Parts of Speech in Computational Linguistics

Antal van den Bosch
KNAW Meertens Institute

Interframework Colloquium, Utrecht, 13 February 2020

PoS in Computational Linguistics

"You shall know a word by the company it keeps" (Firth, J. R. 1957:11).

- Fruitfly status
- Proliferation and standards
- Frog
- Relativity

Fruitfly status

- PoS tagging was fruitfly of CL in 1990s
  - By 2000, "solved"
  - "solved" = 95% accurate or better on main tags
- Generic NLP & machine learning problems first encountered with PoS tagging
  - Machine learning only works under i.i.d. assumption
    - Independent & identically distributed. If test is not like training, forget it.
  - Training on Wall Street Journal corpus, testing on Alice's Adventures in Wonderland
    - Drop of 95% accurate to 75% accurate
    - "rose" is a verb in past tense, "shares" is a plural noun

Proliferation and standards

- Many tagsets for many languages
  - Different historical reasons for existence
  - Reflecting linguistic frameworks, theories
  - Early standardization efforts (Penn Treebank, EAGLES) – between 12 and 50 tags
- Benchmarking / shared task culture
  - Often: WSJ Penn Treebank
  - English-centeredness
- Standardisation
  - Universal PoS tags - https://universaldependencies.org/u/pos/
  - Universal features - https://universaldependencies.org/u/feat/index.html
- Allergy towards PoS tagsets that are biased by a framework

Frog - https://languagemachines.github.io/frog/

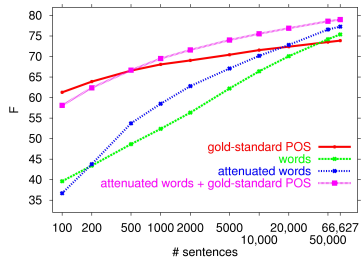
```
1     Marie   Marie   [Marie] SPEC(deeleigen)                          1.000000   B-PER   B-NP   2   su
2     vroeg   vragen  [vraag] WW(pv,verl,ev)                          0.532544   O       B-VP   0   ROOT
3     zich    zich    [zich]  VNW(refl,pron,obl,red,3,getal)          0.999740   O       B-NP   2   se
4     of      of      [of]    VZ(fin)                                 0.996853   O       O      2   svp
5     of      of      [of]    VG(onder)                               0.733333   O       B-SBAR 4   vc
6     hij     hij     [hij]   VNW(pers,pron,nomin,vol,3,ev,masc)      0.999659   O       B-NP   8   su
7     nog     nog     [nog]   BW()                                    0.999930   O       B-ADVP 8   mod
8     zou     zullen  [zal]   WW(pv,verl,ev)                          0.999947   O       B-VP   5   body
9     komen   komen   [kom][en]       WW(inf,vrij,zonder)             0.861549   O       I-VP   8   vc
10    .       .       [.]     LET()                                   0.999956   O       O      9   punct
```

- CGN tagset
  - Van Eynde, Frank. 2004. Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands. Technical report, Centrum voor Computerlinguistiek, KU Leuven, Belgium.
  - "Tot de adjectieven rekenen we niet alleen de prenominaal en de predikatief gebruikte bijvoeglijke voornaamwoorden, maar ook de zelfstandig (of nominaal) gebruikte en de adverbiaal gebruikte."

Relativity

- PoS considered a standard preprocessing task
  - Alongside lemmatization
- But is it really necessary?
  - Even 12 core tags are overkill; sometimes only content / function word distinction matters
  - "implicit linguistics"

Function tagging, learning curves



7

Case study: Overall setup

- "chunking-function tagging", English
- Select input:
  - Gold-standard or predicted PoS
  - Words only
  - Both
- Learn with increasing amounts of training data
  - Which learning curve grows faster?
  - Do they meet or cross? Where?

8

Data (1): Get tree from PTB

```
((S (ADVP-TMP Once)
    (NP-SBJ-1 he)
    (VP was
        (VP held
            (NP *-1)
            (PP-TMP for
                    (NP three months))
            (PP without
                (S-NOM (NP-SBJ *-1)
                       (VP being
                           (VP charged)
))))) .))
```

9

Data (2): Shallow parse

[ADVP Once$^{\text{ADVP-TMP}}$]
[NP he$^{\text{NP-SBJ}}$]
[VP was held$^{\text{VP/S}}$]
[PP for$^{\text{PP-TMP}}$]
[NP three months$^{\text{NP}}$]
[PP without$^{\text{PP}}$]
[VP being charged$^{\text{VP/SNOM}}$]

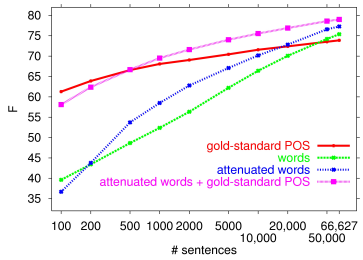
10

Case study: Details

- experiments based on Penn Treebank III
  (WSJ, Brown, ATIS)
  - 74K sentences, 1,637,268 tokens (instances)
  - 62,472 unique words, 874 chunk-tag codes
- 10-fold cross-validation experiments:
  - Split data 10 times in 90% train and 10% test
  - Grow every training set stepwise
- precision-recall on correctly chunked and typed chunks with correct function tags
- memory-based learning (TiMBL)
  - MVDM, k=7, gain ratio feature weights, inverse distance class voting
  - TRIBL level 2 (approximate k-NN)

11

Results: learning curves



12