A Topic Model for Word Sense Disambiguation

Jordan Boyd-Graber
Computer Science
Princeton University
Princeton, NJ 08540
jbg@princeton.edu

David Blei
Computer Science
Princeton University
Princeton, NJ 08540
blei@cs.princeton.edu

Xiaojin Zhu
Computer Science
University of Wisconsin
Madison, WI 53706
jerryzhu@cs.wisc.edu

Abstract

We develop latent Dirichlet allocation with WORDNET (LDAWN), an unsupervised probabilistic topic model that includes word sense as a hidden variable. We develop a probabilistic posterior inference algorithm for simultaneously disambiguating a corpus and learning the domains in which to consider each word. Using the WORDNET hierarchy, we embed the construction of Abney and Light (1999) in the topic model and show that automatically learned domains improve WSD accuracy compared to alternative contexts.

1 Introduction

Word sense disambiguation (WSD) is the task of determining the meaning of an ambiguous word in its context. It is an important problem in natural language processing (NLP) because effective WSD can improve systems for tasks such as information retrieval, machine translation, and summarization. In this paper, we develop latent Dirichlet allocation with WORDNET (LDAWN), a generative probabilistic topic model for WSD where the sense of the word is a hidden random variable that is inferred from data.

There are two central advantages to this approach. First, with LDAWN we automatically learn the context in which a word is disambiguated. Rather than disambiguating at the sentence-level or the document-level, our model uses the other words that share the same hidden topic across many documents.

Second, LDAWN is a fully-fledged generative model. Generative models are modular and can be easily combined and composed to form more complicated models. (As a canonical example, the ubiquitous hidden Markov model is a series of mixture models chained together.) Thus, developing a generative model for WSD gives other generative NLP algorithms a natural way to take advantage of the hidden senses of words.

In general, topic models are statistical models of text that posit a hidden space of topics in which the corpus is embedded (Blei et al., 2003). Given a corpus, posterior inference in topic models amounts to automatically discovering the underlying themes that permeate the collection. Topic models have recently been applied to information retrieval (Wei and Croft, 2006), text classification (Blei et al., 2003), and dialogue segmentation (Purver et al., 2006).

While topic models capture the polysemous use of words, they do not carry the explicit notion of *sense* that is necessary for WSD. LDAWN extends the topic modeling framework to include a hidden meaning in the word generation process. In this case, posterior inference discovers both the topics of the corpus and the meanings assigned to each of its words.

After introducing a disambiguation scheme based on probabilistic walks over the WORDNET hierarchy (Section 2), we embed the WORDNET-WALK in a topic model, where each topic is associated with walks that prefer different neighborhoods of WORDNET (Section 2.1). Then, we describe a Gibbs sampling algorithm for approximate posterior inference that learns the senses and topics that best explain a corpus (Section 3). Finally, we evaluate our system on real-world WSD data, discuss the properties of the topics and disambiguation accuracy results, and draw connections to other WSD algorithms from the research literature.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1024–1033, Prague, June 2007.

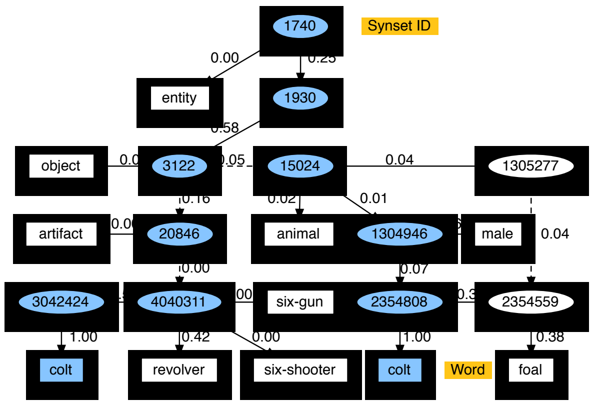

Figure 1: The possible paths to reach the word "colt" in WORDNET. Dashed lines represent omitted links. All words in the synset containing "revolver" are shown, but only one word from other synsets is shown. Edge labels are probabilities of transitioning from synset i to synset j. Note how this favors frequent terms, such as "revolver," over ones like "six-shooter."

2 Topic models and WordNet

The WORDNET-WALK is a probabilistic process of word generation that is based on the hyponomy relationship in WORDNET (Miller, 1990). WORDNET, a lexical resource designed by psychologists and lexicographers to mimic the semantic organization in the human mind, links "synsets" (short for synonym sets) with myriad connections. The specific relation we're interested in, hyponomy, points from general concepts to more specific ones and is sometimes called the "is-a" relationship.

As first described by Abney and Light (1999), we imagine an agent who starts at synset `[entity]`, which points to every noun in WORDNET 2.1 by some sequence of hyponomy relations, and then chooses the next node in its random walk from the hyponyms of its current position. The agent repeats this process until it reaches a leaf node, which corresponds to a single word (each of the synset's words are unique leaves of a synset in our construction). For an example of all the paths that might generate the word "colt" see Figure 1. The WORDNET-WALK is parameterized by a set of distributions over children for each synset s in WORDNET, $\boldsymbol{\beta}_s$.

Symbol	Meaning
K	number of topics
$\beta_{k,s}$	multinomial probability vector over the successors of synset s in topic k
S	scalar that, when multiplied by α_s gives the prior for $\beta_{k,s}$
α_s	normalized vector whose i^{th} entry, when multiplied by S, gives the prior probability for going from s to i
θ_d	multinomial probability vector over the topics that generate document d
τ	prior for θ
z	assignment of a word to a topic
Λ	a path assignment through WORDNET ending at a word.
$\lambda_{i,j}$	one link in a path λ going from synset i to synset j.

Table 1: A summary of the notation used in the paper. Bold vectors correspond to collections of variables (i.e. z_u refers to a topic of a single word, but $\boldsymbol{z}_{1:D}$ are the topics assignments of words in document 1 through D).

2.1 A topic model for WSD

The WORDNET-WALK has two important properties. First, it describes a random process for word generation. Thus, it is a distribution over words and thus can be integrated into any generative model of text, such as topic models. Second, the synset that produces each word is a hidden random variable. Given a word assumed to be generated by a WORDNET-WALK, we can use posterior inference to predict which synset produced the word.

These properties allow us to develop LDAWN, which is a fusion of these WORDNET-WALKs and latent Dirichlet allocation (LDA) (Blei et al., 2003), a probabilistic model of documents that is an improvement to pLSI (Hofmann, 1999). LDA assumes that there are K "topics," multinomial distributions over words, which describe a collection. Each document exhibits multiple topics, and each word in each document is associated with one of them.

Although the term "topic" evokes a collection of ideas that share a common theme and although the topics derived by LDA seem to possess semantic coherence, there is no reason to believe this would be true of the most likely multinomial distributions that could have created the corpus given the assumed generative model. That semantically similar words are likely to occur together is a byproduct of how language is actually used.

In LDAWN, we replace the multinomial topic distributions with a WORDNET-WALK, as described above. LDAWN assumes a corpus is generated by the following process (for an overview of the notation used in this paper, see Table 1).

1. For each topic, $k \in \{1, \ldots, K\}$
   (a) For each synset s, randomly choose transition probabilities $\beta_{k,s} \sim \mathrm{Dir}(S\boldsymbol{\alpha}_s)$.
2. For each document $d \in \{1, \ldots, D\}$
   (a) Select a topic distribution $\theta_d \sim \mathrm{Dir}(\tau)$
   (b) For each word $n \in \{1, \ldots, N_d\}$
      i. Select a topic $z \sim \mathrm{Mult}(1, \theta_d)$
      ii. Create a path $\Lambda_{d,n}$ starting with λ_0 as the root node.
      iii. From children of λ_i:
         A. Choose the next node in the walk $\lambda_{i+1} \sim \mathrm{Mult}(1, \beta_{z,\lambda_i})$
         B. If λ_{i+1} is a leaf node, generate the associated word. Otherwise, repeat.

Every element of this process, including the synsets, is hidden except for the words of the documents. Thus, given a collection of documents, our goal is to perform *posterior inference*, which is the task of determining the conditional distribution of the hidden variables given the observations. In the case of LDAWN, the hidden variables are the parameters of the K WORDNET-WALKS, the topic assignments of each word in the collection, and the synset path of each word. In a sense, posterior inference reverses the process described above.

Specifically, given a document collection $\boldsymbol{w}_{1:D}$, the full posterior is

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{z}_{1:D}, \boldsymbol{\theta}_{1:D}, \boldsymbol{\Lambda}_{1:D} \,|\, \boldsymbol{w}_{1:D}, \tau, S\boldsymbol{\alpha}) \propto \Big(\prod_{k=1}^{K} p(\boldsymbol{\beta}_k \,|\, S\boldsymbol{\alpha}) \prod_{d=1}^{D} p(\theta_d \,|\, \tau) \prod_{n=1}^{N_d} p(\Lambda_{d,n} \,|\, \boldsymbol{\beta}_{1:K}) p(w_{d,n} \,|\, \Lambda_{d,n})\Big), \quad (1)$$

where the constant of proportionality is the marginal likelihood of the observed data.

Note that by encoding the synset paths as a hidden variable, we have posed the WSD problem as a question of posterior probabilistic inference. Further note that we have developed an unsupervised model. No labeled data is needed to disambiguate a corpus. Learning the posterior distribution amounts to simultaneously decomposing a corpus into topics and its words into their synsets.

The intuition behind LDAWN is that the words in a topic will have similar meanings and thus share paths within WORDNET. For example, WORDNET has two senses for the word "colt;" one referring to a young male horse and the other to a type of handgun (see Figure 1).

Although we have no *a priori* way of knowing which of the two paths to favor for a document, we assume that similar concepts will also appear in the document. Documents with unambiguous nouns such as "six-shooter" and "smoothbore" would make paths that pass through the synset `[firearm, piece, small-arm]` more likely than those going through `[animal, animate being, beast, brute, creature, fauna]`. In practice, we hope to see a WORDNET-WALK that looks like Figure 2, which points to the right sense of cancer for a medical context.

LDAWN is a Bayesian framework, as each variable has a prior distribution. In particular, the Dirichlet prior for β_s, specified by a scaling factor S and a normalized vector α_s fulfills two functions. First, as the overall strength of S increases, we place a greater emphasis on the prior. This is equivalent to the need for balancing as noted by Abney and Light (1999).

The other function that the Dirichlet prior serves is to enable us to encode any information we have about how we suspect the transitions to children nodes will be distributed. For instance, we might expect that the words associated with a synset will be produced in a way roughly similar to the token probability in a corpus. For example, even though "meal" might refer to both ground cereals or food eaten at a single sitting and "repast" exclusively to the latter, the synset `[meal, repast, food eaten at a single sitting]` still prefers to transition to "meal" over "repast" given the overall corpus counts (see Figure 1, which shows prior transition probabilities for "revolver").

By setting $\alpha_{s,i}$, the prior probability of transitioning from synset s to node i, proportional to the total number of observed tokens in the children of i, we introduce a probabilistic variation on information content (Resnik, 1995). As in Resnik's definition, this value for non-word nodes is equal to the sum of all the frequencies of hyponym words. Unlike Resnik, we do not divide frequency among all senses of a word; each sense of a word contributes its full frequency to α.

3 Posterior Inference with Gibbs Sampling

As described above, the problem of WSD corresponds to posterior inference: determining the probability distribution of the hidden variables given observed words and then selecting the synsets of the most likely paths as the correct sense. Directly computing this posterior distribution, however, is not tractable because of the difficulty of calculating the normalizing constant in Equation 1.

To approximate the posterior, we use Gibbs sampling, which has proven to be a successful approximate inference technique for LDA (Griffiths and Steyvers, 2004). In Gibbs sampling, like all Markov chain Monte Carlo methods, we repeatedly sample from a Markov chain whose stationary distribution is the posterior of interest (Robert and Casella, 2004). Even though we don't know the full posterior, the samples can be used to form an empirical estimate of the target distribution. In LDAWN, the samples contain a configuration of the latent semantic states of the system, revealing the hidden topics and paths that likely led to the observed data.

Gibbs sampling reproduces the posterior distribution by repeatedly sampling each hidden variable conditioned on the current state of the other hidden variables and observations. More precisely, the state is given by a set of assignments where each word is assigned to a path through one of K WORDNET-WALK topics: u^{th} word w_u has a topic assignment z_u and a path assignment Λ_u. We use $\boldsymbol{z}_{-u}$ and $\boldsymbol{\Lambda}_{-u}$ to represent the topic and path assignments of all words except for u, respectively.

Sampling a new topic for the word w_u requires us to consider all of the paths that w_u can take in each topic and the topics of the other words in the document u is in. The probability of w_u taking on topic i is proportional to

$$p(z_u = i \,|\, \boldsymbol{z}_{-u}) \sum_{\lambda} p(\lambda \,|\, \boldsymbol{\Lambda}_{-u}) \mathbb{1}[w_u \in \lambda], \quad (2)$$

which is the probability of selecting z from θ_d times the probability of a path generating w_u from a path in the i^{th} WORDNET-WALK.

The first term, the topic probability of the u^{th} word, is based on the assignments to the K topics for words other than u in this document,

$$p(z_u = i | \boldsymbol{z}_{-u}) = \frac{n_{-u,i}^{(d)} + \tau_i}{\sum_j n_{-u,j}^{(d)} + \sum_{j=1}^{K} \tau_j}, \quad (3)$$

where $n_{-u,j}^{(d)}$ is the number of words other than u in topic j for the document d that u appears in.

The second term in Equation 2 is a sum over the probabilities of every path that could have generated the word w_u. In practice, this sum can be computed using a dynamic program for all nodes that have unique parent (i.e. those that can't be reached by more than one path). Although the probability of a path is specific to the topic, as the transition probabilities for a synset are different across topics, we will omit the topic index in the equation,

$$p(\Lambda_u = \lambda | \boldsymbol{\Lambda}_{-u},) = \prod_{i=1}^{l-1} \beta_{\lambda_i, \lambda_{i+1}}^{-u}. \quad (4)$$

3.1 Transition Probabilities

Computing the probability of a path requires us to take a product over our estimate of the probability from transitioning from i to j for all nodes i and j in the path λ. The other path assignments within this topic, however, play an important role in shaping the transition probabilities.

From the perspective of a single node i, only paths that pass through that node affect the probability of u also passing through that node. It's convenient to have an explicit count of all of the paths that transition from i to j in this topic's WORDNET-WALK, so we use $T_{i,j}^{-u}$ to represent all of the paths that go from i to j in a topic other than the path currently assigned to u.

Given the assignment of all other words to paths, calculating the probability of transitioning from i to j with word u requires us to consider the prior α and the observations $T_{i,j}$ in our estimate of the expected value of the probability of transitioning from i to j,

$$\beta_{i,j}^{-u} = \frac{T_{i,j}^{-u} + S_i \alpha_{i,j}}{S_i + \sum_k T_{i,k}^{-u}}. \quad (5)$$

As mentioned in Section 2.1, we paramaterize the prior for synset i as a vector α_i, which sums to one, and a scale parameter S.

The next step, once we've selected a topic, is to select a path within that topic. This requires the computation of the path probabilities as specified in Equation 4 for all of the paths w_u can take in the sampled topic and then sampling from the path probabilities.

The Gibbs sampler is essentially a randomized hill climbing algorithm on the posterior likelihood as a function of the configuration of hidden variables. The numerator of Equation 1 is proportional to that posterior and thus allows us to track the sampler's progress. We assess convergence to a local mode of the posterior by monitoring this quantity.

4 Experiments

In this section, we describe the properties of the topics induced by running the previously described Gibbs sampling method on corpora and how these topics improve WSD accuracy.

Of the two data sets used during the course of our evaluation, the primary dataset was SEMCOR (Miller et al., 1993), which is a subset of the Brown corpus with many nouns manually labeled with the correct WORDNET sense. The words in this dataset are lemmatized, and multi-word expressions that are present in WORDNET are identified. Only the words in SEMCOR were used in the Gibbs sampling procedure; the synset assignments were only used for assessing the accuracy of the final predictions.

We also used the British National Corpus, which is not lemmatized and which does not have multi-word expressions. The text was first run through a lemmatizer, and then sequences of words which matched a multi-word expression in WORDNET were joined together into a single word. We took nouns that appeared in SEMCOR twice or in the BNC at least 25 times and used the BNC to compute the information-content analog $\boldsymbol{\alpha}$ for individual nouns (For example, the probabilities in Figure 1 correspond to $\boldsymbol{\alpha}$).

4.1 Topics

Like the topics created by structures such as LDA, the topics in Table 2 coalesce around reasonable themes. The word list was compiled by summing over all of the possible leaves that could have generated each of the words and sorting the words by decreasing probability. In the vast majority of cases, a single synset's high probability is responsible for the words' positions on the list.

Reassuringly, many of the top senses for the present words correspond to the most frequent sense in SEMCOR. For example, in Topic 4, the senses for "space" and "function" correspond to the top senses in SEMCOR, and while the top sense for "set" corresponds to "an abstract collection of numbers or symbols" rather than "a group of the same kind that belong together and are so used," it makes sense given the math-based words in the topic. "Point," however, corresponds to the sense used in the phrase "I got to the point of boiling the water," which is neither the top SEMCOR sense nor a sense which makes sense given the other words in the topic.

While the topics presented in Table 2 resemble the topics one would obtain through models like LDA (Blei et al., 2003), they are not identical. Because of the lengthy process of Gibbs sampling, we initially thought that using LDA assignments as an initial state would converge faster than a random initial assignment. While this was the case, it converged to a state that less probable than the randomly initialized state and no better at sense disambiguation (and sometimes worse). The topics presented in 2 represent words both that co-occur together in a corpus and co-occur on paths through WORDNET. Because topics created through LDA only have the first property, they usually do worse in terms of both total probability and disambiguation accuracy (see Figure 3).

Another interesting property of topics in LDAWN is that, with higher levels of smoothing, words that don't appear in a corpus (or appear rarely) but are in similar parts of WORDNET might have relatively high probability in a topic. For example, "maturity" in topic two in Table 2 is sandwiched between "foot" and "center," both of which occur about five times more than "maturity." This might improve LDA-based information retrieval schemes (Wei and Croft, 2006) .

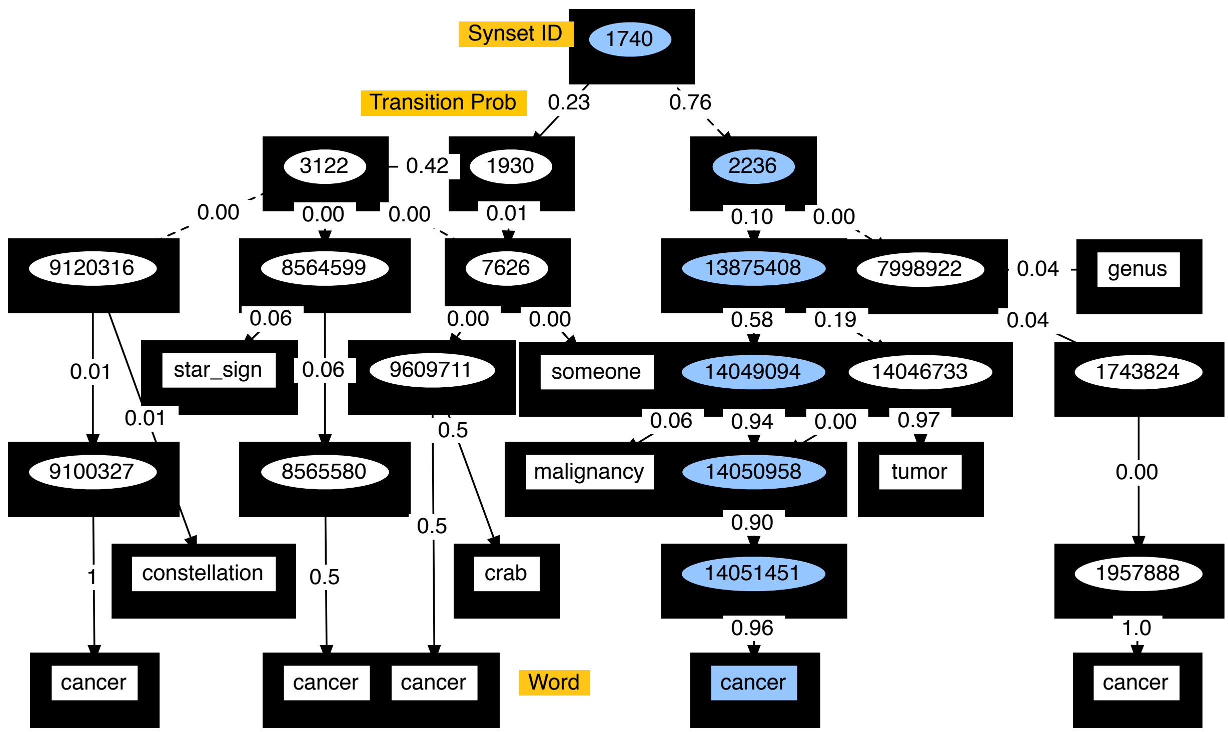

Figure 2: The possible paths to reach the word "cancer" in WORDNET along with transition probabilities from the medically-themed Topic 2 in Table 2, with the most probable path highlighted. The dashed lines represent multiple links that have been consolidated, and synsets are represented by their offsets within WORDNET 2.1. Some words for immediate hypernyms have also been included to give context. In all other topics, the person, animal, or constellation senses were preferred.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
president	growth	material	point	water	plant	music
party	age	object	number	house	change	film
city	treatment	color	value	road	month	work
election	feed	form	function	area	worker	life
administration	day	subject	set	city	report	time
official	period	part	square	land	mercer	world
office	head	self	space	home	requirement	group
bill	portion	picture	polynomial	farm	bank	audience
yesterday	length	artist	operator	spring	farmer	play
court	level	art	component	bridge	production	thing
meet	foot	patient	corner	pool	medium	style
police	maturity	communication	direction	site	petitioner	year
service	center	movement	curve	interest	relationship	show

Table 2: The most probable words from six randomly chosen WORDNET-walks from a thirty-two topic model trained on the words in SEMCOR. These are summed over all of the possible synsets that generate the words. However, the vast majority of the contributions come from a single synset.

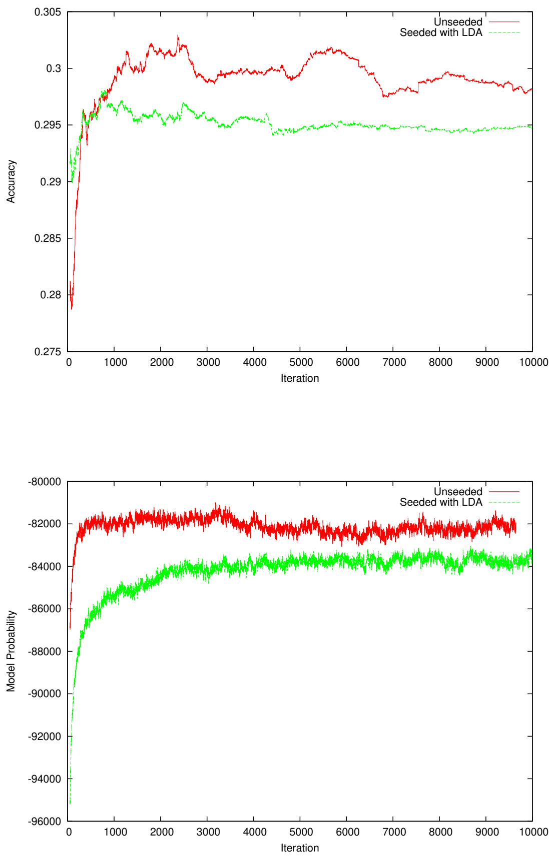

Figure 3: Topics seeded with LDA initially have a higher disambiguation accuracy, but are quickly matched by unseeded topics. The probability for the seeded topics starts lower and remains lower.

4.2 Topics and the Weight of the Prior

Because the Dirichlet smoothing factor in part determines the topics, it also affects the disambiguation. Figure 4 shows the modal disambiguation achieved for each of the settings of $S = \{0.1, 1, 5, 10, 15, 20\}$. Each line is one setting of K and each point on the line is a setting of S. Each data point is a run for the Gibbs sampler for 10,000 iterations. The disambiguation, taken at the mode, improved with moderate settings of S, which suggests that the data are still sparse for many of the walks, although the improvement vanishes if S dominates with much larger values. This makes sense, as each walk has over 100,000 parameters, there are fewer than 100,000 words in SEMCOR, and each

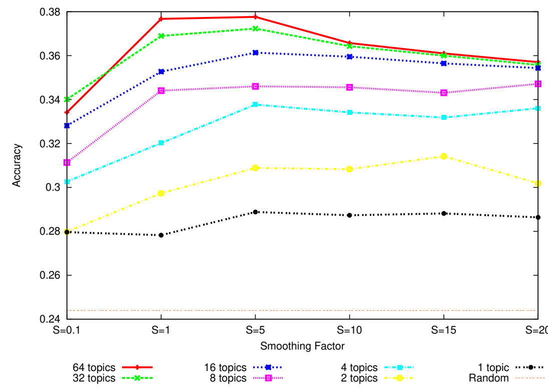

Figure 4: Each line represents experiments with a set number of topics and variable amounts of smoothing on the SEMCOR corpus. The random baseline is at the bottom of the graph, and adding topics improves accuracy. As smoothing increases, the prior (based on token frequency) becomes stronger. Accuracy is the percentage of correctly disambiguated polysemous words in SEMCOR at the mode.

word only serves as evidence to at most 19 parameters (the length of the longest path in WORDNET).

Generally, a greater number of topics increased the accuracy of the mode, but after around sixteen topics, gains became much smaller. The effect of α is also related to the number of topics, as a value of S for a very large number of topics might overwhelm the observed data, while the same value of S might be the perfect balance for a smaller number of topics. For comparison, the method of using a WORDNET-WALK applied to smaller contexts such as sentences or documents achieves an accuracy of between 26% and 30%, depending on the level of smoothing.

5 Error Analysis

This method works well in cases where the delineation can be readily determined from the overall topic of the document. Words such as "kid," "may," "shear," "coach," "incident," "fence," "bee," and (previously used as an example) "colt" were all perfectly disambiguated by this method. Figure 2 shows the WORDNET-WALK corresponding to a medical topic that correctly disambiguates "cancer."

Problems arose, however, with highly frequent words, such as "man" and "time" that have many senses and can occur in many types of documents. For example, "man" can be associated with many possible meanings: island, game equipment, servant, husband, a specific mammal, etc.

Although we know that the "adult male" sense should be preferred, the alternative meanings will also be likely if they can be assigned to a topic that shares common paths in WordNet; the documents contain, however, many other places, jobs, and animals which are reasonable explanations (to LDAWN) of how "man" was generated. Unfortunately, "man" is such a ubiquitous term that topics, which are derived from the frequency of words within an entire document, are ultimately uninformative about its usage.

While mistakes on these highly frequent terms significantly hurt our accuracy, errors associated with less frequent terms reveal that WordNet's structure is not easily transformed into a probabilistic graph. For instance, there are two senses of the word "quarterback," a player in American football. One is position itself and the other is a person playing that position. While one would expect co-occurrence in sentences such as "quarterback is a easy position, so our quarterback is happy," the paths to both terms share only the root node, thus making it highly unlikely a topic would cover both senses.

Because of WordNet's breadth, rare senses also impact disambiguation. For example, the metonymical use of "door" to represent a whole building as in the phrase "girl next door" is under the same parent as sixty other synsets containing "bridge," "balcony," "body," "arch," "floor," and "corner." Surrounded by such common terms that are also likely to co-occur with the more conventional meanings of door, this very rare sense becomes the preferred disambiguation of "door."

6 Related Work

Abney and Light's initial probabilistic WSD approach (1999) was further developed into a Bayesian network model by Ciaramita and Johnson (2000), who likewise used the appearance of monosemous terms close to ambiguous ones to "explain away" the usage of ambiguous terms in selectional restrictions. We have adapted these approaches and put them into the context of a topic model.

Recently, other approaches have created *ad hoc* connections between synsets in WordNet and then considered walks through the newly created graph. Given the difficulties of using existing connections in WordNet, Mihalcea (2005) proposed creating links between adjacent synsets that might comprise a sentence, initially setting weights to be equal to the Lesk overlap between the pairs, and then using the PageRank algorithm to determine the stationary distribution over synsets.

6.1 Topics and Domains

Yarowsky was one of the first to contend that "there is one sense for discourse" (1992). This has lead to the approaches like that of Magnini (Magnini et al., 2001) that attempt to find the category of a text, select the most appropriate synset, and then assign the selected sense using domain annotation attached to WordNet.

LDAWN is different in that the categories are not an *a priori* concept that must be painstakingly annotated within WordNet and require no augmentation of WordNet. This technique could indeed be used with any hierarchy. Our concepts are the ones that best partition the space of documents and do the best job of describing the distinctions of diction that separate documents from different domains.

6.2 Similarity Measures

Our approach gives a probabilistic method of using information content (Resnik, 1995) as a starting point that can be adjusted to cluster words in a given topic together; this is similar to the Jiang-Conrath similarity measure (1997), which has been used in many applications in addition to disambiguation. Patwardhan (2003) offers a broad evaluation of similarity measures for WSD.

Our technique for combining the cues of topics and distance in WordNet is adjusted in a way similar in spirit to Buitelaar and Sacaleanu (2001), but we consider the appearance of a single term to be evidence for not just that sense and its immediate neighbors in the hyponomy tree but for all of the sense's children and ancestors.

Like McCarthy (2004), our unsupervised system acquires a single predominant sense for a domain based on a synthesis of information derived from a textual corpus, topics, and WORDNET-derived similarity, a probabilistic information content measure. By adding syntactic information from a thesaurus derived from syntactic features (taken from Lin's automatically generated thesaurus (1998)), McCarthy achieved 48% accuracy in a similar evaluation on SEMCOR; LDAWN is thus substantially less effective in disambiguation compared to state-of-the-art methods. This suggests, however, that other methods might be improved by adding topics and that our method might be improved by using more information than word counts.

7 Conclusion and Future Work

The LDAWN model presented here makes two contributions to research in automatic word sense disambiguation. First, we demonstrate a method for automatically partitioning a document into topics that includes explicit semantic information. Second, we show that, at least for one simple model of WSD, embedding a document in probabilistic latent structure, i.e., a "topic," can improve WSD.

There are two avenues of research with LDAWN that we will explore. First, the statistical nature of this approach allows LDAWN to be used as a component in larger models for other language tasks. Other probabilistic models of language could insert the ability to query synsets or paths of WORDNET. Similarly, any topic based information retrieval scheme could employ topics that include semantically relevant (but perhaps unobserved) terms. Incorporating this model in a larger syntactically-aware model, which could benefit from the local context as well as the document level context, is an important component of future research.

Second, the results presented here show a marked improvement in accuracy as more topics are added to the baseline model, although the final result is not comparable to state-of-the-art techniques. As most errors were attributable to the hyponomy structure of WORDNET, incorporating the novel use of topic modeling presented here with a more mature unsupervised WSD algorithm to replace the underlying WORDNET-WALK could lead to advances in state-of-the-art unsupervised WSD accuracy.

References

Steven Abney and Marc Light. 1999. Hiding a semantic hierarchy in a markov model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*, pages 1–8.

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

Paul Buitelaar and Bogdan Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001.* Association for Computational Linguistics.

Massimiliano Ciaramita and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with bayesian networks. In *COLING-00*, pages 187–193.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, pages 5228–5235.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, Taiwan.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proc. 15th International Conf. on Machine Learning*, pages 296–304.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2001. Using domain information for word sense disambiguation. In *In Proceedings of 2^{nd} International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, France.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *In 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287.

Rada Mihalcea. 2005. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference*, pages 411–418.

George Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *3rd DARPA Workshop on Human Language Technology*, pages 303–308.

George A. Miller. 1990. Nouns in WordNet: A lexical inheritance system. *International Journal of Lexicography*, 3(4):245–264.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2003. Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 241–257.

Matthew Purver, Konrad Körding, Thomas Griffiths, and Joshua Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of COLING-ACL*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conferences on Artificial Intelligence*, pages 448–453.

Christian Robert and George Casella. 2004. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY.

Xing Wei and Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the Twenty-Ninth Annual International SIGIR Conference*.