Obsah minulých dílů telenovely

Bez ladu a skladu

Unsupervised značkování

Další jazyky

Nejnovější pokroky ve značkování (nejen) češtiny

johanka

21. dubna 2008

Tipovací soutěž

Pro který/é z následujících jazyků bude v průběhu přednášky překonán state-of-the-art?

Čeština	95.68 % (johanka et. al., 2007)
Slovenština	89.36 % (TnT, Brants 2000)
Angličtina	97.33 % (Shen et. al., 2007)

(čeština PDT 2.0 etest, slovenština provizorní etest, angličtina WSJ sekce 21-24)

Morfologická analýza (češtiny)

	Název	Popis	Tag1	Tag2
1	POS	Slovní druh	0.61	0.57
2	SUBPOS	Detailní slovní druh	0.69	0.64
3	GENDER	Jmenný rod	**1.82**	**1.37**
4	NUMBER	Číslo	**1.56**	**1.15**
5	CASE	Pád	**4.03**	**2.62**
6	POSSGENDER	Přivlastňovací rod	0.02	0.02
7	POSSNUMBER	Přivlastňovací číslo	0.01	0.01
8	PERSON	Osoba	0.06	0.05
9	TENSE	Čas	0.05	0.04
10	GRADE	Stupeň	0.29	0.27
11	NEGATION	Negace	0.29	0.28
12	VOICE	Vid	0.05	0.04
15	VAR	Varianta, styl	0.31	0.29

Možnosti a meze taggingu

Nejlepší metoda se neshodne s anotací na 4 % dat – co to může znamenat?

1. Ruční anotace vůbec není elementem morfologické nabídky
   1. Rozklížení verzí morfologie
   2. Úmysl anotátora (nedostatečný recall morfologie)
2. Ruční anotace je elementem morfologické nabídky
   1. Tag je jednoznačný
      1. K jeho určení anotátorovi stačila aktuální věta
      2. Bylo třeba znát širší kontext nebo něco dalšího
   2. Tag je sporný (více anotátorů může mít různý názor)

Chyba v anotaci (anotátor zaklikl něco jiného, než chtěl) může být kombinována s kteroukoli z uvedených možností!

Co jsme provedli?

Označkovali jsme data několika taggery a provedli několikanásobnou re-anotaci dobře vytipovaných tagů.

- Dtest – problémy 5000, placebo 2500, 3 anotátoři
- Pididtest – problémy 667, placebo 333, 5 anotátorů
- Train – problémy 5000, placebo 2500, 3 anotátoři

Problémy: neplatí a = b = m = pdt
Placebo: rovnost platí a zároveň morfologie nabízí více než jeden tag.

Co lze tímto způsobem najít?

- Chybně anotované tagy (může znamenat nejen chybu anotátora, ale i rozjetí verzí morfologie)
- Sporné tagy (dělají problém i anotátorům)
- Bezesporné, leč pro taggery problémové tagy
- Odhad chybovosti jednotlivých anotátorů
- Horní odhad možností taggeru

Odhad chybovosti anotace PDT (1)

	správné	chybné	nejasné
dtest problémy	92.10 %	3.42 %	4.48 %
dtest placebo	99.28 %	0.16 %	0.56 %
dtest vážený	**98.99 %**	**0.37 %**	**0.65 %**
train problémy	89.16 %	5.10 %	5.74 %
train placebo	98.84 %	0.52 %	0.64 %
train vážený	**98.90 %**	**0.50 %**	**0.59 %**

Odhad chybovosti anotace PDT (2)

Absolutní odhad pro celé PDT (i s e-testem):

- 9563 chybných tagů (0.49 %)
- 11780 nejasných tagů (0.60 %)
- 98.91 % dat je tedy zcela v pořádku...
- ...což ovšem stále nezaručuje, že jsou pokryta morfologií

Odhad pro horní mez taggerů:

- 1.56 % sjednocení všech překážek (nejasnosti a chyby anotace, nedostatečnost morfologie)
- tj. měly by jít teoreticky vylepšit až do úspěšnosti 98.44 :)

Značkování mluvené řeči

Specifika přepisů/výsledků rozpoznávání mluvené řeči:

- Text je prasečí (chybějící kapitalizace a interpunkce, přeřeky, chyby rozpoznávadla)
- Občas i doménově specifický (Malach...)

Možnosti:

- Použití stávajících taggerů, tak jak jsou
- Přímé přetrénování na přepisech
- Fikanější přetrénování na kombinovaných datech
- (P.S. Pravidla si fakt neškrtnou ;))

Značkování mluvené řeči - výsledky

V obou případech přepisy – ASR výstup není (zatím) jak evaluovat.

	malach-dtest	dialog
Počet tokenů	71038	46725
Recall morfologie	99.75 %	99.96 %
Feature-based tagger	92.58 %	92.13 %
Morče (ragby) standardní	93.79 %	93.56 %
Ragby jen na Malachu	96.12 %	92.05 %
Ragby malach+PDT	96.30 %	94.21 %
Ragby malach+PDT měňavka	**96.50 %**	**94.22 %**

K čemu to bude dobré, to se teprve uvidí...

Vliv taggingu na parsing

Má vůbec smysl sbírat desetinky?
Jak moc to pomůže navazujícím úlohám a jak moc by pomohl ideální tagger?
A co je vlastně vhodná navazující úloha? :)

Parsing (McDonald, dtest):

tagger	accuracy parseru
Standardní (Feature-based b)	84.303 %
Morče ragby	84.755 %
Morče unsupervised	84.969 %
Anotace místo taggeru	85.767 %

Závěr: křišťálová koule snižuje chybu o 9.33 %, náš nejlepší pokus o 4.24 %.
Překlad: Obo selhal ;)

Házení rukavice do Brna alias hrachu na stěnu

Možnosti porovnání našich nástrojů:

- Na hřišti jednoho z nás (tj. hosté převezmou tagset i trénovací a testovací data) – my můžeme na jejich hřiště a chceme, oni dělají fóry :)
- Na neutrální půdě (tj. na neznámých datech s ad hoc vyhodnocením průniku tagsetů – chtějí oba, neshody v detailech a je třeba netriviálního času a peněz)
- Na aplikaci, tj. např. parseru z minulého slajdu – bez obtíží realizovatelné, my chceme, oni zase nic...

Potřeba oddémonizovat značkování češtiny

Výchozí stav – závěr loňské přednášky a technologické změny v mezičase (vše dtest).

Metoda*	Loni	Mezičas	Letos
M1	95.43 %	95.90 %	95.90 %
M2	**95.43 %**	95.90 %	??
M3	95.87 %	96.02 %	??
M4	**96.09 %**	96.20 %	??

*) M1 Nejlepší metoda snadno trénovatelná, spustitelná, přenositelná...
M2 Nejlepší metoda snadno spustitelná
M3 Nejlepší metoda ryze statistická
M4 Nejlepší metoda vůbec

Nápady, ideologie (1)

Připomenutí nejlepší metody: ze sjednocení výsledků několika taggerů se udělá "morfologická nabídka", volitelně se prořeže pravidly a následně se předhodí závěrečnému taggeru.
Jak se zbavit magie?

- Natrénovat Morče na megadatech označkovaných hydridem – nepomohlo.
- Natrénovat Morče na podmnožině megadat, kde se shodlo vícero taggerů – nepomohlo.
- Přidat jako feature do Morčete slovní třídu (získanou z megadat magickou implementací Davida Klusáčka) – pomohlo nepatrně, prozatím odloženo.

Nápady, ideologie (2)

- (Strategický nápad č. 1) Natrénovat Morče na megadatech označkovaných hybridem vtipně proložených kopiemi PDT trainu....
- (Doladění strategického nápadu) V každé iteraci dát Morčeti jiná trénovací data: vždy nejprve PDT train a za ním přilepený unikátní kus megadat (v řádu jednotek megatokenů).
- ...a následovala už jen hromada experimentů na vyladění nejlepších parametrů :)
- Postup lze samozřejmě zkusit iterovat, tj. udělat hybrid zahrnující jedno či více různých takto vzniklých unsupervised Morčat.

Shrnutí provedeného oddémonizování

Dtest:

	Loni	Mezičas	Letos
M1	95.43 %	95.90 %	95.90 %
M2	95.43 %	95.90 %	96.24 %
M3	95.87 %	96.02 %	96.14 %
M4	96.09 %	96.20 %	96.37 %

Etest:

	Loni	Letos
M1 (transparentní)	95.12 %	95.58 %
M2 (použitelná)	**95.12 %**	**95.98 %**
M3 (statistická)	95.52 %	95.90 %
M4 (nejlepší)	**95.68 %**	**96.10 %**

Slovenčina

- Těžce ve vývoji, a to jak u zdroje (opravy morfologie, anotace), tak následně u nás
- Existuje morfologická analýza, zatím dost chyb, ale v zásadě dosti propracovaná
- Tagset podobně bohatý jako náš, ale formálně jiný (není poziční) – technická adaptace provedena, obsahová by byla těžší
- Docela dost ručně značkovaných dat (provizorně rozdělena na train (993,841 tokens), dtest (108,176) a etest (94,249 tokens))
- Na .sk straně zatím testovány pouze cizí taggery (TnT, SVM..), které navíc neberou v potaz morfologickou nabídku
- U nás zatím přetrénovány a otestovány Feature-based tagger a Morče v (téměř) stejném nastavení jako pro češtinu

Slovenčina – temporární výsledky a výhled

tagger	accuracy (etest)
TnT	89.36 %
Feature-based	91.48 %
Morče	92.17 %

- Momentálně čekáme na opravy na slovenské straně
- Následovat bude přetrénování taggerů, experimenty s kombinacemi a unsupervised metodami a označkování SNK :)

Angličtina – intro

Vyrobit nějaký tagger – triviální.
Vyrobit nejlepší tagger – čiré šílenství! :)

`http://aclweb.org/aclwiki/index.php?title=State_of_the_art`

První liga (etest):

tagger	acc. publikovaná	acc. dosažená
Shen	**97.33 %**	**97.33 %**
Stanford	97.24 %	97.23 %
SVM	97.16 %	97.13 %
Collins (Morče)	97.11 %	97.13 %

Další taggery (rozchozené): TnT, MXPOST, Tree tagger...

Angličtina – technické záležitosti

- Určuje se pouze slovní druh, celkem 45 kategorií (včetně vší interpunkce)
- Všichni používají Penn Treebank 3 WSJ
- Poslední léta se všichni drží rozdělení 0-18 train, 19-21 dtest, 22-24 etest
- Pravděpodobně všichni berou data z parsed a ne z tagged (neměla by se lišit, ale liší se)
- Někdo používá závorky jako závorky, někdo jako -LRB-, -RRB- apod. Je třeba na to pamatovat a buď přetrénovat, nebo obalit konverzemi
- Morfologickou nabídku si dělá každý sám, postupy jsou různé (většinou triviálně na základě trainu, ale např. Shen dává všem všechno)
- Zcela výjimečně jsou taggery schopny požrat nabídku externě vyrobenou

Angličtina – morfologická nabídka

- Defaultně jde pouze o Morče, které ji potřebuje
- Triviálně můžeme získat tabulku možností z trainu nebo z větších dat ojetých někým jiným (ale recall (cca 99.5 % podle typu dat) nám dost ublíží)
- Rovněž triviálně můžeme dát všem všechno (ale bude to neskutečně pomalé)
- Aktuální stav: ručně zpracovány všechny uzavřené třídy, dále (kvůli rychlosti) vytažen slovník několika tisíc nejčastějších slov z obří tabulky dua Čmejrek-Cuřín, zbytku přiřazeny všechny otevřené třídy. Recall (dtest) 99.96 % (chybí prakticky jen chyby anotace), precision není podstatná :) (asi 15.5 %)

Angličtina – kombinace

Sedm a více taggerů == mnoho možností?
Ne tak docela...

- Jelikož neprořezáváme pravidly, je lepší dělat iniciální sjednocení jen ze dvou a dorazit třetím (jiným!)
- Je možné dělat i stromoidní šílenosti, ale neosvědčilo se to (vyzkoušeno cca 8000 možností :))
- Taggery mimo první ligu nemají prakticky žádný přínos
- A hlavně: skoro žádný tagger nelze jen tak vzít a použít pro dorážkový krok (pouze Morče a Tree tagger)

...ovšem Johanka se dobře vdala :)

Angličtina – unsupervised trénování

- ...pročež se nejlépe vydařila kombinace Stanford+Shen doraženo SVM
- (témeř stejně to vyšlo v pořadí Shen+SVM doraženo Stanfordem)
- Označkovali jsme tím North American News, rozdělili na kousky, a od té doby se Morče trénovalo a trénovalo a jestli neskončilo, trénuje se dosud!

Angličtina – výsledky

Etest:

Metoda	accuracy	redukce*
Stanford (2003)	97.24 %	
Shen (2007)	97.33 %	3.26 %
Kombinace	**97.48 %**	5.62 %
Unsupervised Morče	**97.37 %****	1.50 %

*) redukce chyby oproti předchozímu nejlepšímu publikovanému výsledku
**) stav z dnešního rána

Pro Šlezu: (F. Wilcoxon says:) zlepšení hybridu je signifikantní, unsupervised Morčete zatím ne :)