Unsupervised Russian POS Tagging with Appropriate Context

Li Yang, Erik Peterson, John Chen, Yana Petrova, and Rohini Srihari
Janya Inc.
1408 Sweet Home Road, Suite 1
Amherst, NY 14228, USA
lyang,epeterson,jchen,ypetrova,rohini@janya.com

Abstract

While adopting the contextualized hidden Markov model (CHMM) framework for unsupervised Russian POS tagging, we investigate the possibility of utilizing the left, right, and unambiguous context in the CHMM framework. We propose a backoff smoothing method that incorporates all three types of context into the transition probability estimation during the expectation-maximization process. The resulting model with this new method achieves overall and disambiguation accuracies comparable to a CHMM using the classic backoff smoothing method for HMM-based POS tagging from (Thede and Harper, 1999).

1 Introduction

A careful review of the work on unsupervised POS tagging in the past two decades reveals that the hidden Markov model (HMM) has been the standard approach since the seminal work of (Kupiec, 1992) and (Merialdo, 1994) and that researchers sought to improve HMM-based unsupervised POS tagging from a variety of perspectives, including exploring dictionary usage, context utilization, sparsity control and modeling, and parameter and model updates tuned to linguistic features. For example, (Banko and Moore, 2004) and (Goldberg et al., 2008) utilized contextualized HMM (CHMM) to capture rich context. To account for sparsity, (Goldwater and Griffiths, 2007) and (Johnson, 2007) utilized the Dirichlet hyperparameters of the Bayesian HMM. (Berg-Kirkpatrick et al., 2010) integrated the discriminative logistic regression model into the M-step of the standard generative model to allow rich linguistically-motivated features.

Unsupervised systems went beyond the mainstream HMM framework by employing methods such as prototype-driven clustering (Haghighi and Klein, 2006; Abend et al., 2010), Bayesian LDA (Toutanova and Johnson, 2007), integer programming (Ravi and Knight, 2009), and K-means clustering (Lamar et al., 2010).

Despite this large body of work, little effort has been devoted to unsupervised Russian POS tagging. Supervised Russian POS systems emerged in recent years. For example, eleven supervised systems entered the POS track of the 2010 Russian Morphological Parsers Evaluation [1]. Although the top two systems from the 2010 Evaluation achieved near perfect accuracy over the Russian National Corpus, little has been done on unsupervised Russian POS tagging. In this paper, we present our solution to unsupervised Russian POS tagging by adopting the CHMM. Our choice is based on the accuracy and efficiency of CHMM, an identical rationale to that behind (Goldberg et al., 2008).

We aim to achieve two goals. First, we intend to resolve the potential issue of missing useful contextual features by the backoff smoothing scheme in (Thede and Harper, 1999) and (Goldberg et al., 2008) for transition probabilities. Second, we explore the possibility of incorporating unambiguous context into transition probability estimation in an HMM framework. We propose a novel plan to achieve both goals in a unified approach.

In the following, we adopt the CHMM for unsupervised Russian POS tagging in section 2. Section 3 highlights the potential issue of missing useful left context in the backoff scheme by (Thede and Harper, 1999). Section 4 illustrates an updated backoff scheme to resolve this potential issue. This scheme also unifies the left, right, and unambiguous context. The experiments and discussion are presented in section 5. We present conclusions in section 6.

[1] See http://ru-eval.ru/tables_index.html

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 30–34,
Chiang Mai, Thailand, November 8-12, 2011.

2 CHMM for Russian POS Tagging

Our system is built upon the architecture of a contextualized HMM. Like other existing unsupervised HMM-based POS systems, the task of unsupervised POS tagging for us is to construct an HMM to predict the most likely POS tag sequence in the new data, given only a dictionary listing all possible parts-of-speech of a set of words and a large amount of unlabeled text for training.

Traditionally, the transition probability in a second-orderHMMis given by $p(t_i|t_{i-2}t_{i-1})$, and the emission probability by$p(w_i|t_i)$ ((Kriouile, 1990; Banko and Moore, 2004)). The CHMM, such as such as (Banko and Moore, 2004), (Adler, 2007), and (Goldberg et al., 2008),, incorporates more context into the transition and emission probabilities. Here, we adopt the transition probability p(ti—ti1ti+1) of (Adler, 2007) and (Goldberg et al., 2008) and the emission probability p(wi—titi+1) of (Adler, 2007).

Our training corpus consists of all 406,342 words of the plain text for training from the Appen Russian Named Entity Corpus [2], containing textual documents from a variety of sources. We created a POS dictionary for all 61,020 unique tokens in this corpus, using the output from the Russian lemmatizer [3]. The lemmatizer returns the stems of words and a list of POS tags for each word, relying on the morphology dictionary of the AOT Team [4]. Our tag set consists of 17 tags, comparable to those [5] used in Russian National Corpus (RNC), with the only addition of the Punct tag for punctuation marks. We relied on the Appen data because we did not have access to the RNC when our project was being developed. But we hope to be able to train and test out system with the RNC in the future.

3 Parameter Estimation and a Potential Issue

Given the model and resources for training described in section 2, we estimate the model parameters for our CHMM by following the standard EM procedures. During pre-processing, the dictionary is consulted, and a list of potential POS tags is provided for each word/token in the training sequence. In case of unknown words, the morphology analyzer built in the Russian lemmatizer suggests a list of tags. If the morphology analyzer does not make any suggestion, a list of open POS tags are assigned to the unknown words.

The potential POS tags in the training data provide counts to roughly esitimate the initial transition and emission probabilities. (Adler, 2007) initialized transition probabilities using a small portion of the training data. In our work, we initialize the emission probabilities using 20% of the training data with $p(w_i|t_it_{i+1}) = \frac{\#(w_i,t_i,t_{i+1})}{\#(t_i,t_{i+1})}$. During the EM process, we use additive smoothing when estimating $p(w_i|t_it_{i+1})$ (Chen, 1996).

We initialize the transition probabilities $p(t_i|t_{i-1}t_{i+1})$ with a uniform distribution. When re-estimating $p(t_i|t_{i-1}t_{i+1})$, we use the method from (Thede and Harper, 1999) for backoff smoothing in equation (1).

$$\hat{p}(t_i|t_{i-1}t_{i+1}) = \lambda_3 \frac{N_3}{C_2} + (1-\lambda_3)\lambda_2 \cdot \frac{N_2}{C_1} + (1-\lambda_3)(1-\lambda_2) \cdot \frac{N_1}{C_0} \quad (1)$$

The λ coefficients are calculated the same way as in (Thede and Harper, 1999), that is $\lambda_2 = \frac{\log(N_2+1)+1}{\log(N_2+2)}$ and $\lambda_3 = \frac{\log(N_3+1)+1}{\log(N_3+2)}$. The counts, N_i and C_j are modified for our unsupervised CHMM, as shown in Table 1. Note that N_2 captures the counts of the bi-gram t_it_{i+1}, consisting of the current state t_i and its right context t_{i+1}.

(Thede and Harper, 1999) and (Goldberg et al., 2008) show that equation (1) is quite effective in both supervised and unsupervised scenarios. However, in our case where Russian is concerned, there are situations where equation (1) may not give good estimates.

Through RNC's online search tool, we discovered that the word from a specific set of pronouns following the comma is always analyzed as a conjunction, which itself can be followed by a number of possible POS tags. This set includes ambiguous words such as *chto* and *chem*. Although the Appen corpus does not come with POS tags, our Russian linguist observed similar linguistic regularties in the corpus. Some examples regarding *chto* from

$N_1 = N_1^e$	estimated counts of t_{i+1}
$N_2 = N_2^e$	estimated counts of t_it_{i+1}
$N_3 = N_3^e$	estimated counts of $t_{i-1}t_it_{i+1}$
$C_0 = C_0^e$	estimated total # of tags
$C_1 = C_1^e$	estimated counts of t_i
$C_2 = C_2^e$	estimated counts of $t_{i-1}t_{i+1}$

Table 1: Estimated counts as superscript ^{e}.

[2]Licensed from http://www.appen.com.au/

[3]Available at http://lemmatizer.org/en/

[4]See http://aot.ru/

[5]Listed at http://www.ruscorpora.ru

Appen are listed below.

Example 1 *,(Punct) chto(CONJ) na(PREP)*

Gloss comma and/or/that on

Example 2 *,(Punct) chto(CONJ) gotovy(ADJ)*

Gloss comma and/or/that ready

In the preceding examples, the comma to the left of chto provides for a useful clue. However, a potential issue arises when we estimate p(ti-1titi+1) using equation (1). That is, when the trigram $t_{i-1}t_it_{i+1}$ is rare and the first term of the equation is very small, the second term will affect $\hat{p}(t_{i-1}t_it_{i+1})$ more. The count, N_2, in the second term is for the bi-gram (*chto-CONJ, right word-POS*), right word-POS) but not for (*left word-comma, chto-CONJ*). Therefore, the useful clue in the latter bi-gram is missed. To resolve this, one cannot simply switch to the left context in N_2 because there are cases where the right context provides more of a clue. For example, observed from the Russian National Corpus, adjectival pronouns are only followed by a noun or an adjective and a noun, where the right context of adjectival pronouns are more important for disambiguating the adjectival pronouns. Several more examples from the Appen data where the left or right context contributing to disambiguation are listed in the Appendix.

4 Incorporating All Three Types of Context

Several systems made use of the information provided in unambiguous POS tag sequence. (Brill, 1995) learned rules from the context of unambiguous words. (Mihalcea, 2003) created equivalence classes from unambiguous words for training. We expected the assumption that unambiguous context helps with disambiguation to hold for Russian as well.

$N_1 = N_1^u$, # of unambiguous counts of t_{i+1}
$N_2^L = N_2^{uL}$, # of unamb. bi-gram $t_{i-1}t_i$ w left context t_{i-1}
$N_2^R = N_2^{uR}$, # of unamb. bi-gram t_it_{i+1} w right context t_{i+1}
$N_3 = N_3^u$, # of unamb. tri-gram $t_{i-1}t_it_{i+1}$
$C_0 = C_0^u$, total # of unamb. tags
$C_1 = C_1^u$, # of unamb. t_i
$C_2 = C_2^u$, # of unamb. bi-gram of $t_{i-1}t_{i+1}$

Table 2: Counts of unambiguous tri-grams, bi-grams, and unigrams. The superscript ^{u} stands for unambiguous counts.

$N_1^u \leftarrow N_1^e$	estimated counts of t_{i+1}
$N_2^{uL} \leftarrow N_2^{eL}$	estimated counts of $t_{i-1}t_i$
$N_2^{uR} \leftarrow N_2^{eR}$	estimated counts of t_it_{i+1}
$N_3^u \leftarrow N_3^e$	estimated counts of $t_{i-1}t_it_{i+1}$
$C_0^u \leftarrow C_0^e$	estimated total # of tags
$C_1^u \leftarrow C_1^e$	estimated counts of t_i
$C_2^u \leftarrow C_2^e$	estimated counts of $t_{i-1}t_{i+1}$

Table 3: Replacement plan for unambiguous counts

In the Appen training corpus, 84% of the words/tokens have a unique POS tag, based on our dictionary and the Russian lemmatizer. We can easily spot examples in the corpus where unambiguous context helps with disambiguation. Again, in our earlier example, *,(Punct) chto(CONJ) na(PREP)*, the unambiguous left context ',' reveals that *chto* is a CONJ instead of a PRON. To take advantage of the unambiguous context, we collect the counts for all unambiguous tri-gram and bi-gram sequences in the Appen training corpus and integrate these counts into equation (2) through the equivalence in Table 2.

$$\begin{aligned}\hat{p}(t_i|t_{i-1}t_{i+1}) &= \lambda_3\frac{N_3}{C_2} \\ &\quad +(1-\lambda_3)\lambda_2\cdot\frac{N_2^L}{C_1^L}\times\frac{N_2^R}{C_1^R} \\ &\quad +(1-\lambda_3)(1-\lambda_2)\cdot\frac{N_1}{C_0} \quad (2)\end{aligned}$$

where $\lambda_2 = \frac{\log(N_2^L+1)+1}{\log(N_2^L+2)} \times \frac{\log(N_2^R+1)+1}{\log(N_2^R+2)}$, and $\lambda_3 = \frac{\log(N_3+1)+1}{\log(N_3+2)}$. λ_2 incorporates both the left and right context. The unambiguous counts are defined in Table 2.

Now that the new backoff smoothing plan combines both the left and right unambiguous bi-gram counts, we extend this plan to cover the cases where the unambiguous tri/bi/uni-grams are not available, by replacing them with the estimated counts from Table 1. Table 3 displays the scheme for replacing an unambiguous count with an estimated count from the EM process.

5 Experiments and Results

We designed three experiments to test three combinations of the context, in addition to experimenting with a traditional second-order HMM. The Appen corpus contains a development set and an

Model & setting(s)	Overall Accuracy	Disamb. Accuracy
2nd-order HMM	94.88%	63.42%
CHMM_left_context	95.72%	69.42%
CHMM_right_context	96.05%	71.78%
CHMM_unique_ ←_left/right context	96.06%	71.85%

Table 4: Experiments, overall and disambiguation accuracies over test data

evaluation set. We passed both sets through the Russian lemmatizer to obtain POS tags for the data and had the tags manually corrected by a Russian linguist. Thus, we have created both development and evaluation data. 14% of words/tokens in both development and evaluation data have multiple POS tags. Table 4 summarizes our experimental settings and results over the evaluation data.

The second-order HMM was trained with the traditional transition probability $p(t_i|t_{i-2}t_{i-1})$ and emission probability $p(w_i|t_i)$. It gained an overall accuracy of 94.88%, and was able to correctly disambiguate 63.42% of the ambiguous words/tokens.

All three CHMM models were trained with the emission probability $p(w_i|t_it_{i+1})$ initialized with 20% of the unlabeled training corpus. Model *CHMM_left_context* considered the left context bigram $t_{i-1}t_i$ when calculating the second term in equation (1). Model *CHMM_right_context* considered the right context bi-gram t_it_{i+1} when calculating the same term. Model *CHMM_unique_* ← *_left/right* unified both unambiguous context counts and estimated counts for left and right context from the EM process, using equation (2).

All CHMM models achieved accuracies 1% higher than the HMM, while the disambiguation accuracies from the former three are $7-9\%$ higher than the latter. This shows that the CHMM models capture more useful context information for Russian POS tagging than the traditional HMM. At the same time, the overall and disambiguation accuracies between *CHMM_right_context* and *CHMM_unique_* ← *_left/right* are comparable. Error analyses indicate that a backoff scheme for emission probabilities is also needed to incorporate the left context.

6 Conclusion and Future Work

We adopted the CHMM to unsupervised Russian POS tagging. The CHMM models using either the left or right context were able to outperform the traditional second-order HMM. To resolve the potential issue of missing out the left context with the classic smoothing scheme in (Thede and Harper, 1999), we experimented with an approach to unifying the information provided in the left, right, and unambiguous contexts. The results from the latter were comparable to a CHMM with the classic backoff smoothing method in (Thede and Harper, 1999), although we expected a more significant improvement. We plan to investigate a backoff scheme for emission probabilities where we will incorporate the left context as well, while currently we only rely on additive smoothing for emission probabilities.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions. Our work was partially funded by the Air Force Research Laboratory/RIEH in Rome, New York through contracts FA8750-09-C-0038 and FA8750-10-C-0124.

References

Omri Abend, Roi Reichart, and Ari Rappoport. 2010. Improved unsupervised pos induction through prototype discovery. In *Proceedings of the 48th ACL*.

Meni Adler. 2007. *Hebrew Morphological Disambiguation*. Ph.D. thesis, University of the Negev.

Michele Banko and Robert C. Moore. 2004. Part of speech tagging in context. In *Proceedings of the 20th international conference on Computational Linguistics*.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Ct, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of NAACL 2010*.

Eric Brill, 1995. *Very Large*, chapter Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging, pages 1–13. Kluwer Academic Press.

Stanley F. Chen. 1996. *Building Probabilistic Models for Natural Language*. Ph.D. thesis, Harvard University.

Yoav Goldberg, Meni Adler, and Michael Elhadad. 2008. Em can find pretty good pos taggers (when given a good start). In *Proceedings of ACL-08: HLT*.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th ACL*.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the main conference on HLT-NAACL*.

Mark Johnson. 2007. Why doesnt em find good hmm pos-taggers. In *n EMNLP*.

Abdelaziz Kriouile. 1990. Some improvements in speech recognition algorithms based on hmm. In *Acoustics, Speech, and Signal Processing*.

Julian Kupiec. 1992. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6:225–242.

Michael Lamar, Yariv Maron, and Elie Bienenstock. 2010. Latent descriptor clustering for unsupervised pos induction. In *EMNLP 2010*.

Bernard Merialdo. 1994. Tagging english text with a probabilistic model. *Computational Linguistics*, 20:155–171.

Rada Mihalcea. 2003. The role of non-ambiguous words in natural language disambiguation. In *Proceedings of the Conference on RANLP*.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of ACL-IJCNLP 2009,*, pages 504–512.

Scott M. Thede and Mary P. Harper. 1999. A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the 37th Annual Meeting of the ACL*.

Kristina Toutanova and Mark Johnson. 2007. A bayesian lda-based model for semi-supervised part-of-speech tagging. In *Proceedings of NIPS*.

Appendix: Linguistic Patterns Observed in Appen

In Section 3, we illustrated how the left context helped to disambiguate *chto*. In the following we present several more examples from the Appen corpus illustrating the helpful left or right context. While the patterns our Russian linguist observed are common in both the RNC and Appen, the counts and statistics regarding each pattern are unavailable for reporting because the RNC was then inaccessible to us and Appen was not tagged with POS tags.

Examples 3 through 7 show that the left context of *chem*, *poka*, and *kak* helps to disambiguate them as conjuctions.

Example 3 *,(Punct) chem(CONJ) v(PREP) stolitse(NOUN)*

Gloss comma and/than in capital

Example 4 *,(Punct) poka(CONJ) eta(PRONOUN)*

Gloss comma yet this

Example 5 *,(Punct) poka(CONJ) Sovet(NOUN)*

Gloss comma yet council

Example 6 *,(Punct) kak(CONJ) dva(NUMERAL) neudachnika(NOUN)*

Gloss comma as two losers

Example 7 *,(Punct) kak(CONJ) on(PRONOUN)*

Gloss comma as he

The next examples show that the right context determines the adjectival tag, *PRONOUN_P*, of the pronouns.

Example 8 *obekty(NOUN) svoey(PRONOUN_P) sistemy(NOUN)*

Gloss units their/they system

Example 9 *esli(CONJ) mnogie(PRONOUN_P) mnogie(NOUN)*

Gloss if many/various emigrants