POS induction with distributional and morphological information using a distance-dependent Chinese restaurant process

Kairit Sirts
Institute of Cybernetics at
Tallinn University of Technology
sirts@ioc.ee

Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology
jacobe@gatech.edu

Micha Elsner
Department of Linguistics
The Ohio State University
melsner0@gmail.com

Sharon Goldwater
ILCC, School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

Abstract

We present a new approach to inducing the syntactic categories of words, combining their distributional and morphological properties in a joint nonparametric Bayesian model based on the distance-dependent Chinese Restaurant Process. The prior distribution over word clusterings uses a log-linear model of morphological similarity; the likelihood function is the probability of generating vector word embeddings. The weights of the morphology model are learned jointly while inducing part-of-speech clusters, encouraging them to cohere with the distributional features. The resulting algorithm outperforms competitive alternatives on English POS induction.

1 Introduction

The morphosyntactic function of words is reflected in two ways: their distributional properties, and their morphological structure. Each information source has its own advantages and disadvantages. Distributional similarity varies smoothly with syntactic function, so that words with similar syntactic functions should have similar distributional properties. In contrast, there can be multiple paradigms for a single morphological inflection (such as past tense in English). But accurate computation of distributional similarity requires large amounts of data, which may not be available for rare words; morphological rules can be applied to any word regardless of how often it appears.

These observations suggest that a general approach to the induction of syntactic categories should leverage both distributional and morphological features (Clark, 2003; Christodoulopoulos et al., 2010). But these features are difficult to combine because of their disparate representations. Distributional information is typically represented in numerical vectors, and recent work has demonstrated the utility of continuous vector representations, or "embeddings" (Mikolov et al., 2013; Luong et al., 2013; Kim and de Marneffe, 2013; Turian et al., 2010). In contrast, morphology is often represented in terms of sparse, discrete features (such as morphemes), or via pairwise measures such as string edit distance. Moreover, the mapping between a surface form and morphology is complex and nonlinear, so that simple metrics such as edit distance will only weakly approximate morphological similarity.

In this paper we present a new approach for inducing part-of-speech (POS) classes, combining morphological and distributional information in a non-parametric Bayesian generative model based on the *distance-dependent Chinese restaurant process* (ddCRP; Blei and Frazier, 2011). In the ddCRP, each data point (word type) selects another point to "follow"; this chain of following links corresponds to a partition of the data points into clusters. The probability of word w_1 following w_2 depends on two factors: 1) the *distributional* similarity between all words in the proposed partition containing w_1 and w_2, which is encoded using a Gaussian likelihood function over the word embeddings; and 2) the *morphological* similarity between w_1 and w_2, which acts as a prior distribution on the induced clustering. We use a log-linear model to capture suffix similarities between words, and learn the feature weights by iterating between sampling and weight learning.

We apply our model to the English section of the the Multext-East corpus (Erjavec, 2004) in order to evaluate both against the coarse-grained and

fine-grained tags, where the fine-grained tags encode detailed morphological classes. We find that our model effectively combines morphological features with distributional similarity, outperforming comparable alternative approaches.

2 Related work

Unsupervised POS tagging has a long history in NLP. This paper focuses on the POS induction problem (i.e., no tag dictionary is available), and here we limit our discussion to very recent systems. A review and comparison of older systems is provided by Christodoulopoulos et al. (2010), who found that imposing a one-tag-per-word-type constraint to reduce model flexibility tended to improve system performance; like other recent systems, we impose that constraint here. Recent work also shows that the combination of morphological and distributional information yields the best results, especially cross-linguistically (Clark, 2003; Berg-Kirkpatrick et al., 2010). Since then, most systems have incorporated morphology in some way, whether as an initial step to obtain prototypes for clusters (Abend et al., 2010), or as features in a generative model (Lee et al., 2010; Christodoulopoulos et al., 2011; Sirts and Alumäe, 2012), or a representation-learning algorithm (Yatbaz et al., 2012). Several of these systems use a small fixed set of orthographic and/or suffix features, sometimes obtained from an unsupervised morphological segmentation system (Abend et al., 2010; Lee et al., 2010; Christodoulopoulos et al., 2011; Yatbaz et al., 2012). Blunsom and Cohn's (2011) model learns an n-gram character model over the words in each cluster; we learn a log-linear model, which can incorporate arbitrary features. Berg-Kirkpatrick et al. (2010) also include a log-linear model of morphology in POS induction, but they use morphology in the likelihood term of a parametric sequence model, thereby encouraging all elements that share a tag to have the same morphological features. In contrast, we use *pairwise morphological similarity* as a prior in a non-parametric clustering model. This means that the membership of a word in a cluster requires only morphological similarity to some other element in the cluster, not to the cluster centroid; which may be more appropriate for languages with multiple morphological paradigms. Another difference is that our non-parametric formulation makes it unnecessary to know the number of tags in advance.

3 Distance-dependent CRP

The ddCRP (Blei and Frazier, 2011) is an extension of the CRP; like the CRP, it defines a distribution over partitions ("table assignments") of data points ("customers"). Whereas in the regular CRP each customer chooses a table with probability proportional to the number of customers already sitting there, in the ddCRP each customer chooses another *customer* to follow, and sits at the same table with that customer. By identifying the connected components in this graph, the ddCRP equivalently defines a prior over clusterings.

If c_i is the index of the customer followed by customer i, then the ddCRP prior can be written

$$P(c_i = j) \propto \begin{cases} f(d_{ij}) & \text{if } i \neq j \\ \alpha & \text{if } i = j, \end{cases} \quad (1)$$

where d_{ij} is the distance between customers i and j and f is a decay function. A ddCRP is *sequential* if customers can only follow previous customers, i.e., $d_{ij} = \infty$ when $i > j$ and $f(\infty) = 0$. In this case, if $d_{ij} = 1$ for all $i < j$ then the ddCRP reduces to the CRP.

Separating the distance and decay function makes sense for "natural" distances (e.g., the number of words between word i and j in a document, or the time between two events), but they can also be collapsed into a single similarity function. We wish to assign higher similarities to pairs of words that share meaningful suffixes. Because we do not know which suffixes are meaningful *a priori*, we use a maximum entropy model whose features include all suffixes up to length three that are shared by at least one pair of words. Our prior is then:

$$P(c_i = j|\mathbf{w}, \alpha) \propto \begin{cases} e^{\mathbf{w}^\top \mathbf{g}(i,j)} & \text{if } i \neq j \\ \alpha & \text{if } i = j, \end{cases} \quad (2)$$

where $g_s(i, j)$ is 1 if suffix s is shared by ith and jth words, and 0 otherwise.

We can create an infinite mixture model by combining the ddCRP prior with a likelihood function defining the probability of the data given the cluster assignments. Since we are using continuous-valued vectors (word embeddings) to represent the distributional characteristics of words, we use a multivariate Gaussian likelihood. We will marginalize over the mean μ and covariance Σ of each cluster, which in turn are drawn from Gaussian and inverse-Wishart (IW) priors respectively:

$$\Sigma \sim IW(\nu_0, \Lambda_0) \qquad \mu \sim \mathcal{N}(\mu_0, {}^{\Sigma}/_{\kappa_0}) \quad (3)$$

The full model is then:

$$P(\mathbf{X},\mathbf{c},\boldsymbol{\mu},\boldsymbol{\Sigma}|\Theta,\mathbf{w},\alpha) \quad (4)$$
$$= \prod_{k=1}^{K} P(\Sigma_k|\Theta)p(\mu_k|\Sigma_k,\Theta)$$
$$\times \prod_{i=1}^{n} (P(c_i|\mathbf{w},\alpha)P(\mathbf{x}_i|\mu_{z_i},\Sigma_{z_i})),$$

where Θ are the hyperparameters for $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and z_i is the (implicit) cluster assignment of the ith word $\mathbf{x}_i$. With a CRP prior, this model would be an infinite Gaussian mixture model (IGMM; Rasmussen, 2000), and we will use the IGMM as a baseline.

4 Inference

The Gibbs sampler for the ddCRP integrates over the Gaussian parameters, sampling only follower variables. At each step, the follower link c_i for a single customer i is sampled, which can implicitly shift the entire block of n customers fol(i) who follow i into a new cluster. Since we marginalize over the cluster parameters, computing $P(c_i = j)$ requires computing the likelihood $P(\text{fol}(i), \mathbf{X}_j|\Theta)$, where $\mathbf{X}_j$ are the k customers already clustered with j. However, if we do *not* merge fol(i) with $\mathbf{X}_j$, then we have $P(\mathbf{X}_j|\Theta)$ in the overall joint probability. Therefore, we can decompose $P(\text{fol}(i), \mathbf{X}_j|\Theta) = P(\text{fol}(i)|\mathbf{X}_j,\Theta)P(\mathbf{X}_j|\Theta)$ and need only compute the change in likelihood due to merging in fol(i):[1]:

$$P(\text{fol}(i)|\mathbf{X}_j,\Theta) = \pi^{-nd/2}\frac{\kappa_k^{d/2}|\Lambda_k|^{\nu_k/2}}{\kappa_{n+k}^{d/2}|\Lambda_{n+k}|^{\nu_{n+k}/2}}$$
$$\times \prod_{i=1}^{d}\frac{\Gamma\left(\frac{\nu_{n+k}+1-i}{2}\right)}{\Gamma\left(\frac{\nu_k+1-i}{2}\right)}, \quad (5)$$

where the hyperparameters are updated as $\kappa_n = \kappa_0 + n$, $\nu_n = \nu_0 + n$, and

$$\mu_n = \frac{\kappa_0\mu_0 + \bar{x}}{\kappa_0 + n} \quad (6)$$
$$\Lambda_n = \Lambda_0 + Q + \kappa_0\mu_0\mu_0{}^T - \kappa_n\mu_n\mu_n^T, \quad (7)$$

where $Q = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$.

Combining this likelihood term with the prior, the probability of customer i following j is

$$P(c_i = j|\mathbf{X},\Theta,\mathbf{w},\alpha)$$
$$\propto P(\text{fol}(i)|\mathbf{X}_j,\Theta)P(c_i = j|\mathbf{w},\alpha). \quad (8)$$

Our non-sequential ddCRP introduces cycles into the follower structure, which are handled in the sampler as described by Socher et al. (2011). Also, the block of customers being moved around can potentially be very large, which makes it easy for the likelihood term to swamp the prior. In practice we found that introducing an additional parameter a (used to exponentiate the prior) improved results—although we report results without this exponent as well. This technique was also used by Titov and Klementiev (2012) and Elsner et al. (2012).

Inference also includes optimizing the feature weights for the log-linear model in the ddCRP prior (Titov and Klementiev, 2012). We interleave L-BFGS optimization within sampling, as in Monte Carlo Expectation-Maximization (Wei and Tanner, 1990). We do not apply the exponentiation parameter a when training the weights because this procedure affects the follower structure only, and we do not have to worry about the magnitude of the likelihood. Before the first iteration we initialize the follower structure: for each word, we choose randomly a word to follow from amongst those with the longest shared suffix of up to 3 characters. The number of clusters starts around 750, but decreases substantially after the first sampling iteration.

5 Experiments

Data For our experiments we used the English word embeddings from the Polyglot project (Al-Rfou' et al., 2013)[2], which provides embeddings trained on Wikipedia texts for 100,000 of the most frequent words in many languages.

We evaluate on the English part of the Multext-East (MTE) corpus (Erjavec, 2004), which provides both coarse-grained and fine-grained POS labels for the text of Orwell's "1984". Coarse labels consist of 11 main word classes, while the fine-grained tags (104 for English) are sequences of detailed morphological attributes. Some of these attributes are not well-attested in English (e.g. gender) and some are mostly distinguishable via semantic analysis (e.g. 1st and 2nd person verbs). Many tags are assigned only to one or a few words. Scores for the fine-grained tags will be lower for these reasons, but we argue below that they are still informative.

Since Wikipedia and MTE are from different domains their lexicons do not fully overlap; we

[1] http://www.stats.ox.ac.uk/~teh/research/notes/GaussianInverseWishart.pdf

[2] https://sites.google.com/site/rmyeid/projects/polyglot

Wikipedia tokens	1843M
Multext-East tokens	118K
Multext-East types	9193
Multext-East & Wiki types	7540

Table 1: Statistics for the English Polyglot word embeddings and English part of MTE: number of Wikipedia tokens used to train the embeddings, number of tokens/types in MTE, and number of types shared by both datasets.

take the intersection of these two sets for training and evaluation. Table 1 shows corpus statistics.

Evaluation With a few exceptions (Biemann, 2006; Van Gael et al., 2009), POS induction systems normally require the user to specify the number of desired clusters, and the systems are evaluated with that number set to the number of tags in the gold standard. For corpora such as MTE with both fine-grained and coarse-grained tages, previous evaluations have scored against the coarse-grained tags. Though coarse-grained tags have their place (Petrov et al., 2012), in many cases the distributional and morphological distinctions between words are more closely aligned with the fine-grained tagsets, which typically distinguish between verb tenses, noun number and gender, and adjectival scale (comparative, superlative, etc.), so we feel that the evaluation against fine-grained tagset is more relevant here. For better comparison with previous work, we also evaluate against the coarse-grained tags; however, these numbers are not strictly comparable to other scores reported on MTE because we are only able to train and evaluate on the subset of words that also have Polyglot embeddings. To provide some measure of the difficulty of the task, we report baseline scores using K-means clustering, which is relatively strong baseline in this task (Christodoulopoulos et al., 2011).

There are several measures commonly used for unsupervised POS induction. We report greedy one-to-one mapping accuracy (1-1) (Haghighi and Klein, 2006) and the information-theoretic score V-measure (V-m), which also varies from 0 to 100% (Rosenberg and Hirschberg, 2007). In previous work it has been common to also report many-to-one (m-1) mapping but this measure is particularly sensitive to the number of induced clusters (more clusters yield higher scores), which is variable for our models. V-m can be somewhat sensitive to the number of clusters (Reichart and Rappoport, 2009) but much less so than m-1 (Christodoulopoulos et al., 2010). With different number of induced and gold standard clusters the 1-1 measure suffers because some induced clusters cannot be mapped to gold clusters or vice versa. However, almost half the gold standard clusters in MTE contain just a few words and we do not expect our model to be able to learn them anyway, so the 1-1 measure is still useful for telling us how well the model learns the bigger and more distinguishable classes.

In unsupervised POS induction it is standard to report accuracy on tokens even when the model itself works on types. Here we report also type-based measures because these can reveal differences in model behavior even when token-based measures are similar.

Experimental setup For baselines we use K-means and the IGMM, which both only learn from the word embeddings. The CRP prior in the IGMM has one hyperparameter (the concentration parameter α); we report results for $\alpha = 5$ and 20. Both the IGMM and ddCRP have four hyperparameters controlling the prior over the Gaussian cluster parameters: Λ_0, μ_0, ν_0 and κ_0. We set the prior scale matrix Λ_0 by using the average covariance obtained from a K-means run with $K = 200$. When setting the average covariance as the expected value of the IW distribution the suitable scale matrix can be computed as $\Lambda_0 = E[X](\nu_0 - d - 1)$, where ν_0 is the prior degrees of freedom (which we set to d + 10) and d is the data dimensionality (64 for the Polyglot embeddings). We set the prior mean μ_0 equal to the sample mean of the data and κ_0 to 0.01.

We experiment with three different priors for the ddCRP model. All our ddCRP models are non-sequential (Socher et al., 2011), allowing cycles to be formed. The simplest model, *ddCRP uniform*, uses a uniform prior that sets the distance between any two words equal to one.[3] The second model, *ddCRP learned*, uses the log-linear prior with weights learned between each two Gibbs iterations as explained in section 4. The final model, *ddCRP exp*, adds the prior exponentiation. The α parameter for the ddCRP is set to 1 in all experiments. For *ddCRP exp*, we report results with the exponent a set to 5.

[3] In the sequential case this model would be equivalent to the IGMM (Blei and Frazier, 2011). Due to the non-sequentiality this equivalence does not hold, but we do expect to see similar results to the IGMM.

		Fine types		Fine tokens		Coarse tokens	
Model	**K**	Model	K-means	Model	K-means	Model	K-means
K-means	104 or 11	16.1 / 47.3	-	39.2 / 62.0	-	44.4 / 45.5	-
IGMM, $\alpha = 5$	55.6	41.0 / 45.9	23.1 / 49.5	48.0 / 64.8	37.2 / 61.0	48.3 / 58.3	40.8 / 55.0
IGMM, $\alpha = 20$	121.2	35.0 / 47.1	14.7 / 46.9	50.6 / 67.8	44.7 / 65.5	48.7 / 60.0	48.3 / 57.9
ddCRP uniform	80.4	50.5 / 52.9	18.6 / 48.2	52.4 / 68.7	35.1 / 60.3	**52.1 / 62.2**	40.3 / 54.2
ddCRP learned	89.6	50.1 / 55.1	17.6 / 48.0	51.1 / **69.7**	39.0 / 63.2	48.9 / 62.0	41.1 / 55.1
ddCRP exp, $a = 5$	47.2	**64.0 / 60.3**	25.0 / 50.3	**55.1** / 66.4	33.0 / 59.1	47.8 / 55.1	36.9 / 53.1

Table 2: Results of baseline and ddCRP models evaluated on word types and tokens using fine-grained tags, and on tokens using coarse-grained tags. For each model we present the number of induced clusters K (or fixed K for K-means) and 1-1 / V-m scores. The second column under each evaluation setting gives the scores for K-means with K equal to the number of clusters induced by the model in that row.

Results and discussion Table 2 presents all results. Each number is an average of 5 experiments with different random initializations. For each evaluation setting we provide two sets of scores—first are the 1-1 and V-m scores for the given model, second are the comparable scores for K-means run with the same number of clusters as induced by the non-parametric model.

These results show that all non-parametric models perform better than K-means, which is a strong baseline in this task (Christodoulopoulos et al., 2011). The poor performace of K-means can be explained by the fact that it tends to find clusters of relatively equal size, although the POS clusters are rarely of similar size. The common noun singular class is by far the largest in English, containing roughly a quarter of the word types. Non-parametric models are able to produce cluster of different sizes when the evidence indicates so, and this is clearly the case here.

From the token-based evaluation it is hard to say which IGMM hyperparameter value is better even though the number of clusters induced differs by a factor of 2. The type-base evaluation, however, clearly prefers the smaller value with fewer clusters. Similar effects can be seen when comparing IGMM and ddCRP uniform. We expected these two models perform on the same level, and their token-based scores are similar, but on the type-based evaluation the ddCRP is clearly superior. The difference could be due to the non-sequentiality, or becuase the samplers are different—IGMM enabling resampling only one item at a time, ddCRP performing blocked sampling.

Further we can see that the ddCRP uniform and learned perform roughly the same. Although the prior in those models is different they work mainly using the the likelihood. The ddCRP with learned prior does produce nice follower structures within each cluster but the prior is in general too weak compared to the likelihood to influence the clustering decisions. Exponentiating the prior reduces the number of induced clusters and improves results, as it can change the cluster assignment for some words where the likelihood strongly prefers one cluster but the prior clearly indicates another.

The last column shows the token-based evaluation against the coarse-grained tagset. This is the most common evaluation framework used previously in the literature. Although our scores are not directly comparable with the previous results, our V-m scores are similar to the best published 60.5 (Christodoulopoulos et al., 2010) and 66.7 (Sirts and Alumäe, 2012).

In preliminary experiments, we found that directly applying the best-performing English model to other languages is not effective. Different languages may require different parametrizations of the model. Further study is also needed to verify that word embeddings effectively capture syntax across languages, and to determine the amount of unlabeled text necessary to learn good embeddings.

6 Conclusion

This paper demonstrates that morphology and distributional features can be combined in a flexible, joint probabilistic model, using the distance-dependent Chinese Restaurant Process. A key advantage of this framework is the ability to include arbitrary features in the prior distribution. Future work may exploit this advantage more thoroughly: for example, by using features that incorporate prior knowledge of the language's morphological structure. Another important goal is the evaluation of this method on languages beyond English.

Acknowledgments: KS was supported by the Tiger University program of the Estonian Information Technology Foundation for Education. JE was supported by a visiting fellowship from the Scottish Informatics & Computer Science Alliance. We thank the reviewers for their helpful feedback.

References

Omri Abend, Roi Reichart, and Ari Rappoport. 2010. Improved unsupervised pos induction through prototype discovery. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 1298–1307.

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Thirteenth Annual Conference on Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Alexandre B. Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590.

Chris Biemann. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 7–12.

David M Blei and Peter I Frazier. 2011. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.

Phil Blunsom and Trevor Cohn. 2011. A hierarchical pitman-yor process hmm for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics*, pages 865–874.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2011. A Bayesian mixture model for part-of-speech induction using multiple features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the European chapter of the ACL*.

Micha Elsner, Sharon Goldwater, and Jacob Eisenstein. 2012. Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics*.

Tomaž Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *LREC*.

A. Haghighi and D. Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2010. Simple type-level unsupervised pos tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 853–861.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Thirteenth Annual Conference on Natural Language Learning*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 746–751.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, May.

Carl Rasmussen. 2000. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA. MIT Press.

Roi Reichart and Ari Rappoport. 2009. The nvi clustering evaluation measure. In *Proceedings of the Ninth Annual Conference on Natural Language Learning*, pages 165–173.

A. Rosenberg and J. Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–42.

Kairit Sirts and Tanel Alumäe. 2012. A hierarchical Dirichlet process model for joint part-of-speech and morphology induction. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 407–416.

Richard Socher, Andrew L Maas, and Christopher D Manning. 2011. Spectral chinese restaurant processes: Nonparametric clustering based on similarities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 698–706.

Ivan Titov and Alexandre Klementiev. 2012. A bayesian approach to unsupervised semantic role induction. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 678–687, Singapore.

Greg CG Wei and Martin A Tanner. 1990. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Mehmet Ali Yatbaz, Enis Sert, and Deniz Yuret. 2012. Learning syntactic categories using paradigmatic representations of word context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 940–951.