A Hierarchical Dirichlet Process Model for Joint Part-of-Speech and Morphology Induction

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

Tanel Alumäe
Institute of Cybernetics at
Tallinn University of Technology
tanel.alumae@phon.ioc.ee

Abstract

In this paper we present a fully unsupervised nonparametric Bayesian model that jointly induces POS tags and morphological segmentations. The model is essentially an infinite HMM that infers the number of states from data. Incorporating segmentation into the same model provides the morphological features to the system and eliminates the need to find them during preprocessing step. We show that learning both tasks jointly actually leads to better results than learning either task with gold standard data from the other task provided. The evaluation on multilingual data shows that the model produces state-of-the-art results on POS induction.

1 Introduction

Nonparametric Bayesian modeling has recently become very popular in natural language processing (NLP), mostly because of its ability to provide priors that are especially suitable for tasks in NLP (Teh, 2006). Using nonparametric priors enables to treat the size of the model as a random variable with its value to be induced during inference which makes its use very appealing in models that need to decide upon the number of states.

The task of unsupervised parts-of-speech (POS) tagging has been under research in numerous papers, for overview see (Christodoulopoulos et al., 2010). Most of the POS induction models use the structure of hidden Markov model (HMM) (Rabiner, 1989) that requires the knowledge about the number of hidden states (corresponding to the number of tags) in advance. According to our considerations, supplying this information is not desirable for two opposing reasons: 1) it injects into the system a piece of knowledge which in a truly unsupervised setting would be unavailable; and 2) the number of POS tags used is somewhat arbitrary anyway because there is no common consensus of what should be the true number of tags in each language and therefore it seems unreasonable to constrain the model with such a number instead of learning it from the data.

Unsupervised morphology learning is another popular task that has been extensively studied by many authors. Here we are interested in learning concatenative morphology of words, meaning the substrings of the word corresponding to morphemes that, when concatenated, will give the lexical representation of the word type. For the rest of the paper we will refer to this task as (morphological) segmentation.

Several unsupervised POS induction systems make use of morphological features (Blunsom and Cohn, 2011; Lee et al., 2010; Berg-Kirkpatrick et al., 2010; Clark, 2003; Christodoulopoulos et al., 2011) and this approach has been empirically proved to be helpful (Christodoulopoulos et al., 2010). In a similar fashion one could think that knowing POS tags could be useful for learning morphological segmentations and in this paper we will study this hypothesis.

In this paper we will build a model that combines POS induction and morphological segmentation into one learning problem. We will show that the unsupervised learning of both of these tasks in the same

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 407–416,
Montréal, Canada, June 3-8, 2012. ©2012 Association for Computational Linguistics

model will lead to better results than learning both tasks separately with the gold standard data of the other task provided. We will also demonstrate that our model produces state-of-the-art results on POS tagging. As opposed to the compared methods, our model also induces the number of tags from data.

In the following, section 2 gives the overview of the Dirichlet Processes, section 3 describes the model setup followed by the description of inference procedures in section 4, experimental results are presented in section 5, section 6 summarizes the previous work and last section concludes the paper.

2 Background

2.1 Dirichlet Process

Let H be a distribution called base measure. Dirichlet process (DP) (Ferguson, 1973) is a probability distribution over distributions whose support is the subset of the support of H:

$$G \sim DP(\alpha, H), \tag{1}$$

where α is the concentration parameter that controls the number of values instantiated by G.

DP has no analytic form and therefore other representations must be developed for sampling. In the next section we describe Chinese Restaurant Process that enables to obtain samples from DP.

2.2 Chinese Restaurant Process

Chinese Restaurant Process (CRP) (Aldous, 1985) enables to calculate the marginal probabilities of the elements conditioned on the values given to all previously seen items and integrating over possible DP prior values.

Imagine an infinitely big Chinese restaurant with infinitely many tables with each table having capacity for infinitely many customers. In the beginning the restaurant is empty. Then customers, corresponding to data points, start entering one after another. The first customer chooses an empty table to sit at. Next customers choose a new table with probability proportional to the concentration parameter α or sit into one of the already occupied tables with probability proportional to the number of customers already sitting there. Whenever a customer chooses an empty table, he will also pick a dish from H to be served on that table. The predictive probability distribution over dishes for the i-th customer is:

$$P(x_i = \phi_k | \mathbf{x}_{-i}, \alpha, H) = \frac{n_{\phi_k} + \alpha}{i - 1 + \alpha} p_H(\phi_k), \tag{2}$$

where $\mathbf{x}_{-i}$ is the seating arrangement of customers excluding the i-th customer and n_{ϕ_k} is the number of customers eating dish ϕ_k and $p_H(\cdot)$ is the probability according to H.

2.3 Hierarchical Dirichlet Process

The notion of hierarchical Dirichlet Process (HDP) (Teh et al., 2006) can be derived by letting the base measure itself to be a draw from a DP:

$$G_0 | \alpha_0, H \sim DP(\alpha_0, H) \tag{3}$$

$$G_j | \alpha, G_0 \sim DP(\alpha, G_0) \quad j = 1 \cdots J \tag{4}$$

Under HDP, CRP becomes Chinese Restaurant Franchise (Teh et al., 2006) with several restaurants sharing the same franchise-wide menu G_0. When a customer sits at an empty table in one of the G_j-th restaurants, the event of a new customer entering the restaurant G_0 will be triggered. Analogously, when a table becomes empty in one of the G_j-th restaurants, it causes one of the customers leaving from restaurant G_0.

3 Model

We consider the problem of unsupervised learning of POS tags and morphological segmentations in a joint model. Similarly to some recent successful attempts (Lee et al., 2010; Christodoulopoulos et al., 2011; Blunsom and Cohn, 2011), our model is type-based, arranging word types into hard clusters. Unlike many recent POS tagging models, our model does not assume any prior information about the number of POS tags. We will define the model as a generative sequence model using the HMM structure. Graphical depiction of the model is given in Figure 1.

3.1 Generative story

We assume the presence of a fixed length vocabulary W. The process starts with generating the lexicon that stores for each word type its POS tag and morphological segmentation.

- Draw a unigram tag distribution from the respective DP;
- Draw a segment distribution from the respective DP;
- For each tag, draw a tag-specific segment distribution from HDP with the segment distribution as base measure;
- For each word type, draw a tag from the unigram tag distribution;
- For each word type, draw a segmentation from the respective tag-specific segment distribution.

Next we proceed to generate the HMM parameters:

- For each tag, draw a bigram distribution from HDP with the unigram tag distribution as base measure;
- For each tag bigram, draw a trigram distribution from HDP with the respective bigram distribution as base measure;
- For each tag, draw a Dirichlet concentration parameter from Gamma distribution and an emission distribution from the symmetric Dirichlet.

Finally the standard HMM procedure for generating the data sequence follows. At each time step:

- Generate the next tag conditioned on the last two tags from the respective trigram HDP;
- Generate the word from the respective emission distribution conditioned on the tag just drawn;
- Generate the segmentation of the word deterministically by looking it up from the lexicon.

3.2 Model setup

The trigram transition hierarchy is a HDP:

$$G^U \sim DP(\alpha^U, H) \tag{5}$$

$$G_j^B \sim DP(\alpha^B, G^U) \quad j = 1 \cdots \infty \tag{6}$$

$$G_{jk}^T \sim DP(\alpha^T, G_j^B) \quad j, k = 1 \cdots \infty, \tag{7}$$

where G^U, G^B and G^T denote the unigram, bigram and trigram context DP-s respectively, α-s are the respective concentration parameters coupled for DP-s of the same hierarchy level. Emission parameters are drawn from multinomials with symmetric Dirichlet priors:

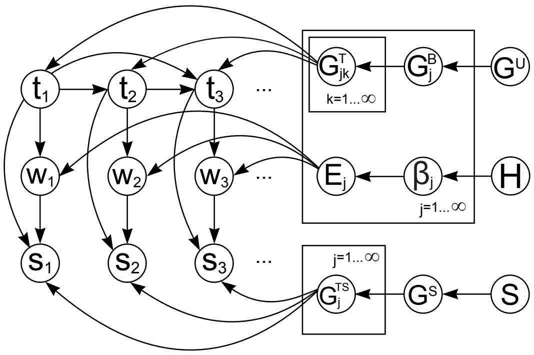

Figure 1: Plate diagram representation of the model. t_i-s, w_i-s and s_i-s denote the tags, words and segmentations respectively. G-s are various DP-s in the model, E_j-s and β_j-s are the tag-specific emission distributions and their respective Dirichlet prior parameters. H is Gamma base distribution. S is the base distribution over segments. Coupled DP concetrations parameters have been omitted for clarity.

$$E_j|\beta_j, H \sim \int Mult(\theta)Dir(\beta_j)d\theta \quad j = 1 \cdots \infty, \tag{8}$$

where each emission distribution has its own Dirichlet concentration parameter β_j drawn from H.

Morphological segments are modelled with another HDP where the groups are formed on the basis of tags:

$$G^S \sim DP(\alpha^S, S) \tag{9}$$

$$G_j^{TS} \sim DP(\alpha^{TS}, G^S) \quad j = 1 \cdots \infty, \tag{10}$$

where G_j^{TS} are the tag-specific segment DP-s and G^S is their common base distribution with S as base measure over all possible strings. S consists of two components: a geometric distribution over the segment lengths and collapsed Dirichlet-multinomial over character unigrams.

4 Inference

We implemented Gibbs sampler to draw new values for tags and Metropolis-Hastings sampler for resampling segmentations. We use a type-based collapsed sampler that draws the tagging and segmentation values for all tokens of a word type in one step and integrates out the random DP measures by using the CRP representation. The whole procedure alternates between three sampling steps:

- Sampling new tag value for each word type;
- Resampling the segmentation for each type;
- Sampling new values for all parameters.

4.1 Tag sampling

The tags will be sampled from the posterior:

$$P(\mathbf{T}|\mathbf{W},\mathbf{S},\mathbf{w},\mathbf{\Theta}), \tag{11}$$

where $\mathbf{W}$ is the set of words in the vocabulary, $\mathbf{T}$ and $\mathbf{S}$ are tags and segmentations assigned to each word type, $\mathbf{w}$ is the actual word sequence, and $\mathbf{\Theta}$ denotes the set of all parameters relevant for tag sampling. For brevity, we will omit $\mathbf{\Theta}$ notation in the formulas below. For a single word type, this posterior can be factored as follows:

$$\begin{aligned} P(T_i = t|\mathbf{T}_{-i},\mathbf{S},\mathbf{W},\mathbf{w}) \sim \\ P(S_i|T_i = t,\mathbf{T}_{-i},\mathbf{S}_{-i})\times \\ P(W_i|T_i = t,\mathbf{T}_{-i},\mathbf{W}_{-i})\times \\ P(\mathbf{w}|T_i = t,\mathbf{T}_{-i},\mathbf{W}), \end{aligned} \tag{12}$$

where $-i$ in the subscript denotes the observations with the i-th word type excluded.

The first term is the segmentation likelihood and can be computed according to the CRP formula:

$$P(S_i|T_i = t,\mathbf{T}_{-i},\mathbf{S}_{-i}) = \prod_{j=1}^{|W_i|}\prod_{s\in S_i}\left(\frac{n_{ts}^{-S_i}}{n_{t\cdot}^{-S_i}+\alpha} + \frac{\alpha(m_s^{-S_i}+\beta P_0(s))}{(n_{t\cdot}^{-S_i}+\alpha)(m_{\cdot}^{-S_i}+\beta)}\right), \tag{13}$$

where the outer product is over the word type count, n_{ts} and m_s denote the number of customers "eating" the segment s under tag t and the number of tables "serving" the segment s across all restaurants respectively, dot represents the marginal counts and α and β are the concentration parameters of the respective DP-s. $-S_i$ in upper index means that the segments belonging to the segmentation of the i-th word type and not calculated into likelihood term yet have been excluded.

The word type likelihood is calculated according to the collapsed Dirichlet-multinomial likelihood formula:

$$P(W_i|T_i = t,\mathbf{T}_{-i},\mathbf{W}_{-i},\mathbf{w}) = \prod_{j=0}^{|W_i|-1}\frac{n_{tW_i}+j+\alpha}{n_{t\cdot}+j+\alpha N} \tag{14}$$

where n_{tW_i} is the number of times the word W_i has been tagged with tag t so far, $n_{t\cdot}$ is the number of total word tokens tagged with the tag t and N is the total number of words in the vocabulary.

The last factor is the word sequence likelihood and covers the transition probabilities. Relevant trigrams are those three containing the current word, and in all contexts where the word token appears in:

$$\begin{aligned} P(\mathbf{w}|T_i = t,\mathbf{T}_{-i},\mathbf{W}) \sim \\ \prod_{c\in C_{W_i}} P(t|t(c_{-2}),t(c_{-1}))\cdot \\ P(t(c_{+1})|t(c_{-1}),t)\cdot \\ P(t(c_{+2})|t,t(c_{+1})) \end{aligned} \tag{15}$$

where C_{W_i} denotes all the contexts where the word type W_i appears in, $t(c)$ are the tags assigned to the context words. All these terms can be calculated with CRP formulas.

4.2 Segmentation sampling

We sample the whole segmentation of a word type as a block with forward-filtering backward-sampling scheme as described in (Mochihashi et al., 2009).

As we cannot sample from the exact marginal conditional distribution due to the dependencies between segments induced by the CRP, we use the Metropolis-Hastings sampler that draws a new proposal with forward-filtering backward-sampling scheme and accepts it with probability $min(1,\frac{P(S_{prop})}{P(S_{old})})$, where S_{prop} is the proposed segmentation and S_{old} is the current segmentation of a word type. The acceptance rate during experiments varied between 94-98%.

For each word type, we build a forward filtering table where we maintain the forward variables $\alpha[t][k]$ that present the probabilities of the last k characters of a t-character string constituting a segment. Define:

$$\alpha[0][0] = 1 \tag{16}$$

$$\alpha[t][0] = 0, \quad t > 0 \tag{17}$$

Then the forward variables can be computed recursively by using dynamic programming algorithm:

$$\alpha[t][k] = p(c_{t-k}^{t}) \sum_{j=0}^{t-k} \alpha[t-k][j], \quad t = 1 \cdots L, \tag{18}$$

where c_m^n denotes the characters $c_m \cdots c_n$ of a string c and L is the length of the word.

Sampling starts from the end of the word because it is known for certain that the word end coincides with the end of a segment. We sample the beginning position k of the last segment from the forward variables $\alpha[t][k]$, where t is the length of the word. Then we set $t = t - k$ and continue to sample the start of the previous to the last segment. This process continues until $t = 0$. The segment probabilities, conditioned on the tag currently assigned to the word type, will be calculated according to the segmentation likelihood formula (13).

4.3 Hyperparameter sampling

All DP and Dirichlet concentration parameters are given vague Gamma(10, 0.1) priors and new values are sampled by using the auxiliary variable sampling scheme described in (Escobar and West, 1995) and the extended version for HDP-s described in (Teh et al., 2006). The segment length control parameter is given uniform Beta prior and its new values are sampled from the posterior which is also a Beta distribution.

5 Results

5.1 Evaluation

We test the POS induction part of the model on all languages in the Multext-East corpora (Erjavec, 2010) as well as on the free corpora from CONLL-X Shared Task[1] for Dutch, Danish, Swedish and Portuguese. The evaluation of morphological segmentations is based on the Morpho Challenge gold segmented wordlists for English, Finnish and Turkish[2]. We gathered the sentences from Europarl corpus[3] for English and Finnish, and use the Turkish text data from the Morpho Challenge 2009[4]. Estonian gold standard segmentations have been obtained from the Estonian morphologically annotated corpus[5].

We report three accuracy measures for tagging: greedy one-to-one mapping (**1-1**) (Haghighi and Klein, 2006), many-to-one mapping (**m-1**) and V-measure (**V-m**) (Rosenberg and Hirschberg, 2007).

Segmentation is evaluated on the basis of standard F-score which is the harmonic mean of precision and recall.

5.2 Experimental results

For each experiment, we made five runs with random initializations and report the results of the median. The sampler was run 200 iterations for burnin, after which we collected 5 samples, letting the sampler to run for another 200 iterations between each two sample. We start with 15 segmenting iterations during each Gibbs iteration to enable the segmentation sampler to burnin to the current tagging state, and gradually reduce this number to one. Segmentation likelihood term for tagging is calculated on the basis of the last segment only because this setting gave the best results in preliminary experiments and it also makes the whole computation less expensive.

The first set of experiments was conducted to test the model tagging accuracy on different languages mentioned above. The results obtained were in general slightly lower than the current state-of-the-art and the number of tags learned was generally bigger than the number of gold standard tags. We observed that different components making up the corpus logarithmic probability have different magnitudes. In particular, we found that the emission probability component in log-scale is roughly four times smaller than the transition probability. This observation motivated introducing the likelihood scaling heuristic into the model to scale the emission probability up. We tried a couple of different scaling factors on Multext-East English corpus and then set its value to 4 for all languages for the rest of the experiments. This improved the tagging results consistently across all languages.

[1] http://ilk.uvt.nl/conll/free_data.html

[2] http://research.ics.tkk.fi/events/morphochallenge2010/datasets.shtml

[3] http://www.statmt.org/europarl/

[4] http://research.ics.tkk.fi/events/morphochallenge2009/datasets.shtml

[5] http://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=eng

POS induction results are given in **Table 1**. When comparing these results with the recently published results on the same corpora (Christodoulopoulos et al., 2011; Blunsom and Cohn, 2011; Lee et al., 2010) we can see that our results compare favorably with the state-of-the-art, resulting with the best published results in many occasions. The number of tag clusters learned by the model corresponds surprisingly well to the number of true coarse-grained gold standard tags across all languages. There are two things to note here: 1) the tag distributions learned are influenced by the likelihood scaling heuristic and more experiments are needed in order to fully understand the characteristics and influence of this heuristic; 2) as the model is learning the coarse-grained tagset consistently in all languages, it might as well be that the POS tags are not as dependent on the morphology as we assumed, especially in inflectional languages with many derivational and inflectional suffixes, because otherwise the model should have learned a more fine-grained tagset.

Segmentation results are presented in **Table 2**. For each language, we report the lexicon-based precision, recall and F-measure, the number of word types in the corpus and and number of word types with gold segmentation available. The reported standard deviations show that the segmentations obtained are stable across different runs which is probably due to the blocked sampler. We give the segmentation results both with and without likelihood scaling heuristic and denote that while the emission likelihood scaling improves the tagging accuracy, it actually degrades the segmentation results.

It can also be seen that in general precision score is better but for Estonian recall is higher. This can be explained by the characteristics of the evaluation data sets. For English, Finnish and Turkish we use the Morpho Challenge wordlists where the gold standard segmentations are fine-grained, separating both inflectional and derivational morphemes. Especially derivational morphemes are hard to learn with pure data-driven methods with no knowledge about semantics and thus it can result in undersegmentation. On the other hand, Estonian corpus separates only inflectional morphemes which thus leads to higher recall. Some difference can also come from the fact that the sets of gold-segmented word types for other languages are much smaller than in Estonian and thus it would be interesting to see whether and how the results would change if the evaluation could be done on all word types in the corpus for other languages as well. In general, undersegmentation is more acceptable than oversegmentation, especially when the aim is to use the resulting segmentations in some NLP application.

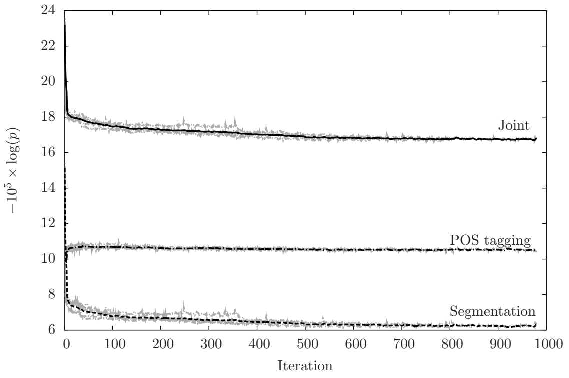

Figure 2: Log-likelihood of samples plotted against iterations. Dark lines show the average over five runs, grey lines in the back show the real samples.

Next, we studied the convergence characteristics of our model. For these experiments we made five runs with random initializations on Estonian corpus and let the sampler run up to 1100 iterations. Samples were taken after each ten iterations. **Figure 2** shows the log-likelihood of the samples plotted against iteration number. Dark lines show the averages over five runs and gray lines in the background are the likelihoods of real samples showing also the variance. We first calculated the full likelihood of the samples (the solid line) that showed a quick improvement during the first few iterations and then stabilized by continuing with only slow improvements over time. We then divided the full likelihood into two factors in order to see the contribution of both tagging and segmentation parts separately. The results are quite surprising. It turned out that the random tagging initializations are very good in terms of probability and as a matter of fact much better than the data can support and thus the tagging likelihood drops quite significantly after the first iteration and then continues with very slow improvements. The matters are totally different with segmentations where the initial random segmentations result in a low likelihood that improves heavily

	Types	1-1	m-1	V-m	Induced	True	Best Pub.		
Bulgarian	15103	**50.3 (0.9)**	**71.9 (3.8)**	54.9 (2.2)	13 (1.6)	12	-	66.5*	**55.6***
Czech	17607	**46.0 (1.0)**	60.7 (1.6)	46.2 (0.7)	12 (0.8)	12	-	**64.2***	**53.9***
Danish	17157	**53.2 (0.2)**	69.5 (0.1)	52.7 (0.4)	14 (0.0)	25	43.2†	**76.2⋆**	**59.0***
Dutch	27313	**60.5 (1.9)**	**74.0 (1.6)**	**59.1 (1.1)**	22 (0.0)	13	55.1†	71.1*	54.7*
English	9196	**67.4 (0.1)**	**79.8 (0.1)**	**66.7 (0.1**	13 (0.0)	12	-	73.3*	63.3*
Estonian	16820	**47.6 (0.9)**	**64.5 (1.9)**	45.6 (1.4)	14 (0.5)	11	-	64.4*	**53.3***
Farsi	11319	**54.9 (0.1)**	**65.3 (0.1)**	**52.1 (0.1)**	13 (0.5)	12	-	-	-
Hungarian	19191	**62.1 (0.7)**	**71.4 (0.3)**	**56.0 (0.6)**	11 (0.9)	12	-	68.2*	54.8*
Polish	19542	**48.5 (1.8)**	**59.6 (1.9)**	**45.4 (1.0)**	13 (0.8)	12	-	-	-
Portuguese	27250	45.4 (1.1)	71.3 (0.3)	55.4 (0.3)	21 (1.1)	16	**56.5†**	**78.5⋆**	**63.9***
Romanian	13822	**44.3 (0.5)**	60.5 (1.7)	46.7 (0.5)	14 (0.8)	14	-	**61.1***	**52.3***
Serbian	16813	**40.1 (0.2)**	60.1 (0.2)	43.5 (0.2)	13 (0.0)	12	-	**64.1***	**51.1***
Slovak	18793	**44.1 (1.5)**	**56.2 (0.8)**	**41.2 (0.6)**	14 (1.1)	12	-	-	-
Slovene	16420	**51.6 (1.5)**	66.8 (0.6)	51.6 (1.0)	12 (0.7)	12	-	**67.9***	**56.7***
Swedish	18473	**50.6 (0.1)**	60.3 (0.1)	55.8 (0.1)	17 (0.0)	41	38.5†	**68.7***	**58.9***

Table 1: Tagging results for different languages. For each language we report median one-to-one (1-1), many-to-one (m-1) and V-measure (V-m) together with standard deviation from five runs where median is taken over V-measure. **Types** is the number of word types in each corpus, **True** is the number of gold tags and **Induced** reports the median number of tags induced by the model together with standard deviation. **Best Pub.** lists the best published results so far (also 1-1, m-1 and V-m) in (Christodoulopoulos et al., 2011)*, (Blunsom and Cohn, 2011)⋆ and (Lee et al., 2010)†.

		Precision	Recall	F1	Types	Segmented
Estonian	without LLS	43.5 (0.8)	59.4 (0.6)	50.3 (0.7)	16820	16820
	with LLS	42.8 (1.1)	54.6 (0.7)	48.0 (0.9)		
English	without LLS	69.0 (1.3)	37.3 (1.5)	48.5 (1.1)	20628	399
	with LLS	59.8 (1.8)	29.0 (1.0)	39.1 (1.3)		
Finnish	without LLS	56.2 (2.5)	29.5 (1.7)	38.7 (2.0)	25364	292
	with LLS	56.0 (1.1)	28.0 (0.6)	37.4 (0.7)		
Turkish	without LLS	65.4 (1.8)	44.8 (1.8)	53.2 (1.7)	18459	293
	with LLS	68.9 (0.8)	39.2 (1.0)	50.0 (0.6)		

Table 2: Segmentation results on different languages. Results are calculated based on word types. For each language we report precision, recall and F1 measure, number of word types in the corpus and number of word types with gold standard segmentation available. For each language we report the segmentation result without and with emission likelihood scaling (without LLS and with LLS respectively).

with the first few iterations and then stabilizes but still continues to improve over time. The explanation for this kind of model behaviour needs further studies and we leave it for future work.

Figure 3 plots the V-measure against the tagging factor of the log-likelihood for all samples. It can be seen that the lower V-measure values are more spread out in terms of likelihood. These points correspond to the early samples of the runs. The samples taken later during the runs are on the right in the figure and the positive correlation between the V-measure and likelihood values can be seen.

Next we studied whether the morphological seg-

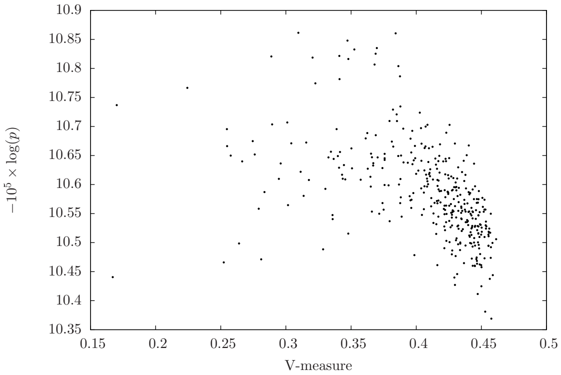

Figure 3: Tagging part of log-likelihood plotted against V-measure

	1-to-1	m-to-1	V-m
Fixed seg	40.5 (1.5)	53.4 (1.0)	37.5 (1.3)
Learned seg	47.6 (0.4)	64.5 (1.9)	45.6 (1.4)
	Precision	**Recall**	**F1**
Fixed tag	36.7 (0.3)	56.4 (0.2)	44.5 (0.3)
Learned tag	42.8 (1.1)	54.6 (0.7)	48.0 (0.9)
Morfessor	51.29	52.59	51.94

Table 3: Tagging and segmentation results on Estonian Multext-East corpus (Learned seg and Learned tag) compared to the semisupervised setting where segmentations are fixed to gold standard (Fixed seg) and tags are fixed to gold standard (Fixed tag). Finally the segmentatation results from Morfessor system for comparison are presented.

mentations and POS tags help each other in the learning process. For that we conducted two semisupervised experiments on Estonian corpus. First we provided gold standard segmentations to the model and let it only learn the tags. Then, we gave the model gold standard POS tags and only learned the segmentations. The results are given in **Table 3**. We also added the results from joint unusupervised learning for easier comparison. Unfortunately we cannot repeat this experiment on other languages to see whether the results are stable across different languages because to our knowledge there is no other free corpus with both gold standard POS tags and morphological segmentations available.

From the results it can be seen that the unsupervised learning results for both tagging and segmentation are better than the results obtained from semisupervised learning. This is surprising because one would assume that providing gold standard data would lead to better results. On the other hand, these results are encouraging, showing that learning two dependent tasks in a joint model by unsupervised manner can be as good or even better than learning the same tasks separately and providing the gold standard data as features.

Finally, we learned the morphological segmentations with the state-of-the-art morphology induction system Morfessor baseline[6] (Creutz and Lagus, 2005) and report the best results in the last row of **Table 3**. Apparently, our joint model cannot beat Morfessor in morphological segmentation and when using the emission likelihood scaling that influences the tagging results favorably, the segmentation results get even worse. Altough the semisupervised experiments showed that there are dependencies between tags and segmentations, the conducted experiments do not reveal of how to use these dependencies for helping the POS tags to learn better morphological segmentations.

6 Related Work

We will review some of the recent works related to Bayesian POS induction and morphological segmentation.

One of the first Bayesian POS taggers is described in (Goldwater and Griffiths, 2007). The model presented is a classical HMM with multinomial transition and emission distributions with Dirichlet priors. Inference is done using a collapsed Gibbs sampler and concentration parameter values are learned during inference. The model is token-based, allowing different words of the same type in different locations to have a different tag. This model can actually be classified as semi-supervised as it assumes the presence of a tagging dictionary that contains the list of possible POS tags for each word type - an assumption that is clearly not realistic in an unsupervised setting.

Models presented in (Christodoulopoulos et al., 2011) and (Lee et al., 2010) are also built on Dirichlet-multinomials and, rather than defining a sequence model, present a clustering model based on features. Both report good results on type basis and use (among others) also morphological features, with (Lee et al., 2010) making use of fixed length suffixes and (Christodoulopoulos et al., 2011) using the suffixes obtained from an unsupervised morphology induction system.

Nonparametric Bayesian POS induction has been studied in (Blunsom and Cohn, 2011) and (Gael et al., 2009). The model in (Blunsom and Cohn, 2011) uses Pitman-Yor Process (PYP) prior but the model itself is finite in the sense that the size of the tagset is fixed. Their model also captures morphological regularities by modeling the generation of words with character n-grams. The model in (Gael et al., 2009) uses infinite state space with Dirichlet Process prior. The model structure is classical HMM consisting

[6] http://www.cis.hut.fi/projects/morpho/

only of transitions and emissions and containing no morphological features. Inference is done by using beam sampler introduced in (Gael et al., 2008) which enables parallelized implementation.

One close model for morphology stems from Bayesian word segmentation (Goldwater et al., 2009) where the task is to induce word borders from transcribed sentences. Our segmentation model is in principle the same as the unigram word segmentation model and the main difference is that we are using blocked sampler while (Goldwater et al., 2009) uses point-wise Gibbs sampler by drawing the presence or absence of the word border between every two characters.

In (Goldwater et al., 2006) the morphology is learned in the adaptor grammar framework (Johnson et al., 2006) by using a PYP adaptor. PYP adaptor caches the numbers of observed derivation trees and forces the distribution over all possible trees to take the shape of power law. In the PYP (and also DP) case the adaptor grammar can be interpreted as PYP (or DP) model with regular PCFG distribution as base measure.

The model proposed in (Goldwater et al., 2006) makes several assumptions that we do not: 1) segmentations have a fixed structure of stem and suffix; and 2) there is a fixed number of inflectional classes. Inference is performed with Gibbs sampler by sampling for each word its stem, suffix and inflectional class.

7 Conclusion

In this paper we presented a joint unsupervised model for learning POS tags and morphological segmentations with hierarchical Dirichlet Process model. Our model induces the number of POS clusters from data and does not contain any hand-tuned parameters. We tested the model on many languages and showed that by introcing a likelihood scaling heuristic it produces state-of-the-art POS induction results. We believe that the tagging results could further be improved by adding additional features concerning punctuation, capitalization etc. which are heavily used in the other state-of-the-art POS induction systems but these features were intentionally left out in the current model for enabling to test the concept of joint modelling of two dependent tasks.

We found some evidence that the tasks of POS induction and morphological segmentation are dependent by conducting semisupervised experiments where we gave the model gold standard tags and segmentations in turn and let it learn only segmentations or tags respectively and found that the results in fully unsupervised setting are better. Despite of that, the model failed to learn as good segmentations as the state-of-the-art morphological segmentation model Morfessor. One way to improve the segmentation results could be to use segment bigrams instead of unigrams.

The model can serve as a basis for several further extensions. For example, one possibility would be to expand it into multilingual setting in a fashion of (Naseem et al., 2009), or it could be extended to add the joint learning of morphological paradigms of the words given their tags and segmentations in a manner described by (Dreyer and Eisner, 2011).

Acknowledgments

We would like to thank the anonymous reviewers who helped to improve the quality of this paper. This research was supported by the Estonian Ministry of Education and Research target-financed research theme no. 0140007s12, and by European Social Funds Doctoral Studies and Internationalisation Programme DoRa.

References

D. Aldous. 1985. Exchangeability and related topics. In *École d'été de Probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer.

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

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 for Computational Linguistics: Human Language Technologies - Volume 1*, pages 865–874.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceed-*

ings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 575–584.

Christos Christodoulopoulos, Sharo Goldwater, and Mark Steedman. 2011. A Bayesian mixture model for PoS induction using multiple features. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 638–647, Edinburgh, Scotland, UK.

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*, pages 59–66.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *In Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 106–113.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet Process mixture model. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 616–627.

Toma Erjavec. 2010. MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Michael D. Escobar and Mike West. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430).

Thomas S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. Beam sampling for the infinite Hidden Markov Model. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1088–1095.

Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. 2009. The infinite HMM for unsupervised PoS tagging. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, pages 678–687.

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, Prague, Czech Republic.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18*, Cambridge, MA.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 320–327, New York City, USA.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19*, pages 641–648.

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

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. 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: Volume 1 - Volume 1*, pages 100–108.

Tahira Naseem, Benjamin Snyder, Jacob Eisenstein, and Regina Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *Journal of Artificial Intelligence Research*, 36:1–45.

Lawrence R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286.

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 (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic.

Yee Whye Teh, Michel I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.