Interset: Reusable Tagset Conversion

Daniel Zeman, Rudolf Rosa

March 26, 2021
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
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!
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>Prague Tags for Czech</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1-----A-----</td>
</tr>
<tr>
<td>AGFS3-----A-----</td>
</tr>
<tr>
<td>P1ZS3FS3-------</td>
</tr>
<tr>
<td>ClXP3---------2</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
</tr>
<tr>
<td>Dg--------3A-----</td>
</tr>
<tr>
<td>RR--6------------</td>
</tr>
<tr>
<td>J,-X---3---------</td>
</tr>
<tr>
<td>TT-----------------</td>
</tr>
<tr>
<td>II-----------------</td>
</tr>
<tr>
<td>X@-----------------</td>
</tr>
<tr>
<td>Z:-----------------</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------------  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 : _
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNMS1----A----</td>
<td>Ncmsny</td>
</tr>
<tr>
<td>AGFS3-----A------</td>
<td>Afpfsd</td>
</tr>
<tr>
<td>P1ZS3FS3-------</td>
<td>Pr3mdsfnayn</td>
</tr>
<tr>
<td>ClXP3----------2</td>
<td>Mcmn3y</td>
</tr>
<tr>
<td>VB-S---1P-AA---</td>
<td>Vmip1smanyn</td>
</tr>
<tr>
<td>Dg-------3A-----</td>
<td>Rgs</td>
</tr>
<tr>
<td>RR--6---------</td>
<td>Sps1</td>
</tr>
<tr>
<td>J,-X---3-------</td>
<td>Css3</td>
</tr>
<tr>
<td>TT------------</td>
<td>Q</td>
</tr>
<tr>
<td>II-------------</td>
<td>I</td>
</tr>
<tr>
<td>X©-------------</td>
<td>X</td>
</tr>
<tr>
<td>Z:------------</td>
<td></td>
</tr>
</tbody>
</table>
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

- 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  СПАБ
PR
CONJ
PART
INTJ
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>
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
...
P66 = 為 wèi
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

RG0S ... inte

PF©000©S ... sig
<table>
<thead>
<tr>
<th>Code</th>
<th>Meaning</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>politiska</td>
</tr>
<tr>
<td>AJ</td>
<td>adjective</td>
<td></td>
</tr>
<tr>
<td>AV</td>
<td>verbs</td>
<td></td>
</tr>
<tr>
<td>HV</td>
<td>the verb hava</td>
<td>har</td>
</tr>
<tr>
<td>IR</td>
<td>punctuation</td>
<td>oberoende</td>
</tr>
<tr>
<td>RGOS</td>
<td></td>
<td>inte</td>
</tr>
<tr>
<td>PFOS</td>
<td></td>
<td>sig</td>
</tr>
</tbody>
</table>
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);
```
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
  - 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
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?
• 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!
• Atomic tagsets (Penn): no choice
• Positional tagsets can encode “impossible” combinations, e.g. a plural accusative adverb

• What is our goal?
  • Just querying attributes? \(\Rightarrow\) Preserve as much info as possible!
  • Use a pre-trained black-box tool? \(\Rightarrow\) 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'],
    ['count', '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

pronotype
- prs
- int
- ind

NNM1-----A-----
pos noun
Substitution Trie

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

**NNMS1-----**
- **pos**: noun
- **polarity**: pos
Substitution Trie

- **pos**
  - noun
  - adj
  - num
  - verb
  - adv
  - adp
  - conj
  - part
  - int
  - punc

- **prontype**
  - prs
  - int
  - ind

- **definiteness**
  - ind
  - def

- **gender**
  - com
  - neut

- **NN**
  - S1

- **pos**
  - noun

- **polarity**
  - pos

- **gender**
  - masc

- **animacy**
  - anim
Google Universal Part-of-Speech Tags

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 → 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|^všechen|^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