Syntactic analysis of a sentence

Targeting real-world and Slovak data

Jaroslav Loebl, supervisor: Marián Šimko

Parsing real-world data

Evaluation of syntactic parsers is usually executed on very well crafted test data, which consists of heavily edited texts (newspaper and books) and containing ale needed prerequisites (such as morphological annotation. Very little interest has been put into parsing text generated by ordinary users of the web – text, which often contains errors that negatively impacts parsing accuracy (Petrov, 2012).

Artificially corrupting data

We conducted number of operations to simulate real-world data:

- Removal of diacritics (D)
- Typographical errors (TY)
- Grammar mistakes (G, phonology assimilation, interchange of *i* and *y*, removal or addition of accent)
- Custom morphology annotation by TreeTagger (MA)

	D	TY	G	MA	LAS	UAS
Baseline	0	0	0	0	65,28%	82,09%
Setup 1	0	0	0	1	61,14%	82,15%
Setup 2	1	0	0	1	49,81%	69,05%
Setup 3	2	0	0	1	62,64%	82,21%
Setup 4	0	1	0	1	53,77%	73,07%
Setup 5	0	2	0	1	56,92%	76,73%
Setup 6	0	0	1	1	54,33%	73,96%
Setup 7	0	0	2	1	56,97%	76,86%

0 – no action taken

1 – deformation was applied (D, TY, G) or custom morphology tagging (MA)

2 – deformation was applied and tool to repair input was used (reconstruction for D and Aspell correction for TY and G)

Improving parsing accuracy

We looked for possible ways of improving the accuracy of certain *deprel* tags (namely predicated and nominal part of predicate respectively). One promising method is extending the morphological features by flags indicating chance of particular *deprel* tag. In order to know where put flags we used association rules mining for extracting the rules by which flags are assigned. Below is shown an example rule which is used to enrich morphological features:

```
IF 'Vform=L' AND 'per=c' IN features THEN add 'pred_candidate'
```

This assigns *pred_candidate* to third person verbs in participle verb form. We built similar rules for nominal predicate as well. Next we conducted two experiments, one for predicates and one for nominal predicates.

Predicate experiment	*Pred* accuracy	LAS
Baseline	37,89%	51,60%
Pred_candidate	65,69%	53,80%

Nominal predicate accuracy	*Pnom* accuracy	LAS
Baseline	39,77%	51,60%
Pred_candidate	47,61%	51,70%

Contributions

- First comprehensive evaluation of existing top-performing parser on Slovak language
- Evaluation of various real-world text imperfections on parsing accuracy -> significant impact on parsing accuracy\
- Improvement of specific, more important *deprel* tags computation
- Web service realization