Dependency parsing with latent refinements of part-of-speech tags

Thomas Müller[†], Richard Farkas[§], Alex Judea[‡], Helmut Schmid[†], and Hinrich Schütze[†]

[†]Center for Information and Language Processing, University of Munich, Germany
[§]Department of Informatics, University of Szeged, Hungary
[‡]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany
muellets@cis.lmu.de

Abstract

In this paper we propose a method to increase dependency parser performance without using additional labeled or unlabeled data by refining the layer of predicted part-of-speech (POS) tags. We perform experiments on English and German and show significant improvements for both languages. The refinement is based on generative split-merge training for Hidden Markov models (HMMs).

1 Introduction

Probabilistic Context-free Grammars with latent annotations (PCFG-LA) have been shown (Petrov et al., 2006) to yield phrase structure parsers with state-of-the-art accuracy. While Hidden Markov Models with latent annotations (HMM-LA) (Huang et al., 2009), stay somewhat behind the performance of state-of-the-art discriminative taggers (Eidelman et al., 2010). In this paper we address the question of whether the resulting latent POS tags are linguistically meaningful and useful for upstream tasks such as syntactic parsing. We find that this is indeed the case, leading to a procedure that significantly increases the performance of dependency parsers. The procedure is attractive because the refinement of predicted part-of-speech sequences using a coarse-to-fine strategy (Petrov and Klein, 2007) is fast and efficient. More precisely, we show that incorporating the induced POS into a state-of-the-art dependency parser (Bohnet, 2010) gives increases in Labeled Attachment Score (LAS): from 90.34 to 90.57 for English and from 87.92 to 88.24 (resp. 88.35 to 88.51) for German without using (resp. with using) morphological features.

2 Related Work

Petrov et al. (2006) introduce generative split-merge training for PCFGs and provide a fully automatic method for training state-of-the-art phrase structure parsers. They argue that the resulting latent annotations are linguistically meaningful. Sun et al. (2008) induce latent sub-states into CRFs and show that noun phrase (NP) recognition can be improved, especially if no part-of-speech features are available. Huang et al. (2009) apply split-merge training to create HMMs with latent annotations (HMM-LA) for Chinese POS tagging. They report that the method outperforms standard generative bigram and trigram tagging, but do not compare to discriminative methods. Eidelman et al. (2010) show that a bidirectional variant of latent HMMs with incorporation of prosodic information can yield state-of-the-art results in POS tagging of conversational speech.

3 Split-Merge Training for HMMs

Split-merge training for HMMs (Huang et al., 2009) iteratively splits every tag into two subtags. Word emission and tag transition probabilities of subtags are then initialized close to the values of the parent tags but with some randomness to break symmetry. Using expectation–maximization (EM) training the parameters can then be set to a local maximum of the training data likelihood. After this split phase, the merge phase reverts splits that only lead to small improvements in the likelihood function in order to increase the robustness of the model. This approach requires an approximation of the gain in likelihood of every split analogous to Petrov et al. (2006) as an exact computation is not feasible.

We have observed that this procedure is not

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 963–967,
October 25-29, 2014, Doha, Qatar.

	Universal Tag	Feature	Tag_0	Tag_1
English	Adjectives	$p(w\|t)$	more (0.05) many (0.03) last (0.03)	new (0.03) other (0.03) first (0.02)
	(ADJ)	$p(u\|t)$	**VERB** (0.32) ADV (0.27) NOUN (0.14)	**DET** (0.39) ADP (0.17) ADJ (0.10)
	Particles	$p(w\|t)$	's (0.93) ' (0.07)	to (0.89) up (0.04) out (0.02) off (0.01)
	(PRT)	$p(b\|t)$	**POS** (1.00)	**TO** (0.89) **RP** (0.10)
	Prepositions	$p(w\|t)$	that (0.11) in (0.10) by (0.09)	of (0.43) in (0.19) for (0.11)
	(ADP)	$p(u\|t)$	**VERB** (0.46) NOUN (0.15) . (0.13)	**NOUN** (0.84) NUM (0.06) ADJ (0.03)
	Pronouns	$p(w\|t)$	its (0.30) their (0.15) his (0.14)	it (0.21) he (0.16) they (0.12)
	(PRON)	$p(b\|t)$	**PRP$** (0.68) PRP (0.26) WP (0.05)	**PRP** (0.87) WP (0.11) PRP$ (0.02)
	Verbs	$p(w\|t)$	be (0.06) been (0.02) have (0.02)	is (0.10) said (0.08) was (0.05)
	(VERB)	$p(u\|t)$	**VERB** (0.38) PRT (0.22) ADV (0.11)	**NOUN** (0.52) **PRON** (0.20) . (0.12)
German	Conjunctions	$p(w\|t)$	daß (0.26) wenn (0.08) um (0.06)	und (0.76) oder (0.07) als (0.06)
	(CONJ)	$p(b\|t)$	**KOUS** (0.58) KON (0.30) KOUI (0.06)	**KON** (0.88) KOKOM (0.10) APPR (0.02)
	Particles	$p(w\|t)$	an (0.13) aus (0.10) ab (0.09)	**nicht** (0.49) **zu** (0.46) Nicht (0.01)
	(PRT)	$p(b\|t)$	PTKVZ (0.92) ADV (0.04) ADJD (0.01)	**PTKNEG** (0.52) **PTKZU** (0.44) PTKA (0.02)
	Pronouns	$p(w\|t)$	sich (0.13) die (0.08) es (0.07)	ihre (0.06) seine (0.05) seiner (0.05)
	(PRON)	$p(b\|t)$	**PPER** (0.33) PRF (0.14) PRELS (0.14)	**PPOSAT** (0.40) PIAT (0.34) PDAT (0.16)
	Verbs	$p(w\|t)$	werden (0.04) worden (0.02) ist (0.02)	ist (0.07) hat (0.04) sind (0.03)
	(VERB)	$p(u\|t)$	NOUN (0.46) VERB (0.22) PRT (0.10)	NOUN (0.49) **.** (0.19) PRON (0.16)

Table 1: Induced sub-tags and their statistics, word forms ($p(w|t)$), treebank tag ($p(b|t)$) and preceding Universal tag probability ($p(u|t)$). Bold: linguistically interesting differences.

only a way to increase HMM tagger performance but also yields annotations that are to a considerable extent linguistically interpretable. As an example we discuss some splits that occurred after a particular split-merge step for English and German. For the sake of comparability we applied the split to the Universal Tagset (Petrov et al., 2011). Table 1 shows the statistics used for this analysis. The Universal POS tag set puts the three Penn-Treebank tags RP (particle), POS (possessive marker) and TO into one particle tag (see "PRT" in English part of the table). The training essentially reverses this by splitting particles first into possessive and non-possessive markers and in a subsequent split the non-possessives into TO and particles. For German we have a similar split into verb particles, negation particles like *nicht* 'not' and the infinitive marker *zu* 'to' ("PRT") in the German part of the table). English prepositions get split by proximity to verbs or nouns ("ADP"). Subordinate conjunctions like *that*, which in the Penn-Treebank annotation are part of the preposition tag IN, get assigned to the sub-class next to verbs. For German we also see a separation of "CONJ" into predominantly subordinate conjunctions (Tag 0) and predominantly coordinating conjunctions (Tag 1). For both languages adjectives get split by predicative and attributive use. For English the predicative sub-class also seems to hold rather atypical adjectives like "such" and "last." For English, verbs ("VERB") get split into a predominantly infinite tag (Tag 0) and a predominantly finite tag (Tag 1) while for German we get a separation by verb position. In German we get a separation of pronouns ("PRON") into possessive and non-possessive; in English, pronouns get split by predominant usage in subject position (Tag 0) and as possessives (Tag 1).

Our implementation of HMM-LA has been released under an open-source licence.[1]

In the next section we evaluate the utility of these annotations for dependency parsing.

4 Dependency Parsing

In this section we investigate the utility of induced POS as features for dependency parsing. We run our experiments on the CoNLL-2009 data sets (Hajič et al., 2009) for English and German. As a baseline system we use the latest version of the mate-tools parser (Bohnet, 2010).[3] It was the highest scoring syntactic parser for German and English in the CoNLL 2009 shared task evaluation. The parser gets automatically annotated lemmas, POS and morphological features as input which are part of the CoNLL-2009 data sets.

In this experiment we want to examine the benefits of tag refinements isolated from the improvements caused by using two taggers in parallel, thus we train the HMM-LA on the automatically tagged POS sequences of the training set and use it to add an additional layer of refined POS to the input data of the parser. We do this by calculating the forward-backward charts that are also used in the E-steps during training — in these charts base

[1] https://code.google.com/p/cistern/

[1] Unlabeled Attachment Score

[3] We use v3.3 of Bohnet's graph-based parser.

	#Tags	μ_{LAS}	$\max_{\text{LAS}}$	σ_{LAS}	μ_{UAS}	$\max_{\text{UAS}}$	σ_{UAS}
English	Baseline	88.43			91.46		
	58	88.52	(88.59)	0.06	91.52	(91.61)	0.08
	73	88.55	(88.61)	0.05	91.54	(91.59)	0.04
	92	88.60	(88.71)	0.08	91.60	(91.72)	0.08
	115	88.62	(88.73)	0.07	91.58	(91.71)	0.08
	144	88.60	(88.70)	0.07	91.60	(91.71)	0.07
German (no feat.)	Baseline	87.06			89.54		
	85	87.09	(87.18)	0.06	89.61	(89.67)	0.04
	107	87.23	(87.36)	0.09	89.74	(89.83)	0.08
	134	87.22	(87.31)	0.09	89.75	(89.86)	0.09
German (feat.)	Baseline	87.35			89.75		
	85	87.33	(87.47)	0.11	89.76	(89.88)	0.09
	107	87.43	(87.73)	0.16	89.81	(90.14)	0.17
	134	87.38	(87.53)	0.08	89.75	(89.89)	0.08

Table 2: LAS and UAS[1] mean (μ), best value ($\max$) and std. deviation (σ) for the development set for English and German dependency parsing with (feat.) and without morphological features (no feat.).

tags of the refined tags are constrained to be identical to the automatically predicted tags.

We use 100 EM iterations after each split and merge phase. The percentage of splits reverted in each merge phase is set to .75.

We integrate the tags by adding one additional feature for every edge: the conjunction of latent tags of the two words connected by the edge.

Table 2 shows results of our experiments. All numbers are averages of five independent runs. For English the smaller models with 58 and 73 tags achieve improvements of ≈.1. The improvements for the larger tag sets are ≈.2. The best individual model improves LAS by .3. For the German experiments without morphological features we get only marginal average improvements for the smallest tag set and improvements of ≈.15 for the bigger tag sets. The average ULA scores for 107 and 134 tags are at the same level as the ULA scores of the baseline with morph. features. The best model improves LAS by .3. For German with morphological features the absolute differences are smaller: The smallest tag set does not improve the parser on average. For the tag set of 107 tags the average improvement is .08. The best model improves LAS by .38. In all experiments we see the highest improvements for tag set sizes of roughly the same size (115 for English, 107 for German). While average improvements are low (esp. for German with morphological features), peak improvements are substantial.

Running the best English system on the test set gives an improvement in LAS from 90.34 to 90.57; this improvement is significant[4] ($p < .02$). For German we get an improvement from 87.92 to 88.24 without and from 88.35 to 88.51 with morphological features. The difference between the values without morphological features is significant ($p < .05$), but the difference between models with morphological features is not ($p = .26$). However, the difference between the baseline system with morphological features and the best system without morphological features is also not significant ($p = .49$).

We can conclude that HMM-LA tags can significantly improve parsing results. For German we see that HMM-LA tags can substitute morphological features up to an insignificant difference. We also see that morphological features and HMM-LA seem to be correlated as combining the two gives only insignificant improvements.

5 Contribution Analysis

In this section we try to find statistical evidence for why a parser using a fine-grained tag set might outperform a parser based on treebank tags only.

The results indicate that an induced latent tag set as a whole increases parsing performance. However, not every split made by the HMM-LA seems to be useful for the parser. The scatter plots in Figure 1 show that there is no strict correlation between tagging accuracy of a model and the resulting LAS. This is expected as the latent induction optimizes a tagging objective function, which does not directly translate into better parsing performance. An example is lexicalization. Most latent models for English create a subtag for the preposition "of". This is useful for a HMM as "of" is frequent and has a very specific context. A lexicalized syntactic parser, however, does not benefit from such a tag.

[4] Approx. randomization test (Yeh, 2000) on LAS scores

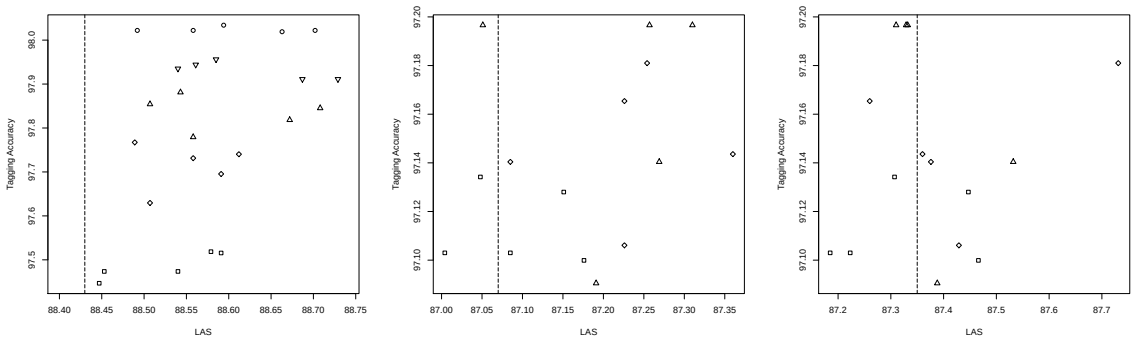

Figure 1: Scatter plots of LAS vs tagging accuracy for English (left) and German without (middle) and with (right) morphological features. English tag set sizes are 58 (squares), 73 (diamonds), 92 (triangles), 115 (triangles pointing downwards) and 144 (circles). German tag set sizes are 85 (squares), 107 (diamonds) and 134 (triangles). The dashed lines indicate the baselines.

We base the remainder of our analysis on the results of the baseline parser on the English development set and the results of the best performing latent model. The best performing model has a LAS score of 88.73 vs 88.43 for the baseline, a difference of .3. If we just look at the LAS of words with incorrectly predicted POS we see a difference of 1.49. A look at the data shows that the latent model helps the parser to identify words that might have been annotated incorrectly. As an example consider plural nouns (NNS) and two of their latent subtags NNS_1 and NNS_2 and how often they get classified correctly and misclassified as proper nouns (NNPS):

	NNS	NNPS
NNS	**2019**	**104**
NNS_1	90	72
NNS_2	**1100**	13
...	...	...

We see that NNS_1 is roughly equally likely to be a NNPS or NNS while NNS_2 gives much more confidence of the actual POS being NNS. So one benefit of HMM-LA POS tag sets are tags of different levels of confidence.

Another positive effect is that latent POS tags have a higher correlation with certain dependency relations. Consider proper nouns (NNP):

	NAME	NMOD	SBJ
NNP	**962**	**662**	**468**
NNP_1	10	27	**206**
NNP_2	24	50	**137**
...	...	...	...

We see that NNP_1 and NNP_2 are more likely to appear in subject relations. NNP_1 contains surnames; the most frequent word forms are *Keating*, *Papandreou* and *Kaye*. In contrast, NNP_2 contains company names such as *Sony*, *NBC* and *Keystone*. This explains why the difference in LAS is twice as high for NNPs as on average.

For German we see similar effects and the anticipated correlation with morphology. The 5 determiner subtags, for example, strongly correlate with grammatical case:

	Nom	Gen	Dat	Acc
ART	**1185**	**636**	**756**	**961**
ART_1	**367**		7	38
ART_2	11	28	**682**	21
ART_3	6	**602**	7	3
ART_4	39		43	**429**
ART_5	**762**	6	17	**470**

6 Conclusion and Future Work

We have shown that HMMs with latent annotations (HMMLA) can generate latent part-of-speech tagsets are linguistically interpretable and can be used to improve dependency parsing. Our best systems improve an English parser from a LAS of 90.34 to 90.57 and a German parser from 87.92 to 88.24 when not using morphological features and from 88.35 to 88.51 when using morphological features . Our analysis of the parsing results shows that the major reasons for the improvements are: the separation of POS tags into more and less trustworthy subtags, the creation of POS subtags with higher correlation to certain dependency labels and for German a correlation of tags and morphological features such as case.

7 Future Work

The procedure works well in general. However, not every split is useful for the parser; e.g., as

discussed above lexicalization increases HMM accuracy, but does not help an already lexicalized parser. We would like to use additional information (e.g., from the dependency trees) to identify useless splits. The different granularities of the hierarchy induced by split-merge training are potentially useful. However, the levels of the hierarchy are incomparable: a child tag is in general not a subtag of a parent tag. We think that coupling parents and children in the tag hierarchy might be one way to force a consistent hierarchy.

Acknowledgments

We would like to thank the anonymous reviewers for their comments. The first author is a recipient of the Google Europe Fellowship in Natural Language Processing, and this research is supported in part by this Google Fellowship and by DFG (grant SFB 732). Most of this work was conducted while the authors worked at the Institute for Natural Language Processing of the University of Stuttgart.

References

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING*.

Vladimir Eidelman, Zhongqiang Huang, and Mary Harper. 2010. Lessons learned in part-of-speech tagging of conversational speech. In *Proceedings of EMNLP*.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL*.

Zhongqiang Huang, Vladimir Eidelman, and Mary Harper. 2009. Improving a simple bigram hmm part-of-speech tagger by latent annotation and self-training. In *Proceedings of NAACL*.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of ACL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *ArXiv:1104.2086v1*.

Xu Sun, Louis-Philippe Morency, Daisuke Okanohara, and Jun'ichi Tsujii. 2008. Modeling latent-dynamic in shallow parsing: a latent conditional model with improved inference. In *Proceedings of COLING*.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING*.