DZ Interset

aneb

rádoby univerzální převod morfologických značek mezi různými sadami a jazyky

Daniel Zeman

ÚFAL MFF, Univerzita Karlova, Praha

Grant MSM 0021620838

Proč převádět značky

- Nástroj, který se značkami pracuje (např. parser)
  - Něco o nich předpokládá (nejsou to pouhé řetězce)
- Mezijazyková adaptace parserů (Zeman a Resnik 2008)
  - Pro Stanfordský a Charniakův parser
- Můj parser pro CoNLL 2007
- Lingvistovy dotazy na korpus?

Lze ušetřit práci?

- Během dvou let jsem psal hned několik převodních procedur:
  - Ruština RDT → PDT, arabština Buckwalter → PADT → PDT, švédština MAMBA → PAROLE → dánské PAROLE → anglické Penn, CoNLL (ar, bg, cs, da, de, en, pt, sv, zh…) → PDT
- Vždy šlo o desítky **`if-then-else`**, cílené na jeden konkrétní pár sad značek
  - Takže např. PDT a Penn jsem přepisoval do Perlu už několikrát, *což je hróóózná otrava*

Pozor, nápad!

- Univerzální sada atributů: INTERSET
- (Téměř) vše lze převést do intersetu
- Hodnoty atributů lze interpretovat jako značku (převést interset do tagsetu)

- Jednou zpracovanou sadu značek lze pak převádět kamkoliv
- Něco jako MT přes interlingvu

Související práce

- EAGLES, PAROLE, MULTEXT
  - spíše snaha standardizovat značky, než se poprat s existujícími značkami
  - navíc poměrně eurocentrické
- Hajdarábád: všechny indické jazyky
  - indoárijské i drávidské!
- GOLD ontologie
  - definuje lingvistické pojmy (stejný pojem může v různých jazycích označovat různé věci)
- Články o tom, že univerzální sada **neexistuje**

Nutná omezení

- Motivace je technická, ne lingvistická!
- K univerzálnosti se chceme blížit, ale nikdy jí nedosáhneme
  - Trvalý prostor pro rozšiřování
  - Některé jevy jsou příliš okrajové, ignorovat?
- Část informace ztratíme, nic nepřidáme
  - Cílová sada neumí informaci zachytit
  - I dvě sady pro jeden jazyk mohou být velmi rozdílné
  - Neděláme tagger! Co není na vstupu, to nevíme!

Zemanova Značková Zoo

NNMS1-----A----	Josef
AGFS3-----A----	následující
P1ZS3FS3-------	jejímuž
ClXP3---------2	stě
VB-S---1P-AA---	jsem
Dg-------3A----	nejméně
RR--6----------	v
J,-X---3-------	aby
TT-------------	jen
II-------------	ejhle
X@-------------	noor
Z:-------------	,

Pražské značky (PDT)

NNMS1-----A----	NMS1A
AGFS3-----A----	AVGFS3A
P1ZS3FS3-------	PSEFSZS3
ClXP3---------2	CGXP3-2
VB-S---1P-AA---	VPS1A
Dg-------3A----	DG3A
RR--6----------	R6
J,-X---3-------	JVX3
TT-------------	T
II-------------	I
X@-------------	NOMORPH
Z:-------------	ZIP

Pražské značky (CoNLL)

NNMS1-----A----	N N Gen=M|Num=S|Cas=…
AGFS3-----A----	A G Gen=F|Num=S|Cas=…
P1ZS3FS3-------	P 1 Gen=Z|Num=S|Cas=…
ClXP3---------2	C l Gen=X|Num=P|Cas=…
VB-S---1P-AA---	V B Num=S|Per=1|Ten=…
Dg-------3A----	D g Gra=3|Neg=A
RR--6----------	R R Cas=6
J,-X---3-------	J , Num=X|Per=3
TT-------------	T T _
II-------------	I I _
X@-------------	X @ _
Z:-------------	Z : _

Multext East

NNMS1-----A----	Ncmsny
AGFS3-----A----	Afpfsd
P1ZS3FS3-------	Pr3mdsfnayn
ClXP3---------2	Mcmn3y
VB-S---1P-AA---	Vmip1smanyn
Dg-------3A----	Rgs
RR--6----------	Spsl
J,-X---3-------	Css3
TT-------------	Q
II-------------	I
X@-------------	X
Z:-------------	

Brněnské značky (DESAM)

NNMS1-----A----	k1gMnSc1eA
AGFS3-----A----	k2gFnSc3eA
P1ZS3FS3-------	k3gUnSc3p3hFxR
ClXP3---------2	k4gXnPc3xC
VB-S---1P-AA---	k5gXnSp1mIaIeA
Dg-------3A----	k6d3eAxD
RR--6----------	k7c6
J,-X---3-------	k8p3xS
TT-------------	k9
II-------------	k0
X@-------------	
Z:-------------	

Pennsylvánské značky

CC CD DT EX FW IN JJ JJR JJS LS MD NN NNS
NNP NNPS PDT POS PRP PRP$ RB RBR RBS RP
SYM TO UH VB VBD VBG VBN VBP VBZ WDT WP
WP$ WRB . , : $ # `` '' -LRB- -RRB-

EX = existential *there*

FW = foreign word

IN = preposition or subordinating conjunction

TO = *to*

UH = interjection…

Značky z Brown Corpusu

ABL ABN ABX AP AP$ AP+AP AT BE BED BED* BEDZ
BEDZ* BEG BEM BEM* BEN BER BER* BEZ BEZ*
CC CD CD$ CS DO DO* DO+PPSS DOD DOD* DOZ
DOZ* DT DT$ DT+BEZ DT+MD DTI DTS DTS+BEZ
DTX EX EX+BEZ EX+HVD EX+HVZ EX+MD FW-* FW-
AT FW-AT+NN FW-BE FW-BER FW-BEZ FW-CC FW-
CD FW-CS FW-DT FW-DT+BEZ FW-DTS FW-HV FW-
IN FW-IN+AT FW-IN+NN FW-IN+NP FW-JJ FW-JJR
FW-JJT FW-NN FW-NN$ FW-NNS FW-NP FW-NPS
FW-NR FW-OD FW-PN FW-PP$ FW-PPL FW-PPL+VBZ
FW-PPO FW-PPO+IN FW-PPS FW-PPSS FW-PPSS+HV
FW-QL FW-RB FW-RB+CC FW-TO+VB FW-UH FW-VB…

Ruský závislostní korpus

S ЕД МУЖ ИМ	NNMS1-----A----
S МН РОД ОД *(их)*	PSXXXXP3-------
A МН ИМ	AAXP1----1A----
NUM ВИН	ClXX4----------
V НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л	VB-P---3P-AA---
ADV СРАВ	Dg-------2A----
PR	RR--6----------
CONJ	J^-------------
PART	TT-------------
INTJ	II-------------

Stuttgart-Tübingen Tagset

ADJA ADJD ADV APPR APPRART APPO APZR ART CARD FM ITJ KOUI KOUS KON KOKOM NN NE PDS PDAT PIS PIAT PIDAT PPER PPOSS PPOSAT PRELS PRELAT PRF PWS PWAT PWAV PAV PTKZU PTKNEG PTKVZ PTKANT PTKA TRUNC VVFIN VVIMP VVINF VVIZU VVPP VAFIN VAIMP VAINF VAPP VMFIN VMINF VMPP XY $, $. $(

- Podobně jako v Penn hlavně slovní druhy, jen trochu jemnější
  - Žádná morfologie (rod, číslo, pád, stupeň i osobu němčina rozlišuje!)
  - Substantivní a atributivní zájmena (**S** vs. **AT**)
  - Adposition = Präposition, Postposition, Zirkumposition

Anncorra (Hajdarábád)

NN NST NNP PRP DEM VM VAUX JJ RB PSP RP CC WQ QF QC QO CL INTF INJ NEG UT SYM *C RDP ECH UNK

- Ambice: společné pro všechny indické jazyky (indoárijské i drávidské!)
- Žádná morfologie, přestože jí ind. jazyky mají spoustu
  - Hierarchická sada, morfologii lze prý dodat na konec. A nechtějí kazit skóre taggerům ☺
- Vyšli z Penn sady, upravili si ji
  - Nové slovní druhy, např. postpozice, „quotative“…
  - Vyházeli i zbytky morfologie (množné číslo, stupňování)

Arabština Tima Buckwaltera

```
<token_Arabic>وبالفالوجة
  <voc>wabiAlfAlwjp</voc>
  <pos>wa/CONJ+bi/PREP+AlfAlwjp/NOUN_PROP</p
  os>
  </token_Arabic>
<token_Arabic>مثال
  <voc>mivAlu</voc>
  <pos>mivAl/NOUN+u/CASE_DEF_NOM</pos>
  </token_Arabic>
```

- Značkování prorůstá tokenizací

Arabština Oty Smrže (PADT)

```
N-------1D
Z-------1-
A-----FP2D
S----3MP1-
VIS-------
```

```
NNXX1-----A----
NNXX1-----A----
AAFP2----1A----
PPMP1--3-------
VcXX---XP-AA---
```

Čínština Rocling (Sinica)

Na = common noun
Nb = proper noun
Nc = location noun
Nd = time noun
Nf = classifier
Nh = pronoun
Ne = determiner or cardinal number
Ng = postposition
P = preposition

P01 = 為 wèi, 承蒙 chéngméng, 深為 shēnwèi
P02 = 被 bèi
P03 = 為了 wèile, 為 wèi
P04 = 給 gěi
P06 = 由 yóu
P07 = 把 bǎ, 將 jiāng
P08 = 拿著 názhe, 拿 ná
…
P66 = 為 wèi

PAROLE dánština a švédština

NCCPU==I … *historikere*

NCNPU==D … *Charta_77-folkene*

ANP[CN]PU=[DI]U … *russiske*

AC---U=-- … *5.000*

VADR=----A- … *har*

VAPR=[SP][CN][DI]A-U … *gældende*

RGU … *af*

PP3[CN][SP]U-YU … *sig*

NCUPN@DS … *konflikterna (substantiv utrum pluralis bestämd nominativ)*

AQP0PN0S … *politiska*

DF@NS@S … *det*

MC00G0S … *fyras (gt. gen.)*

V@IPAS … *har*

AP000N0S … *oberoende*

SPS … *av*

RG0S … *inte*

PF@00O@S … *sig*

Švédština MAMBA a PAROLE

NN … noun

PN … proper noun

VN … gerund

AJ … adjective

AV BV FV GV HV KV MV QV SP SV VV WV … verbs

HV … the verb *hava*

I? IC IG IK IP IQ IR IS IT IU … punctuation

NCUPN@DS … *konflikterna (substantiv utrum pluralis bestämd nominativ)*

AQP0PN0S … *politiska*

DF@NS@S … *det*

MC00G0S … *fyras (gt. gen.)*

V@IPAS … *har*

AP000N0S … *oberoende*

SPS … *av*

RG0S … *inte*

PF@00O@S … *sig*

Interset: současný stav

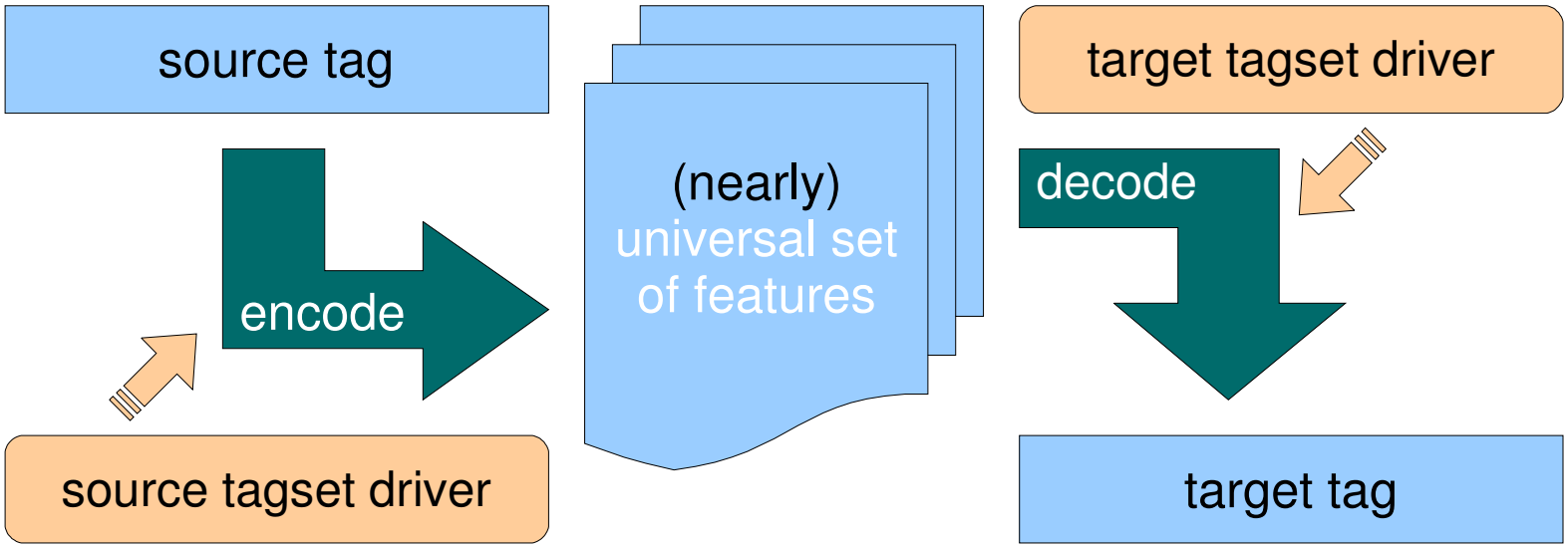
pos	noun	adj	num	verb	adv	prep	conj	part	int	punc		
subpos	prop	class	pdt	det	art	digit	roman	card	ord	...		
prontype	prs	rcp	int	rel	dem	neg	ind	tot				
punctype	peri	qest	excl	quot	brck	comm	colo	semi	dash	symb	root	
puncside	ini	fin										
synpos	subst	attr	adv	pred								
poss	poss											
reflex	reflex											
negativeness	pos	neg										
definiteness	ind	def	red									
gender	masc	fem	com	neut								
animateness	anim	inan										
number	sing	dual	plu									
case	nom	gen	dat	acc	voc	loc	ins					
prepcase	npr	pre										
degree	pos	com	sup	abs								
person	1	2	3									
politeness	inf	pol										
possgender	masc	fem	com	neut								
possnumber	sing	dual	plu									
subcat	intr	tran										
verbform	fin	inf	sup	part	trans	ger						
mood	ind	imp	cnd	sub	jus							
tense	past	pres	fut									
subtense	aor	imp	ppq									
aspect	imp	perf										
voice	act	pass										
foreign	foreign											
abbr	abbr											
hyph	hyph											
style	arch	form	norm	coll								
typo	typo											
variant	short	long	0	1	2	3	4	5	6	7	8	9
tagset	*cs::pdt*											
other	*{ obscure_feature_1 => [0, 7,351.2, [„a“, „b“]] }*											

Disjunkce hodnot

- Někdy značka říká, že např. rod je *mužský* nebo *střední*.
- Interset umí reprezentovat pole alternativních hodnot.
  - Komplikuje se tím práce s ním (viz později)
- Nelze zachytit alternativní kombinace hodnot
  - Např. že jde
    - buď o ženský rod v jednotném čísle
    - nebo o střední rod v množném čísle
    - ale ne ženský v množném nebo střední v jednotném

Ovladač sady značek

- Modul v Perlu s těmito funkcemi:
  - **`decode()`** … převede značku do intersetu
  - **`encode()`** … vygeneruje značku z intersetu
  - **`list()`** … seznam známých značek (volitelné)



Použití ovladače

- Vlastní konverzní skript závisí na formátu dat
- Kromě konverze lze použít i k dotazům na rysy
- Uvnitř konverzního skriptu přibližně toto:

```
use tagset::cs::pdt;
    tagset::en::penn;
while(<>)
{
  chomp;
  my $fs = tagset::en::penn::encode($_);
  my $tgt = tagset::cs::pdt::decode($fs);
  print("$tgt\n");
}
```

Do cílové sady se nevejde vše

- Využijeme pouze atributy, které lze reprezentovat. Ostatní zahodíme.
- Pozor! Může vzniknout značka mimo cílovou sadu.
  - Švédština umí: pos = noun & gender = com | neut
  - A taky: prontype = prs & gender = masc | fem | com | neut
  - Z češtiny máme: pos = noun & gender = masc
  - Buď změnit podstatné jméno na zájmeno, nebo mužský rod na společný. Co má mít přednost?

Vadí značka mimo sadu?

- Atomické značky (Penn) nedávají na výběr.
- Poziční značka zakóduje i „nemožné“ kombinace, např. sloveso v 6. pádě.
- Je-li cílem dotaz na konkrétní atributy, nevadí to. Zachovat co nejvíc informace.

- Je-li cílem práce s programem, do kterého nevidíme, asi do něj nechceme pustit data, na jaká není zvyklý.

Algoritmus znásilňování (rysů)

- Potřebujeme pro každou sadu:
  - Seznam všech možných značek
- K tomu stanovíme centrálně:
  - Priority atributů
  - Pro každou hodnotu pořadí náhradníků
    - Nejčastěji chceme hodnotu jen vynulovat
    - Ale třeba *duál* nejdřív zkusíme nahradit *plurálem*

Algoritmus znásilňování (rysů)

- Dekódovat všechny značky ze seznamu
- Vybudovat trie pro povolené kombinace hodnot
- Při kódování se jich držet
- Není-li hodnota povolená, znásilnit ji (najít nejlepší náhradu)
- Komplikace s poli alternativních hodnot

```
    ],
    "number" =>
    [
        ["sing"],
        ["dual", "plu"],
        ["plu"]
    ],
    "possnumber" =>
    [
```

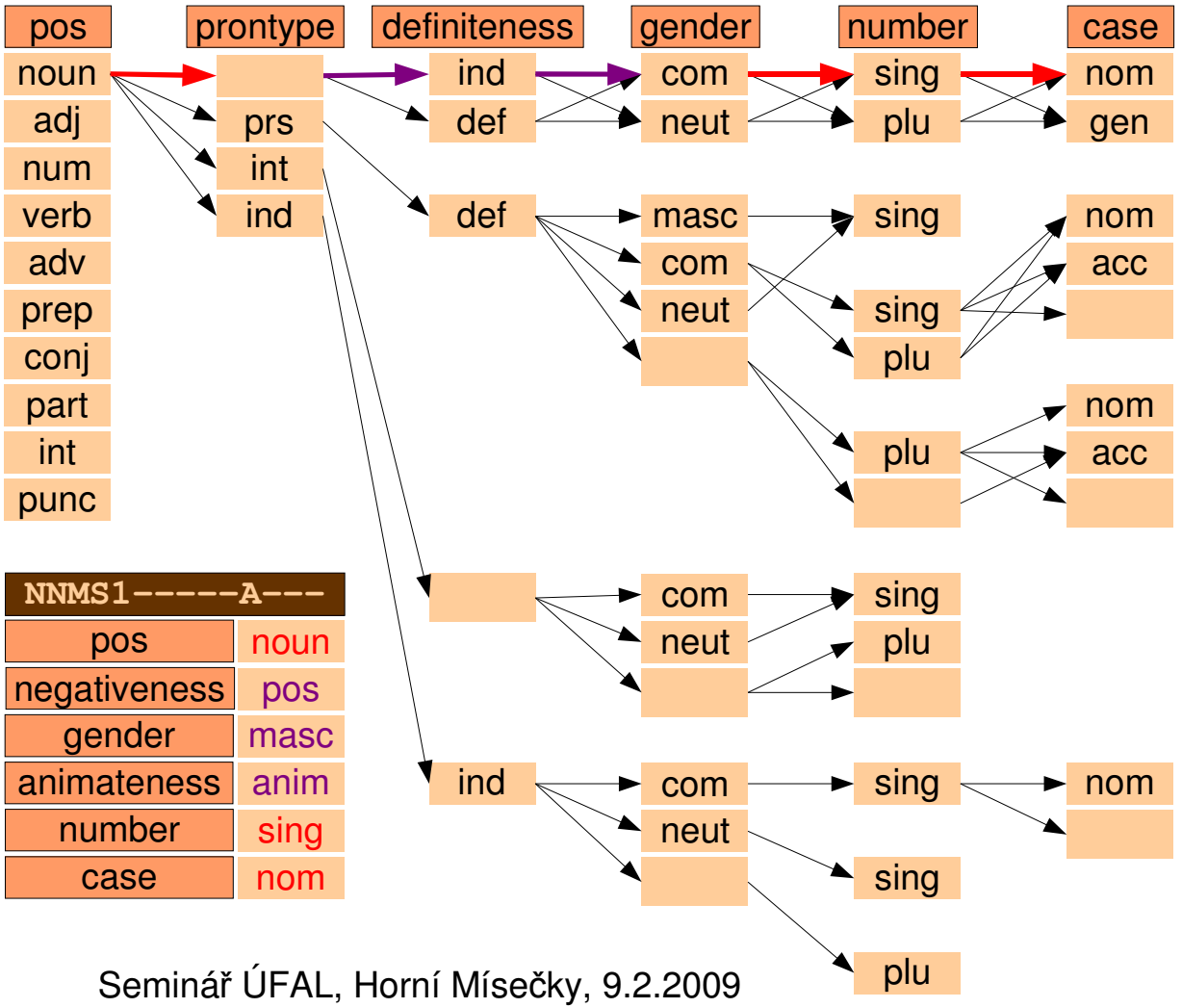
0 → sing, dual, plu

sing → 0, dual, plu

dual → plu, 0, sing

plu → 0, sing, dual

Příklad: cs → sv



Časová náročnost

Sada / Ovladač	Počet značek	Z toho využilo „other“	Čas na implementaci
ar::conll	241	21	13 h
bg::conll	528	247	35 h
cs::conll	4854	775	6 h
cs::pdt	4288	209	18 h
da::conll	143	6	7 h
de::conll	54	1	10 min
de::stts	54	1	4 h
en::conll	45	2	45 min
en::penn	45	2	3 h
pt::conll	657	260	28 h
sv::conll	41	12	20 min
sv::hajic	156	17	8 h
sv::mamba	41	12	3 h
zh::conll	294	294	21 h

Kolik značek přežije?

	ar	bg	csc	csp	da	de	en	pt	svh	svm	zh
ar	241	42	68	54	29	17	15	55	33	12	11
bg	65	528	104	94	64	32	25	87	50	15	11
csc	68	46	4854	4288	44	21	26	125	56	14	11
csp	66	42	4288	4288	42	20	24	120	54	13	11
da	25	46	55	54	143	24	24	49	71	14	11
de	14	16	17	16	17	54	20	29	18	15	10
en	16	17	28	26	22	20	45	24	28	17	11
pt	54	34	113	108	51	30	27	657	46	15	10
svh	33	34	63	62	62	22	28	46	156	17	11
svm	14	15	15	14	15	17	17	15	16	41	10
zh	10	9	10	10	10	11	9	11	10	9	294

Table 2:

Převádíme z řádků do sloupců.

Úspěšnost DZ Parseru na CoNLL

Lang	Year	P(orig)	P(cnv)	Signif
ar	2006	64.3	67.6	yes
ar	2007	59.8	66.9	yes
bg	2006	68.0	71.3	yes
cs	2006	56.1	71.4	yes
cs	2007	58.7	74.0	yes
da	2006	68.3	69.8	yes
de	2006	69.5	67.7	yes
en	2007	63.8	67.3	yes
pt	2006	73.5	76.4	yes
sv	2006	71.0	73.5	yes
zh	2006	69.0	68.0	no
zh	2007	66.1	63.5	yes

Table 3:

Reference

- Daniel Zeman, Philip Resnik: *Cross-Language Parser Adaptation between Related Languages*. In: Proceedings of IJCNLP 2008 Workshop on NLP for Less Privileged Languages. Hajdarábád, Indie, 2008.