PUTOP: Turning Predominant Senses into 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

Abstract

We extend on McCarthy et al.'s predominant sense method to create an unsupervised method of word sense disambiguation that uses automatically derived topics using Latent Dirichlet allocation. Using topic-specific synset similarity measures, we create predictions for each word in each document using only word frequency information. It is hoped that this procedure can improve upon the method for larger numbers of topics by providing more relevant training corpora for the individual topics. This method is evaluated on SemEval-2007 Task 1 and Task 17.

1 Generative Model of WSD

Word Sense Disambiguation (WSD) is the problem of labeling text with the appropriate semantic labels automatically. Although WSD is claimed to be an essential step in information retrieval and machine translation, it has not seen effective practical application because the dearth of labeled data has prevented the use of established supervised statistical methods that have been successfully applied to other natural language problems.

Unsupervised methods have been developed for WSD, but despite modest success have not always been well understood statistically (Abney, 2004). Unsupervised methods are particularly appealing because they do not require expensive sense-annotated data and can use the ever-increasing amount of raw text freely available. This paper expands on an effective unsupervised method for WSD and embeds it into a topic model, thus allowing an algorithm trained on a single, monolithic corpora to instead hand-pick relevant documents in choosing a disambiguation. After developing this generative statistical model, we present its performance on a number of tasks.

1.1 The Intersection of Syntactic and Semantic Similarity

McCarthy et al. (2004) outlined a method for learning a word's most-used sense given an untagged corpus that ranks each sense ws_i using a distributional syntactic similarity γ and a WordNet-derived semantic similarity α. This process for a word w uses its distributional neighbors N_w, the possible senses of not only the word in question, S_w, and also those of the distributionally similar words, S_{n_j}. Thus, $P(ws_i) =$

$$\sum_{n_j \in N_w} \gamma(w, n_j) \frac{wnss(ws_i, n_j)}{\sum_{ws_j \in S_w} wnss(ws_j, n_j)}, \quad (1)$$

where $wnss(s, c) =$

$$\max_{a \in S_c} \alpha(a, s). \quad (2)$$

One can view finding the appropriate sense as a search in two types of space. In determining how good a particular synset ws_i is, α guides the search in the semantic space and γ drives the search in the syntactic space. We consider all of the words used in syntactically similar contexts, which we call "corroborators," and for each of them we find the closest meaning to ws_i using a measure of semantic similarity α, for instance a WordNet-based similarity measure such as Jiang-Conrath (1997). Each of the neighboring words' contributions is weighted by the syntactic probability, as provided by Lin's distributional similarity measure (1998), which rates two words to be similar if they enter into similar syntactic constructions.

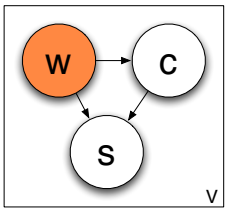

Figure 1: A reinterpretation of McCarthy et al.'s predominant sense method as a generative model. Note that this model has no notion of context; a synset is assigned in an identical manner for all of the words in a vocabulary.

One can think of this process as a generative model, even though it was not originally posed in such a manner. For each word w in the vocabulary, we generate one of the neighbor corroborators according to the Lin similarity, $\gamma(c, w)$, between the two words. We then generate a synset s for that word proportional to the maximum semantic similarity between s and any synset that contains the corroborator c (see Figure 1).

Our aim in this paper is to extend the method of McCarthy et al. using topic models. It is hoped that allowing the method to in effect "choose" the contexts that it uses will improve its ability to disambiguate sentences.

1.2 Using Topic Models to Partition a Document's Words

Topic models like Latent Dirichlet allocation (LDA) (Blei et al., 2003) assume a model of text generation where each document has a multinomial distribution over topics and each word comes from one of these topics. In LDA, each topic is a multinomial distribution, and each document has a multinomial distribution over topics drawn from a Dirichlet prior that selects the topic for each word in a document. Previous work has shown that such a model improves WSD over using a single corpus (Boyd-Graber et al., 2007), and we use this insight to develop an extension of McCarthy's method for multiple topics.

Although describing the statistical background and motivations behind topic models are beyond the scope of this paper, it suffices to note that the topics induced from a corpus provide a statistical grouping of words that often occur together and a probabilistic assignment of each word in a corpus to topics. Thus, one topic might have terms like "government," "president," "govern," and "regal," while another topic might have terms like "finance," "high-yield," "investor," and "market." This paper assumes that the machinery for learning these distributions can, given a corpus and a specified number of topics, return the topic distributions most likely to have generated the corpus.

1.3 Defining the Model

While the original predominant senses method used Lin's thesaurus similarity method alone in generating the corroborator, we will also use the probability of that word being part of the same topic as the word to be disambiguated. Thus the process of choosing the "corroborator" is no longer identical for each word; it is affected by its topic, which changes for every document. This new generative process can be thought of as a modified LDA system that, after selecting the word generated by the topic, continues on by generating a corroborator and a sense for the original word:

For each document $d \in \{1 \dots D\}$:

1. Select a topic distribution $\theta_d \sim \text{Dir}(\tau)$
2. For each word in the document $n \in \{1 \dots N\}$:
   (a) Select a topic $z_n \sim \text{Mult}(1, \theta_d)$
   (b) Select a word from that topic $w_n \sim \text{Mult}(1, \beta_z)$
   (c) Select a "corroborator" c_n also proportional to how important it is to the topic and its similarity to w
   (d) Now, select a synset s_n for that word based on a distribution $p(s_n|w_n, c_n, z_n)$

The conditional dependencies for generating a synset are shown in Figure 2. Our goal, like McCarthy et al.'s, is to determine the most likely sense for each word. This amounts to posterior inference, which we address by marginalizing over the unobserved variables (the topics and the corroborators), where $p(ws_i) =$

$$p(s|w) = \int_\theta \sum_z \sum_c p(s|w, c, z)p(c|z, w)p(z|w, \theta). \tag{3}$$

In order to fully specify this, we must determine the distribution from which the corroborator is drawn and the distribution from which the synset is drawn.

Ideally, we would want a distribution that for a single topic would be identical to McCarthy et al.'s

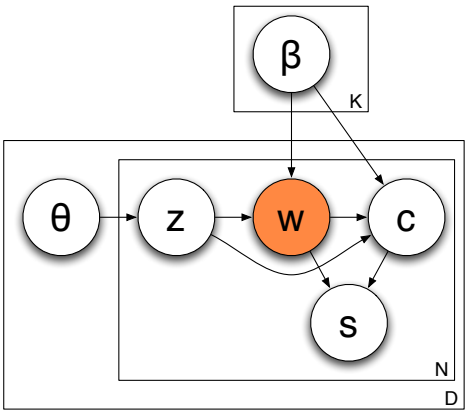

Figure 2: Our generative model assumes that documents are divided into topics and that these topics generate both the observed word and a "corroborator," a term similar in usage to the word. Next, a sense that minimizes the semantic distance between the corroborator and the word is generated.

method but would, as more topics are added, favor corroborators in the same topic as the number of topics increases. In McCarthy et al.'s method, the probability of the corroborator given a word w is proportional to the Lin similarity $\gamma(w, c)$ between the word and the corroborator. Here, the probability of a corroborator c is

$$p(c|z, w) \propto \frac{\beta_{z,c}}{\beta_c^0}\gamma(w, c), \quad (4)$$

where $\beta_{z,c}$ is the multinomial probability of word c in the z^{th} topic, and β_c^0 is the multinomial probability of the word with a single topic (i.e. background word probability).

Before, the corroborator was weighted simply based on its syntactic similarity to the word w, now we also weight that contribution by how important (or unimportant) that word is to the topic that w has been assigned to. This has the effect of increasing the probability of words pertinent to the topic that also have high syntactic similarity. Thus, whenever the syntactic similarity captures polysemous usage, we hope to be able to separate the different usages. Note, however, that since for a single topic the β term cancels out and the procedure is equivalent to McCarthy et al.

We adapt the semantic similarity in much the same way to make it topic specific. Because the Jiang-Conrath similarity measure uses an underlying term frequency to generate a similarity score, we use the topic term frequency instead of the undivided term frequency. Thus, the probability of a sense is proportional to semantic similarity between it and the closest sense among the senses of a corroborator with respect to this topic-specific similarity (c.f. the global similarity in Equation 2). The probability of selecting a synset s given the corroborator c and a topic z then becomes

$$p(s|w, c, z) \propto \max_{s' \in S(c)} \alpha_z(s, s'). \quad (5)$$

This new dependence on the topic happens because we recompute the information content used by Jiang-Conrath with the distribution over words implied by each topic. We then use the similarity implied by that similarity for α_z. Following the lead of McCarthy, for notational ease, this becomes defined as $wnss$ in Equation 8.

1.4 Choosing a Synset

The problem of choosing a synset then is reduced to finding the synset with the highest probability under this model. The model is also designed so that the task of learning the assignment of topics to words and documents is not affected by this new machinery for corroborators and senses that we've added onto the model. Thus, we can use the variational inference method described in (Blei et al., 2003) as a foundation for the problem of synset inference.

Taking $p(z|w)$ as a given (i.e. determined by running LDA on the corpus), the probability for a synset s given a word w then becomes

$$p(s|w, z) = \sum_z \sum_c p(s|w, c, z)p(c|z)p(z|w), \quad (6)$$

whose terms have been described in the previous section. With all of the normalization terms, we now see that $p(s|w, z)$ becomes

$$\sum_z \sum_c \frac{\frac{\beta_{z,c}}{\beta_c^0}\gamma(w, c)}{\sum_{c'} \frac{\beta_{z,c}}{\beta_c^0}\gamma(w, c')} \frac{wnss(s, c, z)}{\sum_{s' \in S_w} wnss(s', c, z)}. \quad (7)$$

and $wnss(s, c, z)$ now becomes, for the z^{th} topic,

$$\max_{a \in S(c)} \alpha_z(a, s). \quad (8)$$

Thus, we've now assigned a probability to each of the possible senses a word can take in a document.

1.5 Intuition

For example, consider the word "fly," which has two other words that have high syntactic similarity (in our formulation, γ) with the terms "fly_ball" and "insect." Both of these words would, given the semantic similarity provided by WORDNET, point to a single sense of "fly;" one of them would give a higher value, however, and thus all senses of the word "fly" would be assigned that sense. By separately weighting these words by the topic frequencies, we would hope to choose the sports sense in topics that have a higher probability of the terms like "foul_ball," "pop_fly," and "grounder" and the other sense in the contexts where insect has a higher probability in the topic.

2 Evaluations

This section describes three experiments to determine the effectiveness of this unsupervised system. The first was used to help understand the system, and the second two were part of the SemEval 2007 competition.

2.1 SemCor

As an initial evaluation, we learned LDA topics on the British National corpus with paragraphs as the underlying "document" (this allowed for a more uniform document length). These documents were then used to infer topic probabilities for each of the words in SemCor (Miller et al., 1993), and the model described in the previous section was run to determine the most likely synset. The results of this procedure are shown in Table 1. Accuracy is determined as the percentage of words for which the most likely sense was the one tagged in the corpus.

While the method does roughly recreate McCarthy et al.'s result for a single topic, it only offers a one percent improvement over McCarthy et al. on five topics and then falls below McCarthy for all greater numbers of topics tried. Thus, for all subsequent experiments we used a five topic model trained on the BNC.

2.2 SemEval-2007 Task 1: CLIR

Using IR metrics, this disambiguation scheme was evaluated against another competing platform and an algorithm provided by the Task 1 (Agirre et al., 2007) organizers. Our system had the best results of any expansion scheme considered (0.30) , although none of the expansion schemes did better than using no expansion (0.36). Although our technique also yielded a better score than the other competing platform for cross-language queries (0.17), it did not surpass the first sense-heuristic (0.26), but this is not surprising given that our algorithm does not assume the existence of such information. For an overview of Task 1 results, see Table 2.

Topics	All	Nouns
1	.393	.467
5	.397	.478
25	.387	.456
200	.359	.420

Table 1: Accuracy on disambiguating words in SemCor

Task	PUTOP
Topic Expansion	0.30
Document Expansion	0.15
English Translation	0.17
SensEval 2	0.39
SensEval 3	0.33

Table 2: Performance results on Task 1

2.3 SemEval-2007 Task 17: All-Words

Task 17 (Pradhan et al., 2007) asked participants to submit results as probability distributions over senses. Because this is also the output of this algorithm, we submitted the probabilities to the contest before realizing that the distributions are very close to uniform over all senses and thus yielded a precision of 0.12, very close to the random baseline. Placing a point distribution on the argmax with our original submission to the task, however, (consistent with our methodology for evaluation on SemCor), gives a precision of 0.39.

3 Conclusion

While the small improvement over the single topic suggests that topic techniques might have traction in determining the best sense, the addition is not appreciable. In a way the failure of the technique is en-

couraging in that it affirms the original methodology of McCarthy et al. in finding a single predominant sense for each word. While the syntactic similarity measure indeed usually offers high values of similarity for words related to a single sense of a word, the similarity for words related to other senses, which we had hoped to strengthen by using topic features, are on par with words observed because of noise.

Thus, for a word like "bank," words like "firm," "commercial_bank," "company," and "financial_institution" are the closest in terms of the syntactic similarity, and this allows the financial senses to be selected without any difficulty. Even if we had corroborating words for another sense in some topic, these words are absent from the syntactically similar words. If we want the meaning similar to that of "riverbank," the word with the most similar meaning, "side," had a syntactic similarity on par with the unrelated words "individual" and "group." Thus, interpretations other than the dominant sense as determined by the baseline method of McCarthy et al. are hard to find.

Because one topic is equivalent to McCarthy et al.'s method, this means that we do no worse on disambiguation. However, contrary to our hope, increasing the number of topics does not lead to significantly better sense predictions. This work has not investigated using a topic-based procedure for determining the syntactic similarity, but we feel that this extension could provide real improvement to the unsupervised techniques that can make use of the copious amounts of available unlabeled data.

References

Steven Abney. 2004. Understanding the yarowsky algorithm. *Comput. Linguist.*, 30(3):365–395.

Eneko Agirre, Oier Lopez de Lacalle, Arantxa Otegi, German Rigau, and Piek Vossen. 2007. The Senseval-2007 Task 1: Evaluating WSD on cross-language information retrieval. In *Proceedings of SemEval-2007*. Association for Computational Linguistics.

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

Jordan L. Boyd-Graber, David M. Blei, and Jerry Zhu. 2007. Probabalistic walks in semantic hierarchies as a topic model for WSD. In *Proc. EMNLP 2007*.

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. Morgan Kaufmann, San Francisco, CA.

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.

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

Sameer Pradhan, Martha Palmer, and Edward Loper. 2007. The Senseval-2007 Task 17: English fine-grained all-words. In *Proceedings of SemEval-2007*. Association for Computational Linguistics.