Part-of-Speech Tagging by Means of Shallow Parsing, ILP and Active Learning

Miloslav Nepil, Luboš Popelínský, Eva Žáčková
Masaryk University Brno, Czechia
{nepil,popel}@fi.muni.cz

Oni <vlastní> auto.	*They <own> a car.*
Zničili jejich <vlastní> auto.	*They destroyed their <own> car.*

- lack of annotated data: 164 000 stems vs. about 200 000 different words in corpora
- unannotated Czech National Corpus: 140 000 000 words
- **active learning:** some control over the choice of examples

Structure of disambiguation rules

```
remove(Left, Word, Right, Tag) :- <set of conditions>
```

- `Word` contains a word form to be disambiguated, together with all its tags which remain possible
- `Tag` determines a corresponding tag which should be removed
- `Left` and `Right` represent ambiguously tagged left and right contexts

Figure 1: Example of a rule

```
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)]).
```

Method

DIS shallow parser, hand-coded rules, ILP & active learning

1. Employ the DIS shallow parser.
   For the remaning ambiguities apply the following algorithm.
2. Put all the manually-written rules to the rule set.
3. $I = 0$.
4. Apply the rule set to the $Sample_I$.
5. Label the remaining examples of $Sample_I$
   Use these examples for learning new rules.
   Append the new rules to the rule set.
6. $I++$.
7. if $I < 4$ goto 3

Refinement:

- if a rule cover more than 5% of negative examples on the next sample -> remove it

Problems to be solved:

- substantive - adjective ambiguity
- pronoun - verb ambiguity

Data source:

- Prague Dependency Treebank
  41647 items (word positions)

ambiguously annotated with `ajka` morphological analyser
each word was labeled with all possible tags for given word
used a full tag set for Czech that contained about 1600 different tags.

52% of words had more than one tag
14.9% of words contained at least two part-of-speech tags (different word category)

Table 1: Results for substantive-adjective ambiguity

Sample	#ambiguities before	DIS	rules	RECALL	#err.	ACCURACY	# newly learned rules	Set of rules
0.	182	65	63	65.4%	0	100.0%	6	pl1
1.	216	63	17	80.4%	2	99.0%	6	pl2
2.	257	92	47	81.7%	1	99.5%	3	pl3
3.	174	40	4	97.7%	1	99.4%	2	pl4
4.	160	52	0	100.0%	2	98.8%	-	-

Table 2: Results for pronoun-verb ambiguity

Sample	#ambiguities before	DIS	rules	RECALL	#err.	ACCURACY	# newly learned rules	Set of rules
0.	93	83	36	61.3%	0	100.0%	8	pl1
1.	102	86	20	80.4%	1	98.8%	4	pl2
2.	91	74	8	91.2%	0	100.0%	2	pl3
3.	83	64	7	91.6%	3	96.1%	2	pl4
4.	91	76	2	97.8%	3	96.6%	-	-

Passive and active learning

Table 3: Substantive-adjective ambiguity

	#examples to label	#rules learned	RECALL	ACCURACY
passive	250	21	95.0%	100.0%
active	131	17	100.0%	98.8%

Table 4: Pronoun-verb ambiguity

	#examples to label	#rules learned	RECALL	ACCURACY
passive	307	12	94.5%	97.7%
active	71	16	97.8%	96.6%

Active learning and ILP

- smaller number of training examples – 52%, 23%
- decrease of the training time – 1/6

without significant decrease of recall or accuracy