Reusable Tagset Conversion Using Tagset Drivers

Daniel Zeman
Univerzita Karlova
Ústav formální a aplikované lingvistiky
Malostranské náměstí 25
CZ-11800 Praha
zeman@ufal.mff.cuni.cz

https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset
(downloads and documentation)

Abstract

Part-of-speech or morphological tags are important means of annotation in a vast number of corpora. However, different sets of tags are used in different corpora, even for the same language. Tagset conversion is difficult, and solutions tend to be tailored to a particular pair of tagsets. We propose a universal approach that makes the conversion tools reusable. We also provide an indirect evaluation in the context of a parsing task.

Types of tags

Atomic tags. Examples: Penn Treebank (English), Stuttgart-Tübingen Tagset (German), Mamba (Swedish), Sinica Treebank (Chinese).

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

Structured tags (compressed representation of feature values). Examples: Prague Dependency Treebank (Czech), Russian Dependency Treebank (Russian).

```
NNMS1-----A----
AGFS3----1A----
VB-S---1P-AA---
VPS1A
V B Num=S|Per=1|Ten=P|Neg=A|Voi=A
S ЕД МУЖ ИМ
```

word class, subclass, gender, number, case, possessor's gender, possessor's number, person, tense, degree of comparison, negativeness, voice, style

AGFS3----1A----

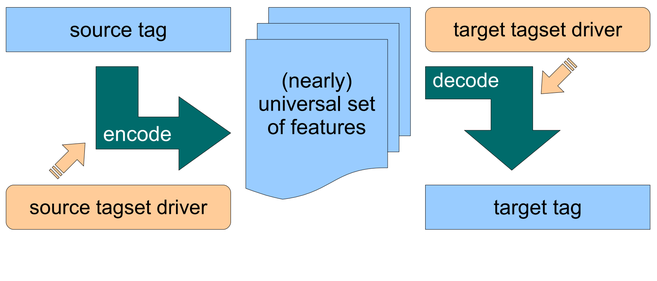
Morpheme tags. Part of speech, morphology and tokenization are mixed together. Example: Buckwalter morphological analysis of Arabic.

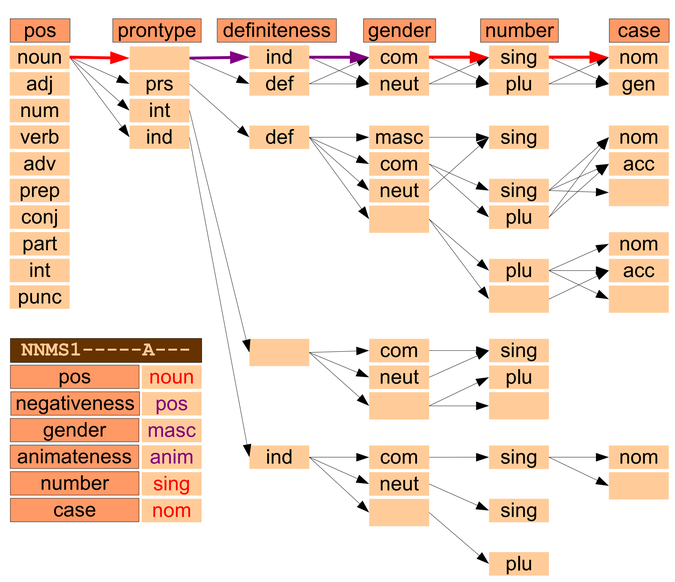
```
<token_Arabic>وبالفالوجة
  <voc>wabiAlfAlwjp</voc>
  <pos>wa/CONJ+bi/PREP+AlfAlwjp/NOUN_PROP</pos>
</token_Arabic>

<token_Arabic>مثال
  <voc>mivAlu</voc>
  <pos>mivAl/NOUN+u/CASE_DEF_NOM</pos>
</token_Arabic>
```

Universal Set of Features
The key idea of our approach is to have a feature structure capable of storing all or most information from any tagset. The structure contains all features whose values are usually encoded in tags. The role of this universal set *("Interset")* is similar to the role of Interlingua in Interlingua-based machine translation or the role of Unicode among character sets. The Interset serves as an intermediate step on the way from tagset A to tagset B. The interaction between the Interset and tagsets A and B, respectively, is described in what we call *tagset drivers.* Once we write the drivers, we can do the two-way conversion A to B and B to A, plus the conversion between one of these tagsets and any other tagset that has been defined so far.

Drivers are implemented as Perl modules.
A simple conversion script could look like this:

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





Partial trie of feature value combinations permitted in Swedish sv::hajic tagset. **Bold** arrows show its traversal during encoding of the Czech tag `NNMS1-----A----` (**purple** arrows indicate replacing of a forbidden value by a default).

Difficult phenomena

Endemic word classes are roofed by more common parts of speech. It makes encoding safer when the target tagset does not know the word class.

Nouns (Common Nouns, Proper Nouns, Substantive **Pronouns** (Personal, Demonstrative, Interrogative, Relative, Indefinite, Negative))
Adjectives (Adjectives, **Determiners** (Articles), Predeterminers, Attributive Pronouns (Possessive, Demonstrative, Interrogative, Relative, Indefinite, Negative), Attributive Numerals (Ordinals))
Particles (all kinds of stuff, see separate frame)

The prontype feature distinguishes pronouns from nouns, determinative pronouns from adjectives, WH and indefinite from concrete adverbs etc.

Participles: in some tagsets verbs, in others adjectives

Numerals: cardinal numbers usually separate; ordinals classified as adjectives in some tagsets; various other kinds of numerals (multiplicative, generic, indefinite)

Different approaches in tagsets could lead to different feature values!

We *recommend* that participles are verbs (in other words, you should not set verbform if you do not set pos = "verb"). However, we do not prohibit any such combination of feature values.

Particles

unclassified particle (Czech `TT`, English `RP`, Swedish `Q--------`)
interrogative particle (Arabic `FI` هل *hal*, Bulgarian `Tn` *ли li*)
affirmative particle (Bulgarian `Ta` *да da*)
negative particle (Arabic `FN` لا *lā*, Bulgarian `Tn` *не ne*, German `PTKNEG` *nicht*)
response particle (German `PTKANT` *ja* = "yes", *nein* = "no", *doch* = "yes", *danke* = "thank you"…)
auxiliary particle (Bulgarian `Tx` *да da* = "to", *ще šte* = "will")
modal particle (Bulgarian `Tm` *май maj* = "possibly")
verbal particle (Bulgarian `Tv` *нека neka* = "let")
emphasis particle (Bulgarian `Te` *даже daže* = "even")
gradable particle (Bulgarian `Tg` *най naj* = "most")
unique POS (Danish `U`, covering the words *at* = infinitival "to", *som*, *der*)
infinitive mark (German `PTKZU` *zu*, Swedish `IM` *att*, English `TO` *to* – includes prepositional occurrences of *to*)
separated verbal prefix (German `PTKVZ`, *vor* in *stellen Sie sich vor*)
adjectival particle (German `PTKA`, *am* in *am besten*, *zu* in *zu groß*)
existential *there* in English (`EX`)
measure word, quantifier (Chinese `DM`)
genitive particle *de* in Chinese (`DE` 的 and 得)
Chinese particles 了 *le* (perfect), 著 *zhe*, 起 *qǐ*, 過 *guò* (`Di`)
Chinese particles 了 *le*, 的 *de*, 來 *lái* (`Ta`)
Chinese particles 而已 *éryǐ*, 沒有 *méiyǒu*, 也罷 *yěba*, 沒有 *méiyǒu*, 好了 *hǎole* (`Tb`)
Chinese particles 呢 *ne*, 吧 *ba*, 啊 *a*, 囉 *luō* (`Tc`)
Chinese particles 嗎 *ma*, 否 *fǒu* (`Td`)

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“]] }*											

List of features and values currently used in the universal feature set ("Interset").

Tagset/Driver	Number of tags	Other feature	Approximate implementation time
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

Table 1: Overview of tagset drivers. The "other" column shows tags that make the decoder set the "other" feature.

	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: Number of tags resulting from conversion from drivers named in row headers to drivers named in column headers.

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: Accuracy of the parser on various CoNLL data sets, using original and converted tags. The last column indicates whether the change was statistically significant, using the McNemar's test with $p \leq 0.05$.