Part-of-speech tagging by means of ILP and active learning

Miloslav Nepil and Luboš Popelínský

Faculty of Informatics, Masaryk University
Botanická 68a, CZ-602 00 Brno, Czech Republic
Email: {nepil,popel}@fi.muni.cz

1 Introduction

Morphological tagging in inflectional languages like Czech is quite a difficult problem that can be hardly solved without employing (semi-)automatic taggers. Several morphological taggers have been developed (or proposed) for Czech, based on either statistical methods [5, 8] or machine learning techniques [9, 10]. Here we focus on part-of-speech (POS) tagging. For instance, in the two sentences in Fig. 1, the word form *own* can be either adjective or verb. Our goal

(oni) <vlastní> auto. *(They <own> a car.)*
<Zničili jejich <vlastní> auto. *(They destroyed their <own> car.)*

Fig. 1. Example of ambiguous words

is to find the correct POS for a given word form if we know the words in its context.

To build a tagger one usually needs a representative set of texts. The principal difficulty for Czech lies in the fact that annotated corpora (i.e. unambiguously tagged ones) are too small. Compare 164 000 stems of Czech words that `ajka` [12], a morphological analyser of Czech, is able to recognise (each of them can have a number of both prefixes and suffixes) with 132 000 different word forms in DESAM corpus [8]. On the other hand it is easy to collect unannotated data. E.g. Czech National Corpus [2] contains over 140 million words. If we knew a criterion how to choose the most informative examples from these data sources, annotation cost would be low.

This situation is quite common for different NLP tasks. It has been already shown – in text categorisation [6], information extraction [11], semantic parsing [11] or part-of-speech tagging [4] – that *active learning* [3] can help. Active learning is defined as any form of learning in which the learning program has some control over the choice of examples on which it is trained. There are two types of active learning. In the first case the learner constructs examples, usually membership queries [1], and asks an oracle to label them. The second type, *selective sampling* [4] is employed here. It consists of two steps. First, the learner examines many examples and selects only the most informative ones for labeling.

In the second step, an oracle is asked to label (membership to one of predefined classes) of the selected examples.

We have investigated a variant of committee-based sampling [4]. An unlabeled example is presented to $N \geq 2$ classifiers, members of the committee. If the members disagree in label prediction, the example is assumed to be informative and the actual label is requested from the teacher. In our task – part-of-speech tagging of Czech – we use a committee of relational learners and combine manually-crafted rules with those induced automatically by inductive logic programming (ILP) [7] techniques.

2 Method

The method combines two different techniques, manually-crafted rules and ILP. First, the learning set is split into N parts of approximately same size.

- As some of ambiguities can be easily removed, it is better to write manually simple rules that have high accuracy (preferably 100%) for each word category in the $Sample_1$.
- Ambiguously tagged words that remain in $Sample_i$ after application of the previous tagger(s) are supposed to be informative. New classifier is being built by means of ILP. The newly induced rules are added to the old ones (that have been built in $Step_1 \ldots Step_{i-1}$) and this classifier is now used for disambiguation of $Sample_i$.

These two steps are repeated until accuracy and/or recall do not change significantly or until all the learning data have been exploited.

3 Experiments

3.1 Data source

We used a subset of PTB corpus [5] that contained 33612 items (word positions) ambiguously annotated with `ajka` morphological analyser [12]. It means that each word was labeled with all possible tags for given word. We used a full tag set for Czech that contained about 1600 different tags. About 52% of words had more than one tag, and 14.9% of words contained at least two part-of-speech tags (different word category). The data set was split into 4 samples of approximately equal size. Three of them were used for learning, the fourth one for testing.

3.2 Algorithm

We tested our method on two of the most frequent part-of-speech ambiguities in Czech, substantive-adjective ambiguity (15.8% of all ambiguities) and substantive-verb ambiguities (6.5%). For each of these ambiguities two rule sets were learned (built). Each set of rules has been learned (or built) to remove just one tag, substantive, adjective or verb. An example of a rule is in Fig. 2. Given words and their possible tags in the left and the right context, if the set of conditions holds then `Tag` will be removed from the set of possible tags for `Word`. Conditions have a form `cn(C[,SubsetOfC],F)` where `C` is either a context or a particular word (including possible tags), and `F` is a first-order condition on `C`. For instance, `F` can test presence of words, lemmata and/or tags in a context. In

```
remove(L,W,R,k1) :- cn(W, e(k1) & e(k2&g:G&n:N&c:C)),
                    cn(R, first(1), [e(k1&g:G&n:N&c:C)]).
```

Fig. 2. Example of a disambiguation rule

the example in Fig. 2, tag `k1` will be removed from the set of possible tags for the word `W` in contexts `L,R` if `W` can have a tag `k1` or `k2` and `W` can have the same gender `G`, the number `N` and the case `C` as the neighbouring substantive (if any) on the right.

First we wrote by hand simple rules for each word category. We included only such rules that displayed 100% accuracy on the first of four data samples. Using these manually-crafted rules about a half of all ambiguities in part-of-speech were removed. The rest of data has been labeled and used for learning new rules with ILP system Aleph[1] (default settings).

In the following two paragraphs we bring first results. Then we discuss two improvements of our method.

3.3 Substantive-adjective ambiguity

Table 1. Results for substantive-adjective ambiguity

Sample	#ambiguities before	after	RECALL	#err.	ACCURACY	Set of rules
0.	182	82	54.9%	0	100.0%	pl1
1.	216	33	84.7%	10	95.4%	pl2
2.	257	34	86.8%	13	94.4%	pl3
3.	174	8	95.4%	17	90.2%	pl4

Results for this ambiguity are displayed in Tab. 1. The second and the third columns contain a number of ambiguities before and after application of a set of the rules. The next column (RECALL) informs about a relative number of cases for which just one of two tags was removed. The last but one column (ACCURACY) displays the relative number of correctly removed tags. In the first line there are values reached with manually-coded rules (pl1). From 182 ambiguous words 100 cases were correctly disambiguated, i.e. recall has reached 54.9%. The remaining 82 cases were labeled and used for learning two sets of rules with

[1] http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.html

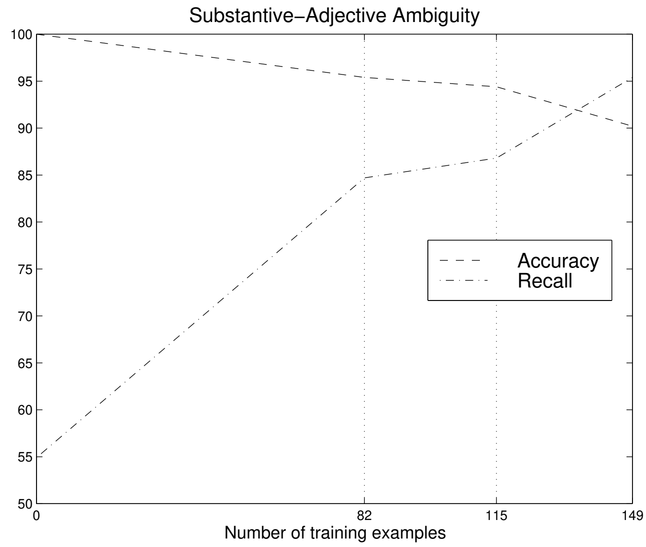

Fig. 3. Learning curves for substantive-adjective ambiguity

Aleph, one for removing tag `k1` (substantive), one for removing the other tag `k2` (adjective). New set of rules, named pl2, contained the set pl1 (5 manually-coded rules) plus 11 newly learned rules. Set of rules pl2 was used for disambiguation of Sample 2, etc. At the end, set pl4 contained 23 rules. It can be observed that after 3 learning cycles with Aleph recall overcame 95% and the accuracy did not decrease below 90%. Fig. 3 presents active learning curves for substantive-adjective ambiguity.

Comparison with passive learning is in Tab. 2. The passive learning experiment was carried out in the following way. First, manually-crafted rules were used to remove ambiguities in all samples 0,1,2,3. The remaining ambiguous words were labeled and all cases from samples 0,1,2 were used for learning rules with Aleph. The learned rules were tested on Sample 3. Results are in the first line of Tab. 2.

Table 2. Substantive-adjective ambiguity: passive and active learning

	#examples to label	#rules learned	RECALL	ACCURACY
passive	174	26	66.2%	89.8%
active	149	18	95.4%	90.2%

Comparing with active learning, the active learning needed less examples to label: 149 vs. 174 examples for passive learning. Also the number of induced rules is much smaller (18 vs. 26 rules). The most important fact is that recall is almost 30% higher than for passive learning without decrease of accuracy.

3.4 Substantive-verb ambiguity

Even better results were obtained for substantive-verb ambiguity (see Tab. 3). Fig. 4 presents active learning curves for this kind of ambiguity. Here recall reached 97.4% and accuracy 94.8% with only 69 training examples and much shorter time needed, as compared to passive learning (see Tab. 4).

Table 3. Results for substantive-verb ambiguity

Sample	#ambiguities before	after	RECALL	#err.	ACCURACY	learning time(sec)
0.	83	41	50.6%	0	100.0%	237.1
1.	70	11	84.3%	5	92.9%	37.5
2.	111	17	84.7%	5	95.5%	52.4
3.	77	2	97.4%	4	94.8%	-

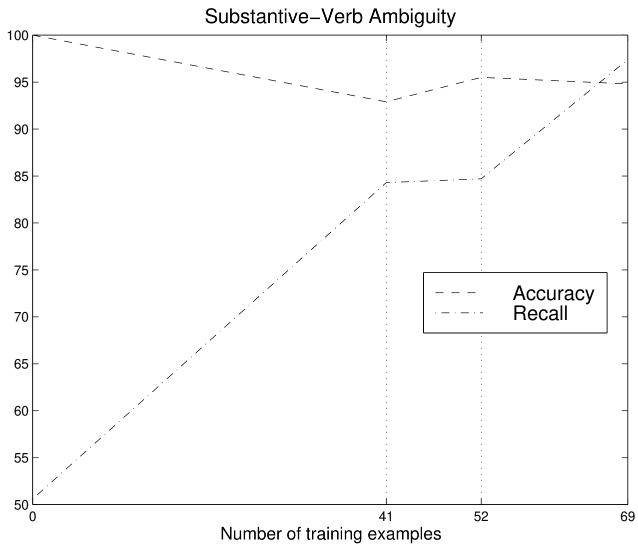

Fig. 4. Learning curves for substantive-verb ambiguity

Table 4. Substantive-verb ambiguity: passive and active learning

	#examples to label	#rules learned	RECALL	ACCURACY	training time(sec)
passive	184	21	97.8%	95.5%	2105.5
active	69	17	97.4%	94.8%	327.0

3.5 Improvements

Accuracy in the experiment of Section 3.3 was lower than expected. To increase it, we removed (in each learning cycle) all rules that covered more than 5% negative examples. Results for substantive-adjective ambiguity are in the right part of Tab. 5. With this simple refinement the accuracy increased to 92.5%. It can be proved that using this modification accuracy is converging but recall

Table 5. Original vs. refined theories

	orig.				refined			
Sample	ambi	RECALL	#err.	ACCURACY	ambi	RECALL	#err.	ACCURACY
1.	33	84.7%	10	95.4%	45	79.2%	6	97.2%
2.	34	86.8%	13	94.4%	32	87.5%	15	94.2%
3.	8	95.4%	17	90.2%	30	82.8%	13	92.5%

not (contrary to the original method). Even though accuracy remains quite high recall is now lower. More sophisticated method of theory refinement needs to be employed, i.e. we should replace an overgeneral rule with its specialisation.

We also tried to decrease the number of examples which needed to be labeled. For each example, a set of subgoals that held for the example was induced using saturation step in Aleph. Then, a distance $d(i,j)$ between the examples i and j was computed as a fraction

$$d(i,j) = \frac{|Sat_i \# Sat_j|}{|Sat_i \cup Sat_j|}$$

where Sat_i, Sat_j are the sets of subgoals for the examples i, j, and the sign # stands for symmetric difference. After that, we removed every new example i for which there had been already a labeled example j with the distance $d(i,j) < 0.3$. Decrease of the number of examples for labeling was about 20% and accuracy decreased only slightly.

4 Conclusion

We described a new method of active learning in ILP. We have shown that for two subtasks of part-of-speech tagging in Czech this method beats passive learning both in learning time and in compactness of the resulting tagger (number of rules) without decrease of recall and accuracy. Our approach is general enough to be applied on morphological disambiguation of other languages than Czech. Considering the fact that Czech is highly inflectional and highly ambiguous, we suppose that our method would perform no worse on other languages.

Nevertheless, our task has not been completely solved yet. Since the accuracy achieved so far is not satisfactory, we also outlined two improvements of our method that can result in increase of accuracy and in decrease of the number of training examples.

Although this contribution was concerned with part-of-speech tagging, in the near future we will tackle other morphological categories (gender, number, case, etc.) as well. Future work will also focus on finding optimal ratio between cardinality of learning samples and the number of learning cycles. In addition, we may consider to explore some other active learning approaches different from committee-based sampling.

Acknowledgement

We thank Karel Pala and Eva Žáčková for fruitful discussions.

References

1. Angluin D.: Queries and Concept Learning. *Machine Learning* 2, 4, April 1988, 319-342
2. Čermák F.: Czech National Corpus: A Case in Many Contexts. *Int. J. of Corpus Linguistics*, 2,2, 181-197, 1997.
3. Cohn D., Atlas L., Ladner R.: Improving Generalization with Active Learning. *Machine Learning*, 15, 201–221, 1994.
4. Dagan I., Engelson S.: Selective sampling in natural language learning. In *IJCAI-95 Workshop On New Approaches to Learning for Natural Language Processing*, 1995.
5. Hajič J., Hladká B.: Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of EACL 1998.*
6. Liere R., Tadepalli P.: Active learning with commitees for text categorization. In *Proceedings of the 14th National Conference on Artificial Intelligence, Providence*, pages 591–596, 1997.
7. Muggleton S., De Raedt L.: Inductive Logic Programming: Theory And Methods. J. Logic Programming 1994:19,20:629-679.
8. Pala K., Rychlý P., and Smrž P.: DESAM – Annotated Corpus for Czech. In *Plášil F., Jeffery K.G.(eds.): Proceedings of SOFSEM'97, Milovy, Czech Republic. LNCS 1338*, pages 60–69. Springer, 1997.
9. Pavelek T. and Popelínský L.: Mining Lemma Disambiguation Rules from Czech Corpora. In *Principles of Knowledge Discovery in Databases: Proceedings of PKDD'99 Conference, LNAI 1704*, pages 498–503. Springer, 1999.
10. Pavelek T., Popelínský L. and Ptáčník T.: On Disambiguation in Czech Corpora. FIMU-RS-2000-07, Faculty of Informatics MU Brno 2000. `http://www.fi.muni.cz/informatics/reports/rep2000/brief.html`
11. Thompson C.A., Califf M.E., Mooney R.J.: Active Learning for Natural Language Parsing and Information Extraction. In *Proceedings of 16th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pages 406–414, 1999.
12. Sedláček R. and Smrž P.: Automatic Processing of Czech Inflectional and Derivative Morphology. In *Proceedings of the Fourth International Conference TSD 2001*, LNAI 1902, Pilsen, Czech Republic, September 2001, Springer-Verlag.