Part-of-Speech Tagging using Conditional Random Fields: Exploiting Sub-Label Dependencies for Improved Accuracy

Miikka Silfverberg[a] **Teemu Ruokolainen**[b] **Krister Lindén**[a] **Mikko Kurimo**[b]

[a] Department of Modern Languages, University of Helsinki,
firstname.lastname@helsinki.fi

[b] Department of Signal Processing and Acoustics, Aalto University,
firstname.lastname@aalto.fi

Abstract

We discuss part-of-speech (POS) tagging in presence of large, fine-grained label sets using conditional random fields (CRFs). We propose improving tagging accuracy by utilizing dependencies within sub-components of the fine-grained labels. These sub-label dependencies are incorporated into the CRF model via a (relatively) straightforward feature extraction scheme. Experiments on five languages show that the approach can yield significant improvement in tagging accuracy in case the labels have sufficiently rich inner structure.

1 Introduction

We discuss part-of-speech (POS) tagging using the well-known conditional random field (CRF) model introduced originally by Lafferty et al. (2001). Our focus is on scenarios, in which the POS labels have a *rich inner structure*. For example, consider

PRON+1SG	V+NON3SG+PRES	N+SG	
I	like	ham	,

where the *compound* labels PRON+1SG, V+NON3SG+PRES, and N+SG stand for pronoun first person singular, verb non-third singular present tense, and noun singular, respectively. Fine-grained labels occur frequently in morphologically complex languages (Erjavec, 2010; Haverinen et al., 2013).

We propose improving tagging accuracy by utilizing dependencies within the *sub-labels* (PRON, 1SG, V, NON3SG, N, and SG in the above example) of the compound labels. From a technical perspective, we accomplish this by making use of the fundamental ability of the CRFs to incorporate arbitrarily defined feature functions. The newly-defined features are expected to alleviate data sparsity problems caused by the fine-grained labels. Despite the (relative) simplicity of the approach, we are unaware of previous work exploiting the sub-labels to the extent presented here.

We present experiments on five languages (English, Finnish, Czech, Estonian, and Romanian) with varying POS annotation granularity. By utilizing the sub-labels, we gain significant improvement in model accuracy given a sufficiently fine-grained label set. Moreover, our results indicate that exploiting the sub-labels can yield larger improvements in tagging compared to increasing model order.

The rest of the paper is organized as follows. Section 2 describes the methodology. Experimental setup and results are presented in Section 3. Section 4 discusses related work. Lastly, we provide conclusions on the work in Section 5.

2 Methods

2.1 Conditional Random Fields

The (unnormalized) CRF model (Lafferty et al., 2001) for a sentence $x = (x_1, \ldots, x_{|x|})$ and a POS sequence $y = (y_1, \ldots, y_{|x|})$ is defined as

$$p\,(y\,|\,x; \boldsymbol{w}) \propto \prod_{i=n}^{|x|} \exp\Big(\boldsymbol{w}\cdot\boldsymbol{\phi}(y_{i-n}, \ldots, y_i, x, i)\Big)\,, \tag{1}$$

where n denotes the model order, $\boldsymbol{w}$ the model parameter vector, and $\boldsymbol{\phi}$ the feature extraction function. We denote the tag set as $\mathcal{Y}$, that is, $y_i \in \mathcal{Y}$ for $i \in 1 \ldots |x|$.

2.2 Baseline Feature Set

We first describe our baseline feature set $\{\phi_j(y_{i-1}, y_i, x, i)\}_{j=1}^{|\boldsymbol{\phi}|}$ by defining *emission* and *transition* features. The emission feature set associates properties of the sentence position i with

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 259–264,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

the corresponding label as

$$\{\chi_j(x,i)\mathbb{1}(y_i = y'_i) \,|\, j \in 1 \ldots |\mathcal{X}| \,, \forall y'_i \in \mathcal{Y}\} \,, \tag{2}$$

where the function $\mathbb{1}(q)$ returns one if and only if the proposition q is true and zero otherwise, that is

$$\mathbb{1}(y_i = y'_i) = \begin{cases} 1 & \textbf{if} \quad y_i = y'_i \\ 0 & \textbf{otherwise} \end{cases} \,, \tag{3}$$

and $\mathcal{X} = \{\chi_j(x,i)\}_{j=1}^{|\mathcal{X}|}$ is the set of functions characterizing the word position i. Following the classic work of Ratnaparkhi (1996), our $\mathcal{X}$ comprises simple binary functions:

1. Bias (always active irrespective of input).

2. Word forms $x_{i-2}, \ldots, x_{i+2}$.

3. Prefixes and suffixes of the word form x_i up to length $\delta_{suf} = 4$.

4. If the word form x_i contains (one or more) capital letter, hyphen, dash, or digit.

Binary functions have a return value of either zero (inactive) or one (active). Meanwhile, the transition features

$$\begin{aligned}\{\mathbb{1}(y_{i-k} = y'_{i-k}) \ldots \mathbb{1}(y_i = y'_i) \,| \\ y'_{i-k}, \ldots, y'_i \in \mathcal{Y} \,, \forall k \in 1 \ldots n\}\end{aligned} \tag{4}$$

capture dependencies between adjacent labels irrespective of the input x.

2.2.1 Expanded Feature Set Leveraging Sub-Label Dependencies

The baseline feature set described above can yield a high tagging accuracy given a conveniently simple label set, exemplified by the tagging results of Collins (2002) on the Penn Treebank (Marcus et al., 1993). (Note that conditional random fields correspond to discriminatively trained hidden Markov models and Collins (2002) employs the latter terminology.) However, it does to some extent overlook some beneficial dependency information in case the labels have a rich sub-structure. In what follows, we describe expanded feature sets which explicitly model the sub-label dependencies.

We begin by defining a function $\mathcal{P}(y_i)$ which partitions any label y_i into its sub-label components and returns them in an unordered set. For example, we could define $\mathcal{P}(\text{PRON+1+SG}) = \{\text{PRON, 1, SG}\}$. (Label partitions employed in the experiments are described in Section 3.2.) We denote the set of all sub-label components as $\mathcal{S}$.

Subsequently, instead of defining only (2), we additionally associate the feature functions $\mathcal{X}$ with all sub-labels $s \in \mathcal{S}$ by defining

$$\{\chi_j(x,i)\mathbb{1}(s \in \mathcal{P}(y_i)) \,|\, \forall j \in 1 \ldots |\mathcal{X}| \,, \forall s \in \mathcal{S}\} \,, \tag{5}$$

where $\mathbb{1}(s \in \mathcal{P}(y_i))$ returns one in case s is in $\mathcal{P}(y_i)$ and zero otherwise. Second, we exploit *sub-label transitions* using features

$$\begin{aligned}\{\mathbb{1}(s_{i-k} \in \mathcal{P}(y_{i-k})) \ldots \mathbb{1}(s_i \in \mathcal{P}(y_i)) \,| \\ \forall s_{i-k}, \ldots, s_i \in \mathcal{S} \,, \forall k \in 1 \ldots m\} \,.\end{aligned} \tag{6}$$

Note that we define the sub-label transitions up to order m, $1 \leq m \leq n$, that is, an nth-order CRF model is not obliged to utilize sub-label transitions all the way up to order n. This is because employing high-order sub-label transitions may potentially cause overfitting to training data due to substantially increased number of features (equivalent to the number of model parameters, $|\boldsymbol{w}| = |\boldsymbol{\phi}|$). For example, in a second-order ($n = 2$) model, it might be beneficial to employ the sub-label emission feature set (5) and first-order sub-label transitions while discarding second-order sub-label transitions. (See the experimental results presented in Section 3.)

In the remainder of this paper, we use the following notations.

1. A standard CRF model incorporating (2) and (4) is denoted as CRF(n,-).

2. A CRF model incorporating (2), (4), and (5) is denoted as CRF(n,0).

3. A CRF model incorporating (2), (4), (5), and (6) is denoted as CRF(n,m).

2.3 On Linguistic Intuition

This section aims to provide some intuition on the types of linguistic phenomena that can be captured by the expanded feature set. To this end, we consider an example on the plural number in Finnish.

First, consider the plural nominative word form *kissat* (*cats*) where the plural number is denoted by the 1-suffix *-t*. Then, by employing the features (2), the suffix *-t* is associated solely with the compound label NOMINATIVE+PLURAL. However, by incorporating the expanded feature set (5), *-t* will also be associated to the sub-label PLURAL. This can be useful because, in Finnish, also adjectives and numerals are inflected according to number and denote the plural number with the suffix *-t* (Hakulinen et al., 2004, §79). Therefore, one can exploit *-t* to predict the plural number also in words such as *mustat* (*plural of black*) with a compound analysis ADJECTIVE+PLURAL.

Second, consider the number agreement (congruence). For example, in the sentence fragment *mustat kissat juoksevat (black cats are running)*, the words *mustat* and *kissat* share the plural number. In other words, the analyses of both *mustat* and *kissat* are required to contain the sub-label PLURAL. This short-span dependency between sub-labels will be captured by a first-order sub-label transition feature included in (6).

Lastly, we note that the feature expansion sets (5) and (6) will, naturally, capture any short-span dependencies within the sub-labels irrespective if the dependencies have a clear linguistic interpretation or not.

3 Experiments

3.1 Data

For a quick overview of the data sets, see Table 1.

Penn Treebank. The English Penn Treebank (Marcus et al., 1993) is divided into 25 sections of newswire text extracted from the Wall Street Journal. We split the data into training, development, and test sets using the sections 0-18, 19-21, and 22-24, according to the standardly applied division introduced by Collins (2002).

Turku Depedency Treebank. The Finnish Turku Depedendency Treebank (Haverinen et al., 2013) contains text from 10 different domains. The treebank does not have default partition to training and test sets. Therefore, from each 10 consecutive sentences, we assign the 9th and 10th to the development set and the test set, respectively. The remaining sentences are assigned to the training set.

Multext-East. The third data we consider is the multilingual Multext-East (Erjavec, 2010) corpus, from which we utilize the Czech, Estonian and Romanian sections. The corpus corresponds to translations of the novel *1984* by George Orwell. We apply the same data splits as for Turku Dependency Treebank.

lang.	train.	dev.	test	tags	train. tags
Eng	38,219	5,527	5,462	45	45
Rom	5,216	652	652	405	391
Est	5,183	648	647	413	408
Cze	5,402	675	675	955	908
Fin	5,043	630	630	2,355	2,141

Table 1: Overview on data. The training (train.), development (dev.) and test set sizes are given in sentences. The columns titled *tags* and *train. tags* correspond to total number of tags in the data set and number of tags in the training set, respectively.

3.2 Label Partitions

This section describes the employed compound label splits. The label splits for all data sets are submitted as data file attachments. All the splits are performed *a priori* to model learning, that is, we do not try to optimize them on the development sets.

The POS labels in the Penn Treebank are split in a way which captures relevant inflectional categories, such as tense and number. Consider, for example, the split for the present tense third singular verb label $\mathcal{P}(\text{VBZ}) = \{\text{VB}, \text{Z}\}$.

In the Turku Dependency Treebank, each morphological tag consists of sub-labels marking word-class, relevant inflectional categories, and their respective values. Each inflectional category, such as case or tense, combined with its value, such as nominative or present, constitutes one sub-label. Consider, for example, the split for the singular, adessive noun $\mathcal{P}(\text{N+CASE\_ADE+NUM\_SG}) = \{\text{POS\_N}, \text{CASE\_ADE}, \text{NUM\_SG}\}$.

The labeling scheme employed in the Multext-East data set represents a considerably different annotation approach compared to the Penn and Turku Treebanks. Each morphological analysis is a sequence of feature markers, for example Pw3–r. The first feature marker (P) denotes word class and the rest (w, 3, and r) encode values of inflectional categories relevant for that word class. A feature marker may correspond to several different values depending on word class and its position in the analysis. Therefore it becomes rather difficult to split the labels into similar pairs of inflectional category and value as we are able to do for the Turku Dependency Treebank. Since the interpretation of a feature marker depends on its position in the analysis and the word class, the markers have to be numbered and appended with the word class marker. For example, consider the split $\mathcal{P}(\text{Pw3–r}) = \{0 : \text{P}, 1 : \text{Pw}, 2 : \text{P3}, 5 : \text{Pr}\}$.

3.3 CRF Model Specification

We perform experiments using first-order and second-order CRFs with zeroth-order and first-order sub-label features. Using the notation introduced in Section 2, the employed models are CRF(1,-), CRF(1,1), CRF(2,-), CRF(2,0), and CRF(2,1). We do not report results using CRF(2,2) since, based on preliminary experiments, this model overfits on all languages.

The CRF model parameters are estimated using the averaged perceptron algorithm (Collins, 2002). The model parameters are initialized with a zero vector. We evaluate the latest averaged parameters on the held-out development set after each pass over the training data and terminate training if no improvement in accuracy is obtained during three last passes. The best-performing parameters are then applied on the test instances.

We accelerate the perceptron learning using beam search (Zhang and Clark, 2011). The beam width, b, is optimized separately for each language on the development sets by considering $b = 1, 2, 4, 8, 16, 32, 64, 128$ until the model accuracy does not improve by at least 0.01 (absolute).

Development and test instances are decoded using Viterbi search in combination with the tag dictionary approach of Ratnaparkhi (1996). In this approach, candidate tags for known word forms are limited to those observed in the training data. Meanwhile, word forms that were unseen during training consider the full label set.

3.4 Software and Hardware

The experiments are run on a standard desktop computer (Intel Xeon E5450 with 3.00 GHz and 64 GB of memory). The methods discussed in Section 2 are implemented in C++.

3.5 Results

The obtained tagging accuracies and training times are presented in Table 2. The times include running the averaged perceptron algorithm and evaluation of the development sets. The column labeled *it.* corresponds to the number of passes over the training data made by the perceptron algorithm before termination. We summarize the results as follows.

First, compared to standard feature extraction approach, employing the sub-label transition features resulted in improved accuracy on all languages apart from English. The differences were statistically significant on Czech, Estonian, and Finnish. (We establish statistical significance (with confidence level 0.95) using the standard 1-sided Wilcoxon signed-rank test performed on 10 randomly divided, non-overlapping subsets of the complete test sets.) This results supports the intuition that the sub-label features should be most useful in presence of large, fine-grained label sets, in which case the learning is most affected by data sparsity.

Second, on all languages apart from English, employing a first-order model with sub-label features yielded higher accuracy compared to a second-order model with standard features. The differences were again statistically significant on Czech, Estonian, and Finnish. This result suggests that, compared to increasing model order, exploiting the sub-label dependencies can be a preferable approach to improve the tagging accuracy.

Third, applying the expanded feature set inevitably causes some increase in the computational cost of model estimation. However, as shown by the running times, this increase is not prohibitive.

4 Related Work

In this section, we compare the approach presented in Section 2 to two prior systems which attempt to utilize sub-label dependencies in a similar manner.

Smith et al. (2005) use a CRF-based system for tagging Czech, in which they utilize expanded emission features similar to our (5). However, they do not utilize the full expanded transition features (6). More specifically, instead of utilizing a single chain as in our approach, Smith et al. employ five parallel structured chains. One of the chains models the sequence of word-class labels such as noun and adjective. The other four chains model gender, number, case, and lemma sequences, respectively. Therefore, in contrast to our approach, their system does not capture cross-dependencies between inflectional categories, such as the dependence between the word-class and case of adjacent words. Unsurprisingly, Smith et al. fail to achieve improvement over a generative HMM-based POS tagger of Hajič (2001). Meanwhile, our system outperforms the generative trigram tagger HunPos (Halácsy et al., 2007) which is an im-

model	it.	time (min)	acc.	OOV.
English				
CRF(1, -)	8	9	97.04	88.65
CRF(1, 0)	6	17	97.02	88.44
CRF(1, 1)	8	22	97.02	88.82
CRF(2, -)	9	15	97.18	88.82
CRF(2, 0)	11	36	97.17	89.23
CRF(2, 1)	8	27	97.15	89.04
Romanian				
CRF(1, -)	14	29	97.03	85.01
CRF(1, 0)	13	68	96.96	84.59
CRF(1, 1)	16	146	97.24	85.94
CRF(2, -)	7	19	97.08	85.21
CRF(2, 0)	18	99	97.02	85.42
CRF(2, 1)	12	118	97.29	86.25
Estonian				
CRF(1, -)	15	28	93.39	78.66
CRF(1, 0)	17	66	93.81	80.44
CRF(1, 1)	13	129	93.77	79.37
CRF(2, -)	15	30	93.48	77.13
CRF(2, 0)	13	53	93.78	79.60
CRF(2, 1)	16	105	94.01	79.53
Czech				
CRF(1, -)	6	28	89.28	70.90
CRF(1, 0)	10	112	89.94	74.44
CRF(1, 1)	10	365	90.78	76.83
CRF(2, -)	19	91	89.81	72.44
CRF(2, 0)	13	203	90.35	76.37
CRF(2, 1)	24	936	91.00	77.75
Finnish				
CRF(1, -)	10	80	87.37	59.29
CRF(1, 0)	13	249	88.58	63.46
CRF(1, 1)	12	474	88.41	62.63
CRF(2, -)	11	106	86.74	56.96
CRF(2, 0)	13	272	88.52	63.46
CRF(2, 1)	12	331	88.68	63.62

Table 2: Results.

proved open-source implementation of the well-known TnT tagger of Brants (2000). The obtained HunPos results are presented in Table 3.

	Eng	Rom	Est	Cze	Fin
HunPos	96.58	96.96	92.76	89.57	85.77

Table 3: Results using a generative HMM-based HunPos tagger of Halacsy et al. (2007).

Ceauşu (2006) uses a maximum entropy Markov model (MEMM) based system for tagging Romanian which utilizes transitional behavior between sub-labels similarly to our feature set (6). However, in addition to ignoring the most informative emission-type features (5), Ceauşu embeds the MEMMs into the tiered tagging framework of Tufis (1999). In tiered tagging, the full morphological analyses are mapped into a coarser tag set and a tagger is trained for this reduced tag set. Subsequent to decoding, the coarser tags are mapped into the original fine-grained morphological analyses. There are several problems associated with this tiered tagging approach. First, the success of the approach is highly dependent on a well designed coarse label set. Consequently, it requires intimate knowledge of the tag set and language. Meanwhile, our model can be set up with relatively little prior knowledge of the language or the tagging scheme (see Section 3.2). Moreover, a conversion to a coarser label set is necessarily lossy (at least for OOV words) and potentially results in reduced accuracy since recovering the original fine-grained tags from the coarse tags may induce errors. Indeed, the accuracy 96.56, reported by Ceauşu on the Romanian section of the Multext-East data set, is substantially lower than the accuracy 97.29 we obtain. These accuracies were obtained using identical sized training and test sets (although direct comparison is impossible because Ceauşu uses a non-documented random split).

5 Conclusions

We studied improving the accuracy of CRF-based POS tagging by exploiting sub-label dependency structure. The dependencies were included in the CRF model using a relatively straightforward feature expansion scheme. Experiments on five languages showed that the approach can yield significant improvement in tagging accuracy given sufficiently fine-grained label sets.

In future work, we aim to perform a more fine-grained error analysis to gain a better understanding where the improvement in accuracy takes place. One could also attempt to optimize the compound label splits to maximize prediction accuracy instead of applying a priori partitions.

Acknowledgements

This work was financially supported by Langnet (Finnish doctoral programme in language studies) and the Academy of Finland under the grant no 251170 (Finnish Centre of Excellence Program (2012-2017)). We would like to thank the anonymous reviewers for their useful comments.

References

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, pages 224–231.

A. Ceausu. 2006. Maximum entropy tiered tagging. In *The 11th ESSLI Student session*, pages 173–179.

Michael Collins. 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, volume 10, pages 1–8.

Tomaž Erjavec. 2010. Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial combination of rules and statistics: A case study in czech tagging. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 268–275.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja Riitta Heinonen, and Irja Alho. 2004. *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura, Helsinki, Finland.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2013. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation.*

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

M.P. Marcus, M.A. Marcinkiewicz, and B. Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing*, volume 1, pages 133–142. Philadelphia, PA.

Noah A. Smith, David A. Smith, and Roy W. Tromble. 2005. Context-based morphological disambiguation with random fields. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 475–482.

Dan Tufis. 1999. Tiered tagging and combined language models classifiers. In *Proceedings of the Second International Workshop on Text, Speech and Dialogue*, pages 28–33.

Yue Zhang and Stephen Clark. 2011. Syntactic processing using the generalized perceptron and beam search. *Computational Linguistics*, 37(1):105–151.