Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa

March 20, 2020
Part-of-Speech Tagset Conversion

- See also NPFL094 (Computational Morphology and Syntax) in Winter
- There: focus on linguistic diversity
- Here: focus on
  - Technical aspects
  - Different expressivity
  - Different granularity
Why Convert Tags?

- For a tool that uses tags (parser)
  - The meaning of the tags is significant (they are not just strings)
  - Or the tool has been trained on a particular tagset

- For a linguist who works with corpora
  - Reduce need to learn new tags
How to Convert Tags?

- Look at source tags only
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (Interset, Universal Dependencies)
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (Interset, Universal Dependencies)

- Look at source tags + words
How to Convert Tags?

- Look at source tags only
  - Conversion tailored to a pair of tagsets
  - Reusable “interlingua” (Interset, Universal Dependencies)

- Look at source tags + words

- Look at source tags + words + context
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

In alignment with this perspective, at IIIT Hyderabad, all Indian languages, including Indo-Aryan, Dravidian, and English, are considered. A notable resource is the Gold Ontology, which defines linguistic terms and acknowledges that the same term may denote different things in different languages. Other frameworks, such as Interset, Google UPOS, and Universal Dependencies, have contributed to the discussion of whether a universal tagset exists. These resources and frameworks demonstrate the complexity and diversity in linguistic research and development.
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- IIIT Hyderabad: all Indian languages
  - Indo-Aryan
  - Dravidian
  - English!

- Gold Ontology
  - Defines linguistic terms
  - The same term may denote different things in different languages

- Interset, Google UPOS, Universal Dependencies
  - Papers claiming that universal tagset does not exist
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- IIIT Hyderabad: all Indian languages
  - Indo-Aryan
  - Dravidian
  - English!

- Gold Ontology
  - Defines linguistic terms
  - The same term may denote different things in different languages
Related Work

- **EAGLES, PAROLE, MULTEXT**
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- **IIIT Hyderabad: all Indian languages**
  - Indo-Aryan
  - Dravidian
  - English!

- **Gold Ontology**
  - Defines linguistic terms
  - The same term may denote different things in different languages

- **Interset, Google UPOS, Universal Dependencies**
Related Work

- EAGLES, PAROLE, MULTEXT
  - Rather wanted to standardize tags
  - Not to work with the tags that are already there
  - Very euro-centric

- IIIT Hyderabad: all Indian languages
  - Indo-Aryan
  - Dravidian
  - English!

- Gold Ontology
  - Defines linguistic terms
  - The same term may denote different things in different languages

- Interset, Google UPOS, Universal Dependencies

- Papers claiming that universal tagset does not exist
Josef následující jejímuž stě jsem nejméně v aby jen ejhle noor ,
<table>
<thead>
<tr>
<th>Tag</th>
<th>Conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1----A----</td>
<td>NMS1A</td>
</tr>
<tr>
<td>AGFS3----A----</td>
<td>AVGFS3A</td>
</tr>
<tr>
<td>P1ZS3FS3------</td>
<td>PSEFSZS3</td>
</tr>
<tr>
<td>ClXP3--------2</td>
<td>CGXP3-2</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
<td>VPS1A</td>
</tr>
<tr>
<td>Dg--------3A--</td>
<td>DG3A</td>
</tr>
<tr>
<td>RR--6----------</td>
<td>R6</td>
</tr>
<tr>
<td>J,-X---3------</td>
<td>JVX3</td>
</tr>
<tr>
<td>TT------------</td>
<td>T</td>
</tr>
<tr>
<td>II------------</td>
<td>I</td>
</tr>
<tr>
<td>X@------------</td>
<td>NOMORPH</td>
</tr>
<tr>
<td>Z:------------</td>
<td>ZIP</td>
</tr>
</tbody>
</table>
Prague Tags for CoNLL 2006 Shared Task

NNMS1------A----- N N Gen=M|Num=S|Cas=1...
AGFS3------A----- A G Gen=F|Num=S|Cas=3...
P1ZS3FS3------- P 1 Gen=Z|Num=S|Cas=3...
ClXP3---------2 C 1 Gen=X|Num=P|Cas=3...
VB-S----1P-AA--- V B Num=S|Per=1|Ten=P...
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 : _
NNMS1-----A----- Ncmsny
AGFS3-----A----- Afpfsd
P1ZS3FS3------- Pr3mdsfnayn
ClXP3--------2 Mcmn3y
VB-S---1P-AA--- Vmip1smayn
Dg---------3A---- Rgs
RR--6-------- Sps1
J,-X---3------- Css3
TT------------- Q
II------------- I
X©------------- X
Z:-------------
Majka Tagset from Brno

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:------------

Interset: Reusable Tagset Conversion
Penn Treebank Tags for English

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…
Brown Corpus Tags for English

SynTagRus Tags for Russian

S ЕД МУЖ ИМ
S МН РОД ОД
A МН ИМ
NUM ВИН
V НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л
ADV СПАВ
PR
CONJ
PART
INTJ

NNMS1------A-----
PSXXXXP3--------
AAXP1------1A-----
CXXX4----------
VB-P---3P-AA---
Dg---------2A-----
RR--6---------
J~-------------
TT-------------
II-------------
Like in Penn TB: parts of speech only, but slightly more fine-grained

No morphology (German has gender, number, case, degree, person...)

“Substantive” vs. “attributive” pronouns (S vs. AT)

Adposition = Präposition, Postposition, Zirkumposition
• Ambition: common tagset for all Indian languages (IE and Dravidian!)
• No morphology (although the languages are rich on morphology)
  ▪ Hierarchical tagset, morphology can be added at the end
  ▪ And they “do not want to decrease tagging accuracy” (!)
• Cloned from Penn tagset and modified
  ▪ New categories, e.g. postposition, “quotative”
  ▪ Removed traces of morphology, e.g. plural, comparative, superlative
Tagging is interwined with tokenization.

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

<token_Arabic>
  <voc>mivAlu</voc>
  <pos>mivAl/NOUN+u/CASE_DEF_NOM</pos>
</token_Arabic>
Interset: Reusable Tagset Conversion
<table>
<thead>
<tr>
<th>Code</th>
<th>Chinese Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>common noun</td>
</tr>
<tr>
<td>Nb</td>
<td>proper noun</td>
</tr>
<tr>
<td>Nc</td>
<td>location noun</td>
</tr>
<tr>
<td>Nd</td>
<td>time noun</td>
</tr>
<tr>
<td>Nf</td>
<td>classifier</td>
</tr>
<tr>
<td>Nh</td>
<td>pronoun</td>
</tr>
<tr>
<td>Ne</td>
<td>determiner or cardinal number</td>
</tr>
<tr>
<td>Ng</td>
<td>postposition</td>
</tr>
<tr>
<td>P</td>
<td>preposition</td>
</tr>
<tr>
<td>P01</td>
<td>為 wèi, 承蒙 chéngméng, 深為 shēnwèi</td>
</tr>
<tr>
<td>P02</td>
<td>被 bèi</td>
</tr>
<tr>
<td>P03</td>
<td>為了 wèile, 為 wèi</td>
</tr>
<tr>
<td>P04</td>
<td>給 gěi</td>
</tr>
<tr>
<td>P06</td>
<td>由 yóu</td>
</tr>
<tr>
<td>P07</td>
<td>把 bǎ, 將 jiāng</td>
</tr>
</tbody>
</table>
PAROLE Danish and Swedish

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
MC00G0S ... fyras (gt. gen.)
V@IPAS ... har
AP000N0S ... Oberoende
RGOS ... inte
PF@000@S ... sig
<table>
<thead>
<tr>
<th>Tagset</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>noun</td>
<td>konflikterna</td>
</tr>
<tr>
<td>PN</td>
<td>proper noun</td>
<td>(substantiv utrum pluralis bestämd nominativ)</td>
</tr>
<tr>
<td>VN</td>
<td>gerund</td>
<td></td>
</tr>
<tr>
<td>AJ</td>
<td>adjective</td>
<td>politiska</td>
</tr>
<tr>
<td>AV BV FV GV HV KV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MV QV SP SV VV WV</td>
<td>verbs</td>
<td></td>
</tr>
<tr>
<td>HV</td>
<td>the verb</td>
<td>har</td>
</tr>
<tr>
<td>I? IC IG IK IP IQ</td>
<td>punctuation</td>
<td></td>
</tr>
<tr>
<td>IR IS IT IU</td>
<td>punctuation</td>
<td>oberoende</td>
</tr>
<tr>
<td>RG0S</td>
<td></td>
<td>inte</td>
</tr>
<tr>
<td>PF@000@S</td>
<td></td>
<td>sig</td>
</tr>
<tr>
<td>pos</td>
<td>noun</td>
<td>adj</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>-----</td>
</tr>
</tbody>
</table>
| subpos   | prop | class | pdt | det | art | digit | roman | card | ord | ...
| prontype | prs  | rep | int | rel | dem | neg | ind | tot |
| punctype | peri | gest | 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 |
| posnumber | sing | dual | plu |
| subcat   | intr | tran |
| verbf orm | fin | inf | sup | part | trans | ger |
| mood     | ind | imp | end | sub | jus |
| tense    | past | pres | fut |
| sub tense | 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 |

Interset: Reusable Tagset Conversion
- Reference:

- CPAN Perl libraries:
  - cpanm install Lingua::Interset

```perl
use Lingua::Interset::Converter;
my $c = new Lingua::Interset::Converter ('from' => 'cs::multext', 'to' => 'cs::pdt');
...
my $target_tag = $c->convert ($source_tag);
```
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature
- Problem: multiple combinations of values
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature

- Problem: multiple combinations of values
  - EITHER (gender = feminine AND number = singular)
  - OR (gender = neuter AND number = plural)
  - BUT NOT (feminine plural or neuter singular)

Interset cannot represent this (only two separate feature structures)
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature

- Problem: multiple combinations of values
  - EITHER (gender = feminine AND number = singular)
  - OR (gender = neuter AND number = plural)
Multi-values (disjunction)

- A tag may say: gender = masculine OR neuter
- Interset allows multiple values of a feature

- Problem: multiple combinations of values
  - EITHER (gender = feminine AND number = singular)
  - OR (gender = neuter AND number = plural)
  - BUT NOT (feminine plural or neuter singular)

- Interset cannot represent this (only two separate feature structures)
Tagset Drivers

- A (Perl) module with the following functions:
  - `decode()` ... converts a tag to Interset
  - `encode()` ... generates a tag from Interset
  - `list()` ... lists known tags in the tagset (optional)
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com/neut
  - and also: personal pronoun, gender=masc/fem/com/neut
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
  - Either change noun to pronoun
  - or change gender=masc to gender=com
Not Everything Fits in the Target Tagset

- Throw away information that cannot be represented
- Warning! May generate “unexpected” tag
  - Swedish knows: noun, gender=com|neut
  - and also: personal pronoun, gender=masc|fem|com|neut
  - From Czech: noun, gender=masc
  - Either change noun to pronoun
  - or change gender=masc to gender=com
  - What has higher priority?
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb
- What is our goal?
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb

- What is our goal?
- Just querying attributes? ⇒ Preserve as much info as possible!
Does It Matter?

- Atomic tagsets (Penn): no choice
- Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb

- What is our goal?
- Just querying attributes? ⇒ Preserve as much info as possible!
- Use a pre-trained black-box tool? ⇒ Don’t give it data that it doesn’t expect!
Enforcing Defaults

- Need the list of known target tags

- Centrally for all tagsets:
  - Priorities of features
  - For every feature value, ordered list of substitutes
    - Typically, empty value is the best substitute
    - But: number = dual is better substituted by plural!

```
[ { "num": 0 } ],
"number" => [
  [ ["sing"],
  ["dual", "plur"],
  ["plur"]
  ],
"possnumber" =>
[ ["sing", "dual", "plur"]
```

0 → sing, dual, plur; sing → 0, dual, plur; dual → plur, 0
Enforcing Defaults

- Decode all known target tags
- Construct trie for known feature-value combinations
- Follow path in trie when encoding
- If a value is not allowed, find the best substitute

(It is more complex when multi-values come into play.)
Substitution Trie

Interset: Reusable Tagset Conversion
Substitution Trie

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

prontype

NNMS1-----

pos
noun
polarity
pos

Interset: Reusable Tagset Conversion
Substitution Trie

- **pos**: noun, adj, num, verb, adv, adp, conj, part, int, punc
- **pronotype**: prs, int, ind
- **definiteness**: ind, def

Interset: Reusable Tagset Conversion
Substitution Trie

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

prontype
prs
int
ind

definiteness
ind
def

gender
com
neut

number
sing
plur

case	nom
gen

NNMS
pos
noun

polarity
pos

gender
masc

animacy
anim

number
sing

case
nom

Interset: Reusable Tagset Conversion
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough

- Good idea
- But it must be applied well!
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough

- Good idea
- But it must be applied well!

- pronoun $\rightarrow$ PRON
  - determiners, numerals, adverbs
Google Universal Part-of-Speech Tags

- Just the POS category. No morphology
- For many tools this is enough

- Good idea
- But it must be applied well!

- pronoun → PRON
  - determiners, numerals, adverbs
- similar for numerals in Danish
- similar for nominal/adjectival verb forms
Lemma-based Re-tagging

```perl
my $lemma = $node->lemma();
# Fix Interset features of pronominal words.
if($node->is_pronominal())
{
    # Indefinite pronouns and determiners cannot be distinguished by their PDT tag (PZ
    if($lemma =~ m/((ně|lec|ledas?|kde|bůhví|kdoví|nevím|málo|sotva)?(kdo|cos?)(si|ko
    
    $node->iset())->set('pos', 'noun');
}
elsif($lemma =~ m/((jaký|který)|(jaký|který)$|^((každý|všechnen|sám|žádný|some|taký
    
    $node->iset())->set('pos', 'adj');
}
# Pronouns čí, něčí, čísí, číkoli, ledačí, kdečí, bůhvíčí, nevímčí, nicí should have
elsif($lemma =~ m/((ně|lec|ledas?|kde|bůhví|kdoví|nevím|ni)?čí|čí(si|koliv?))/$)
{
    $node->iset())->set('pos', 'adj');
    $node->iset())->set('poss', 'poss');
```
Universal Dependencies: UPOS and Features

- **UPOS** = extended version of Google universal tags
- **Features** = extended Interset
  - (now it is the target representation rather than something intermediate)
    - “Universal” feature + set of values
    - Language-specific value of universal feature
    - Language-specific (or treebank-specific) feature + set of values
A Grain of Salt: Even UD Can Be Used Inconsistently!

- https://lindat.mff.cuni.cz/services/pmltq/
  - Find two UD treebanks of related languages
  - Where the “same word” does not get the same UPOS category
A Grain of Salt: Even UD Can Be Used Inconsistently!

- https://lindat.mff.cuni.cz/services/pmltq/
  - Find two UD treebanks of related languages
  - Where the “same word” does not get the same UPOS category