POS TAGGERY

Petr Kříž, 50836@mail.muni.cz

POS TAGGING

- Základní pojmy
- Tagy Brownova korpusu
- Tagy Penn Treebank
- Čtyři systémy POS tagging
- Jejich srovnání

CO JE POS?

- Part of speech
- Tradiční gramatická kategorie
- V angličtině je jich osm: nouns, pronouns, verbs, adjectives, conjunctions, adverbs, interjections, prepositions
- Netradiční: determiners

POS jemnější členění

- Noun
  - basic S/P
  - possessive S/P
  - proper S/P
  - possessive proper S/P
- Verb
  - basic/infinitiv
  - past tense
  - gerund/present participle
  - past participle
  - singular present
    - 3rd person
    - non 3rd person

POS jemnější členění (2)

- Adjective
  - Basic
  - Comparative
  - superlative
    - semantic (top)
    - morphological (biggest)
- Adverb
  - Basic
  - Comparative
  - superlative
  - nominal (here, down, downstairs)

POS jemnější členění (3)

- Pronoun
  - Personal
    - possessive (my)
    - 2nd possessive (mine)
    - subjective (I, you)
    - objective (me, him)
    - reflexive (myself)
  - Nominal
    - Basic (anything, everyone)
    - Possessive (everyone’s)

POS jemnější členění (4)

- Preposition
- Conjunction
  - Coordinating
  - Subordinating
- Interjections

Netradiční POS

- Determiners – modifikátory substantiv, lze je syntakticky odlišit od adjektiv. Odkazují na vnější vlastnosti jako výskyt, množství, vlastnictví apod. (zapojení substantiva do kontextu)
- Členění
  - articles (a, the)
  - demonstratives (this, those)
  - quantifiers (all, few, no)
  - possessive (my, our)
  - cardinal numbers (one, two)
- Post-determiners – enough, indeed

POS TAGGING

- Part-of-speech tagging
- Také grammatical tagging
- Přiřazování správných gramatický značek jednotlivým slovům v textu
- Využívá nástrojů morfologické analýzy – POS tagging tak lze chápat jako její zjednoznačnění
- Výběr na základě formy věty a pozice zkoumaného tvaru v rámci této věty, popř. lze zapojit i širší kontext.

Brown corpus

- The Brown Corpus of Standard American English, Henry Kucera, W. Nelson Francis, Brown University, Rhode Island, 1967
- 1 milión slov, první větší korpus
- Následujících sedm let byl značkován, i automaticky (Greene and Rubin), velký počet chyb
- 81 značek (PsOS v širším smyslu slova)
- Značkovací sada se stala základem pro další vývoj

BROWN TAGSET

- Determiners
  - General
    - DT – singular determiner (this)
    - DTS – plural determiner (these)
    - DTI – singular i plural (some, any)
    - DTX – double conjunction (either)
  - Articles (a, the, no)
  - Pre-determiners
    - ABL – pre-qualifier (quite, rather)
    - ABN – pre-quantifier (half, all)
    - AXN – pre-quantifier (both)

BROWN TAGSET (2)

- Determiners (pokr.)
  - Qualifiers
    - QL – qualifier (very, fairly)
    - QLP – post-qualifier (enough, indeed)
  - Numerals
    - CD – cardinal number
    - OD – ordinal number
- Conjunctions – coordinating conj. (and, or)
  – subordinating conj. (if)

BROWN TAGSET (3)

- Adjective
  - JJ – basic
  - JJR – comparative
  - Superlative
    - JJS – sémantický (top, chief)
    - JJT – morfologický (biggest)
- Adverb
  - RB – basic
  - RBR – comparative
  - RBT – superlative

BROWN TAGSET (4)

- Adverb (pokr.)
  - RN – nominal (here, then, indoors)
  - RP – particle (about, off, up)
- Pronoun
  - Nominal
    - Basic – PN
    - Possessive – PN$
  - Personal
    - Possessive – PP$ (my)
    - Second possessive – PP$$ (mine)

BROWN TAGSET (5)

- Pronoun (pokr.)
  - Personal
    - Reflexive – PPL (myself), PPLS (ourselves)
    - Objective – PPO (me, him)
    - Nominative
      - 3rd person – PPS (he, she, it, one)
      - Non 3d person – PPSS (I, we)
- Noun
  - Basic – NN
  - Possessive – NN$

BROWN TAGSET (6)

- Noun (pokr.)
  - Plural – NNS
  - Possessive plural – NNS$
  - Proper
    - Basic – NP
    - Possessive – NP$
    - Plural possessive – NPS$
  - NR – adverbial (home, today, west)

BROWN TAGSET (7)

- Verb
  - Tvary “be”
    - BE (be), BED (were), BEDZ (was), BEG (being)
    - BEM (am), BEN (been), BER (are), BEZ (is)
  - Tvary “do”
    - DO (do), DOD (did), DOZ (does)
  - Tvary “have”
    - HV (have), HVD (had), HVG (having), HVN (had)

BROWN TAGSET (8)

- Verb (pokr.)
  - Basic – VB
  - Past tense – VBD
  - Present participle/gerund – VBG
  - Past participle – VBN
  - 3rd singular present – VBZ
  - Modal auxiliary – MD (can, should)
- Wh- word
  - Schopny tvořit věty nahrazující členy daného typu.
  - Wh- determiner – WDT (what, which)

BROWN TAGSET (9)

- Wh- word (pokr.)
  - Wh- pronoun
    - Possessive – WP$ (whose)
    - Objective – WPO (whom, which, that)
    - Nominative – WPN (who, which, that)
  - Wh- qualifier – WQF (how)
  - Wh- adverb – WRB (how, where, when)

BROWN TAGSET (10)

- Samostatné značky
  - Existential there – EX
  - Foreign word – FW
  - Infinitive marker – TO
  - Prepositions – IN
  - Interjection – UH
  - Cited word – NC

BROWNůV VLIV

- CLAWS na začátku používal jeho značkování, pak se postupně vyvinul
- Ovlivnil Penn Treebank tagset, jehož minimálně jako základ používá Brillův tagger, TreeTagger i NLProcessor

PENN TREEBANK

- University of Pennsylvania, 1989
- Věty v korpusu anotovány syntaktickými stromy.
- 36 tagů
- Jednodušší než Brown tagset

PENN TREEBANK TAGSET

- Rozdíly vzhledem k Brown tagset
- Kategorie DT (determiner) a PDT (pre-determiner) slučují několik kategorií z Browna
- Přibyl LS (List item marker – 1.)
- Přibyl NPS (noun proper plural)
- POS (possesive ending) – výrazná redukce possessive kategorií

PENN TREEBANK TAGSET (2)

- Přibylo PP – personal pronoun (I, me, myself, mine), jiné než osobní zde není (možná wh- p.)
- PP$ – possessive pronoun (my, your), mine ne, to je značkováno jako PP
- RBS – adverb superlative, v Brownovi RBT
- SYM – symbol (matematické operátory, fyzikální veličiny)
- TO – na rozdíl od Browna, v němž značí to pouze to v infinitivu, značí zde i předložku

TREETAGGER

- Univerzita ve Stuttgartu
- Implementuje několik jazyků (Angličtina, Francouzština, Italština, Španělština, Bulharština, Ruština, Portugalština, Galicijština)
- Pro angličtinu používá tagset Penn Treebank rozšířenou o VH* tagy. Klasické VB* se stává značkou pro tvary slovesa "to be", VH* je značkou pro tvary slovesa to have, V* slouží k označkování všech zbývající sloves.
- Tečka má značku SENT (sentence).

TREETAGGER – ukázka výstupu

No	UH	No	
one	CD	one	
would	MD	would	
have	VB	have	
believed	VBN	believe	
in	IN	in	
the	DT	the	
last	JJ	last	
years	NNS	year	
of	IN	of	
the	DT	the	
nineteenth	JJ	nineteenth	
century	NN	century	
that	IN	that	
this	DT	this	
world	NN	world	
was	VBD	be	
being	VBG	be	
watched	VBN	watch	
keenly	RB	keenly	
and	CC	and	
closely	RB	closely	
by	IN	by	
intelligences	NNS	intelligence	
greater	JJR	great	
than	IN	than	
man	NN	man	
's	POS	's	

BRILLůV TAGGER

- Eric Brill, John Hopkins University, Baltimore, USA
- 1996
- Rule Based Tagger
- Supervised POS tagger
- Jiný Brillův tagger – unsupervised

BRILLůV TAGGER – ukázka výstupu

No/DT one/PRP would/MD have/VB believed/VBN in/IN the/DT last/JJ years/NNS of/IN the/DT nineteenth/JJ century/NN that/IN this/DT world/NN was/VBD being/VBG watched/VBN keenly/RB and/CC closely/RB by/IN intelligences/NNS greater/JJR than/IN man's/NNS and/CC yet/RB as/RB mortal/JJ as/IN his/PRP$ own;/NN

NLPROCESSOR

- Následník systému vyvinutého na univerzitě v Edinburghu, UK
- V současnosti vyvíjen firmou Infogistics
- Zvládá plaintext i XML
- Používá rozšiřitelný lexikon
- 96-98% accuracy pro slova z lexikonu
- 88-92% accuracy pro neznámá slova

NLPROCESSOR (2)

- Umí několik funkcí:
  - Tokenizace
  - Capitalized word normalization
  - Sentence segmentation
  - POS tagging
  - Syntactic chunking
- Přístupnost
  - Není volně ke stažení
  - Lze si ho vyzkoušet přes webové rozhraní (max. 300 slov)

NLPROCESSOR – tagset

- Používá Modified Penn Treebank tagset
- Pouze minimální změny
- PP $\rightarrow$ PRP (personal pronoun), PP$ $\rightarrow$ PRP$, odstraňuje SYM (symbol), TO značí nejen to před infinitivem, ale i předložku

NLPROCESSOR – ukázka výstupu

- <P><S>([No_DT one_NN]) ((would_MD have_VB believed_VBN)) in_IN ([the_DT last_JJ years_NNS]) of_IN ([the_DT nineteenth_JJ
- century_NN]) that_IN ([this_DT world_NN]) ((was_VBD being_VBG watched_VBN keenly_RB)) and_CC closely_RB by_IN
- ([intelligences_NNS]) greater_JJR than_IN ([man_NN])'s_POS and_CC yet_RB as_RB mortal_JJ as_IN his_PRP$ own_JJ ;

CLAWS

- Vyvíjeno nepřetržitě od 1980 v ústavu UCREL na univerzitě v Lancasteru, UK (konference Corpus Linguistics každé 2 roky).
- Značkování BNC (všech 100 mil. slov): přesnost 96-97% (tagset C5), 1.5% chyb, 3.3% nejednoznačností nerozhodnuto

CLAWS – tagset

- CLAWS1 tagset
  - 1980
  - 132 tagů, vznikl úpravou BROWNova tagsetu
- CLAWS2 tagset
  - 1983-86
  - 166 tagů
- C5 tagset
  - 60 tagů
  - Pro potřeby značkování velkých objemů dat (BNC)
  - Snaha zachovávat velikost tagů rovnou třem

CLAWS – tagset (2)

- C6 tagset
  - Opět zvětšeno, na 160 tagů
- C7 tagset
  - Podobný C6
  - Zachování určité zpětné kompatibility k C5 – C7 je zjemnění C5 (mapování n:1)
- C8 tagset
  - Ještě více tagů, zjemnění hlavně v oblasti determiners, zájmen a pomocných sloves
  - Rezignace na pevnou velikost tagu

CLAWS – C5 tagset

- Adjective: JJ → AJ
  - AJ0 – unmarked
  - AJC – comparative
  - AJS – superlative
- Adverbia: RB → AV
  - AV0 – unmarked
  - AVP - adverb particle (up, off)
  - AVQ - wh-adverb (when, how, why)
- Conjunctions: CC → CJ
  - CJC – coordinating conjunction
  - CJS – subordinating conjunction
  - CJT – “that”

CLAWS – C5 tagset (2)

- Cardinal numbers: CD → CRD
- Ordinal numbers: OD → ORD
- Prepositions: IN → PRP
  - Možná kolize s personal pronoun
  - PRF – “of”
  - PRP – basic
- Substantiva: NN
  - NN0 – neutral number noun (data, aircraft)
  - NN1 – singular number

CLAWS – C5 tagset (3)

- Substantiva (pokr.)
  - NN2 – plural
  - NN3 – proper
- Pronoun: PN
  - PNI – indefinite pronoun (none)
  - PNP – personal pronoun (you, mine)
  - PNQ – wh-pronoun (who)
  - PNX – reflexive pronoun (itself)
- Determiners: DT
  - DPS – possessive determiner (your)
  - DT0 – general determiner (these, some)
  - DTQ – wh- determiner (whose, which)

CLAWS – C5 tagset (4)

- Verb:
  - Be – VB
    - VBB (am, are), VBD (was, were), VBG (doing)
    - VBI (infinitive do), VBN (been), VBZ (is, 's)
  - Do – VD
    - VDB, VDD, VDG, VDI, VDN, VDZ
  - Have – VH
    - VHB, VHD, VHG, VHI, VHN, VHZ
  - General – VV
    - VVB, VVD, VVG, VVI, VVN, VVZ
  - modal – VM0
- Punctuation – PU
  - PUL (left bracket), PUN (general mark)
  - PUQ (quotation mark), PUR (right bracket)

CLAWS – C5 tagset (5)

- Article – AT0
- Existential there – EX0
- Interjection – UH $\rightarrow$ ITJ
- Neoznačeno – NULL
- Není v lexikonu – UNC (unclassified)
- The possessive – POS (‘s nebo ‘)
- To před infinitivem – TO0
- XX0 – negace (not, n’t)
- ZZ0 – písmeno abecedy

CLAWS – ukázka výstupu

No_PNI21 one_PNI22 would_VM0 have_VHI believed_VVN in_PRP the_AT0 last_ORD years_NN2 of_PRF the_AT0 nineteenth_ORD century_NN1 that_CJT this_DT0 world_NN1 was_VBD being_VBG watched_VVN keenly_AV0 and_CJC closely_AV0 by_PRP intelligences_NN2 greater_AJC than_CJS man_NN1 's_POS and_CJC yet_AV0 as_AV0 mortal_AJ0 as_CJS his_DPS own_DT0 ;

RASP

- Robust Accurate Statistical Parsing
- University of Sussex, Británie (část pracovníků z Cambridge)
- CLAWS2 tagset
- Podpora rychlého doménového učení
- Sloučení statistického přístupu s inkrementální indukcí gramatických pravidel – proto robustní (vůči změnám gramatiky)
- Závislostní parsing, typy závislostí

RASP – ukázka výstupu

(|No:1_AT| |one:2_PN1| |would:3_VM|
|have:4_VH0| |believe+ed:5_VVN| |in:6_II|
|the:7_AT| |last:8_MD| |year+s:9_NNT2|
|of:10_IO| |the:11_AT| |nineteenth:12_MD|
|century:13_NNT1| |that:14_CST|
|this:15_DD1| |world:16_NN1|
|be+ed:17_VBDZ| |be+ing:18_VBG|
|watch+ed:19_VVN| |keenly:20_RR|
|and:21_CC| |closely:22_RR| |by:23_II|

MINIPAR

- The University of Manitoba, Kanada
- přesnost: 88% precision, 80% recall (testováno na korpusu Susanne, v úvahu se berou i závislosti)
- Tagger, závislostní parser
- Lemmatizace
- Vlastní tagset – jednodušší než Penn Treebank a CLAWS

MINIPAR – ukázka výstupu

```
> (
E0      (()       fin C    *       )
1   (No           ~ U              2       lex-mod         (gov no
    one))
2   (one          no one N         5       s       (gov believe in))
3   (would        ~ Aux            5       aux     (gov believe in))
4   (have         ~ have           5       have    (gov believe in))
5   (believed     believe in V     E0      i       (gov fin))
6   (in           ~ U              5       lex-mod         (gov
    believe in))
E2      (()       no one N         5       subj    (gov believe in)
        (antecedent 2))
```

SROVNÁNÍ ZNAČEK – no one

“No one would have believed”	No	One
Brill’s tagger	DT	PRP
TreeTagger	UH	CD
NLProcessor	DT	NN
CLAWS	PNI21	PNI22
RASP	AT	PN1
MINIPAR	U	N

SROVNÁNÍ – čas

Real time/ User time	Celý text (63.292 slov)	První odstavec (231 slov)	Akce navíc
Brill's tagger	4.595 s	0.750 s	-
TreeTagger	1.139 s	0.247 s	-
NLProcessor	c. 8 m.	1.944 s	chunking
CLAWS	2 m. 26 s	0.804 s	-
RASP	15 m. 51 s	8.601 s	parsing
MINIPAR	43.947 s	0.649 s	parsing

SROVNÁNÍ – 50 slov (1)

- Tři skupiny: Brown/Penn Treebank tagset (Brill, TreeTagger, Infogistics), CLAWS tagset (CLAWS, RASP), MINIPAR
- Prvních 50 slov Války světů H. G. Wellse

SROVNÁNÍ – 50 slov (2)

- Brown/Penn Treebank
- „own“ v „as mortal as his own“ vyhodnotil Brill jako nn, zbylí dva jako jj
- „that“ v „that as men busied themselves“ vyhodnotil Brill jako dt, TT jako wdt (wh-determiner), IG jako in
- „scrutinised“ v „they were scrutinised and studied“ vyhodnotil Brill chybně jako vbd, ostatní správně jako vbn

SROVNÁNÍ – 50 slov (3)

- CLAWS tagset skupina
- První „as“ v „yet as mortal as his own“ vyhodnotil CLAWS jako rg (degree adverb), zatímco RASP jako csa („as“ jako spojka)
- „that“ v „that as men busied“ vyhodnotil CLAWS jako dd1 (singular determiner) zatímco RASP jako cst („that“ jako spojka)
- „studied“ v „they were scrutinised and studied“ vyhodnotil RASP jako vvd, správnější je zřejmě volba CLAWS, který to vyhodnotil jako vvn („scrutinised“ přitom oba vyhodnotily jako vvn)

SROVNÁNÍ – 50 slov (4)

- MINIPAR
- „studied“ v „they were scrutinised and studied“ vyhodnotil MINIPAR jako a
- MINIPAR vyhodnotil „in“ v „would have believed in the last years“ jako u (interjection), „last“ v „the last years“ jako postdet, „nineteenth“ jako n
- „and“ vyhodnocuje jako u
- „his“ (v „as his own“), interpunkce, obě z „were scrutinised“ vyhodnocena jako u

SROVNÁNÍ – 50 slov (5)

- Skupiny mezi sebou
- CLAWS/RASP vyhodnotil „that“ v „that this world was“ jako cst („that“ jako spojka), zatímco BrillI/TT/IG jako in, podobně „than“ v „greater than man“ C/R „than“ jako spojka, B/TT/IG jako in, MINIPAR jako prep
- „as“ v „yet as mortal as his own“ – C/R první hodnotí rozdílně (adverbium, spojka), druhé jako spojku; B/TT/IG hodnotí první jako adverbium, druhé jako spojku; MINIPAR hodnotí oba jako předložku
- „own“ v „as his own“ C/R hodnotí jako da (after determiner), B/TT/IG jako nn/jj, MINIPAR jako n

VYHODNOCENÍ SROVNÁNÍ

- CLAWS a RASP nejpodrobnější
- MINIPAR nejméně podrobný a často nevyhodnocuje (značka u) (ale i s parsingem poměrně rychlý)

ODKAZY (20.5.2007)

- Brill Tagger: http://www.cs.jhu.edu/~brill/
- TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
- Infogistics NLProcessor: http://www.infogistics.com/textanalysis.html
- CLAWS: http://www.comp.lancs.ac.uk/ucrel/claws/
- RASP: http://www.informatics.susx.ac.uk/research/nlp/rasp/
- MINIPAR: http://www.cs.ualberta.ca/~lindek/minipar.htm