Validating the Quality of Full Morphological Annotation

Tagging of Training Data Finally Found Helpful

Drahomíra "johanka" Spoustová, Pavel Pecina,
Jan Hajič, Miroslav Spousta

Institute of Formal and Applied Linguistics, Charles University, Prague

Motivation

Supervised machine-learning methods rely on extent and quality of annotated data ⇒ all data should be checked for the quality of annotation.

Our aim is to:

- Check quality of the corpus annotation in a cost effective way
- Find out whether there is still a room for improvement of current POS taggers

Czech Part of Speech Tagging

- Every token is assigned a set of morpho-syntactic features: *tag*.
- Currently 13 features are used: Part of Speech, Detailed Part of Speech, Gender, Number, Case, Possessor's Gender, Possessor's Number, Person, Tense, Degree of comparison, Negation, Voice, Style
- Out of all possible combinations about 4,200 are currently used.
- 2 mil. tokens of manually anotated data (Prague Depedency Treebank 2.0)
- Several different POS taggers are currently available for Czech:
    - HMM tagger
    - Feature-Based (maximum entropy) tagger
    - Averaged perceptron tagger

Task: Validation of Quality of the POS Tagging

- Simple way: re-annotate all data by several annotators
  - Expensive
- Common way: re-annotate randomly selected sample of the data only
  - Ineffective
- Best way: re-annotate carefully selected sample of the data only
  - How to carefully select the data for re-annotation?

Data selection

How to carefully select the data for re-annotation? We split our data (train and devel-test set) into several parts:

- Trivial data: non-ambiguous tokens
- Easy data: (at least) three taggers were run on the *train* and *dtest* data sets; easy data are those positions, where all taggers agree with each other
- Problematic data: the rest (at least one tagger does not agree with the rest)

For the re-annotation, we select shuffled random sample of the *easy* $(\frac{1}{3})$ and *problematic* data $(\frac{2}{3})$.

Data selection: PDT 2.0

All three taggers were run on the PDT 2.0 *train* and *dtest* data sets.

data	*train*		*dtest*	
trivial	679,061	44.12 %	86,922	43.11 %
easy	788,573	51.23 %	95,604	47.41 %
problematic	71,607	4.65 %	19,125	9.48 %
total	1,539,241	100 %	201,651	100 %

- For re-annotation we selected 25 % of *problematic* data of the *dtest* set (5,000 tokens) and only half the size from the *easy* data (2,500 tokens).
- The same amounts of tokens were sampled from the *train* data set.
- Entire set of 15,000 tokens was independently annotated by three human annotators

Inter-annotator agreement

- We measured agreement for both *easy* and *problematic* data samples (in %)

	data	*size*	*A1*	*A2*	*A3*	*voted*	*tagger*
dtest	easy	2,500	97.00	99.04	98.36	99.32	100.00
	problematic	5,000	88.48	92.86	88.46	92.48	52.58
	all (weighted est.)	—	97.60	98.94	98.24	**99.04**	**95.95**
train	easy	2,500	97.64	98.52	97.92	98.92	100.00
	problematic	5,000	86.66	90.12	81.46	89.88	62.30
	all (weighted est.)	—	98.17	98.78	98.07	**98.98**	**98.25**

- Voted: at least two annotators have to agree on the tag, in case of draw, A2 is used
- Agreement of human re-annotation (A1 – A3) and the current reference annotation is high
- Room for improvement of the taggers: 99.04 % – 95.95 % = 3.09 %.

Detailed Analysis

Following results of the annotation, we can further distinguish several classes of tokens:

Correct annotation At least two annotators agree with the reference annotation (we eliminate error of single annotator)

Incorrect annotation All three annotators agree with each other, the reference annotation differs (the reference tag is probably wrong)

Vague annotation All other cases: either multiple tags equally correct or error of multiple annotators. We cannot distinguish these two cases, so we only know the upper limit of a number of vague tags.

data		*all*	*correct*	*incorrect*	*vague*
dtest	easy	2,500	2,482	4	14
	problematic	5,000	4,605	171	224
train	easy	2,500	2,471	13	15
	problematic	5,000	4,458	255	287

Detailed Analysis: PDT 2.0

- Estimation of *correct*, *incorrect* and *vague* classes in the PDT 2.0 *train* and *dtest* data sets according to results of the annotation:

data		*size*	*correct*	*incorrect*	*vague*
dtest	easy	95,604	99.28	0.16	0.56
	problematic	19,125	92.10	3.42	4.48
	all (weighted)	201,651	**98.99**	**0.37**	**0.65**
train	easy	788,573	98.84	0.52	0.64
	problematic	71,607	89.16	5.10	5.74
	all (weighted)	1,539,241	**98.90**	**0.50**	**0.59**

- Extending our estimation to the entire PDT 2.0 (including *etest*) we can conclude that
  - 1,939,314 tokens (98.91 %) are annotated with correct tags
  - 9,563 tokens (0.49 %) are annotated with incorrect tags
  - (up to) 11,780 tokens (0.60 %) are vague tags (undecidable ambiguities, foreign words etc.).

Conclusion

- New method of validation of the corpus quality
  - Only 0.49 % incorrect tags in PDT 2.0
- Detection of large subset of *problematic* and *vague* tags with minimal costs
  - By re-annotation of 5 000 tokens of *problematic* data we found 255 incorrect tags. If chosen randomly (out of non-trivial data), we need re-annotation of 28 000 tokens for the same amount of incorrect tags.
- Estimation of room for improvement of current POS taggers:
  - Current taggers can be improved (up to 99 % accuracy)