than, say, the verb synsets, where the subsumption hierarchy is organized according how specific the verb's manner of action is.

The choice of the SemEval2010 verb data set was motivated by the fact that SemEval2007 verb data is dominated by the most frequent sense for many target verbs, with 11 out of 65 verbs only having one sense in the combined test and training data. All verbs in the SemEval2010 verb data set have at least two senses in the data provided. The implications of this work are two-fold: (1) we confirm independently on a different data set that parametric and non-parametric models perform comparably, and outperform the current state-of-the-art methods using the baseline bag-of-words feature set (2) we show that integrating populated syntactic and ontology-based features directly into the generative model consistently leads to statistically significant improvement in accuracy. Our system outperforms both the bag-of-words baselines and the best-performing system in the SemEval2010 competition.

The remainder of the paper is organized as follows. In Section 2, we review the relevant related work. Sections 3 and 4 give the details on how the models are defined and trained, and describe the incorporated feature classes. Section 5 describes the data used to conduct the experiments. Finally, in Section 6, we describe the evaluation methods and present and discuss the experimental results.

2 Related Work

Over the past twenty years, a number of unsupervised methods for word sense induction have been developed, both for clustering contexts and for clustering word senses based on their distributional similarity (Hindle, 1990; Pereira et al., 1993; Schütze, 1998; Grefenstette, 1994; Lin, 1998; Pantel and Lin, 2002; Dorow and Widdows, 2003; Agirre et al., 2006).

One of the recent evaluations of the state of the art in word sense induction was conducted at SemEval2010 (Manandhar et al., 2010). The participant systems focused on a variety of WSI improvements including feature selection/dimensionality reduction techniques (Pedersen, 2010), experiments with bigram and cooccurrence features (Pedersen, 2010) and syntactic features (Kern et al., 2010), and increased scalability (Jurgens and Stevens, 2010).

Following the success of topic modeling in information retrieval, Boyd-Graber et al. (2007) developed an extension of the LDA model for word sense disambiguation that used WordNet walks to generate sense assignments for lexical items. Their model treated synset paths as hidden variables, with the as-

sumption that words within the same topic will share synset paths within WordNet, i.e. each topic will be associated with walks that prefer different "neighborhoods"of WordNet. One problem with their approach is that it relies fully on the integrity of WordNet's organization, and has no way to disprefer certain segments of WordNet, nor the ability to reorganize or redefine the senses it identifies for a given lexical item.

Brody and Lapata (2009) have proposed another adaptation of the LDA generative topic model to the WSI task. Text segments that contain instances of the target word are treated as documents in the classical IR setup for the LDA. The target word's senses are then similar to the hidden topics and are associated with a probability distribution over context features.

LDA assumes that each instance has been produced by a process that generates each of its context features by picking a sense of the target word from a known set of senses and then picking a feature for the context based on a sense-specific underlying probability distribution over context features. Importantly, the same prior distribution is assumed for all the features of an instance. However for many feature classes, for example, words vs. part-of-speech tags, this is false. Thus these algorithms do not immediately adapt well to being given features from many classes.

Brody and Lapata (2009) used part-of-speech and word n-grams as well as syntactic dependencies in addition to bag-of-words features, and used a multi-layer LDA model to handle the different classes separately in different "layers", bringing them together when necessary in a weighted combination. Their best model, however, showed very similar performance to the LDA model using only bag-of-words features. Yao and Van Durme (2011) reproduced some of their LDA experiments using HDP, a non-parametric model that induces the number of topics from data, over bag-of-words context representation.

3 Methods

We applied the LDA model (Brody and Lapata, 2009) and the the HDP model (Yao and Durme, 2011) over a set of features that included populated syntactic dependencies as well as knowledge-enriched syntactic features. Note that unlike the model proposed by Boyd et al (2007), which relies fully on the on the pre-existing sense structure reflected in WordNet, under this setup, we will only incorporate the relevant information from the ontology, while allowing the senses themselves to be derived empirically from the distributional context patterns. The assumption here is that if any semantic features prove relevant for a particular target word, i.e. if they correlate well with other features characterizing the word's context patterns, they will be strongly associated with the corresponding topic.

In reality, the topics modeled by LDA and HDP may not correspond directly to senses, but may represent some subsense or supersense. In fact, the induced topics are more likely to correspond to the sense-selecting patterns, rather than the senses per se, and quite frequently the same sense may be expressed with multiple patterns. We describe how we deal with this in Section 6.1.

3.1 Model Description

The LDA model is more formally defined as follows: Consider one target word with M instances and K senses, and let the context of instance j be described by some set of N_j features from a vocabulary of size V. These may be the words around the target or could be any properties of the instance. LDA assumes that there are M probability distributions $\theta_{\mathbf{j}} = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jK})$, with θ_{jk} = the probability of generating sense k for instance j, and K probability distributions $\phi_{\mathbf{k}} = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{kV})$, with ϕ_{kf} = the probability of generating feature f from sense k. This makes the probability of generating the corpus where the features for instance j are $f_{j1}, f_{j2}, \ldots, f_{jN_j}$:

$$P(\text{corpus}) = \prod_{j=1}^{M} \prod_{i=1}^{N_J} \sum_{k=1}^{K} \theta_{jk} \phi_{k f_{ji}}$$

The goal of LDA for WSI is to obtain the distribution $\theta_{\mathbf{j}^*}$ for an instance j^* of interest, as this gives each sense's probability of being picked to generate some feature in the instance, which corresponds to the probability of being the correct sense for the target word in this context.

The corpus generation process for HDP is similar to that of LDA, but obtains the document-specific

sense distribution (corresponding to LDA's θ_j) via a Dirichlet Process whose base distribution is determined via another Dirichlet Process, allowing for an unfixed number of senses because the draws from the resulting sense distribution are not limited to a preset range. The concentration parameters of both Dirichlet Processes are determined via hyperparameters.

3.2 Model Training

LDA

Our process for training an LDA model uses Gibbs sampling to assign topics to each feature in each instance, utilizing GibbsLDA++ (Phan and Nguyen, 2007). Initially topics are assigned randomly and during each subsequent iteration, assignments are made by sampling from the probability distributions resulting from the last iteration. Following the previous work in applying topic-modeling to WSI, we use hyperparameters $\alpha = 0.02, \beta = 0.1$ (Brody and Lapata, 2009). We train the model using 2000 iterations of Gibbs sampling (GibbsLDA++ default). To obtain θ for an instance of interest, the inference mode initializes the training corpus with the assignments from the model and initializes new test documents with random assignments. We then run 20 iterations of Gibbs sampling on this augmented corpus. 5 models are trained for each target using the same parameters and data. This is done to reduce the effect of randomization in the training algorithms on our results. Although the randomization is also present in the inference algorithms and we do not perform more than one inference run per model.

HDP

The HDP training and inference procedures are similar to LDA, but using Gibbs sampling on topic and table assignment in a Chinese Restaurant Process. We use Chong Wang's program for HDP (Wang and Blei, 2012) running the Gibbs sampling for 1000 iterations during training and another 1000 during inference (the defaults), and using the hyperparameters suggested in previous work (Yao and Durme, 2011) of $H = 0.1, \alpha_0 \sim Gamma(0.1, 0.028), \gamma \sim Gamma(1, 0.1)$.

This software does not directly produce θ values but instead produces all assignments of words to topics. This output is used to compute

$$\theta_{jk} = \frac{\text{count(words in document j labeled k)}}{\text{count(words in document j)}}.$$

Since new topics can appear during inference, we smooth these probabilities with additive smoothing using a parameter of 0.02 to avoid the case where all words are labeled with unseen topics, which would make prediction of a sense using our evaluation methods impossible.

4 Features

We used three types of features: bag-of-words with different window sizes, populated syntactic features, and ontology-populated syntactic features. Instead of using a multi-layered LDA model, we attempt to mitigate the effects of using multiple classes of features by choosing extra features whose distributions are sufficiently similar to the bag-of-words features. We describe these classes in more detail below.

Preprocessing done on the data includes: (1) tokenization, (2) identifying stopwords, (3) stemming tokens, (4) detecting sentence boundaries, (5) tagging tokens with their parts of speech, and (6) obtaining collapsed dependencies within sentences including the target words. For tokenization, sentence boundary detection, and part-of-speech tagging, we use OpenNLP (OpenSource, 2010). We remove the stop words and stem using the Snowball stemmer. For collapsed syntactic dependencies we use the Stanford Dependency Parser (Klein and Manning, 2003).

Bag of Words Following previous literature (Brody and Lapata, 2009), we use a 20 word window (excluding stopwords) for BOW features. In our experiments, a smaller window size failed to produce better performance.

Ontology-Based Populated Syntactic Features To capture syntactic information, we use populated dependency relations. We populate these relations with semantic information from WordNet (Miller et al., 1990) as follows. For each syntactic dependency between the target word and the context word, we locate all synsets for the context word. We then traverse the WordNet hierarchy upwards from each of these synsets, and include a feature for each node

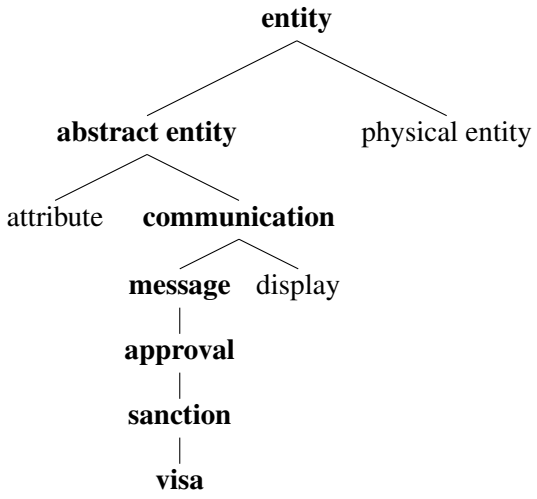

Figure 1: WordNet hierarchy path for "visa".

we visit. We use collapsed relations produced by the Stanford Dependency Parser (Klein and Manning, 2003).

For example, consider the path up the hierarchy for the word "visa", given in Figure 1. If the noun "visa" is found in direct object position of the target verb, traversing the tree to the root would produce features such as *noun-approval_dobj*, etc.

5 Data

We evaluate our methods on the 50 verb targets from the SemEval2010 dataset. The evaluation data is split into 5 mapping/test set pairs, with 60% for mapping (2179 instances) and 40% for testing (1451 instances). Each split is created randomly and independently each time, and 3354 out of 3630 instances appear in a test set at least once.

We train our topic models on unlabeled data from SemEval2010, which contains a total of 162,862 instances for all verbs. The targets "happen" and "regain" have the most and fewest instances with 11,286 and 266 respectively. We use this data to train our topic models. We limit each target to 50,000 instances for training HDP models, in order to maintain reasonable processing time.

6 Results

We show the comparisons of our systems with (1) the most-frequent-sense (MFS) (MFS in the mapping set predicted for all instances in the test set), (2) BOW baseline models, and (3) the best-performing system from SemEval2010. Since HDP performs better overall, we chose the HDP model to experiment with syntactic and ontological features. For completeness, we include results for the WordNet-populated syntactic features with the LDA model.

6.1 Evaluation Measures

Following the established practice in SemEval competitions and subsequent work (Agirre and Soroa, 2007; Manandhar et al., 2010; Brody and Lapata, 2009; Yao and Durme, 2011), we conduct supervised evaluation. A small amount of labeled data is used to map the induced topics to real-world senses; for a description of the method see (Agirre and Soroa, 2007). The resulting mapping is probabilistic; for topics $1, \ldots, K$ and senses $1, \ldots, S$, we compute the KS values

$$P(s|k) = \frac{\text{count(instances predicted k, labeled s)}}{\text{count(instances predicted k)}}.$$

Then given θ_{j^*}, we can make a better prediction for instance j^* than just assigning the most likely sense to its most likely topic. Instead, we compute

$$\text{argmax}_{s=1}^{S} \sum_{k=1}^{K} \theta_{j^*k} P(s|k),$$

the sense with the highest probability of being correct for this instance, given the topic probabilities and the KS mapping probabilities.

The supervised metrics traditionally reported include precision, recall, and F-score, but since our WSI system makes a prediction for every instance, we report accuracy throughout this section.

6.2 Cross-Validation

We use cross-validation on the mapping set to select the best system configuration. We use leave-one-out or 50-fold cross-validation, whichever has fewer folds for a given target word. The system configurations that we compare vary with respect to the following: (1) topic modeling algorithm (HDP or LDA), (2) included feature classes (bag-of-words with different window sizes, populated syntactic features, ontology-populated syntactic features), and (3) number of topics (i.e. senses) for the LDA model. The best configuration is then tested on the

Configuration	CV acc.
HDP, 20w +WN1h	**72.5%**
HDP, 20w +WN1h-limited	70.8%
HDP, 20w +Synt	71.3%
HDP, 20w (baseline)	**69.7%**
LDA, 5 senses, 20w +WN1h	71.2%
LDA, 5 senses, 20w	71.2%
LDA, 12 senses, 20w +WN1h	72.2%
LDA, 12 senses, 20w	70.2%

Table 1: Cross-validation accuracies using the SemEval2010 mapping sets.

evaluation data. Table 1 shows cross-validation results for some of the relevant configurations on the SemEval2010 dataset.

Since the evaluation data has 5 different mapping sets, one for each 60/40 split, we do cross-validation on each and average the results. We perform this process for each of our 5 trained models and again average the results.

The best HDP configuration outperforms the LDA configurations with low numbers of topics. This configuration combines the 20 closest non-stopwords bag-of-words (20w) with WordNet-populated syntactic dependencies (+WN1h) and achieves 72.5% accuracy. We evaluate two other configurations using HDP as well: 20w +WN1h-limited, which is 20w +WN1h minus those features from WordNet within 5 hops of the hierarchy's root; and 20w +Synt, which is the 20 closest non-stopwords bag-of-words plus syntactic dependencies 1 hop away from the target word populated with the stemmed token appearing there. As shown in Table 1, WordNet-based populated features do introduce some gain with respect to the syntactic features populated only at the word level. Interestingly, removing the top-level WordNet-based features, and therefore making the possible restrictions on the semantics of the dependent nouns more specific, does not lead to performance improvement.

Each topic produced by the model is a distribution over all feature types, and is comprised by a mix of bag-of-words and ontology-populated syntactic features. Each node on the path from a given synset to the root generates its own ontological feature, so when many nodes that activate the same sense have a common hypernym, that hypernym is likely to "float to the top" - become more strongly associated with the corresponding topic.

To illustrate this, consider the following two senses of the verb 'cultivate': "prepare the soil for crops" and "teach or refine". Topic 1 generated by the HDP 20w +WN1h model corresponds to the first sense and is associated with examples about cultivating land, earth, grassland, waste areas. Topic 5 generated by the same model corresponds to the second sense and is associated with examples about cultivating knowledge, understanding, habits, etc. One of the top-scoring features for Topic 1 is *location_dobj* which corresponds to the direct object position being occupied by one of the 'location' synsets, with direct hyponym nodes for 'region' and 'space' contributing the most. For topic 5, *cognition_dobj* is selected as one of the top features, with direct hyponyms for 'ability', 'process', and 'information' contributing the most.

In this best configuration, HDP produces an average of 18.6 topics, far more than the number of real-world senses. We investigated the possibility that its improvement over LDA might be due to this larger number of topics, testing the same feature combination on LDA with 12 topics. This does produce a similar accuracy, 72.2%, and the simpler bag-of-words features with 12 topics yield an accuracy drop to 70.2%, similar to the drop seen between HDP 20w +WN1h and HDP 20w.

6.3 Evaluation Set Results

For the five SemEval2010 test sets, senses are assigned slightly differently than in cross-validation. Instead of averaging over five models trained per target, for each instance, we predict the sense assigned by the majority of these models.

Table 2 shows the comparison of the configuration with the best cross-validation accuracy (HDP, 20w +WN1h) against the following: (1) MSF baseline, (2) the baseline bag-of-words model (3) the results obtained on this data set by the best-performing SemEval2010 system using supervised evaluation, Duluth-Mix-Narrow-Gap from the University of Minnesota Duluth (Manandhar et al., 2010). The HDP model with knowledge-enriched features obtains the best accuracy of 73.3%. For comparison, we also show results for the LDA model with 12 topics that performed well in cross-validation.

System	Accuracy
MFS	66.7 %
HDP, 20w +WN1h	**73.3%**
HDP, 20w (baseline)	71.2%
LDA, 12 senses, 20w +WN1h	72.5%
LDA, 12 senses, 20w	71.1%
Duluth-Mix-Narrow-Gap	68.6%

Table 2: Test set accuracies, SemEval2010 verbs

The improvements obtained by the best configuration are statistically significant by paired two-tailed t-test, treating each of the 3354 distinct test instances as separate samples. We consider a system's prediction on one such instance to be the sense it predicted in the majority of the test sets in which the instance appears. Significance levels are as follows:

- The best HDP configuration (20w +WN1h) vs. Duluth-Mix-Narrow-Gap: $p < 0.0001$
- The best HDP configuration (20w +WN1h) vs. HDP 20w: $p < 0.001$
- 12-sense LDA configuration 20w +WN1h vs. Duluth-Mix-Narrow-Gap: $p < 0.0001$
- 12-sense LDA configuration 20w +WN1h vs. 12-sense LDA 20w: $p < 0.05$.

7 Conclusion

We have presented a system that uses an adaptation of two Bayesian topic modeling algorithms to the task of word sense induction. Both the parametric and the non-parametric versions, when enriched with WordNet-based populated syntactic features, outperform the baseline bag-of-words models as well as the current state of the art in the WSI task for verbs. The next step for this system is an improved integration of knowledge-based features that would not require assuming a similar distribution on different feature types.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Carpuat and D. Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proc. of EMNLP-CoNLL*, pages 61–72.

Y. S. Chan, H. T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proc. of ACL*, pages 33–40, Prague, Czech Republic, June.

B. Dorow and D. Widdows. 2003. Discovering corpus-specific word-senses. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages Conference Companion pp. 79–82, Budapest, Hungary, April.

Christiane Fellbaum. 2010. Wordnet. *Theory and Applications of Ontology: Computer Applications*, pages 231–243.

David Jurgens and Keith Stevens. 2010. Hermit: Flexible clustering for the semeval-2 wsi task. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 359–362, Uppsala, Sweden, July. Association for Computational Linguistics.

Roman Kern, Markus Muhr, and Michael Granitzer. 2010. Kcdc: Word sense induction by using grammatical dependencies and sentence phrase structure. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 351–354, Uppsala, Sweden, July. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

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(SemEval)*, pages 63–68, Uppsala, Sweden, July. Association for Computational Linguistics.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

OpenSource. 2010. Opennlp: $http$: $//opennlp.sourceforge.net/$.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD02*.

Ted Pedersen. 2010. Duluth-wsi: Senseclusters applied to the sense induction task of semeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 363–366, Uppsala, Sweden, July. Association for Computational Linguistics.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. Gibbslda++: A c/c++ implementation of latent dirichlet allocation (lda).

P. Resnik. 2006. Word sense disambiguation in NLP applications. In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Springer.

J. Ruppenhofer, M. Ellsworth, M. Petruck, C. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Chong Wang and David M. Blei. 2012. A Split-Merge MCMC Algorithm for the Hierarchical Dirichlet Process, January.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric bayesian word sense induction. In *Graph-based Methods for Natural Language Processing*, pages 10–14. The Association for Computer Linguistics.