Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM

William M. Darling
School of Computer Science
University of Guelph
wdarling@uoguelph.ca

Michael J. Paul
Dept. of Computer Science
Johns Hopkins University
mpaul@cs.jhu.edu

Fei Song
School of Computer Science
University of Guelph
fsong@uoguelph.ca

Abstract

Unsupervised part-of-speech (POS) tagging has recently been shown to greatly benefit from Bayesian approaches where HMM parameters are integrated out, leading to significant increases in tagging accuracy. These improvements in unsupervised methods are important especially in specialized social media domains such as Twitter where little training data is available. Here, we take the Bayesian approach one step further by integrating semantic information from an LDA-like topic model with an HMM. Specifically, we present *Part-of-Speech LDA* (POSLDA), a syntactically and semantically consistent generative probabilistic model. This model discovers POS specific topics from an unlabelled corpus. We show that this model consistently achieves improvements in unsupervised POS tagging and language modeling over the Bayesian HMM approach with varying amounts of side information in the noisy and esoteric domain of Twitter.

1 Introduction

The explosion of social media in recent years has led to the need for NLP tools like part-of-speech (POS) taggers that are robust enough to handle data that is becoming increasingly "noisy." Unfortunately, many NLP systems fail at out-of-domain data and struggle with the informal style of social text. With spelling errors, abbreviations, uncommon acronyms, and excessive use of slang, systems that are designed for traditional corpora such as news articles may perform poorly when given difficult input such as a Twitter feed (Ritter et al., 2010).

Recognizing the limitations of existing systems, Gimpel et al. (2011) develop a POS tagger specifically for Twitter, by creating a training corpus as well as devising a tag set that includes parts of speech that are uniquely found in online language, such as emoticons (smilies). This is an important step forward, but a POS tagger tailored to Twitter cannot tackle the social Web as a whole. Other online communities have their own styles, slang, memes, and other idiosyncrasies, so a system trained for one community may not apply to others.

For example, the 140-character limit of Twitter encourages abbreviations and word-dropping that may not be found in less restrictive venues. The first-person subject is often assumed in "status messages" that one finds in Twitter and Facebook, so the pronominal subject can be dropped, even in English (Weir, 2012), leading to messages like "Went out" instead of "I went out." Not only does Twitter follow these unusual grammatical patterns, but many messages contain "hashtags" which could be considered their own syntactic class not found in other data sources. For these reasons, POS parameters learned from Twitter data will not necessarily fit other social data.

In general, concerns about the limitations of domain-dependent models have motivated the use of sophisticated unsupervised methods. Interest in unsupervised POS induction has been revived in recent years after Bayesian HMMs are shown to increase accuracy by up to 14 percentage points over basic maximum-likelihood estimation (Goldwater and Griffiths, 2007). Despite falling well short of the accuracy obtained with supervised taggers, unsupervised approaches are preferred in situations where there is no access to

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–9,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

large quantities of training data in a specific domain, which is increasingly common with Web data. We therefore hope to continue improving accuracy with unsupervised approaches by introducing semantics as an additional source of information for this task.

The ambiguities of language are amplified through social media, where new words or spellings of words are routinely invented. For example, “ow” on Twitter can be a shorthand for “how,” in addition to its more traditional use as an expression of pain (ouch). While POS assignment is inherently a problem of **syntactic** disambiguation, we hypothesize that the underlying **semantic** content can aid the disambiguation task. If we know that the overall content of a message is about police, then the word “cop” is likely to be a noun, whereas if the context is about shopping, this could be slang for acquiring or stealing (verb). The HMM approach will often be able to tag these occurrences appropriately given the context, but in many cases the syntactic context may be limited or misleading due to the noisy nature of the data. Thus, we believe that semantic context will offer additional evidence toward making an accurate prediction.

Following this intuition, this paper presents a semantically and syntactically coherent Bayesian model that uncovers POS-specific sub-topics within general semantic topics, as in latent Dirichlet allocation (LDA) (Blei et al., 2003), which we call **part-of-speech LDA**, or POSLDA. The resulting posterior distributions will reflect specialized topics such as “verbs about dining” or “nouns about politics”. To the best of our knowledge, we also present the first experiments with unsupervised tagging for a social media corpus. In this work, we focus on Twitter because the labeled corpus by Gimpel et al. (2011) allows us to quantitatively evaluate our approach. We demonstrate the model’s utility as a predictive language model by its low perplexity on held-out test data as compared to several related topic models, and most importantly, we show that this model achieves statistically significant and consistent improvements in unsupervised POS tagging accuracy over a Bayesian HMM. These results support our hypothesis that semantic information can directly improve the quality of POS induction, and our experiments present an in-depth exploration of this task on informal social text.

The next section discusses related work, which is followed by a description of our model, POSLDA. We then present POS tagging results on the Twitter POS dataset (Gimpel et al., 2011). Section 5 describes further experiments on the POSLDA model and section 6 includes a discussion on the results and why POSLDA can do better on POS tagging than a vanilla Bayesian HMM. Finally, section 7 concludes with a discussion on future work.

2 Related Work

Modern unsupervised POS tagging originates with Merialdo (1993) who trained a trigram HMM using maximum likelihood estimation (MLE). Goldwater and Griffiths (2007) improved upon this approach by treating the HMM in a Bayesian sense; the rows of the transition matrix are random variables with proper Bayesian priors and the state emission probabilities are also random variables with their own priors. The posterior distribution of tags is learned using Gibbs sampling and this model improves in accuracy over the MLE approach by up to 14 percentage points.

In the “Topics and Syntax” model (or HMMLDA), the generative process of a corpus is cast as a composite model where syntax is modeled with an HMM and semantics are modeled with LDA (Griffiths et al., 2005). Here, one state of an HMM is replaced with a topic model such that the words with long-range dependencies (“content” words) will be drawn from a set of topics. The remaining states are reserved for “syntax” words that exhibit only short-range dependencies. Griffiths et al. (2005) briefly touch on POS tagging with their model, but its superiority to a plain Bayesian HMM is not shown and the authors note that this is partially because all semantic-like words get assigned to the single semantic class in their model. This misses the distinction between at least nouns and verbs, but many other semantic-dependent words as well. If more variation could be provided in the semantic portion of the model, the POS tagging results would likely improve.

3 Part-of-Speech LDA (POSLDA)

In their canonical form, topic models do not capture local dependencies between words (i.e. syntactic relations), but they do capture long-range

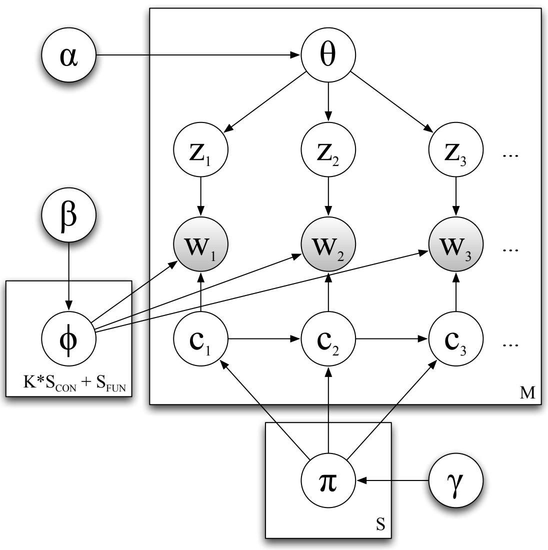

Figure 1: Graphical model depiction of POSLDA.

context such as the overall topical content or gist of a document. Conversely, under an HMM, words are assumed completely independent of their broader context by the Markov assumption. We seek to bridge these restrictions with our unified model, Part-of-Speech LDA (POSLDA).

Under this model, each word token is now associated with two latent variables: a semantic topic z and a syntactic class c. We posit that the topics are generated through the LDA process, while the classes are generated through an HMM. The observed word tokens are then dependent on both the topic and the class: rather than a single multinomial for a particular topic z or a particular class c, there are distributions for each topic-class pair (z, c) from which we assume words are sampled.

We denote the set of classes $\mathcal{C} = \mathcal{C}_{\text{CON}} \cup \mathcal{C}_{\text{FUN}}$, which includes the set of content or "semantic" classes $\mathcal{C}_{\text{CON}}$ for word types such as nouns and verbs that depend on the current topic, and functional or "syntactic-only" classes $\mathcal{C}_{\text{FUN}}$. If a word is generated from a functional class, it does not depend on the topic. This allows our model to accommodate functional words like determiners which appear independently of the topical content of a document.

We use the same notation as LDA, where θ is a document-topic distribution and ϕ is a topic-word distribution. Additionally, we denote the HMM transition rows as π, which we assume is drawn from a Dirichlet with hyperparameter γ. Denote $S = |\mathcal{C}|$ and $K = |\mathcal{Z}|$, the numbers of classes and topics, respectively. There are S_{FUN} word distributions $\phi^{(\text{FUN})}$ for function word classes and $K \times S_{\text{CON}}$ word distributions $\phi^{(\text{CON})}$ for content word classes. A graphical model depiction of POSLDA is shown in Figure 1.

Thus, the generative process of a corpus can be described as:

1. Draw $\pi \sim \text{Dirichlet}(\gamma)$
2. Draw $\phi \sim \text{Dirichlet}(\beta)$
3. For each document $d \in \mathcal{D}$:
   (a) Draw $\theta_d \sim \text{Dirichlet}(\alpha)$
   (b) For each word token $w_i \in d$:
       i. Draw $c_i \sim \pi_{c_{i-1}}$
       ii. If $c_i \notin \mathcal{C}_{\text{CON}}$:
           A. Draw $w_i \sim \phi^{(\text{FUN})}_{c_i}$
       iii. Else:
           A. Draw $z_i \sim \theta_d$
           B. Draw $w_i \sim \phi^{(\text{CON})}_{c_i, z_i}$

In topic models, it is generally true that common function words may overwhelm the word distributions, leading to suboptimal results that are difficult to interpret. This is usually accommodated by data pre-processing (e.g. stop word removal), by backing off to "background" word models (Chemudugunta et al., 2006), or by performing term re-weighting (Wilson and Chew, 2010). In the case of POSLDA, these common words are naturally captured by the functional classes.

3.1 Relations to Other Models

The idea of having multinomials for the cross products of topics and classes is related to multifaceted topic models where word tokens are associated with multiple latent variables (Paul and Girju, 2010; Ahmed and Xing, 2010). Under such models, words can be explained by a latent topic as well as a second underlying variable such as the perspective or dialect of the author, and words may depend on both factors. In our case, the second variable is the part-of-speech – or functional purpose – of the token.

We note that POSLDA is a generalization of many existing models. POSLDA becomes a Bayesian HMM when the number of topics $K = 1$; the original LDA model when the number of classes $S = 1$; and the HMMLDA model of Griffiths et al. (2005) when the number of content word classes $S_{\text{CON}} = 1$. The beauty of these generalizations is that one can easily experiment with any of these models by simply altering the model parameters under a single POSLDA implementation.

3.2 Inference

As with many complex probabilistic models, exact posterior inference is intractable for POSLDA. Nevertheless, a number of approximate inference techniques are at our disposal. In this work, we use collapsed Gibbs sampling to sample the latent class assignments and topic assignments ($\mathbf{c}$ and $\mathbf{z}$), and from these we can compute estimates of the multinomial parameters for the topics (ϕ), the document-topic portions (θ), and the HMM transition matrix (π). Under a trigram version of the model – which we employ for all our experiments in this work – the sampling equation for word token i is as follows:

$$p(c_i, z_i | \mathbf{c_{-i}}, \mathbf{z_{-i}}, \mathbf{w}) \propto$$

$$\begin{cases} \rho_{c_i} \times \frac{n_{z_i}^{(d)} + \alpha_{z_i}}{n_{\cdot}^{(d)} + \alpha_{\cdot}} \frac{n_w^{(c_i, z_i)} + \beta}{n_{\cdot}^{(c_i, z_i)} + W\beta} & c_i \in S_{\text{CON}} \\ \rho_{c_i} \times \frac{n_w^{(c_i)} + \beta}{n_{\cdot}^{(c_i)} + W\beta} & c_i \in S_{\text{FUN}} \end{cases}$$

where

$$\rho_{c_i} = \frac{n_{(c_{i-2}, c_{i-1}, c_i)} + \gamma_{c_i}}{n_{(c_{i-2}, c_{i-1})} + \gamma_{\cdot}} \cdot \frac{n_{(c_{i-1}, c_i, c_{i+1})} + \gamma_{c_i}}{n_{(c_{i-1}, c_i)} + \gamma_{\cdot}} \cdot \frac{n_{(c_i, c_{i+1}, c_{i+2})} + \gamma_{c_i}}{n_{(c_i, c_{i+1})} + \gamma_{\cdot}}$$

Although we sample the pair (c_i, z_i) jointly as a block, which requires computing a sampling distribution over $S_{\text{FUN}} + K \times S_{\text{CON}}$, it is also valid to sample c_i and z_i separately, which requires only $S + K$ computations. In this case, the sampling procedure would be somewhat different. Despite the lower number of computations per iteration, however, the sampler is likely to converge faster with our blocked approach because the two variables are tightly coupled. The intuition is that a non-block-based sampler could have difficulty escaping local optima because we are interested in the most probable *pair*; a highly probable class c sampled on its own, for example, could prevent the sampler from choosing a more likely pair (c', z).

4 POS Tagging Experiments

To demonstrate the veracity of our approach, we performed a number of POS tagging experiments using the POSLDA model. Our data is the recent Twitter POS dataset released at ACL 2011 by Gimpel et al. (2011) consisting of approximately 26,000 words across 1,827 tweets. This dataset provides a unique opportunity to test our unsupervised approach in a domain where it would likely be of most use – one that is novel and therefore lacking large amounts of training data. We feel that this sort of specialized domain will become the norm – particularly in social media analysis – as user generated content continues to grow in size and accessibility. The Twitter dataset uses a domain-dependent tag set of 25 tags that are described in (Gimpel et al., 2011).

For our experiments, we follow the established form of Merialdo (1993) and Goldwater and Griffiths (2007) for unsupervised POS tagging by making use of a tag dictionary to constrain the possible tag choices for each word and therefore render the problem closer to disambiguation. Like Goldwater and Griffiths (2007), we employ a number of dictionaries with varying degrees of knowledge.

We use the full corpus of tweets[1] and construct a tag dictionary which contains the tag information for a word only when it appears more than d times in the corpus. We ran experiments for $d = 1, 2, 3, 5, 10,$ and ∞ where the problem becomes POS clustering. We report both tagging accuracy and the variation of information (VI), which computes the information lost in moving from one clustering C to another C': $VI(C, C') = H(C) + H(C') - 2I(C, C')$ (Meilă, 2007). This can be interpreted as a measure of similarity between the clusterings, where a smaller value indicates higher similarity.

We run our Gibbs sampler for 20,000 iterations and obtain a maximum a posteriori (MAP) estimate for each word's tag by employing simulated annealing. Each posterior probability $p(c, z|\cdot)$ in the sampling distribution is raised to the power of $\frac{1}{\tau}$ where τ is a temperature that approaches 0 as the sampler converges. This approach is akin to

[1] The Twitter POS dataset consists of three subsets of tweets: development, training, and testing. Because we are performing fully unsupervised tagging, however, we combine these three subsets into one.

Accuracy	1	2	3	5	10	∞
random	62.8	49.6	45.2	40.2	35.0	
BHMM	78.4	65.4	59.0	51.8	44.0	
POSLDA	**80.9**	**67.5**	**62.0**	**55.9**	**47.6**	
VI						
random	2.34	3.31	3.56	3.81	4.05	5.86
BHMM	1.41	2.47	2.84	3.22	3.61	5.07
POSLDA	**1.30**	**2.34**	**2.66**	**2.98**	**3.35**	**4.96**
Corpus stats						
% ambig.	54.2	67.9	72.2	76.4	80.4	100
tags / token	2.62	5.91	7.19	8.59	10.3	25

Table 1: POS tagging results on Twitter dataset.

K	Accuracy	σ
1 (HMM)	78.6	0.23
5	80.0	0.06
10	80.9	0.17
15	80.1	0.10
20	80.2	0.21
25	80.1	0.25
30	80.2	0.15
35	80.1	0.12
40	79.9	0.20
45	80.1	0.12

Table 2: POS tagging results as K varies on Twitter dataset.

bringing a system from an arbitrary state to one with the lowest energy, thus viewing the Gibbs sampling procedure as a random search whose goal is to identify the MAP tag sequence – a technique that is also employed by Goldwater and Griffiths (2007). Finally, we run each experiment 5 times from random initializations and report the average accuracy and variation of information.

4.1 Results for Twitter Dataset

In our experiments, we use 8 content classes that correspond to the following parts-of-speech: noun, proper noun, proper noun + possessive, proper noun + verbal, verb, adjective, adverb, and other abbreviations / foreign words. We chose these classes because intuitively they are the types of words whose generative probability will depend on the given latent topic. As the Twitter POS data consists of 25 distinct tags, this leaves 17 remaining classes for function words. In this section, we report results for $K = 10$ topics. We will discuss the effect of varying K in section 4.2. We set symmetric priors with $\alpha = 1.0/K = 0.1$, $\beta = 0.5$, and $\gamma = 0.01$.

As is demonstrated in Table 1, our POSLDA model shows marked improvements over a random tag assignment and, more importantly, the Bayesian HMM approach described by Goldwater and Griffiths (2007). It does so for every setting of d on both accuracy and variation of information. For $d = 1$ our method outperforms the BHMM by 2.5 percentage points. With higher values of d, however, POSLDA increases its improvement over the BHMM to up to 4.1 percentage points. The increase in tagging accuracy as d increases suggests that our method may be particularly suitable for domains with little training data.[2] For $d = \infty$, where we are performing POS *clustering*, our model improves the variation of information by 0.11. Each of these improvements over the Bayesian HMM is statistically significant with $p \ll 0.01$. Despite the clear improvements in POS tagging accuracy and clustering that we demonstrate in this section, we trained our POSLDA model with a "blind" topic setting of $K = 10$. In the following section, we will investigate how this parameter affects the achievable results with our technique.

4.2 Topic Variance

In the previous section we set the number of topics *a priori* to $K = 10$. However, it is well known in topic modeling research that different datasets exhibit different numbers of "inherent" topics (Blei et al., 2003). Therefore, a POSLDA model fit with the "correct" number of topics will likely achieve higher accuracy in POS tagging. A standard approach to tuning the number of topics to fit a topic model is to try a number of different topics and choose the one that results in the lowest perplexity on a held-out test set (Claeskens and Hjort, 2008). Here, we can choose the optimal K more directly by trying a number of different values and choosing the one that maximizes the POS tagging accuracy.

For this experiment, we again make use of the Twitter POS dataset (Gimpel et al., 2011). We use the same setup as that described above with simulated annealing, 20,000 iterations, and a tag dic-

[2] The differences in tagging accuracy in terms of percentage points between POSLDA and the BHMM for $d = \{1, 2, 3, 5, 10\}$ are $\Delta_a = \{2.5, 2.1, 3.0, 4.1, 3.6\}$, respectively. For clustering, the increases in VI are even more clear as d increases. They are $\Delta_{VI} = \{0.11, 0.13, 0.18, 0.24, 0.26\}$.

tionary with $d = 1$. As before, we set $\alpha = 1.0/K$, $\beta = 0.5$, and $\gamma = 0.01$. We perform experiments with $K = \{1, 5, 10, \ldots, 40, 45\}$, where $K = 1$ corresponds to the Bayesian HMM. The results averaged over 3 runs are tabulated in Table 2 with the associated standard deviations (σ), and shown graphically in Figure 2.

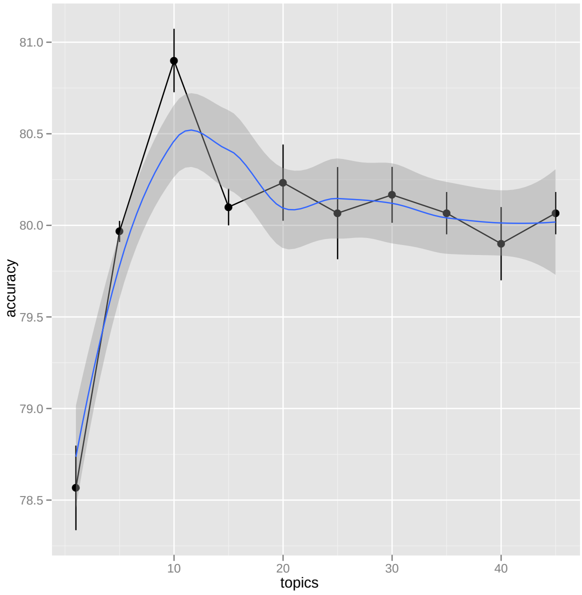

Figure 2: Number of topics K vs. POS tagging accuracy on the Twitter dataset. The average accuracies, along with their standard errors, are shown in black, while a smoothed curve of the same data is shown in blue.

As we expect, the tagging accuracy depends on the number of topics specified by the model. In fact, the accuracy improves by nearly a full percentage point from both the previous and next topic settings when we hit a critical point at $K = 10$. When $K = 1$ the model reduces to the Bayesian HMM and our accuracy suffers. It steadily increases until we hit the critical point and then drops off again but plateaus at a level that is approximately 1.5 percentage points higher than the BHMM. This shows that determining an appropriate setting for the number of topics is essential for the best possible tagging accuracy using POSLDA. Nevertheless, even with a "blind" setting within a large range of topic values (here from $K = 5$ to at least $K = 45$), we see marked improvements over the baseline system that does not include any semantic topic information.

5 Model Evaluation

In this section we present further experiments on the raw output of POSLDA to demonstrate its capabilities beyond simply POS tagging. We show the model's ability both qualitatively and quantitatively to capture the semantic (or "content") and syntactic (or "functional") axes of information prevalent in a corpus made up of social media data. We begin qualitatively with topic interpretability when the model is learned given a collection of unannotated Twitter messages, and then present quantitative results on the ability of POSLDA as a predictive language model in the Twitter domain.

5.1 Topic Interpretability

Judging the interpretability of a set of topics is highly subjective, and there are understandably various differing approaches of evaluating topic cohesiveness. For example, Chang et al. (2009) look at "word intrusion" where a user determines an *intruding* word from a set of words that does not thematically fit with the other words, and "topic intrusion" where a user determines whether the learned document-topic portion θ_d appropriately describes the semantic theme of the document. In this section, we are most interested in subjectively demonstrating the low incidence of "word intrusion" both in terms of semantics (theme) and syntax (part-of-speech). We do not conduct formal experiments to demonstrate this, but we subjectively show that our model learns semantic and syntactic word distributions that are likely robust towards problems of word intrusion and that are therefore "interpretable" for humans examining the learned posterior word distributions.

Table 3 shows three topics – manually labelled as "party", "status update", and "politics" – learned from the relatively small Twitter POS dataset. We set the number of topics $K = 20$, the number of classes $S = 25$, and the number of content word classes $S_{\text{CON}} = 8$, following our earlier POS tagging experiments. We show the top five words from three POS-specific topics labelled manually as *noun*, *verb*, and *adjective*. Given the relatively small size of the dataset, the short length of the documents, and the esoteric language and grammar use, the interpretability of the topics is reasonable. All three topics assign high probability to words that one would

PARTY			STATUS UPDATE			POLITICS		
noun	*verb*	*adj*	*noun*	*verb*	*adj*	*noun*	*verb*	*adj*
party	gets	awesome	day	is	nice	anything	say	late
man	is	old	pm	looking	nasty	truth	has	real
shit	knew	original	school	so	last	face	wait	high
men	were	fake	today	have	hard	city	cant	republican
person	wasnt	drunk	body	got	tired	candidate	going	important

Table 3: Example topics learned from the Twitter POS dataset with POSLDA.

CONJ	DET	PREP	RP
and	the	to	to
but	a	of	it
or	my	in	up
n	your	for	away
in	this	on	in
yet	that	with	on
plus	is	at	around
nd	some	NUMBER	out
an	an	if	over
to	his	from	off

Table 4: Example topic-independent function class distributions ($\mathcal{C}_{\text{FUN}}$) learned from the Twitter POS dataset with POSLDA.

expect to have high importance with one or two outliers. More importantly, however, the POS-specific topics also generally reflect their syntactic roles. Each of the verbs is assuredly (even without the proper context) a verb (with the single outlier being the word "so"), and the same thing for the nouns. The adjectives seem to fit as well; though many of the words could be considered nouns depending on the context, it is clear how given the topic each of the words could very well act as an adjective. A final point worth mentioning is that, unlike LDA, we do not perform stopword removal. Instead, the POSLDA model has pushed stopwords to their own *function* classes (rather than content) freeing us from having to perform pre- or post-processing steps to ensure interpretable topics. The top words in four of these topic-independent function classes, learned from the Twitter POS dataset, are shown in Table 4.[3] These function word distributions are even more cohesive than the content word distributions, showing that the standard stopwords have been accounted for as we expect in their respective function classes.

5.2 Predictive Language Modeling

While we have demonstrated that our model can achieve improved accuracy in POS tagging for Twitter data, it can also be useful for other kinds of language analysis in the social media domain. In the following experiments, we test the POSLDA model quantitatively by determining its ability as a predictive language model. Following a standard practice in topic modeling research (Blei et al., 2003; Griffiths et al., 2005), we fit a model to a training set and then compute the perplexity of a held-out test set. For this experiment, we use the Twitter POS *training* dataset described earlier (16,348 words across 999 tweets). We then perform testing on the Twitter POS *testing* dataset (8,027 words across 500 tweets). We compare the perplexity – a monotonically decreasing function of the log likelihood – to LDA, a Bayesian HMM, and HMMLDA. Finally, we use Minka's fixed-point method (Wallach, 2008) to optimize the hyperparameters α and β.

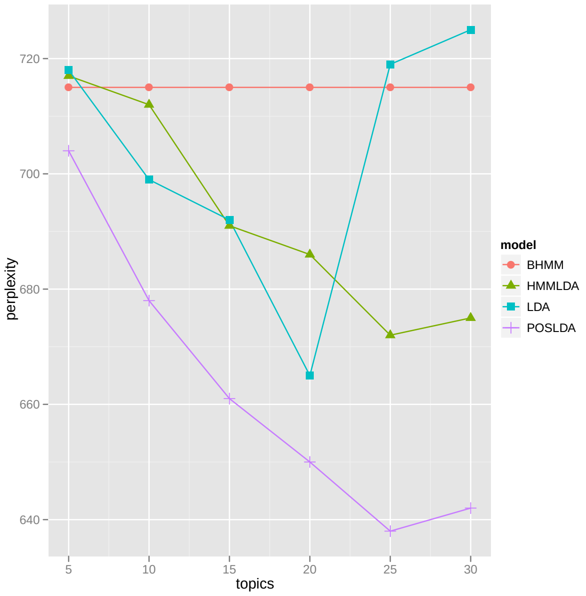

Figure 3: Perplexity of POSLDA and other probabilistic models.

[3] Note that we make use of the tag dictionary when learning these word distributions.

Figure 3 shows the perplexity on the held-out Twitter test set for models trained with $K = \{5, 10, 15, 20, 25, 30\}$. The Bayesian HMM is not affected by the number of topics and is able to beat the HMMLDA model at $K = 5$. It also achieves lower perplexity than the LDA model at $K = 5, 25$, and 30. Our POSLDA model, however, achieves the lowest perplexity of all tested models at all topic settings that we tested. This demonstrates that POSLDA is a good candidate for both language modeling and for further latent probabilistic model-based analysis of Twitter data.

6 Discussion

In the previous section we demonstrated both qualitatively and quantitatively that our model captures two sources of information from unstructured texts: thematic (or semantics) and functional (or syntactic). An important question to consider is why – as we demonstrated in section 4 – learning this sort of information improves our ability to perform unsupervised POS tagging. One reason is discussed in the introduction: semantic information can help disambiguate the POS for a word that typically serves a different function depending on the topic that it is normally associated with. This phenomenon likely plays an important role in the accuracy improvements that we observe. However, another feature of the model is the distinction between "content" POS classes and "function" POS classes. The former will depend on the current topic while the latter are universal across thematic space. This will also represent an improvement over the bare HMM because words that depend on the current topic – typically nouns, verbs, adjectives, and adverbs – will be forced to these classes due to their long-range thematic dependencies while words with only short-range dependencies will be pushed to the function POS classes. This latter type of words – conjunctions, determiners, etc. – naturally do not depend on themes so as they are pushed to the function-only POS classes, and so one step of disambiguation has already been performed. This is the same behaviour as in the HMMLDA model by Griffiths et al. (2005), but here we are able to perform proper POS tagging because there is more than just a single content word class and we are therefore able to discern between the topic-dependent parts-of-speech.

7 Conclusions and Future Work

In this paper, we have shown that incorporating semantic topic information into a Bayesian HMM can result in impressive increases in accuracy for unsupervised POS tagging. Specifically, we presented POSLDA – a topic model consistent across the axes of both semantic and syntactic meanings. Using this model to perform unsupervised POS tagging results in consistent and statistically significant increases in POS tagging accuracy and decreases in variation of information when performing POS clustering. These improvements are demonstrated on a novel release of data from the microblogging social network site Twitter. This type of dataset is of particular interest because unsupervised POS tagging will likely be most important in specialized idiosyncratic domains with atypical features and small amounts of labelled training data. Crucially, we showed that even with the inconsistent and at times strange use of grammar, slang, and acronyms, the syntactic portion of the model demonstrably improves not only the predictive ability of the model in terms of perplexity, but also the accuracy in unsupervised POS tagging. This is important because in general tweets are far from being representative of "proper" grammar. Nevertheless, there clearly exists some adherence to syntactic structure as the use of the HMM within our model improves word prediction and POS tagging.

This work represents the first – to our knowledge – application of latent thematic information to the unsupervised POS tagging task.[4] However, due to the encouraging results, there are a number of future research directions that present themselves from this work. One immediate task is to extend POSLDA to a nonparametric Bayesian model. Section 4.2 shows how varying the number of topics K in the model can affect the tagging accuracy by up to a full percentage point. A nonparametric version of the model would free us from having to perform the initial model selection step to get the best accuracy. Another avenue for future work is to infuse more structure into the model such as word morphology.

[4] There has been some work done to include semantic information collected separately in a *supervised* POS tagging approach (Toutanova and Johnson, 2008).

Acknowledgments

William Darling is supported by an NSERC Doctoral Postgraduate Scholarship, and Michael Paul is supported by an NSF Graduate Research Fellowship. The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Amr Ahmed and Eric P. Xing. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1140–1150, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248.

G. Claeskens and N.L. Hjort. 2008. *Model Selection and Model Averaging*. Cambridge University Press.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

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. Association for Computational Linguistics.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press.

Donna Harman. 1992. Overview of the first text retrieval conference (trec-1). In *TREC*, pages 1–20.

M. Meilă. 2007. Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May.

Bernard Merialdo. 1993. Tagging english text with a probabilistic model. *Computational Linguistics*, 20:155–171.

Michael J. Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 354–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kristina Toutanova and Mark Johnson. 2008. A bayesian lda-based model for semi-supervised part-of-speech tagging. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1521–1528. MIT Press, Cambridge, MA.

Hanna Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *NIPS*.

Hanna M. Wallach. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Andrew Weir. 2012. Left-edge deletion in english and subject omission in diaries. *English Language and Linguistics*.

Andrew T. Wilson and Peter A. Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 465–473, Stroudsburg, PA, USA. Association for Computational Linguistics.