PU-BCD: Exponential Family Models for the Coarse- and Fine-Grained All-Words Tasks

Jonathan Chang
Princeton University
Department of Electrical Engineering
jcone@princeton.edu

Miroslav Dudík, David M. Blei
Princeton University
Department of Computer Science
{mdudik,blei}@cs.princeton.edu

Abstract

This paper describes an exponential family model of word sense which captures both occurrences and co-occurrences of words and senses in a joint probability distribution. This statistical framework lends itself to the task of word sense disambiguation. We evaluate the performance of the model in its participation on the SemEval-2007 coarse- and fine-grained all-words tasks under a variety of parameters.

1 Introduction

This paper describes an *exponential family model* suited to performing word sense disambiguation. Exponential family models are a mainstay of modern statistical modeling (Brown, 1986) and they are widely and successfully used for example in text classification (Berger et al., 1996). In statistical machine learning research, a general methodology and many algorithms were developed for *undirected graphical model* representation of exponential families (Jordan, 2004), providing a solid basis for efficient inference.

Our model differs from other probabilistic models used for word sense disambiguation in that it captures not only word-sense co-occurrences but also contextual sense-sense co-occurrences, thereby breaking the naïve Bayes assumption. Although spare in the types of features, the model is extremely expressive. Our model has parameters that control for word-sense interaction and sense-sense similarity, allowing us to capture many of the salient features of word and sense use. After fitting the parameters of our model from a labeled corpus, the task of word sense disambiguation immediately follows by considering the *posterior distribution* of senses given words.

We used this model to participate in SemEval-2007 on the coarse- and fine-grained all-words tasks. In both of these tasks, a series of sentences are given with certain words tagged. Each competing system must assign a sense from a sense inventory to the tagged words. In both tasks, performance was gauged by comparing the output of each system to human-tagged senses. In the fine-grained task, precision and recall were simply and directly computed against the golden annotations. However, in the coarse-grained task, the sense inventory was first clustered semi-automatically with each cluster representing an equivalence class over senses (Navigli, 2006). Precision and recall were computed against equivalence classes.

This paper briefly derives the model and then explores its properties for WSD. We show how common algorithms, such as "dominant sense" and "most frequent sense," can be expressed in the exponential family framework. We then proceed to present an evaluation of the developed techniques on the SemEval-2007 tasks in which we participated.

2 The model

We describe an exponential family model for word sense disambiguation. We posit a joint distribution over words $\mathbf{w}$ and senses $\mathbf{s}$.

2.1 Notation

We define a *document* d to be a sequence of words from some lexicon $\mathcal{W}$; for the participation in this contest, a document consists of a sentence. Associated with each word is a *sense* from a lexicon $\mathcal{S}$. In

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 272–276,
Prague, June 2007. ©2007 Association for Computational Linguistics

this work, our sense lexicon is the synsets of WordNet (Fellbaum and Miller, 2003), but our methods easily generalize to other sense lexicons, such as VerbNet (Kipper et al., 2000).

Formally, we denote the sequence of words in a document d by $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,n_d})$ and the sequence of synsets by $\mathbf{s}_d = (s_{d,1}, s_{d,2}, \ldots, s_{d,n_d})$, where n_d denotes the number of words in the document. A *corpus* $\mathcal{D}$ is defined as a collection of documents. We also write $w \in s$ if w can be used to represent sense s.

2.2 An exponential family of words and senses

We turn our attention to an exponential family of words and senses. The vector of parameters $\boldsymbol{\eta} = (\boldsymbol{\kappa}, \boldsymbol{\lambda})$ consists of two blocks capturing dependence on word-synset co-occurrences, and synset co-occurrences.

$$p_{\boldsymbol{\eta},n}(\mathbf{s}, \mathbf{w}) = \exp\{\textstyle\sum_i \kappa_{w_i,s_i} + \sum_{i,j} \lambda_{s_i,s_j}\} / Z_{\boldsymbol{\eta},n} \ . \quad (1)$$

The summations are first over all positions in the document, $1 \le i \le n$, and then over all pairs of positions in the document, $1 \le i, j \le n$. We discuss parameters of our exponential model in turn.

Word-sense parameters $\boldsymbol{\kappa}$ Using parameters $\boldsymbol{\kappa}$ alone, it is possible to describe an arbitrary context independent distribution between a word and its assigned synset.

Sense co-occurrence parameters $\boldsymbol{\lambda}$ Parameters $\boldsymbol{\lambda}$ are the only parameters that establish the dependence of sense on its context. More specifically, they capture co-occurrences of synset pairs within a context. Larger values favor, whereas smaller values disfavor each pair of synsets.

3 Parameter estimation

With the model in hand, we need to address two problems in order to use it for problems such as WSD. First, in *parameter estimation*, we find values of the parameters that explain a labeled corpus, such as SemCor (Miller et al., 1993). Once the parameters are fit, we use *posterior inference* to compute the posterior probability distribution of a set of senses given a set of unlabeled words in a context, $p(\mathbf{s} \mid \mathbf{w})$. This distribution is used to predict the senses of the words.

In this section, it will be useful to introduce the notation $\tilde{p}(s, w)$ to denote the empirical probabilities of observing the word-sense pair s, w in the entire corpus:

$$\tilde{p}(s, w) = \textstyle\sum_{d,i} \delta(s_{d,i}, s)\delta(w_{d,i}, w) / \sum_d n_d \ ,$$

where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise. Similarly, we will define $\tilde{p}(s)$ to denote the empirical probability of observing a sense s over the entire corpus:

$$\tilde{p}(s) = \textstyle\sum_{d,i} \delta(s_{d,i}, s) / \sum_d n_d \ .$$

3.1 Word-sense parameters $\boldsymbol{\kappa}$

Fallback Let $\kappa^{\text{WN}}_{w,s} = 0$ if $w \in s$ and $\kappa^{\text{WN}}_{w,s} = -\infty$ otherwise. This simply sets to zero the probability of assigning a word w to a synset s when $w \notin s$ while making all $w \in s$ equally likely as an assignment to s. This forces the model to rely entirely on $\boldsymbol{\lambda}$ for inference. If $\boldsymbol{\lambda}$ is also set to $\mathbf{0}$, this then forces the system to fall back onto its arbitrary tie-breaking mechanism such as choosing randomly or choosing the first sense.

Most-frequent synset One approach to disambiguation is the technique of choosing the most frequently occurring synset which the word may express. This can be implemented within the model by setting $\kappa_{w,s} = \kappa^{\text{MFS}}_{w,s} \equiv \ln \tilde{p}(s)$ if $w \in s$ and $-\infty$ otherwise.

MLE Given a labeled corpus, we would like to find the corresponding parameters that maximize likelihood of the data. Equivalently, we would like to maximize the log likelihood

$$\mathcal{L}(\boldsymbol{\eta}) = \sum_d \left[\textstyle\sum_i \kappa_{w_{d,i},s_{d,i}} + \sum_{i,j} \lambda_{s_{d,i},s_{d,j}} - \ln Z_{\boldsymbol{\eta},n_d} \right] \ . \quad (2)$$

In this section, we consider a simple case when it is possible to estimate parameters maximizing the likelihood exactly, i.e., the case where our model depends only on word-synset co-occurrences and is parametrized solely by $\boldsymbol{\kappa}$ (setting $\boldsymbol{\lambda} = \mathbf{0}$).

Using Eq. (1), with $\boldsymbol{\lambda} = \mathbf{0}$, we obtain

$$p_{\boldsymbol{\kappa}}(\mathbf{s}_{\mathcal{D}}, \mathbf{w}_{\mathcal{D}}) = \frac{\exp\{\sum_{d,i} \kappa_{w_{d,i},s_{d,i}}\}}{\prod_d Z_{\boldsymbol{\kappa},n_d}} \ .$$

Thus, $p_\kappa(\mathbf{s}_\mathcal{D}, \mathbf{w}_\mathcal{D})$ can be viewed as a multinomial model with $\sum_d n_d$ trials and $|\mathcal{S}|$ outcomes, parametrized by $\kappa_{w,s}$. The maximum likelihood estimates in this model are $\hat{\kappa}_{w,s} \equiv \ln \tilde{p}(s, w)$.

This setting of the parameters corresponds precisely to the *dominant-sense* model (McCarthy et al., 2004). The resulting model is thus

$$p_{\boldsymbol{\kappa},n}(\mathbf{s}, \mathbf{w}) = \prod_i \tilde{p}(s_i, w_i) \ . \tag{3}$$

3.2 Sense co-occurrence parameters $\boldsymbol{\lambda}$

Unlike $\boldsymbol{\kappa}$, it is impossible to find a closed-form solution for the maximum-likelihood settings of $\boldsymbol{\lambda}$. Therefore, we turn to intuitive methods.

Observed synset co-occurrence One natural ad hoc statistic to use to compute the parameters λ are the empirical sense co-occurrences. In particular, we may set

$$\lambda_{s_i,s_j} = \lambda^{\text{SF}}_{s_i,s_j} \equiv \ln \tilde{p}(s_i, s_j) \ . \tag{4}$$

We will observe in section 5 that the performance of $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{SF}}$ actually degrades the performance of the system, especially when combined with $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}$. This can be understood as a by-product of an unsympathetic interaction between $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$. In other words, $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ overlap; by favoring a sense pair the model will also implicitly favor each of the senses in the pair.

Discounted observed synset co-occurrence As we noted earlier, the combination $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}, \boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{SF}}$ actually performs worse than $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}, \boldsymbol{\lambda} = \mathbf{0}$. In order to cancel out the aforementioned overlap effect, we attempt to compute the number of co-occurrences beyond what the *occurrences* themselves would imply. To do so, we set

$$\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{DSF}} \equiv \ln \frac{\tilde{p}(s_i, s_j)}{\tilde{p}(s_i)\tilde{p}(s_j)} \ , \tag{5}$$

a quantity which finds an analogue in the notion of *mutual information*. We will see shortly that such a setting of $\boldsymbol{\lambda}$ will allow sense co-occurrence to improve disambiguation performance.

4 Word Sense Disambiguation

Finally, we describe how to perform WSD using the exponential family model. Our goal is to assign a synset s_i to every word w_i in an unlabeled document d of length n. In this setting, the synsets are hidden variables. Thus, we assign synsets according to their posterior probability given the observed words:

$$\hat{\mathbf{s}} = \operatorname*{argmax}_{\mathbf{s} \in \mathcal{S}^n} \frac{p_{\boldsymbol{\eta},n}(\mathbf{s}, \mathbf{w})}{\sum_{\mathbf{s}'} p_{\boldsymbol{\eta},n}(\mathbf{s}', \mathbf{w})} \ ,$$

where the sum is over all possible sequences of synsets. This combinatorial sum renders exact inference computationally intractable. We discuss how to obtain the sense assignment using approximate inference.

4.1 Variational Inference

To approximate the posterior over senses, we use *variational inference* (Jordan et al., 1999). In variational inference, one first chooses a family of distributions for which inference is computationlly tractable. Then the distribution in that family which best approximates the posterior distribution of interest is found.

For our purposes, it is convenient to select q from the family of factorized multinomial distributions:

$$q(\mathbf{s}) = \prod_i q_i(s_i) \ ,$$

where each $q_i(s_i)$ is a multinomial distribution over all possible senses. Observe that finding $\hat{\mathbf{s}}$ is much simpler using $q(\mathbf{s})$: one can find the argmax of each individual q_i independently.

It can be shown that the multinomial which minimizes the KL-divergence must satisfy:

$$q_i(s_i) \propto \exp\left\{ \kappa_{w_i,s_i} + \sum_{j \neq i} \sum_{s_j} q_j(s_j) \lambda_{s_i,s_j} \right\} \tag{6}$$

a system of transcendental equations which can be solved iteratively to find q. This q is then used to efficiently perform inference and hence disambiguation.

5 Evaluation

This section evaluates the performance of the model and the techniques described in the previous sections with respect to the coarse- and fine-grained all-words tasks at SemEval-2007.

In order to train the parameters, we trained our model in a supervised fashion on SemCor (Miller et

	$\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{WN}}$	$\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{MFS}}$	$\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}$
$\boldsymbol{\lambda} = \mathbf{0}$	52.0%	45.8%	51.2%
$\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{SF}}$	48.8%	45.3%	52.5%
$\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{DSF}}$	47.0%	44.6%	**54.2%**

Table 1: Precision for the fine-grained all-words task. The results corresponding to the bolded value was submitted to the competition.

al., 1993) with Laplace smoothing for parameter estimates. We utilized the POS tagging and lemmatization given in the coarse-grained all-words test set. Wherever a headword was tagged differently between the two test sets, we produced an answer only for the coarse-grained test and not for the fine-grained one. This led to responses on only 93.9% of the fine-grained test words. Of the 6.1% over which no response was given, 5.3% were tagged as "U" in the answer key.

In order to break ties between equally likely senses, for the fine-grained test, the system returned the first one returned in WordNet's sense inventory for that lemma. For the coarse-grained test, an arbitrary sense was returned in case of ties.

The precision results given in this section are over polysemous words (of all parts of speech) for which our system gave an answer and for which the answer key was not tagged with "U."

5.1 Fine-grained results (Task 17)

The fine-grained results over all permutations of the parameters mentioned in Section 3 are given in Table 1. Note here that the baseline number of $\boldsymbol{\lambda} = \mathbf{0}, \boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{WN}}$ given in the upper-left is equivalent to simply choosing the first WordNet sense. Notably, such a simple configuration of the model outperforms all but two other of the other parameter settings.

When any sort of nonzero sense co-occurrence parameter is used with $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{WN}}$, the performance degrades dramatically, to 48.8% and 47.0% for $\boldsymbol{\lambda}^{\text{SF}}$ and $\boldsymbol{\lambda}^{\text{DSF}}$ respectively. Since the discounting scheme was devised to positively interact with $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}$, it is no surprise that it does poorly when $\boldsymbol{\kappa}$ is not set in such a way. And as mentioned previously, naïvely setting $\boldsymbol{\lambda}$ to $\boldsymbol{\lambda}^{\text{SF}}$ improperly conflates $\boldsymbol{\lambda}$ and $\boldsymbol{\kappa}$, yielding a poor result.

When $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{MFS}}$ is used, the precision is even lower, dropping to 45.8% when no sense co-occurrence information is used. And similarly to $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\text{WN}}$, any nonzero $\boldsymbol{\lambda}$ significantly degrades performance. This seems to indicate the most-frequent synset, as predicted by our earlier analysis, is an inferior technique.

Finally, when $\boldsymbol{\kappa} = \hat{\boldsymbol{\kappa}}$ is used (i.e. dominant sense), the precision is 51.2%, slightly lower than but nearly on par with that of the baseline. When sense co-occurrence parameters are added, the performance increases. For $\boldsymbol{\lambda}^{\text{SF}}$, a precision of 52.5% is achieved; a precision above the baseline. But again, because of the interaction between $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$, here we expect it to be possible to improve upon this performance.

And indeed, when $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{\text{DSF}}$, the highest value of the entire table, 54.2% is achieved. This is a significant improvement over the baseline and demonstrates that our intuitively appealing mutual information discounting mechanism allows for $\boldsymbol{\kappa}$ and $\boldsymbol{\lambda}$ to work cooperatively.

5.2 Coarse-grained results (Task 7)

In order to perform the coarse-grained task, our system first determined the set of sense equivalence classes. We denote a sense equivalence class by $\overline{k}$, where k is some sense key member of the class. The equivalence classes were created according to the following constraints:

- Each sense key k may only belong to one equivalence class $\overline{k}$.
- All sense keys referring to the same sense s must belong in the same class.
- All sense keys clustered together must belong in the same class.

Once the clustering is complete, we can proceed exactly as we did in the previous sections, while replacing all instances of s with $\overline{k}$. Thus, training in this case was performed on a SemCor where all

the senses were mapped back to their corresponding sense equivalence classes.

The model fared considerably worse on the coarse-grained all-words task. The precision of the system as given by the scorer was 69.7% and the recall 62.8%. These results, while naturally much higher than those for the fine-grained test, are low by coarse-grained standards. While the gold standard was not available for comparison for these results, there are two likely causes of the lower performance on this task.

The first is that ties were not adjudicated by choosing the first WordNet sense. Instead, an arbitrary sense was chosen thereby pushing cases in which the model is unsure from the baseline to the much lower random precision rate. The second is the same number of documents are mapped to a smaller number of "senses" (i.e. sense equivalence classes), the number of parameters is greatly reduced. Therefore, the expressive power of each parameter is diluted because it must be spread out across all senses within the equivalence class.

We believe that both of these issues can be easily overcome and we hope to do so in future work. Furthermore, while the model currently captures the most salient features for word sense disambiguation, namely word-sense occurrence and sense-sense co-occurrence, it would be simple to extend the model to include a larger number of features (e.g. syntactic features).

6 Conclusion

In summary, this paper described our participation in the the SemEval-2007 coarse- and fine-grained all-words tasks. In particular, we described an exponential family model of word sense amenable to the task of word sense disambiguation. The performance of the model under a variety of parameter settings was evaluated on both tasks and the model was shown to be particularly effective on the fine-grained task.

7 Acknowledgments

The authors would like to thank Christiane Fellbaum, Daniel Osherson, and the members of the CIMPL group for their helpful contributions. This research was supported by a grant from Google Inc. and by NSF grant CCR-0325463.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Lawrence D. Brown. 1986. *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA.

Christiane Fellbaum and George A. Miller. 2003. Morphosemantic links in WordNet. *Traitement automatique de langue*.

Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Michael I. Jordan. 2004. Graphical models. *Statistical Science*, 19(1):140–155.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence table of contents*, pages 691–696.

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

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

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *COLING-ACL 2006*, pages 105–112, July.