IB047

Automatické značkování

Pavel Rychlý

pary@fi.muni.cz

22. dubna 2022

Automatické značkování

- vstup text
- výstup text + morfologické značky, příp. základní tvary
- různé přístupy
  - pravidlové
  - statistické
  - neuronové sítě
- trénování na označkovaných datech
- s pomocí externích zdrojů (morfologické databáze), velkých (neoznačkovaných) korpusů
- vyhodnocení na nazávislých datech

Vyhodnocení značkování

Porovnání proti *pravdě* (Gold Standard)

```
Všechny   DET     PRON    <<
tři       NUM     NUM
světy     NOUN    NOUN
si        PRON    PRON
vzájemně  ADV     ADV
trvale    ADV     ADV
povídají  VERB    VERB
a         CCONJ   CCONJ
ovlivňují VERB    VERB
se        PRON    ADP     <<
```

- 10 tokenů, 2 chyby
- úspěšnost (accuracy): 8/10 = 80 %
- chybovost (error rate): 2/10 = 20 %
  $accuracy = \frac{correct}{alltokens}$ $errorrate = 1 - accuracy$

Vyhodnocení značkování

Při možnosti více značek pro jeden token

- precision – přesnost

$$precision = \frac{tp}{tp + fp}$$

- recall – pokrytí

$$recall = \frac{tp}{tp + fn}$$

- accuracy – úspěšnost

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

- každý token jedna značka: $acc = prec = rec$

Trénovací data

- ruční tvorba časově náročná
  - školení anotátorů
  - řešení neshod mezi anotátory
  - měření mezianotátorské shody (IAA: inter-annotator agreement)
  - kontrola anotátorů
- velký zdroj: Universal Dependences (vi P4)
- stovky tisíc až miliony tokenů
  DE 3,5M, CZ 1,5M, RU 1,5M, JA 1,3M
- z podstaty jen omezené domény
- velká kvalitní data jsou důležtější než algoritmy

Statistické značkování

- pravděpodobnosti značek, slov, ...
- odhad pravděpodobností z trénovacích dat

	počet výskytů	pravděpodobnost
se	16289	
se PRON	14966	$P(PRON\|se) = 14966/16289 = 0.919$
se ADP	1323	$P(ADP\|se) = 1323/16289 = 0.081$

- volíme nejpravděpodobnější značku

Volba nejpravděpodobnějšího

- předpokládejme: $P(PRON|se) = 0.9$, $P(ADP|se) = 0.1$
- volba PRON: $acc = 0.90$
- volba dle rozložení (9:1):
  - má být PRON: 90/100, z toho 81 správně
  - má být ADP: 10/100, z toho 1 správně
  - celkem $acc = 0.82$
- při generování náhodných vět naopak chceme variabilitu – generujeme dle rozložení

Vyhlazování pravděpodobností

- (ne-)nulová pravděpodobnost pro neviděné jevy
- snížení posti pro časté jevy, určení posti pro neviděné jevy
- Good-Turing
  $N = \sum_{r=1}^{max} rN_r$
  $p_0 = N_1/N$
  $p_r = \frac{(r+1)S(N_{r+1})}{rS(N_r)}$

Pravidlové značkování

- pravidla: *slovo není VERB pokud je předchozí slovo "the"*
- hlavně dříve:
  ruční vytváření + případné ověřování v korpusu
- automatiké učení pravidel (Brillův tagger)
- většinou méně robustní

Neuronové sítě

- formou učení jsou podobné statistickým metodám
- velký rozvoj zhruba od roku 2014
- využití jednoduchých nástrojů pro word embedings
  mapování *slovo* $\rightarrow$ *300D vektor čísel*
- velký pokrok, zejména pro navazující úlohy
  – bez explicitního značkování

Kombinování přístupů

- ořezávací pravidla + dořešení víceznačností statistikou
- použití ručního slovníku $\approx$ pravidla
- hlasování/váhování různých přístupů

`desamb.sh`:

```
tecky.pl | majka -p -f majka.w-lt \
| guesser.pl | remove.pl remove.znacky \
| disna d | statdesam.pl
```


Využití neznačkovaných dat

KernelTagger

- most probable PoS tag for annotated words
- derive a PoS tag from 5 most similar words (kernel trick)
- word similarities from a big corpus

Word Similarity Computation

- context: one preceding and one following word
- logDice salience $D(w_a, c)$ of word w_a and context c.
- count only contexts with $D(w_a, c) > 0$
- similarinty of words w_a and w_b:

$$sim(w_a, w_b) = \frac{\sum_c min(D(w_a, c), D(w_b, c))}{\sum_c D(w_a, c) + \sum_c D(w_b, c)}$$

- Sketch Engine Thesaurus
- word embeddings simmilarity

Využití word embedings

- word embedings = vnoření slov do vektorového prostoru
- slovo $\rightarrow (0.3, 0.1, -0.2, ...)$
- může obsahovat mnoho informací, které explicitně nejsou vidět
- neuronové sítě mohou využí, přestože neznají jejich význam
- využití předtrénovaných modelů
- zatím se nevyužívají morfologické databáze pro trénování