Part-of-Speech Tagging: The Power of the Linear SVM-based Filtration Method for Russian Language

Anton Kazennikov
IQMEN LLC

MorphoRuEval Shared Task Goals

- Extended POS Tagging
  - POS
  - Morphological features
- Lemmatization

- Different text sources
  - News
  - Social media texts
  - Fiction texts

Shared Task Tagset

#	Category	Features
1	POS	NOUN, PROPN, ADJ, PRON, NUM, VERB, ADV, DET, CONJ*, ADP, PART*, H*, INT*, PUNCT*
2	Case	Num, Gen, Dat, Acc, Loc, Ins
3	Number	Sing, Plur
4	Gender	Masc, Fem, Neut
5	Animacy	Anim*, Inan*
6	Tense	Past, Notpast
7	Person	1, 2, 3
8	VerbForm	Inf, Fin, Conv
9	Mood	Ind, Imp
10	Variant	Short/Brev
11	Degree	Pos, Cmp
12	NumForm	Digit

* - non-evaluated features

Training Corpora

Corpus	*Tokens*	*Unique Lemmas*	*Unique Feature sets*	*Unique words*
GICR	1M	43k	303	115k
SynTagRus	0.9M	43k	250	104k
RNC	1.2M	53k	557	127k
OpenCorpora	0.4M	42.5k	337	79k

Proposed approach

- Knowledge-based + Machine Learning
- Unified solution for both tasks
- Based on n*ormalizing substitution*

  <WF ending, NF Form ending, Feature set>

- Two stages:

  1) Generate candidate parses for each word
  2) Select optimal parse in left-to-right manner

Candidate Generation

- Close to classical morphological analysis: based on AOT dictionary
- Guesser: FSA on reversed words (split stem + ending)

Algorithm:

1) Collect parses from AOT Dictionary
2) Collect parses from corpus-based dictionary (GICR)
3) Collect parses from hand-generated dictionary (~50 entries)
4) If no parses found, then guess them

AOT Dictionary

- Based on A. A. Zaliznyak's morphological dictionary
- 175k entries
- 4.6M wordforms

Entry

- Entry-wide features
- Stem
- Paradigm

Paradigm – a sorted list of endings

- First ending – normal form ending

Ending

- Ending string
- Ending features

AOT Dictionary Conversion

Preprocessing

- Feature-based mapping
- Paradigm transformations
  - Verb splitting (participles → Adj)
  - Adj/Cmp → Adv
  - Immutable nouns processing
- Induces some tolerable noise

Conversion procedure

- Build partially converted dictionary
- Filter through GICR+SynTagRus parses
- Keep only unambiguous substs
- Converts 1.6M of 4.6M wordforms
- Various corpora/AOT mismatches:
  - Some nouns have Anim/Inan ambiguity - “Земля” vs. “земля”
  - Gender-No Gender: мр-жр - “пьяница”
  - Proper names vs. locations

Parse Filtering

sent[i] – i-th word

parses[i] – set of parses for i-th word

```
for i=0 to len(sent):
    context = extract_context(sent, parses, i)
    scores = score(parses[i], context)
    max_parse = argmax(parses[i], scores)
    sent[i].feats = max_parse.feats
    sent[i].lemma = max_parse.lemma
```

Parse Filtering: Scorer construction

- Score – sum of scores of each features
- Multi-class SVM Classifier on context features
- Trained on each group of features
- Scores POS and morphological features separately
- Hash kernel for model size reduction

dot(class,w,x) = sum(w_class[i] x[i])

→

dot(class,w,x) = sum(w[hash(i, class)] x[i])

Parse Filtering: Context Features

Window of ±3 words around current word features:

- Prefixes, suffixes, wordform itself
- Stems and endings
- POS tag and features combinations
- Ambiguity classes for unparsed context
- Graphical features: capitalization, punctuation etc.
- Bigram and trigram feature schemes
- ~ 150 features for each word

Ambiguity class for unparsed context - a sorted set of candidate features: “gen/acc” for case ambiguity

Shared Task Results on Closed Track

News			
POS	Sent POS	Lemma	Sent Lemma
93,99	64,8	93,01	**56,42**
93,83	**63,13**	**92,96**	54,19
93,71	61,45	89,61	40,22
93,35	55,03	89,23	37,71

Vkontakte			
POS	Sent POS	Lemma	Sent Lemma
92,42	65,85	**91,69**	**61,09**
92,39	**64,08**	90,97	60,21
92,29	63,56	88,65	52,64
91,49	61,44	90,97	48,94

Fiction			
POS	Sent POS	Lemma	Sent Lemma
94,16	65,23	**92,01**	**57,11**
92,87	**60,91**	91,46	55,08
92,16	60,15	90,28	45,18
92,40	56,60	88,84	35,28

Average			
POS	Sent POS	Lemma	Sent Lemma
93,39	65,29	**92,22**	**58,21**
93,08	**62,71**	91,81	56,49
92,64	61,01	89,32	45,18
92,57	58,40	88,47	44,78

Additional Experiments

#	Team	POS	Sent POS	Lemma	Sent Lemma
1	**AOT-2017**	**93,64**	**63,93**	**92,76**	**59,39**
2	C	93,39	65,29		
3	**AOT-2012**	**93,08**	**62,71**	**92,22**	**58,21**
4	**w/o guesser**	**93,00**	**60,68**	**92,21**	**56,59**
5	H	92,64	58,4	80,71	25,01
6	A	92,57	61,01	91,81	56,49
7	**w/o AOT**	**91,13**	**53,48**	**80,72**	**24,54**
8	**w/o AOT, guesser**	**88,43**	**45,60**	**86,74**	**40,46**

Conclusions

- Solid, high-performance approach
- Robustness across different text sources
- Highly sensitive to dictionary quality
- Efforts on corpus and dictionary unification could further improve the performance of the presented approach

Thanks for your attention!
Questions?

E-mail: kazennikov@iqmen.ru

Code and data available at:

https://github.com/kzn/morphoRuEval