Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa

March 17, 2022
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
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!
- Interset: Reusable Tagset Conversion
  - 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
Prague Tags for Czech

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:------------ ,
Prague Tags for Czech

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
<table>
<thead>
<tr>
<th>Tag</th>
<th>Value</th>
<th>Gen</th>
<th>Num</th>
<th>Cas</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1</td>
<td>------</td>
<td>A</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>AGFS3</td>
<td>------</td>
<td>A</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>P1ZS3FS3</td>
<td>------</td>
<td>P 1</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>ClXP3</td>
<td>-------</td>
<td>2</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>VB-S</td>
<td>---1P-AA---</td>
<td>V  B</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Dg</td>
<td>-------</td>
<td>3A</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>RR</td>
<td>--6------</td>
<td>R  R</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>J</td>
<td>-X---3------</td>
<td>J ,</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>TT</td>
<td>---------</td>
<td>T  T</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>II</td>
<td>---------</td>
<td>I  I</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>X@</td>
<td>----------</td>
<td>X  @</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Z:</td>
<td>-----------</td>
<td>Z  :</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
NNMS1------A-----  Ncmsny
AGFS3------A-----  Afpfsd
P1ZS3FS3--------  Pr3mdsfnayn
ClXP3----------2  Mcmn3y
VB-S---1P-AA---  Vmip1smanyn
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:-----------
Penn Treebank Tags for English

- **CC** = Coordinating conjunction
- **CD** = detached colon or dash
- **DT** = determiner
- **EX** = existential “there”
- **FW** = foreign word
- **IN** = preposition or subordinating conjunction
- **TO** = “to”
- **UH** = interjection…

- **RBR** = adjective comparative
- **RBS** = adverbial comparative
- **RP** = possessive ending
- **SYM** = symbol
- **TO** = “to”
- **UH** = interjection…
- **VB** = verb base form
- **VBD** = verb past tense
- **VBG** = verb gerund
- **VBN** = verb past participle
- **VBP** = verb present participle
- **VBZ** = verb 3rd person singular present
- **WDT** = wh-determiner
- **WP** = wh-pronoun
- **WP$** = wh-pronoun that allows possessive
- **WRB** = wh-adverb
- **. , : $ # ` ” ' ” ) ( -LRB- -RRB- }
Brown Corpus Tags for English

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S</strong></td>
<td>ЕД МУЖ ИМ</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>МН РОД ОД</td>
</tr>
<tr>
<td><strong>A</strong></td>
<td>МН ИМ</td>
</tr>
<tr>
<td><strong>NUM</strong></td>
<td>ВИН</td>
</tr>
<tr>
<td><strong>V</strong></td>
<td>НЕСОВ ИЗЪЯВ НЕПРОШ МН 3-Л</td>
</tr>
<tr>
<td><strong>ADV</strong></td>
<td>СПАВ</td>
</tr>
<tr>
<td><strong>PR</strong></td>
<td></td>
</tr>
<tr>
<td><strong>CONJ</strong></td>
<td></td>
</tr>
<tr>
<td><strong>PART</strong></td>
<td></td>
</tr>
<tr>
<td><strong>INTJ</strong></td>
<td></td>
</tr>
<tr>
<td>NNMS1</td>
<td>------A------</td>
</tr>
<tr>
<td>PSXXXP3</td>
<td>--------</td>
</tr>
<tr>
<td>AAXP1</td>
<td>-----1A-----</td>
</tr>
<tr>
<td>C1XX4</td>
<td>----------</td>
</tr>
<tr>
<td>VB-P</td>
<td>---3P- AA------</td>
</tr>
<tr>
<td>Dg</td>
<td>------2A------</td>
</tr>
<tr>
<td>RR</td>
<td>--6----------</td>
</tr>
<tr>
<td>J^</td>
<td>----------------</td>
</tr>
<tr>
<td>TT</td>
<td>----------------</td>
</tr>
<tr>
<td>II</td>
<td>----------------</td>
</tr>
</tbody>
</table>
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 intertwined 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>
N-------1D    NNXX1------A-----
Z-------1-    NNXX1------A-----
A------FP2D   AAFP2----1A-----
S------3MP1-  PPMP1--3--------
VIS--------   VcXX---XP-AA----
Rocling / Sinica Tagset for Chinese

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
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
RG0S ... inte
PF@000@S ... sig
<table>
<thead>
<tr>
<th>Tagset</th>
<th>Swedish Language</th>
<th>Tagset</th>
<th>Swedish Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>noun</td>
<td>NCUPN@DS</td>
<td>konflikterna</td>
</tr>
<tr>
<td>PN</td>
<td>proper noun</td>
<td>(substantiv utrum pluralis bestämd nominativ)</td>
<td></td>
</tr>
<tr>
<td>VN</td>
<td>gerund</td>
<td>AQP0PN0S</td>
<td>politiska</td>
</tr>
<tr>
<td>AV BV</td>
<td>FV GV HV KV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MV QV</td>
<td>SP SV VV WV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HV</td>
<td>the verb hava</td>
<td>V@IPAS</td>
<td>har</td>
</tr>
<tr>
<td>I? IC</td>
<td>IG IK IP IQ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IR IS</td>
<td>IT IU</td>
<td>AP000N0S</td>
<td>oberoende</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RG0S</td>
<td>inte</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PF@000@S</td>
<td>sig</td>
</tr>
<tr>
<td>pos</td>
<td>noun</td>
<td>adj</td>
<td>num</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td>subpos</td>
<td>prop</td>
<td>class</td>
<td>pdt</td>
</tr>
<tr>
<td>pronltype</td>
<td>prs</td>
<td>rep</td>
<td>int</td>
</tr>
<tr>
<td>punctype</td>
<td>peri</td>
<td>gest</td>
<td>excl</td>
</tr>
<tr>
<td>puncside</td>
<td>ini</td>
<td>fin</td>
<td></td>
</tr>
<tr>
<td>synpos</td>
<td>subst</td>
<td>attr</td>
<td>adv</td>
</tr>
<tr>
<td>poss</td>
<td>poss</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reflex</td>
<td>reflex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>negativeness</td>
<td>pos</td>
<td>neg</td>
<td></td>
</tr>
<tr>
<td>definiteness</td>
<td>ind</td>
<td>def</td>
<td>red</td>
</tr>
<tr>
<td>gender</td>
<td>masc</td>
<td>fem</td>
<td>com</td>
</tr>
<tr>
<td>animateness</td>
<td>anim</td>
<td>inan</td>
<td></td>
</tr>
<tr>
<td>number</td>
<td>sing</td>
<td>dual</td>
<td>plu</td>
</tr>
<tr>
<td>case</td>
<td>nom</td>
<td>gen</td>
<td>dat</td>
</tr>
<tr>
<td>prepcase</td>
<td>npr</td>
<td>pre</td>
<td></td>
</tr>
<tr>
<td>degree</td>
<td>pos</td>
<td>com</td>
<td>sup</td>
</tr>
<tr>
<td>person</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>politeness</td>
<td>inf</td>
<td>pol</td>
<td></td>
</tr>
<tr>
<td>possgender</td>
<td>masc</td>
<td>fem</td>
<td>com</td>
</tr>
<tr>
<td>posnumber</td>
<td>sing</td>
<td>dual</td>
<td>plu</td>
</tr>
<tr>
<td>subcat</td>
<td>intr</td>
<td>tran</td>
<td></td>
</tr>
<tr>
<td>verbform</td>
<td>fin</td>
<td>inf</td>
<td>sup</td>
</tr>
<tr>
<td>mood</td>
<td>ind</td>
<td>imp</td>
<td>end</td>
</tr>
<tr>
<td>tense</td>
<td>past</td>
<td>pres</td>
<td>fut</td>
</tr>
<tr>
<td>sub tense</td>
<td>aor</td>
<td>imp</td>
<td>ppq</td>
</tr>
<tr>
<td>aspect</td>
<td>imp</td>
<td>perf</td>
<td></td>
</tr>
<tr>
<td>voice</td>
<td>act</td>
<td>pass</td>
<td></td>
</tr>
<tr>
<td>foreign</td>
<td>foreign</td>
<td></td>
<td></td>
</tr>
<tr>
<td>abbr</td>
<td>abbr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hyph</td>
<td>hyph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>style</td>
<td>arch</td>
<td>form</td>
<td>norm</td>
</tr>
<tr>
<td>typo</td>
<td>typo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>variant</td>
<td>short</td>
<td>long</td>
<td>0</td>
</tr>
</tbody>
</table>
**Interset**

- 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);
```
- 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
  • 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?
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
• 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!

```python
'number' =>
{
  'priority' => 320,
  'values' => ['sing', 'dual', 'tri', 'pauc', 'grpa', 'plur'],
  'replacements' => [
    ['sing'],
    ['dual', 'plur'],
    ['tri', 'plur'],
    ['pauc', 'plur'],
    ['grpa', 'plur'],
    ['plur'],
    ['grpl', 'plur'],
    ['inv'],
    ['ptan', 'plur'],
    ['coll', 'sing'],
    ['plur']
  ]
}
```

0 → sing, dual, tri, pauc, …
sing → 0, dual, tri, pauc, …
dual → plur, 0, sing, tri, …
tri → plur, 0, sing, dual, …
pauc → plur, 0, sing, …
grpa → plur, 0, sing, …
plur → 0, sing, dual, tri, …
grpl → plur, 0, sing, …
inv → 0, sing, dual, tri, …
ptan → plur, 0, sing, …
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

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

prontype
prs
int
ind

definiteness
ind
def
def
ind

NNMS1-----A-----
pos
noun

polarity
pos

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
masc
com
neut
com
neut
com
neut
com
neut

NN
S1------A----
pos
noun
polarity
pos
gender
masc
animacy
anim

Interset: Reusable Tagset Conversion
Substitution Trie

pos
noun
adj
num
verb
adv
adp
conj
part
int
punc

pronotype
prs
int
ind

definiteness
ind
def

gender
com
neut

number
sing
plur

NNMS1

pos
noun

polarity

gender
masc

animacy
anim

number
sing

Intersect: 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!
• 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
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
  }
  $node->iset()->set('pos', 'adj');
}
# Pronouns čí, něčí, čísí, číkoli, ledačí, kdečí, bůhvící, nevímčí, níčí should hav
elsif ($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
• 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
• 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

• http://quest.ms.mff.cuni.cz/cgi-bin/interset/index.pl?tagset=pt::freeling