Nonparametric Bayesian Word Sense Induction

Xuchen Yao[1] and **Benjamin Van Durme**[1,2]
[1]Department of Computer Science
[2]Human Language Technology Center of Excellence
Johns Hopkins University

Abstract

We propose the use of a nonparametric Bayesian model, the Hierarchical Dirichlet Process (HDP), for the task of Word Sense Induction. Results are shown through comparison against Latent Dirichlet Allocation (LDA), a parametric Bayesian model employed by Brody and Lapata (2009) for this task. We find that the two models achieve similar levels of induction quality, while the HDP confers the advantage of automatically inducing a variable number of senses per word, as compared to manually fixing the number of senses *a priori*, as in LDA. This flexibility allows for the model to adapt to terms with greater or lesser polysemy, when evidenced by corpus distributional statistics. When trained on out-of-domain data, experimental results confirm the model's ability to make use of a restricted set of topically coherent induced senses, when then applied in a restricted domain.

1 Introduction

Word Sense Induction (WSI) is the task of automatically discovering latent *senses* for each word *type*, across a collection of that word's *tokens* situated in context. WSI differs from Word Sense Disambiguation (WSD) in that the task does not assume access to some prespecified sense inventory. This amounts to a clustering task: instances of a word are partitioned into the same bin based on whether a system deems them to have the same underlying meaning. A large body of related work can be found in (Schütze, 1998; Pantel and Lin, 2002; Dorow and Widdows, 2003; Purandare and Pedersen, 2004; Bordag, 2006; Niu et al., 2007; Pedersen, 2007; Brody and Lapata, 2009; Li et al., 2010; Klapaftis and Manandhar, 2010).

Brody and Lapata (2009) (B&L herein) showed that the parametric Bayesian model, Latent Dirichlet Allocation (LDA), could be successfully employed for this task, as compared to previous results published for the WSI component of SemEval-2007[1] (Agirre and Soroa, 2007). A deficiency of the LDA model for WSI is that the number of senses needs to be manually specified *a priori*, either separately for each word type, or (as done by B&L) some fixed value that is shared globally across all types.

Nonparametric methods instead have the flexibility of automatically deciding the number of sense cluters (Vlachos et al., 2009; Reisinger and Mooney, 2010). In this work we first independently verify the results of B&L, and then tackle the limitation on fixing the number of senses through the use of the Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), a nonparametric Bayesian model. We show this approach leads to results of similar quality as LDA, when using a bag-of-words context model, in addition to allowing for variability in the number of senses across different words and domains. When trained on a restricted domain corpus for which manually labeled sense data was present, we verify that the model may be tuned to posit a similar number of senses as determined by human judges. When trained on a broader domain collection, we show that the number of induced senses increase, in line with the intuition that a wider set of genres should lead to a greater diversity in underlying meanings. Automatically inducing the proper number of senses has great practical implications, especially in areas that require word sense disambiguation. For instance, inducing more senses for `bank` helps to tell differ-

[1] Klapaftis and Manandhar (2010) and Brody and Lapata (2009) reported the best scores so far on this dataset.

Proceedings of the TextGraphs-6 Workshop, pages 10–14,
Portland, Oregon, USA, 19-24 June 2011.

ent word senses apart for naturally more ambiguous words, and inducing less senses for job helps to prevent assigning too fined-grained senses in case the same words in two similar contexts are mistakenly regarded as carrying different senses.

2 Bayesian Word Sense Induction

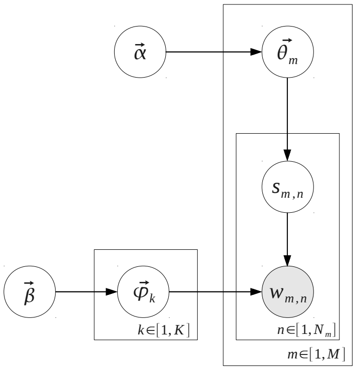

Figure 1: Latent Dirichlet Allocation (LDA) for WSI.

As in prior work including B&L, we rely on the intuition that the senses of words are hinted at by their contextual information (Yarowsky, 1992). From the perspective of a generative process, neighboring words of a target are generated by the target's underlying sense.[2]

Both LDA and HDP define graphical models that generate collections of discrete data. The sense of a target word is first drawn from a distribution and then the context of this word is generated according to that distribution. But while LDA assumes a fixed, finite set of distributions, the HDP draws from an infinite set of distributions generated by a Dirichlet Process. This section details the distinction.

Figure 1 shows the LDA model for word sense induction. The conventional notion of document is replaced by a *pseudo-document*, consisting of every word in an N_m-word window centered on the target item. $w_{m,n}$ is the n-th token of the m-th pseudo-document for target word w. $s_{m,n}$ is the corresponding sense for $w_{m,n}$. Suppose there are K senses for the target word w, then the distribution over a context word $w_{m,n}$ is:

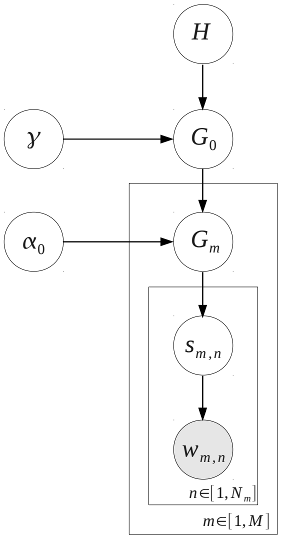

Figure 2: Hierarchical Dirichlet Process (HDP) for WSI.

$$p(w_{m,n}) = \sum_{k=1}^{K} p(w_{m,n} \mid s_{m,n} = k) p(s_{m,n} = k).$$

Let the word distribution given a sense be $p(w_{m,n} \mid s_{m,n} = k) = \vec{\varphi}_k$, which is a vector of length V (vocabulary size) that is generated from a Dirichlet distribution: $\vec{\varphi}_k \sim Dir(\vec{\beta})$. Let the sense distribution given a document be $p(s_{m,n} \mid d = m) = \vec{\theta}_m$, which is a vector of length K that is generated from a Dirichlet distribution: $\vec{\theta}_m \sim Dir(\vec{\alpha})$. The generative story for the data under LDA is then:

```
For $k \in (1, ..., K)$ senses:
  Sample mixture component: $\vec{\varphi}_k \sim Dir(\vec{\beta})$.
For $m \in (1, ..., M)$ pseudo-documents:
  Sample topic components $\vec{\theta}_m \sim Dir(\vec{\alpha})$.
  For $n \in (1, ..., N_m)$ words in pseudo-document $m$:
    Sample sense index $s_{m,n} \sim Mult(\vec{\theta}_m)$.
    Sample word $w_{m,n} \sim Mult(\vec{\varphi}_{s_{m,n}})$.
```

The sense distribution over a word is captured as K mixture components. In the HDP however, we assume the number of active components is unknown, and should be inferred from the data. For each pseudo-document, the sense component $s_{m,n}$ for word $w_{m,n}$ has a nonparametric prior G_m. G_m is nonparametric in the sense that for every new pseudo-document m, a new G_m is sampled from a base distribution G_0. As the corpus grows, there are

[2] For instance, given the word bank with a sense river bank, it is more likely that the neighboring words are river, lake and water than finance, money and loan.

more and more G_m's. However, the mixture component $s_{m,n}$, drawn from G_m, can be shared among pseudo-documents. Thus the number of senses do not simply multiply out as m grows. Both G_0 and G_m's are distributed according to a Dirichlet Process (DP) (Ferguson, 1973). The generative story is:

```
Select base distribution $G_0 \sim DP(\gamma, H)$ which
provides an unlimited inventory of senses.
For $m \in (1, ..., M)$ pseudo-documents:
  Draw $G_m \sim DP(\alpha_0, G_0)$.
For $n \in (1, ..., N_m)$ words in pseudo-document $m$:
  Sample $s_{m,n} \sim G_m$.
  Sample $w_{m,n} \sim Mult(\vec{\varphi}_{s_{m,n}})$.
```

Hyperparameters γ and α_0 are the concentration parameters of the DP, controlling the variability of the distributions G_0 and G_m. In a Chinese restaurant franchise metaphor of the HDP, multiple restaurants (documents) share a set of dishes (senses). Then γ controls the variability of the global sense distribution and α_0 controls the variability of each customer's (word) choice of dishes (senses).[3]

3 Experiment Setting

Model B&L experimented with variations to the LDA model that allowed for generating multiple layers of features, such as smaller (5w) and larger (10w) bag-of-word contexts, and syntactic features. The additional complexity beyond the standard model led to only tenuous performance gains. Normal LDA, when trained on pseudo-documents built from 10 words of surrounding context, performed only slightly below their best reported results.[4] Especially as our goal here was to investigate the sense-specification problem, rather than eking out further improvements in the base WSI evaluation measure, we chose to compare a standard LDA model to HDP, both strictly using a 10 word context.[5]

Test Data Following B&L, we perform WSI on nouns. The evaluation data comes from the WSI task of SemEval-2007 (Agirre and Soroa, 2007). It is derived from the Wall Street Journal portion of the Penn TreeBank (Marcus et al., 1994) and contains 15,852 instances of excerpts on 35 nouns. All the nouns are hand-annotated with their OntoNotes senses (Hovy et al., 2006), with an average of 3.9 senses per word.

Evaluation Method WSI is an unsupervised task that results in sense clusters with no explicit mapping to manually annotated sense data. To derive such a mapping, we follow the *supervised evaluation* strategy of Agirre and Soroa (2007). Annotated senses from SemEval-2007 are partitioned into a standard mapping set (72%), a dev set (14%) and a test set (14%). After an WSI system has tagged the elements in the mapping set with their "cluster IDs", then a cluster to sense derivation is constructed by simply assigning to each cluster the manual sense label that has the highest in-cluster frequency. Once such a mapping has been established, then results on the dev or test set are reported based on treating cluster assignment as a WSD operation.

Training Data As out-of-domain source, we extracted 930K instances of the 35 nouns from the British National Corpus (BNC) (Clear, 1993). As in-domain source we extracted another 930K instances from WSJ in years 87/88/90/94. All pseudo-documents use the ± 10 contextual window.

4 Evaluation

We trained the LDA and HDP models on the WSJ and BNC datasets separately. In their experiments with LDA, B&L iteratively tried 3 up to 9 senses, and then reported the number that led to best results in evaluation (4 senses for WSJ, 8 for BNC). We repeated this approach for LDA, with hyper-parameters $\alpha = 0.02$ and $\beta = 0.1$. For the HDP model, we tuned hyper-parameters on the SemEval-2007 dev set.[6] See Table 1 for results, averaged over 5 runs of LDA and 3 runs of HDP.

We report several findings based on this experiment. First, for the LDA models trained on WSJ and BNC, our F1 measures are 0.8% lower than reported by B&L.[7] Second, based on our own experiment, the HDP model performance is slightly better than that of LDA when training with BNC.

[3] Gibbs sampling (Geman and Geman, 1990) can be applied for inference. Specifically, Teh et al. (2006) describes the posterior sampling in the Chinese restaurant franchise.

[4] F-score of 86.9% (10w), as compared to 87.3% (10w+5w).

[5] We relied on implementations of LDA and HDP respectively from MALLET (McCallum, 2002), and Wang (2010).

[6] Final parameters: $H = 0.1$, $\alpha_0 \sim Gamma(0.1, 0.028)$, $\gamma \sim Gamma(1, 0.1)$.

[7] We consider this acceptable experimental deviation, given the minor variation in respective training data.

WSJ		BNC	
LDA-4s*	86.9	LDA-8s*	84.6
LDA-4s	86.1	LDA-8s	83.8
HDP	86.7	HDP	$85.7^{\triangle}$

Table 1: F-measure when training with WSJ (in-domain) and BNC (out-of-domain). Results with * are taken from B&L. **4** or **8** senses were used per word. $\triangle$: statistically significant against LDA-8s by paired permutation test with $p < 0.001$. The standard baseline, always picking the most frequent sense observed in training, scores 80.9.

	WSJ		BNC	
	Train	Test	Train	Test
LDA	4.0	3.9	8.0	7.4
HDP	5.8	3.9	9.4	4.6

Table 2: The average number of senses the LDA and HDP models output when training with WSJ/BNC and testing on SemEval-2007, which has 3.9 senses per word on average.

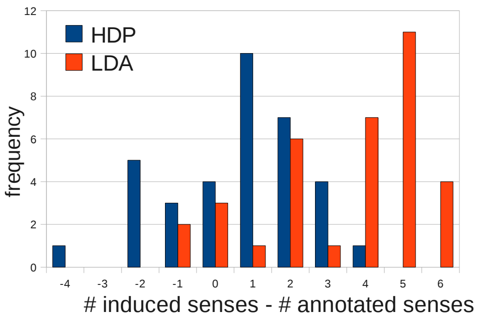

Figure 3: The difference between induced number of senses and annotated senses. The training set is BNC. The test set is SemEval-2007, containing 35 nouns with 3.9 senses. LDA induced 7.4 senses and HDP induced 4.6 senses on average.

WSJ		BNC	
LDA-5.8s	86.0	LDA-9.4s	82.7
LDA-3.9s	85.3	LDA-3.9s	81.4
HDP-5.8s	86.7	HDP-9.4s	$85.7^{\triangle}$

Table 3: F1 measure when training LDA with three other settings: 5.8s, 9.4s and 3.9s. $\triangle$: statistically significant against both LDA-9.4s and LDA-3.9s (for BNC) by paired permutation test with $p < 0.001$.

Third, the HDP model appears to better adapt to data in other domains. When switching the training set from WSJ (in-domain) to BNC (out-of-domain), we, along with B&L, found a 2.3% drop with LDA models. However, with the HDP model, there is only a 1% drop in F1. Moreover, even trained on out-of-domain data, HDP can still better infer the number of senses from the test data, which is illustrated next.

Table 2 shows the number of senses induced from each dataset. When training on WSJ and test on SemEval-2007, HDP induced the correct number of senses (3.9 on average) from test, while LDA did this by assuming 4 senses from the training data. When there is a domain mismatch between training (BNC) and test (SemEval-2007, which comes from the 1989 WSJ), the LDA model preferred far more than the annotated number of senses (7.4 vs. 3.9), largely due to the fact that it assumed 8 senses during training. However, even though the HDP model induced more senses (9.4) when training on the broader coverage BNC set, it still inferred a much reduced average of 4.6 senses on test.

The BNC, being a *balanced corpus*, covers more diverse genres than the WSJ: we would expect it to lead to a more inclusive model of word sense. Figure 3 illustrates this comparison through the difference between sense numbers. For the 35 human-annotated nouns, HDP induced the number of senses mostly within an error of ±2, whereas LDA tended to prefer $3 - 6$ more senses than recognized by annotators (on average the HDP model was off by 1.6 senses, as compared to 3.6 by LDA). Finally, the F1 performance of HDP is 1.9% better than LDA (85.7% vs. 83.8%).

We further evaluated the LDA model by training separately for each of the 35 nouns, first setting as the number of topics the amount induced by HDP (on average, 5.8/9.4 senses for WSJ/BNC), then using the number of senses as used by the human annotators in SemEval-2007 (an average of 3.8). As seen in Table 3, in each of these cases HDP remained the superior model.

5 Conclusion

We proposed the use of a nonparametric Bayesian model (HDP) for word sense induction and compared it with the parametric model by Brody and Lapata (2009), based on LDA. The HDP model confers the advantage of automatically identifying the number of senses, besides having equivalent (or better) performance than the LDA model, verified using the SemEval-2007 dataset. Future work includes large scale sense induction over a larger vocabulary, in tasks such as Paraphrase Acquisition.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 Task 02: Evaluating Word Sense Induction And Discrimination Systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 7–12.

Stefan Bordag. 2006. Word Sense Induction: Triplet-Based Clustering And Automatic Evaluation. In *Proceedings of the 11th EACL*, pages 137–144.

Samuel Brody and Mirella Lapata. 2009. Bayesian Word Sense Induction. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111.

Jeremy H. Clear, 1993. *The British national corpus*, pages 163–187. MIT Press, Cambridge, MA, USA.

Beate Dorow and Dominic Widdows. 2003. Discovering Corpus-Specific Word Senses. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, EACL '03, pages 79–82.

T. S. Ferguson. 1973. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

S. Geman and D. Geman, 1990. *Stochastic Relaxation, Gibbs Distributions, And The Bayesian Restoration Of Images*, pages 452–472.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 57–60.

Ioannis Klapaftis and Suresh Manandhar. 2010. Word Sense Induction & Disambiguation Using Hierarchical Random Graphs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 745–755, October.

Linlin Li, Benjamin Roth, and Caroline Sporleder. 2010. Topic Models For Word Sense Disambiguation And Token-Based Idiom Detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1138–1147.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

Zheng-Yu Niu, Dong-Hong Ji, and Chew-Lim Tan. 2007. I2R: Three Systems For Word Sense Discrimination, Chinese Word Sense Disambiguation, And English Word Sense Disambiguation. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 177–182.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses From Text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 613–619.

Ted Pedersen. 2007. UMND2: SenseClusters applied to the sense induction task of Senseval-4. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 394–397.

Amruta Purandare and Ted Pedersen. 2004. Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In *Proceedings of CoNLL-2004*, pages 41–48.

Joseph Reisinger and Raymond J. Mooney. 2010. A Mixture Model with Sharing for Lexical Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2010)*, pages 1173–1182, MIT, Massachusetts, USA.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Comput. Linguist.*, 24:97–123, March.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 74–82.

Chong Wang. 2010. An implementation of hierarchical dirichlet process (HDP) with split-merge operations.

David Yarowsky. 1992. Word-Sense Disambiguation Using Statistical Models Of Roget's Categories Trained On Large Corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 454–460.