

Kombinované metody značkování

(Kterak růst úspěšnosti kopíroval růst plodu)

johanka

May 13, 2007

Úloha morfologického značkování

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Neformálně: pro každý token z N ($N \geq 1$) značek vybrat právě jednu značku (lemma neřešíme a neměříme)

Dostupné prostředky (obecně):

- statistické taggery: vyberou právě jednu značku, lze užít samostatně i v kombinaci
- disambiguační pravidla: N značek zredukuje na 1 až N (příp. 0 až N) značek (pro danou úlohu lze užít pouze v kombinaci)

Dostupné prostředky (konkrétně)

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- Feature-based tagger
- HMM tagger
- tagger Morče
- (případně další taggery)
- pravidla bezpečná (root)
- pravidla veškerá (bezpečná + heuristická) (disheu1)
- (případně jiná podmnožina pravidel)

Počáteční čísla (dtest)

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

	p	r	f
Morfologie	25.72 %	99.40 %	40.87 %
Bezpečná pravidla (root)	58.76 %	98.90 %	73.72 %
Všechna pravidla (disheu1)	67.36 %	98.24 %	79.92 %

Tagger	accuracy
Feature-based (a)	94.27 %
HMM (b)	95.13 %
Morče (m)	95.43 %

Možnosti kombinace (obecně)

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- jeden tagger + pravidla (už se zkoušelo)
- N taggerů + pravidla (ještě se nezkoušelo)
- N taggerů (mraky lidí zkoušely :))

Sériová kombinace (opakování)

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Pravidla po statistice: nezabírá, vysvětleno loni
Statistika po pravidlech: zabírá, ale ne moc

	–	root	disheu1
a	94.27 %	92.51 %	92.55 %
b	95.13 %	95.48 %	95.30 %
m	95.43 %	95.64 %	95.44 %

Sériová kombinace se zjednoznačeným slovním druhem (taky opakování)

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- 1 (Morfologická analýza)
- 2 Tagger (libovolný)
- 3 Vratíme všechny tagy se stejným subposem, jako má vybraný tag
- 4 Pravidla
- 5 Tagger (ne nutně stejný)

Úspěšnost jednotlivých taggerů při určování SUBPOS:

a	99.31 %
b	99.22 %
m	99.25 %

Číslo

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Řádky: první a druhý krok, sloupce: třetí krok

	a	b	m
a+root	92.81 %	95.68 %	95.78 %
a+disheu1	93.08 %	95.69 %	95.77 %
b+root	92.76 %	95.63 %	95.72 %
b+disheu1	93.02 %	95.64 %	95.71 %
m+root	92.79 %	95.63 %	95.75 %
m+disheu1	93.05 %	95.64 %	95.73 %

Ověření údernosti pravidel

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Po vynechání pravidlového kroku (řádky 1. krok, sloupce 2. krok):

	a	b	m
a	92.96 %	95.18 %	95.42 %
b	92.90 %	95.13 %	95.37 %
m	92.92 %	95.15 %	95.40 %

Nic moc $\Rightarrow$ pravidla něco dělají :)

Sjednocení taggerů

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- 1 (Morfologická analýza)
- 2 Pustíme N taggerů paralelně
- 3 Provedeme sjednocení výsledků a uděláme z něj novou „morfologickou nabídku“ (pro každý token 1 až N tagů)
- 4 Pravidla
- 5 Dorážka jedním taggerem

Volitelné na první pohled: sada pravidel, dorážkový tagger
Volitelné až na druhý pohled :): sada taggerů

Čísla

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

	a	b	m
$(a \cup b) + root$	95.43 %	95.49 %	95.96 %
$(a \cup b) + disheu1$	95.54 %	95.58 %	95.96 %
$(a \cup m) + root$	95.56 %	96.03 %	95.73 %
$(a \cup m) + disheu1$	95.68 %	96.05 %	95.82 %
$(b \cup m) + root$	95.81 %	95.58 %	95.77 %
$(b \cup m) + disheu1$	95.89 %	95.71 %	95.86 %
$(a \cup b \cup m) + root$	95.52 %	95.66 %	95.84 %
$(a \cup b \cup m) + disheu1$	95.69 %	95.80 %	95.95 %

Po vynechání pravidel...

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

	a	b	m
$(a \cup b)$	94.94 %	95.13 %	95.87 %
$(a \cup m)$	95.05 %	95.87 %	95.46 %
$(b \cup m)$	95.56 %	95.13 %	95.48 %
$(a \cup b \cup m)$	94.85 %	95.14 %	95.47 %

⇒ pravidla něco dělají, ale i výsledek bez nich je velmi zajímavý!

Možná vylepšení

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- Optimalizace sady pravidel (jak na výkon, tak na rychlost/kompaktnost)
- Rozšíření sady taggerů (nejlépe o takový tagger, co funguje zase úplně jinak)
- Ale IMHO už to moc nahoru nepůjde

Souhrnná čísla

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Máme k dispozici	d-test	e-test
jeden tagger (m)	95.43 %	95.12 %
dva taggery	—	—
tři taggery	95.87 %	95.52 %
1 tagger (m) + pravidla	95.75 %	95.44 %
2 taggery (b, m) + pravidla	95.86 %	95.49 %
3 taggery + pravidla	96.05 %	95.68 %

Redukce chyby:

	Morče	Sjednocení bez pravidel
Sjednocení bez pravidel	8.20 %	—
Sjednocení s pravidly	11.48 %	3.57 %

Chybovost na jednotlivých pozicích tagu

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

S1 – sjednocení bez pravidel, S2 – sjednocení s pravidly

	a	b	m	S1	S2
1 (POS)	0.61	0.70	0.66	0.57	0.57
2 (SUBPOS)	0.69	0.78	0.75	0.64	0.64
3 (GENDER)	1.82	1.49	1.66	1.39	1.37
4 (NUMBER)	1.56	1.30	1.38	1.18	1.15
5 (CASE)	4.03	3.53	3.08	2.85	2.62
6	0.02	0.03	0.03	0.02	0.02
7	0.01	0.01	0.01	0.01	0.01
8	0.06	0.07	0.08	0.06	0.05
9	0.05	0.08	0.07	0.05	0.04
10	0.29	0.28	0.30	0.26	0.27
11	0.29	0.31	0.33	0.28	0.28
12	0.05	0.08	0.06	0.05	0.04
15	0.31	0.31	0.31	0.28	0.29

Je to k něčemu?

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

Ano, ale...

- Na rozdíl od obvyklého schématu kombinačního experimentu (vyprodukujeme, změříme, publikujeme, zašantročíme) lze tuto metodu užít i v praxi
- A taky jsme tak učinili (oficiální přeznačkování ČNK)
- Ale je to magie :(

Konkrétní magické aspekty:

- Četné konverze formátů a kódování
- Ztrátovost $\Rightarrow$ nemožnost průtokového zpracování
- Některé komponenty nešířitelné, jiné málo udržované...
- Různá rychlost komponent – jiný postup při jednom souboru a jiný při optimalizaci na velká data

Co s tím?

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

1. možnost: polidštit dřevním způsobem (problémové části přepsat a zjednodušit, udělat z toho hezký balíček): pro variantu s pravidly téměř nereálné, pro variantu bez pravidel by to asi šlo (ale já to dělat nebudu ;))

2. možnost: unsupervised metoda – natrénovat něco (asi Morče) na potenciálně nekonečném množství dat označkových kombinovanou metodou s cílem co nejlépe ji emulovat

Výsledky a závěry

Kombinované
metody
značkování

johanka

Intro

Kombinace s
jedním
taggerem

Kombinace s
více taggery

Souhrnná čísla

Přízemní
úvahy

Závěr

- Pěkné číslo :), překonána magická hranice 96 % na d-test
- Dá se v praxi používat, i když je to poněkud náročné
- Více variant pro různé úlohy (varianta bez pravidel horší, ale cca 10x rychlejší $\Rightarrow$ vhodná na hoodně velká data)
- Budoucnost patří unsupervised metodám :)