Evaluating unsupervised learning for natural language processing tasks

Andreas Vlachos
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
vlachos@biostat.wisc.edu

Abstract

The development of unsupervised learning methods for natural language processing tasks has become an important and popular area of research. The primary advantage of these methods is that they do not require annotated data to learn a model. However, this advantage makes them difficult to evaluate against a manually labeled gold standard. Using unsupervised part-of-speech tagging as our case study, we discuss the reasons that render this evaluation paradigm unsuitable for the evaluation of unsupervised learning methods. Instead, we argue that the rarely used *in-context* evaluation is more appropriate and more informative, as it takes into account the way these methods are likely to be applied. Finally, bearing the issue of evaluation in mind, we propose directions for future work in unsupervised natural language processing.

1 Introduction

The development of unsupervised learning methods for natural language processing (NLP) tasks has become an important and popular area of research. The main attraction of these methods is that they can learn a model using only unlabeled data. This is an important advantage, as unlabeled text in digital form is in abundance, while labeled datasets are usually expensive to construct. While methods such as crowdsourcing (Snow et al., 2008) can help reduce this cost, in tasks for which specialist knowledge is required, such as part-of-speech (PoS) tagging or syntactic parsing, labeling datasets in this fashion can be substantially harder.

Nevertheless, the advantage of requiring only unlabeled data to learn a model renders the evaluation of unsupervised learning methods to be more challenging than that of their supervised counterparts. This is primarily because the output of unsupervised methods does not contain labels that would be found in a manually constructed gold standard. Simplistically expressed, no labels for model learning means that there are no labels in the output. As a result, the standard evaluation paradigm of comparing against a gold standard using a performance measure such as accuracy or F-score cannot be used, at least not in the way it would be used in evaluating supervised methods. Since methods are proposed or rejected by researchers, and papers describing these methods are assessed by their peers partly on the basis of such results, the issue of evaluation is an important one.

Before we proceed, it is important to characterize the unsupervised learning methods we are considering, as the term unsupervised is used in multiple ways in the literature. In this work we focus on methods that use only unlabeled data to learn a model and do not involve any form of supervision at any stage. Thus we exclude methods that use seeds such as the dictionaries of PoS tags used by Ravi and Knight (2009) and rules for producing labeled output, e.g. those proposed by Teichert and Daumé III (2009). We also exclude methods for which the data used to learn a model does not contain any of the labels we are learning to predict, but it does contain other information that we use in the learning process. For example, the multilingual PoS induction approach of Das and Petrov (2011) assumes no supervision for the language whose PoS tags are being

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 35–42,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

induced, but it assumes access to a labeled dataset of a different language.

We begin by surveying recent work on unsupervised PoS tagging, focusing on the issue of evaluation (Section 2). While PoS tagging is not the only task for which unsupervised learning methods are popular, its relative simplicity and the variety of evaluation paradigms employed make it a useful case study. Based on this survey, we show that evaluation against a PoS tagging gold standard is not only difficult, but it can be misleading as well. The reason for this is that the unsupervised learning methods used, while they produce output that correlates with PoS tags, perform a different task, namely clustering-based word representation induction (Turian et al., 2010). Instead, we argue that *in-context* evaluation is more appropriate and more informative, as it takes into account the application context in which these methods are intended to be used (Section 3). Finally, bearing the issue of evaluation in mind, we propose some directions for future work in unsupervised learning for NLP (Section 4).

2 The case of unsupervised part-of-speech tagging

PoS tagging is the task of assigning lexical categories such as noun or verb to tokens in a sentence. It is commonly used either as an end-goal or as intermediate processing stage for a downstream task such as syntactic parsing. For languages with substantial amounts of labeled data available such as English, the performance of supervised approaches has reached very high levels.[1] Thus, the research focus has shifted to semisupervised and unsupervised approaches which would allow the processing of languages which do not have similar resources available.

At an abstract level, the unsupervised learning methods applied to PoS tagging take as input tokenized unlabeled sentences, from which they learn a model. These models are either hidden Markov models (HMMs) (Clark, 2003; Goldwater and Griffiths, 2007) or clustering models (Biemann, 2006; Abend et al., 2010). During model learning, state identifiers are assigned to the tokens (Figure 1a). Independently of the learning method and the model, these identifiers are semantically void, i.e. they have no linguistic meaning. Nevertheless, all the studies conclude that there is a strong correlation between the state identifiers assigned and the PoS tags in a labeled gold standard (Figure 1b).

The most common way of assessing the level of correlation achieved is the use clustering evaluation measures. The latter operate on a confusion matrix (Figure 1c), which is constructed by assuming that each cluster consists of all the tokens assigned the same state identifier. Intuitively, all clustering evaluation measures provide definitions for the two desirable properties that a good clustering should possess with respect to a gold standard, homogeneity and completeness. Homogeneity represents the degree to which each cluster contains instances from a single gold standard class, while completeness the degree to which each gold standard class is contained in a single cluster. Note that there tends to be a trade-off between these two properties since, increasing the number of clusters is likely to improve homogeneity but worsen completeness and vice-versa. Therefore, clustering evaluation measures need to balance appropriately between them.

Some authors proposed clustering evaluation techniques that first induce the mapping from state identifiers to gold standard tags automatically and then use supervised measures to compare the mapped output to the gold standard. For example, Gao and Johnson (2008) proposed to induce a many-to-one mapping of state identifiers to PoS tags from one half of the corpus and evaluate on the second half, which is referred to as cross-validation accuracy. However, such techniques evaluate the clustering together with the induced mapping, thus the quality of the latter influences the results obtained. This can be misleading as unsupervised learning methods for PoS tagging induce the clustering, but not the mapping on which they are eventually evaluated.

In order to avoid the mapping induction step, the use of information theoretic measures was proposed instead. These include Variation of Information (VI) (Meilă, 2007), V-measure (Rosenberg and Hirschberg, 2007), and their respective variants NVI (Reichart and Rappoport, 2009) and V-beta (Vlachos et al., 2009). Each of these measures exhibits

[1] According to the ACL wiki, state-of-the-art performance in English is more than 97% per token accuracy.

1	*2*	*3*	*4*	*1*	*5*
There	are	70	children	there	.

(a) Unsupervised PoS tagger output

EX	*VBP*	*CD*	*NNS*	*RB*	*.*
There	are	70	children	there	.

(b) Gold standard

	1	2	3	4	5
EX	1	0	0	0	0
VBP	0	1	0	0	0
CD	0	0	0	1	0
NNS	0	0	1	0	0
RB	1	0	0	0	0
.	0	0	0	0	1

(c) Confusion matrix

Figure 1: Unsupervised PoS tagging evaluation pipeline.

some kind of bias towards certain solutions though, e.g. V-measure favors clusterings with large number of clusters, while VI exhibits the opposite behavior. While these biases might follow some reasonable intuitions, unsurprisingly none is universally accepted as the most appropriate.

In order to avoid these problems, Biemann et al. (2007) proposed to evaluate unsupervised PoS tagging as a source of features for supervised learning approaches to NLP tasks, such as named entity recognition and shallow parsing. The intuition behind this extrinsic evaluation is that if a task relies on discriminating between PoS labels rather than the PoS labels semantics themselves, then the state identifiers obtained by an unsupervised method can be used in the same way as PoS tags obtained from a gold standard or a supervised system. In their experiments they showed that the features obtained from the unsupervised PoS tagger improve the performance in all tasks, and in particular when little training data is available.

Van Gael et al. (2009) evaluated the output of different configurations of their unsupervised PoS tagging approach both by comparing it against a gold standard via clustering evaluation measures and by using it as a source of features for shallow parsing. Table 1 summarizes the results of their experiments. In agreement with Biemann et al. (2007), they found that the features provided by the unsupervised PoS tagger improved shallow parsing performance. However, they observed that the clustering evaluation scores did not correlate with the results of this extrinsic evaluation. In other words, better clustering evaluation scores did not always result in better features for shallow parsing. Van Gael et al. noted that homogeneity correlated better with shallow parsing performance, hypothesizing it is probably worse to assign the same state identifier to tokens that belong to different PoS tags, e.g. verb and adverbs, rather than to generate more than one state identifier for the same PoS. In the same spirit, Christodoulopoulos et al. (2010) used the output of a number of unsupervised PoS tagging methods to extract seeds for the prototype-driven model of Haghighi and Klein (2006). Like Van Gael et al., they also found that better clustering evaluation scores did not result in better seeds.

Given these results, as well as remembering that unsupervised learning methods do not use any label information in model learning, one is entitled to question whether it is reasonable to expect their output to match a particular labeled gold standard. Why not assume that the state identifiers obtained correlate with named entity recognition tags or categorial grammar tags instead of PoS tags, tasks for which sequential models are very common? Even if the state identifiers induced correlate better with PoS tags than with other kinds of annotation, evaluating them using a PoS tagging gold standard and even naming the task unsupervised PoS tagging or induction is probably misleading. We argue that the task performed by the unsupervised PoS tagging methods proposed is more accurately described as clustering-based word representation induction

	homogeneity	completeness	VI	V-measure	V-beta	F-score	accuracy
DP-learned	69.39	51.21	4.19	58.93	55.37	90.98	94.48
DP-fixed	51.80	54.84	3.94	53.27	52.88	89.99	93.89
PY-fixed	62.02	56.25	3.74	59.00	58.79	90.31	94.15
no PoS	-	-	-	-	-	93.81	96.07
supervised PoS	-	-	-	-	-	88.58	93.25

Table 1: Summary the results reported for the three configurations (DP-learned, DP-fixed, PY-fixed) of the unsupervised PoS tagger of Van Gael et al. (2009) and the two baselines (no PoS tags, supervised PoS tags). Except for VI, higher scores mean better performance. The clustering evaluation scores (VI, V-measure, V-beta) are obtained by comparing against a PoS gold standard, while F-score and accuracy scores are obtained by extrinsinc evaluation using shallow parsing.

(Turian et al., 2010), and that this should be taken into account in the evaluation. As further evidence of the relation between the two tasks, note that some of the unsupervised PoS tagging methods applied by Christodoulopoulos et al. (2010) were also used by Turian et al. (2010) for clustering-based word representation induction.

3 *In-context* evaluation

All the papers on unsupervised PoS tagging mentioned in the previous section agree on the fact that its evaluation, at least using clustering evaluation measures, is difficult. This is an important problem for other NLP tasks (e.g. anaphora resolution, word sense induction) in which systems produce clusters that need to be mapped to gold standard classes. In their recent position paper, Guyon et al. (2009) argue that the problem lies in ignoring the *context* in which clustering is performed. They distinguish between two such contexts. The first one is the use of clustering as a *pre-processing* step for a downstream task, in which the evaluation of the latter is used to evaluate the former. The second context is that of data *exploration* in order to assist a human to analyze a large dataset. In this case, performance might not be as straightforward to assess, since it relies on many external factors among which the human computer interaction interface used is likely to be crucial. We cumulatively refer to these evaluation paradigms as *in-context* evaluation.

Returning to unsupervised PoS tagging and NLP, the extrinsic evaluation of Biemann et al. (2007) and Van Gael et al. (2009) falls under the pre-processing paradigm. The approach of Christodoulopoulos et al. (2010) falls between pre-processing and data exploration, as the clusters of tokens produced are semi-automatically processed in order to produce seeds which were then used by the prototype-driven model of Haghighi and Klein (2006).[2] In-context evaluation can be used to assess the performance of unsupervised learning methods for tasks other than clustering-based word representation approaches. For example, topic modeling (Blei et al., 2003) has recently been used and evaluated in approaches to learning models of selectional preferences (Ritter et al., 2010; Ó Séaghdha, 2010).

The issues affecting the evaluation of unsupervised learning methods are not restricted to PoS tagging. Schwartz et al. (2011) discussed similar issues in the context of unsupervised dependency parsing. Note that some of them arise due to the fact unsupervised dependency parsing produces unlabeled directed edges which are interpreted as denoting head-dependent relations. However, there are linguistic phenomena where unless the edges are labeled with a specific interpretation, both directions could be considered correct, e.g. the relation between modal verb and main verb. Even though evaluation against a syntactic parsing gold standard is useful, we argue that in-context evaluation of the output of unsupervised dependency parsers is likely to be more informative and more appropriate.

Despite the criticism against clustering evaluation measures as well as other methods for comparing the

[2] Note that while evaluating in-context, these authors still refer to the task performed as PoS tagging or induction and some of their conclusions are drawn via comparisons against a PoS tagging gold standard.

output of unsupervised learning methods against a gold standard, we argue that they are still useful. The various measures proposed, along with their inherent biases and definitions of clustering quality, provide quantitative analysis of the behavior of unsupervised learning methods by assessing correlations between their output and a gold standard. This can be very useful when developing such methods, as their use is admittedly simpler than the in-context evaluation paradigms discussed. However, they are not as informative as in-context evaluation and they should not be used to draw strong conclusions about the usefulness of a method.

Acknowledging that the evaluation of unsupervised learning for NLP is better performed in-context instead of against a labeled gold standard leads to the use of more appropriate experimental setups. Sometimes unsupervised learning methods are restricted to learning models using the unlabeled gold standard against which they are evaluated subsequently. Thus, they neither take full advantage of nor they demonstrate their main strength, which is that they can use as much data as possible. Using the pre-processing paradigm, clustering-based word representations induced from a large unlabeled dataset would be evaluated according to whether they improve the performance of the downstream task they are evaluated with, whose evaluation is likely to be on a different dataset. This use of clustering-based word representation is sometimes referred to as semi-supervised learning and has been shown to be effective in a variety of tasks, including named entity recognition, shallow parsing and syntactic dependency parsing (Koo et al., 2008; Turian et al., 2010).

The use of large datasets would also help assess the scalability of the unsupervised methods proposed, as the amount of data that can be handled efficiently by an unsupervised method can be as important as the range of linguistic intuitions it can capture. To examine this trade-off, it would be informative to show performance curves with different amounts of data, which should be straightforward to produce under the pre-processing evaluation paradigm. An added benefit is that, as discussed by Ben-Hur et al. (2002), assessing clustering stability using multiple runs and sub-samples of a dataset can help establish whether a particular combination of clustering algorithm and user-defined parameters (including the number of clusters to be discovered) is able to discover an appropriate clustering of the dataset considered.

Avoiding comparisons against a labeled gold standard would also remove the temptation of adapting it to the output of the unsupervised learning method. For example, in unsupervised PoS tagging authors sometimes simplify the gold standard by collapsing the original 45 PoS tags of the Penn treebank to 17, e.g. by removing the distinctions between different noun tags. While such simplifications are linguistically plausible, they substitute one problem for another, as methods are no longer penalized for missing some of the finer distinctions, but they are penalized for making them. Perhaps more importantly, they result in fitting the gold standard to the output of the method being evaluated, which is unlikely to be informative.

Another related issue is that since unsupervised learning methods do not need labeled data, it is a tempting and common practice to learn a model and report results on the same dataset, which usually consists of all the labeled data available and which is used to tune the parameters of the method evaluated. This is equivalent to reporting results for supervised learning methods on the development set, while it is generally accepted that results on a separate test set on which no parameter tuning is allowed provide better performance estimates. The use of the pre-processing evaluation paradigm with a supervised learning approach for the downstream task is likely to result in use the standard distinction between training, development and test set for the evaluation of unsupervised learning methods.

4 Directions for future work

While the previous sections have focused on why unsupervised learning for NLP tasks is hard to evaluate, our intention is not to discourage further research, but to encourage it. Unsupervised learning can help exploit the large amounts of unlabeled text that are available. For this purpose though we need appropriate evaluation, and we argue that in-context evaluation is likely to be more informative than the evaluation against a gold standard.

A potential problem is that in-context evaluation adds an extra layer in the experimental setup, either in the form of a downstream task or of a human-computer interaction study. This can make comparisons between methods harder as there are more experimental conditions to control for and discourage researchers from adopting it. Therefore, it would be useful to have a shared task that would provide an experimental setup that can be re-used. Shared tasks have been beneficial in cases where the existence of multiple datasets and task definitions hindered progress and we would expect them to have a similar effect on unsupervised learning methods.

As different application contexts are likely to benefit from different solutions, this naturally leads to the development of modeling approaches that are adaptable, preferably in ways that enable experts to incorporate their knowledge. This research direction has already been pursued in clustering (Wagstaff and Cardie, 2000; Basu et al., 2006) and more recently in topic modeling (Blei and Mcauliffe, 2008; Andrzejewski et al., 2011). We argue though that the wider adoption of in-context evaluation will help assess their performance and merits in a more informative way. An alternative approach to accommodate for the needs of different application contexts is to induce multiple clusterings simultaneously for the same dataset as proposed by Dasgupta and Ng (2010) in the context of text classification. Such considerations are particularly relevant to NLP applications as language exhibits ambiguity and polysemy, which are rather difficult to capture in a context-independent labeled gold standard.

If in-context evaluation must be avoided, it is advisable to focus on tasks for which most application contexts would agree on the clustering or latent structure that must be discovered, such as the Web People Search (Artiles et al., 2010) task on clustering webpages about persons who share the same name. Even in this case though, in-context evaluation as pre-processing for an information extraction system or as a visualization component in an interface for exploring web pages is still likely to be informative.

Finally, in this paper we considered methods whose output consists of state identifiers which are semantically void. However, obtaining meaningful labels such as those found in a gold standard is a useful and important goal in many NLP tasks. However, this purpose is better served by injecting appropriate supervision to the model, instead of trying to achieve it as an afterthought. Such approaches include the use of PoS dictionaries by sequential tagging models (Haghighi and Klein, 2006; Ravi and Knight, 2009), the use of labeled data from different languages (Snyder et al., 2008; Das and Petrov, 2011) or the (possibly indirect) assignment of labels to topics (Ramage et al., 2009; Zhu et al., 2009). Research in unsupervised learning methods is likely to benefit these partially supervised ones, as they both seek to take advantage of unlabeled data. As the output of such methods uses the same labels as those found in the gold standard, they can be evaluated against a labeled gold standard.

5 Conclusions

In this position paper, we discussed the issue of evaluation of unsupervised learning methods for NLP tasks. Using PoS tagging as our case study, we examined recent attempts of evaluating unsupervised approaches and showed that a lot of confusion is caused due to evaluating their output against a labeled gold standard. Instead, we argue that it is more appropriate to evaluate unsupervised methods in context, either as a pre-processing step for a downstream task or as a tool for data exploration. Following this, we proposed that future work should focus on adapting to and evaluating unsupervised learning methods in the context in which they are intended to be used and that a shared task would facilitate research in this direction. Finally, we hope that the adoption of in-context evaluation will result in the development of improved unsupervised learning methods for NLP tasks, so that researchers and practitioners can exploit the large amounts of textual data available.

Acknowledgments

The author would like thank Mark Craven and Diarmuid Ó Séaghdha for helpful comments and discussions. The author was funded by NIH/NLM grant R01 / LM07050.

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 for Computational Linguistics*, pages 1298–1307.

David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. 2011. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *Proceedings of the 22nd International Joint Conferences on Artificial Intelligence*.

Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, and Enrique Amigó. 2010. WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation*.

Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J. Mooney. 2006. Probabilistic semi-supervised clustering with constraints. In O. Chapelle, B. Schoelkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 73–102. MIT Press.

Asa Ben-Hur, André Elisseeff, and Isabelle Guyon. 2002. A stability based method for discovering structure in clustered data. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 6–17.

Chris Biemann, Claudio Giuliano, and Alfio Gliozzo. 2007. Unsupervised part-of-speech tagging supporting supervised methods. In *Proceedings of the International Conference in Recent Advances in Natural Language Processing*.

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: Student Research Workshop*, pages 7–12.

David Blei and Jon Mcauliffe. 2008. Supervised topic models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*.

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

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: how far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 59–66.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Sajib Dasgupta and Vincent Ng. 2010. Mining clustering dimensions. In *Proceedings of the 27th International Conference on Machine Learning*, pages 263–270.

Jianfeng Gao and Mark Johnson. 2008. A comparison of Bayesian estimators for unsupervised hidden Markov model POS taggers. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 344–352.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751.

Isabelle Guyon, Ulrike Von Luxburg, and Robert C. Williamson. 2009. Clustering: Science or art. In *NIPS 2009 Workshop on Clustering Theory*.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 320–327.

Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, pages 595–603.

Marina Meilă. 2007. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 504–512.

Roi Reichart and Ari Rappoport. 2009. The NVI clustering evaluation measure. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 165–173.

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of* *the Association for Computational Linguistics*, pages 424–434.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association of Computational Linguistics*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein, and Regina Barzilay. 2008. Unsupervised multilingual learning for pos tagging. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1041–1050.

Adam R. Teichert and Hal Daumé III. 2009. Unsupervised part of speech tagging without a lexicon. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*.

Joseph Turian, Lev 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.

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

Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *Proceedings of the EACL workshop on GEometrical Models of Natural Language Semantics*.

Kiri Wagstaff and Claire Cardie. 2000. Clustering with instance-level constraints. In *Proceedings of the 17th International Conference on Machine Learning*, pages 1103–1110.

Jun Zhu, Amr Ahmed, and Eric P. Xing. 2009. MedLDA: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1257–1264.