Dirichlet Processes for Joint Learning of Morphology and PoS Tags

Burcu Can
Department of Computer Engineering
Hacettepe University
Beytepe, Ankara 06800 Turkey
burcu.can@hacettepe.edu.tr

Suresh Manandhar
Department of Computer Science
University of York
Heslington, York, YO10 5GH, UK
suresh.manandhar@york.ac.uk

Abstract

This paper presents a joint model for learning morphology and part-of-speech (PoS) tags simultaneously. The proposed method adopts a finite mixture model that groups words having similar contextual features thereby assigning the same PoS tag to those words. While learning PoS tags, words are analysed morphologically by exploiting similar morphological features of the learned PoS tags. The results show that morphology and PoS tags can be learned jointly in a fully unsupervised setting.

1 Introduction

The morphology of a word is an important indicator that determines its PoS tag, meanwhile the PoS tag of a word helps in identifying the correct morphological segmentation of the word. This relationship between morphology and syntax has been beneficial in both morphology learning with the exploitation of the syntactic features and in PoS tagging with the adoption of morphological features.

There has been a number of research that have performed PoS tagging by making use of morphological information (Clark (2003), Hasan and Ng (2009), Abend et al. (2010), Christodoulopoulos et al. (2011), etc.). There has been also a number of other research that have performed morphological segmentation by adopting syntactic information (Hu et al. (2005), Can and Manandhar (2009), Lee et al. (2011), etc.). However, there is a small number of research that combines two tasks in a single framework.

Sirts and Alumäe (2012) share a similar goal with us in joining PoS tagging and morphological segmentation in a single framework. They use hierarchical Dirichlet process for infinite HMMs to induce both PoS tags and morphological segmentation. Their model is type-based, whereas our model is token based. In our model, we use finite mixture models for PoS tagging and Dirichlet processes for segmentation.

2 Model Definition

The generative story of the model goes as follows:

1. Draw a PoS tag c_i.
2. Generate a word w_i that belongs to c_i.
3. Generate the context $c_{i-1,i+1}$ of the word w_i from c_i.
4. From the possible splits of w_i, generate a suffix m_i conditioned on c_i, such that $w_i = s_i + m_i$, where s_i denotes the stem.

The generative story is summarised as follows:

$$p(c_i, c_{i-1,i+1}, w_i, s, m) = p(c_i)p(c_{i-1,i+1}|c_i) p(w_i|c_i)p(m|c_i)p(s)$$

2.1 PoS Tagging

The model adopts a finite mixture model for PoS tagging (see Figure 1). Each mixture component represents a PoS tag that shares a set of features with other members in the same component. Each mixture component c_i consists of 1. a distribution over contexts and 2. a distribution over words. Each context is a PoS tag pair $< c_{i-1}, c_{i+1} >$ where the previous word w_{i-1} belongs to c_{i-1} and the following word w_{i+1} belongs to c_{i+1}. We employ a token-based approach for PoS tagging due to the significance of the context. The model is

International Joint Conference on Natural Language Processing, pages 1087–1091,
Nagoya, Japan, 14-18 October 2013.

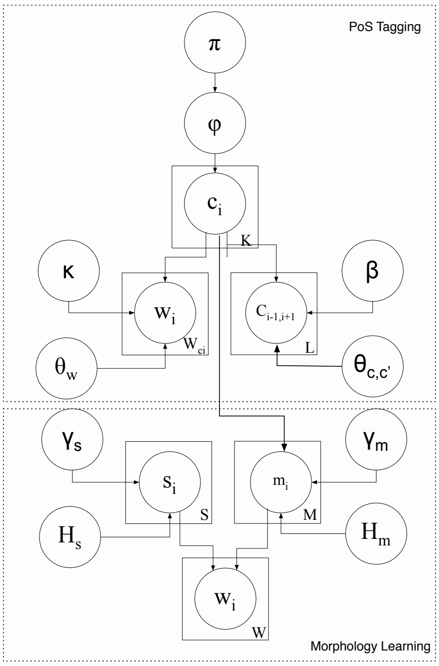

Figure 1: The complete joint model.

defined formally as follows:

$$
\begin{aligned}
c_i &\sim Mult(\phi) && (1)\\
\phi &\sim Dir(\pi) && (2)\\
w_i|c_i &\sim Mult(\theta_w) && (3)\\
\theta_w &\sim Dir(\kappa) && (4)\\
c_{i-1,i+1}|c_i &\sim Mult(\theta_{c,c'}) && (5)\\
\theta_{c,c'} &\sim Dir(\beta) && (6)
\end{aligned}
$$

Class indicators c_i are drawn from a Multinomial distribution with parameters ϕ (which have a Dirichlet prior distribution with hyperparameters π). Each c_i involves a set of words w_i drawn from a Multinomial distribution with parameters θ_w (which have a Dirichlet prior distribution with hyperparameters κ). Each c_i also involves a distribution over contexts $c_{i-1,i+1}$ drawn from a Multinomial distribution with parameters: $\theta_{c,c'}$ (which have a prior distribution with hyperparameters β).

2.2 Morphology Learning

We model morphology using a Dirichlet process (DP) in order to split each word into a stem and a suffix (see Figure 1). Stems are generated by $DP(\gamma_s, H_s)$ with concentration parameter γ_s and base distribution H_s, whereas suffixes are generated by $DP(\gamma_m, H_m)$ with concentration parameter γ_m and base distribution H_m. Hence, the model is defined formally as follows:

$$
\begin{aligned}
s_i &\sim DP(\gamma_s, H_s)\\
m_i|c_i &\sim DP(\gamma_m, H_m)
\end{aligned}
$$

Base distributions are length priors that favour shorter morphs (Creutz and Lagus, 2005):

$$H_x(x_i) = p(c_{ij})^{|x_i|} \tag{7}$$

where x_i is a morph and $|x_i|$ is the length of x_i in letters. Each character has a probability of $p(c_{ij})$, where characters are assumed to be distributed uniformly in the alphabet. We also assume that each morph ends with a special character; i.e. end of morph marker.

Here, $DP(\gamma_s, H_s)$ is a global Dirichlet process where stems may belong to any PoS tag, whereas $DP(\gamma_m, H_m)$ is defined locally for each PoS tag. The reason is that stems are shared amongst different PoS tags. However, words belonging to the same PoS tag usually have similar endings, thereby leading to local distributions.

3 Inference

In our model, we assign values to the hyperparameters $\pi, \kappa, \beta, \gamma_s, \gamma_m$ empirically, and we integrate out the parameters $\phi, \theta_w, \theta_{c,c'}$ by using the Multinomial-Dirichlet conjugacy.

We use Gibbs sampling to infer POS tags, stems and suffixes. We perform inference in two steps: 1. a PoS tag is sampled for the word, 2. a stem and a suffix are sampled for the word.

3.1 Inferring PoS tags

Each word's PoS tag is sampled subject to its context. Let a word be w_i and imagine that it occurs in context $< w_{i-1}, w_{i+1} >$ where w_{i-1} belongs to c_{i-1} and w_{i+1} belongs to c_{i+1}. We define the sampling probability of c_i for w_i as follows:

$$
\begin{aligned}
p(c_i| < w_{i-1}, w_{i+1} >, w_i) &\propto p(< w_{i-1}, w_{i+1} >, w_i|c_i)p(c_i)\\
&\propto p(w_i|c_i)p(< w_{i-1}, w_{i+1} > |c_i)\\
&\qquad p(c_i)
\end{aligned}
$$

We also assume that $< w_{i-1}, w_{i+1} >$ and w_i are independent since it is possible to remove w_i from $< w_{i-1}, w_{i+1} >$ and insert another word instead.

In order to calculate $p(w_i|c_i)$, w_i is removed from the corpus:

$$p(w_i|c_i^{-w_i}, \kappa) = \frac{n_{w_i, c_i^{-w_i}} + \kappa}{N_{c_i}^{-w_i} + W_{c_i^{-w_i}}\alpha} \tag{8}$$

where $c_i^{-w_i}$ denotes the mixture component c_i that excludes w_i, $n_{w_i, c_i^{-w_i}}$ is the number of the word-tag pairs $< w_i, c_i >$, $N_{c_i}^{-w_i}$ is the number of word

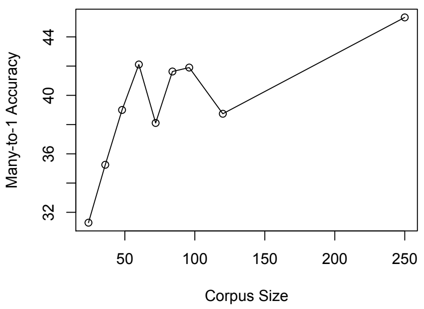

Figure 2: Many-to-1 accuracy scores obtained from corpora of size 24K, 36K, 48K, 60K, 72K, 84K, 96K, 120K, and 250K.

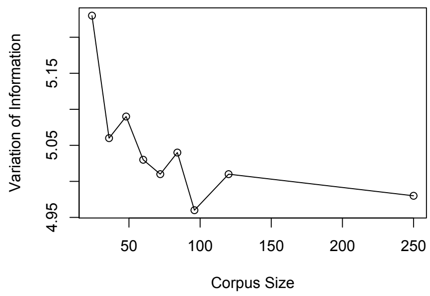

Figure 3: Variation of Information (VI) obtained from corpora of size 24K, 36K, 48K, 60K, 72K, 84K, 96K, 120K and 250K.

tokens having the PoS tag c_i, $W_{c_i^{-w_i}}$ is the number of word types that are tagged with c_i. $p(c_i)$ is computed as follows:

$$p(c_i|\mathbf{c}^{-w_i}, \pi) = \frac{n_{c_i^{-w_i}} + \pi}{N^{-w_i} + K\pi} \tag{9}$$

where N^{-w_i} denotes the number of word tokens in the model excluding w_i, K is the number of class indicators (i.e. number of PoS tags).

In order to mitigate the sparsity within the context probabilities, we use the approximation introduced by Clark (2000):

$$p(< w_{i-1}, w_{i+1} > |c_i) = p(< c_{i-1}, c_{i+1} > |c_i) \; p(w_{i-1}|c_{i-1})p(w_{i+1}|c_{i+1}) \tag{10}$$

where, $p(< c_{i-1}, c_{i+1} > |c_i)$ is computed such that:

$$p(< c_{i-1}, c_{i+1} > |c_x, c_y, c_z, c_i, \beta) = \frac{n_{c_{i-1}, c_i, c_{i+1}} + \beta}{k_{c_i} + L\beta} \tag{11}$$

Here, c_x is $c_i^{-<c_{i-1},c_{i+1}>}$, c_y is $c_{i-1}^{-<c_{i-2},c_i>}$, c_z is $c_{i+1}^{-<c_i,c_{i+2}>}$, k_{c_i} is the number of contexts in c_i, and L denotes the possible number of different contexts in the model (i.e. $K * K$).

3.2 Inferring Morphology

Two latent variables are inferred for morphology: stems and suffixes. The sampling probability for morphology is defined as follows:

$$p(w_i = s_i + m_i|\mathbf{s}^{-i}, \mathbf{m}_{c_i}^{-i}) = p(s_i|\mathbf{s}^{-i})p(m_i|\mathbf{m}_{c_i}^{-i}) \tag{12}$$

where $\mathbf{s}^{-i}$ is the set of stems excluding s_i, $\mathbf{m}_{c_i}^{-i}$ is the set of suffixes assigned with c_i excluding m_i. The conditional probability of a stem is:

$$p(s_i|\mathbf{s}^{-i}, \gamma_s, H_s) = \frac{f^{s^{-i}} + \gamma_s H_s(s_i)}{T^{s^{-i}} + M^{s^{-i}}\gamma_s} \tag{13}$$

where $f^{s^{-i}}$ is the frequency of the stem type s_i already generated, $T^{s^{-i}}$ is the number of all stems in the model, and $M^{s^{-i}}$ is the number of stem types generated excluding s_i. Similarly, the conditional probability of a suffix is computed as follows:

$$p(m_i|\mathbf{m}_{c_i}^{-m_i}, \gamma_m) = \frac{f_{c_i}^{m^{-i}} + \gamma_m H_m(m_i)}{T_{c_i}^{m^{-i}} + M^{m^{-i}}\gamma_m} \tag{14}$$

where $f_{c_i}^{m^{-i}}$ is the frequency of the suffix type m_i already generated in c_i, $T_{c_i}^{m^{-i}}$ is the number of all suffixes assigned with PoS tag c_i, and $M^{m^{-i}}$ is the number of suffix types already generated excluding m_i.

In the algorithm, initially each word is assigned a PoS tag and split randomly. The algorithm goes through each word by sampling a PoS tag, a stem, and a suffix. All constituents of the respective word (tag, stem, suffix, context, contexts of adjacent words) are removed from the model beforehand. This process is repeated for a number of iterations until a convergence is ensured.

4 Experiments & Evaluation

We used small portions of the Penn WSJ treebank (Marcus et al., 1993) for the experiments. We manually set the hyperparameters and concentration parameters for each experiment: $\pi = 10^{-6}, \beta = 10^{-6}, \kappa = 10^{-6}, \gamma_s = 10^{-6}, \gamma_m = 10^{-6}$. These values were set empirically through several experiments. We also inserted a special character at the end of each sentence and assigned it a distinct PoS tag. No other words could be assigned this tag.

4.1 PoS Tagging Results

In our experiments we fixed the number of PoS tags to 45, which is the number of PoS tags in

	V-measure	Many-to-one
Christ.1[1]	48.6	57.8
Joint	41.11	59.67
Clark[2]	63.8	68.8
Christ.2 (Best Pub.)[3]	67.7	72.0

[1] Christodoulopoulos et al. (2011)
[2] Clark (2003)
[3] Christodoulopoulos et al. (2010)

Table 1: PoS tagging scores.

	missing	extra	wrong	correct
Joint	0.72%	28.55%	10.13%	60.60%
Morfessor	15.07%	7.23%	10.22%	67.48%

Table 2: Morphological segmentation scores.

Penn WSJ treebank. We applied many-to-one accuracy by assigning each result tag a gold standard tag having the highest frequency among the words assigned with this result tag (see Figure 2). Second, we applied one-to-one accuracy which have similar results with many-to-one scores.

We also measured the variation of information (VI) (Rosenberg and Hirschberg, 2007) (see Figure 3). Although there is not a smooth decrement in VI measure, it improves with the larger datasets in average[1].

Results show that determiners, modal verbs, prepositions, pronouns, conjunctions, and numbers are discovered generally correctly. The most common error type is due the confusion of nouns and adjectives. Normally, nouns are distributed over several PoS tags. Verbs and adverbs are also generally confused and spread over different tags.

We report our results with a comparison to other systems in Table 1 by using a dataset of 250K words. We use a small portion of Penn WSJ treebank for the comparison. The dataset involves 250K words where the number of word types is 20957. The other systems are also tested on a small portion of WSJ involving 16850 word types, which is reported in Christodoulopoulos et al. (2011).

Our system outperforms Christodoulopoulos et al. (2011) with the many-to-one evaluation, whereas Christodoulopoulos et al. (2011) perform better than our system based on V-measure evaluation. It should be noted that Clark (2003) and Christodoulopoulos et al. (2010) are both type-based.

[1]Although, Figure 3 shows that results for 36k words are better than results for 48k words, this could be due to the particular choice of training sets we used.

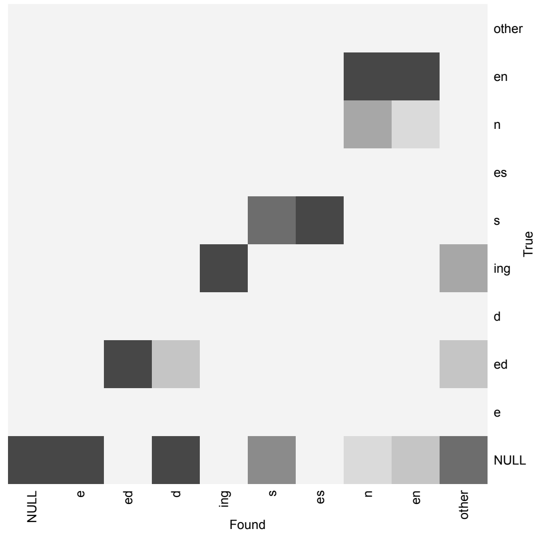

Figure 4: Confusion matrix shows the correlation between found morphs and true morphs. The shades reflect the number of matchings.

4.2 Morphological Segmentation Results

We performed the evaluation of morphological segmentation on verbs. We adopted some heuristics that strip off common verb endings such as *-ed, -d, -ing, -s, -es* from verbs in order to build the gold standard. Irregular verbs are introduced exceptionally and left as they are.

The results obtained from the 96K setting were used for the evaluation. We ran Morfessor Baseline (Creutz and Lagus, 2002; Creutz and Lagus, 2005; Creutz and Lagus, 2007) on the verbs in the same dataset. Table 2 gives the scores where *missing types* refers to the case that gold standard suggests a suffix but no suffix is identified in the results, *extra suffixes* means that gold standard does not identify any suffixes but the results contain suffixes, *wrong suffixes* implies that both gold standard and results identify suffixes but they are not the same, and *correct types* means that both gold standard and results contain suffixes and they match. Our model identifies 12257 suffix types, whereas Morfessor Baseline identifies 2309 due to undersegmentation. In addition, confusion matrix that depicts the result morphs against true morphs is given in Figure 4.

5 Conclusion

We proposed a model that jointly learns PoS tags and morphology. The results show that learning PoS tags and morphology can be performed cooperatively.

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*, ACL '10, pages 1298–1307, Stroudsburg, PA, USA. Association for Computational Linguistics.

Burcu Can and Suresh Manandhar. 2009. Clustering morphological paradigms using syntactic categories. In *Working Notes for the CLEF 2009 Workshop*, September.

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*, EMNLP '10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 Simon Clark. 2000. Inducing syntactic categories by context distribution clustering. pages 91–94.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 59–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*, MPL '02, pages 21–30, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. *Technical Report A81*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4:3:1–3:34, February.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Kazi Saidul Hasan and Vincent Ng. 2009. Weakly supervised part-of-speech tagging for morphologically-rich, resource-scarce languages. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 363–371, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yu Hu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005. Using morphology and syntax together in unsupervised learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, PMHLA '05, pages 20–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2011. Modeling syntactic context improves morphological segmentation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Empirical Methods in Natural Language Processing*.

Kairit Sirts and Tanel Alumäe. 2012. A hierarchical dirichlet process model for joint part-of-speech and morphology induction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 407–416, Stroudsburg, PA, USA. Association for Computational Linguistics.