Cross-Language Parser Adaptation

Dan Zeman

Philip Resnik

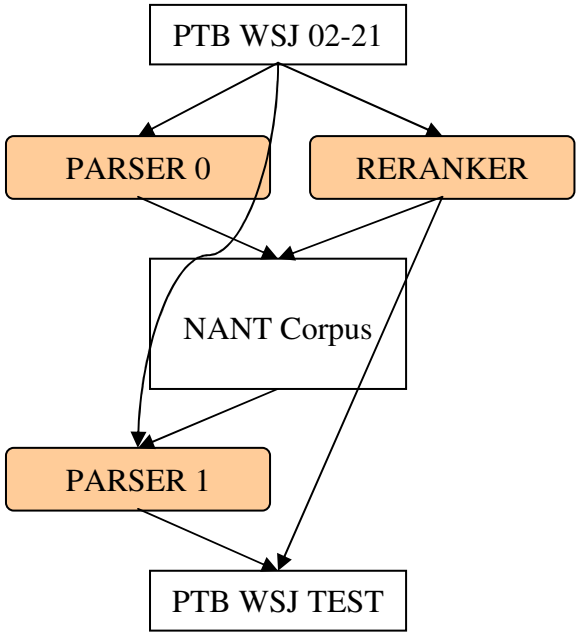
Parser adaptation

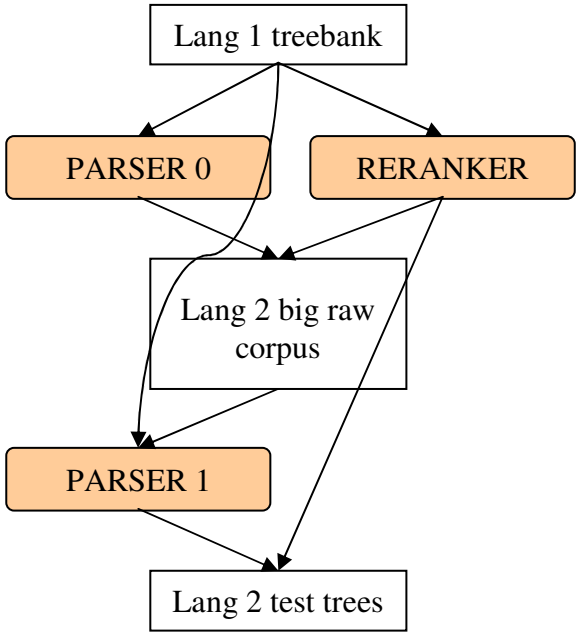
- Idea:
  - Related languages L1 and L2
  - Have L1 treebank and morphology
  - Only morphology for L2
  - Train parser on L1 morph features, apply it to L2

Short detour: self-training

- McClosky, Charniak & Johnson 2006
- Idea:
  - Charniak parser + Johnson reranker (2005)
  - Train both on Penn WSJ
  - Run both over **HUGE** corpus
  - Train parser (not reranker) on parsed HUGE
  - Voilà, new parser is better on Penn WSJ!

Self-training vs. adaptation





What do we need?

- N-best parser (Stanford, Charniak) & reranker (Johnson)
- Two related languages L1 and L2
- Treebank of L1
- Big raw corpus of L2
- Morphological analyzer of L2
- Small test treebank of L2

What do we need?

- N-best parser (Stanford, Charniak) & reranker (Johnson)
- Two related languages L1 and L2
- Treebank of L1
- Big raw corpus of L2
- Morphological analyzer of L2
- Small test treebank of L2
- There are free dependency treebanks from CoNLL 2006 shared task: bg, da, ja, nl, pt, sl, sv

What do we need?

- N-best parser (Stanford, Charniak) & reranker (Johnson)
- Two related languages L1 and L2
- Treebank of L1
- Big raw corpus of L2
- Morphological analyzer of L2
- Small test treebank of L2
- The Europarl / Acquis corpus: EU proceedings, large parallel text in 21 European languages

What do we need?

- N-best parser (Stanford, Charniak) & reranker (Johnson)
- Two related languages DA and SV
- Treebank of L1: Danish Dependency Treebank
- Big raw corpus of L2: Swedish Acquis
- Swedish tagger by Jan Hajič
- Small test treebank of L2: Talbanken05

Most frequent da / sv words

• i	0.024	• och	0.027
• og	0.024	• att	0.027
• at	0.021	• i	0.021
• er	0.017	• är	0.018
• en	0.014	• som	0.017
• til	0.013	• en	0.015
• af	0.013	• det	0.013
• det	0.012	• av	0.012
• på	0.012	• på	0.011

Aligned sentences example 1

- Denne forordning træder i kraft den 1. marts 1986 med forbehold af ikrafttrædelse af traktaten vedrørende Spaniens og Portugals tiltrædelse.

- Denna förordning träder i kraft den 1 mars 1986 under förutsättning att Anslutningsakten för Spanien och Portugal träder i kraft.

Aligned sentences example 2

- Bestemmelserne i denne aftale kan ændres og revideres helt eller delvis efter fælles overenskomst mellem parterne.

- Bestämmelserna i detta avtal får ändras eller revideras helt eller delvis efter gemensam överenskommelse mellan parterna.

Aligned sentences example 3

- 1. Enhver kontraherende part kan opsige denne konvention ved skriftlig henvendelse til depositaren.

- 1. En fördragsslutande part fåt säga upp denna konvention genom skriftlig notifikation till depositarien.

Treebank preparation

- Dependencies to constituents: flattest structure possible:
  - loves/V (John/N, Mary/N)
  - (VP (N John) (V loves) (N Mary))

  ≠ (S (NP (N John)) (VP (V loves) (NP (N Mary))))

Treebank preparation

- Dependencies to constituents: flattest structure possible.
- DA/SV tags converted to Penn tag set
- Nonterminal labels:
  - derived from POS tags
  - than translated to Penn nonterminal set
- Goal: make the parser feel it works with Penn Treebank (rather than teaching it new tables)

Treebank “normalization”

Danish

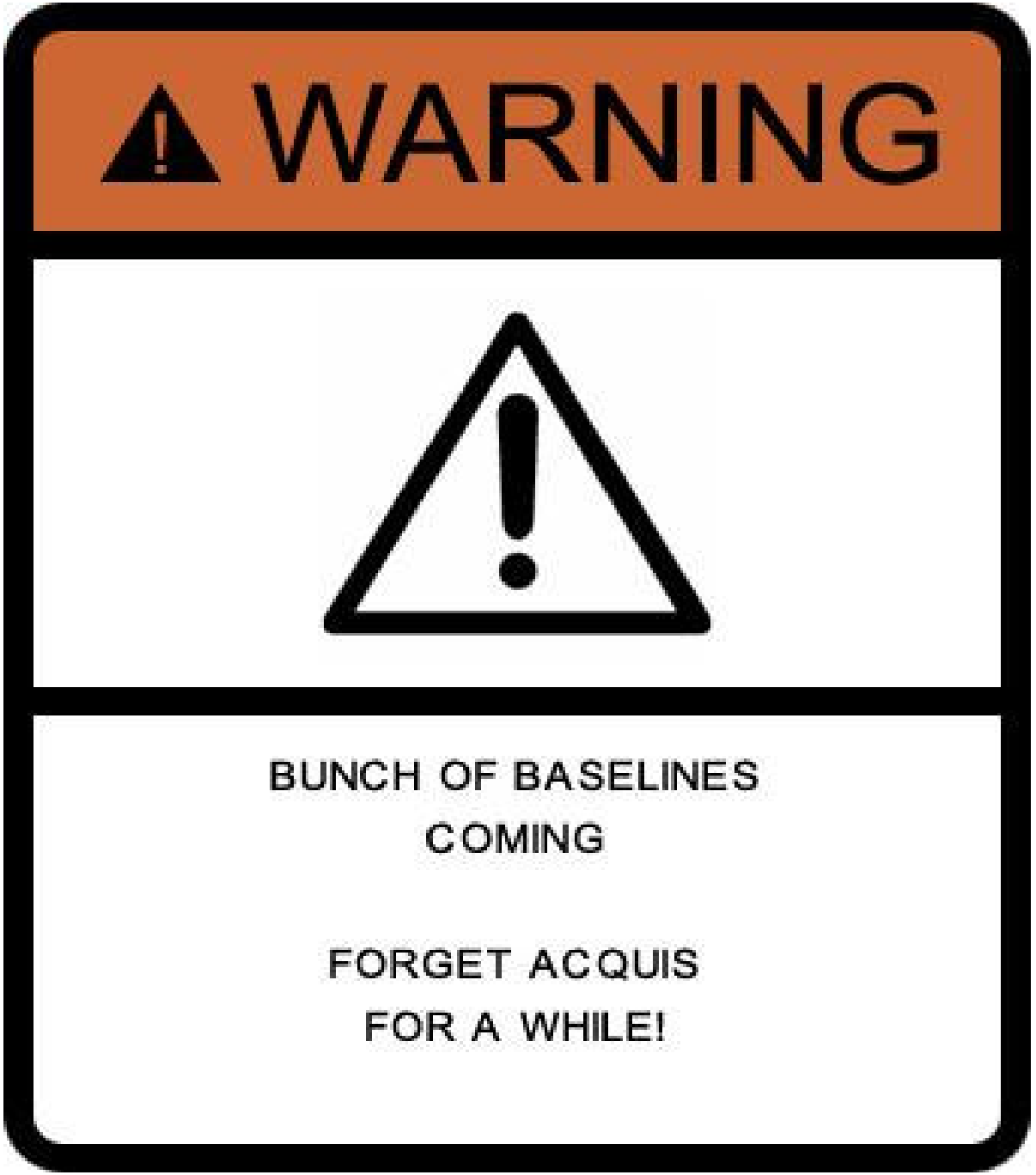
- DET governs ADJ, ADJ governs NOUN
- NUM governs NOUN
- GEN governs NOM
  Ruslands vej
  Russia’s way
- COORD: last member on conjunction, everything else on first member

Swedish

- NOUN governs both DET and ADJ
- NOUN governs NUM
- NOM governs GEN
  års inkomster
  year’s income
- COORD: member on previous member, commas and conjs on next member

Standard evaluation method

- labeled constituent precision P
- labeled constituent recall R
- F = 2PR/(P+R)
- punctuation not evaluated
- sentences over 40 words not evaluated!


WARNING
BUNCH OF BASELINES
COMING
FORGET ACQUIS
FOR A WHILE!

Parsing Danish treebank

- CoNLL test: 322 sents, 5852 words
- CoNLL training: 5190 sents, 94386 words
  - 4900 sents my training
  - 290 sents my devtest
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	78.16
Brown	78.24

Parsing Swedish treebank

- CoNLL test: 389 sents, 5656 words
- CoNLL training: 11042 sents, 191467 words
  - 10700 sents my training
  - 342 sents my devtest
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	77.81
Brown	78.74

Parsing Swedish with Danish parser

- Trained on Danish training data
- Parse Swedish test data
- No morphology tweaking so far!
  - Most words are UNKNOWN
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	43.28
Brown	41.84

Morphology

- What if we feed the parser with tags instead of words?
  - *Ændringer i listen i bilaget offentliggøres og meddeles på samme måde.*
  - **NNS IN NN IN NN VB CC VB IN DT NN**
  - **NNS IN NN MD VB CC VB IN DT NN**
  - *Förändrigar i förteckningen skall offentliggöras och meddelas på samma sätt.*

Delexicalized parsing

- Trained on Danish training data (tags only)
- Parse Swedish test data (tags from Hajič tagger)
- Restuff Swedish trees with original words
- All data in hybrid Swedish-Danish Hajič-like tagset
- Following are results on CoNLL test

Parser	da-da	sv-sv	da-sv
Charniak	79.62	76.07	65.50
Brown	80.20	77.01	66.40

Glosses

- Acquis is a parallel corpus
  - more than 430,000 sentences
- GIZA++ & lexical weighting generate da-sv glossary
- Always use highest weighted gloss
- Translate Swedish word-by-word to Danish
- Many unknown words are now known!

Excerpt from sv-da glossary

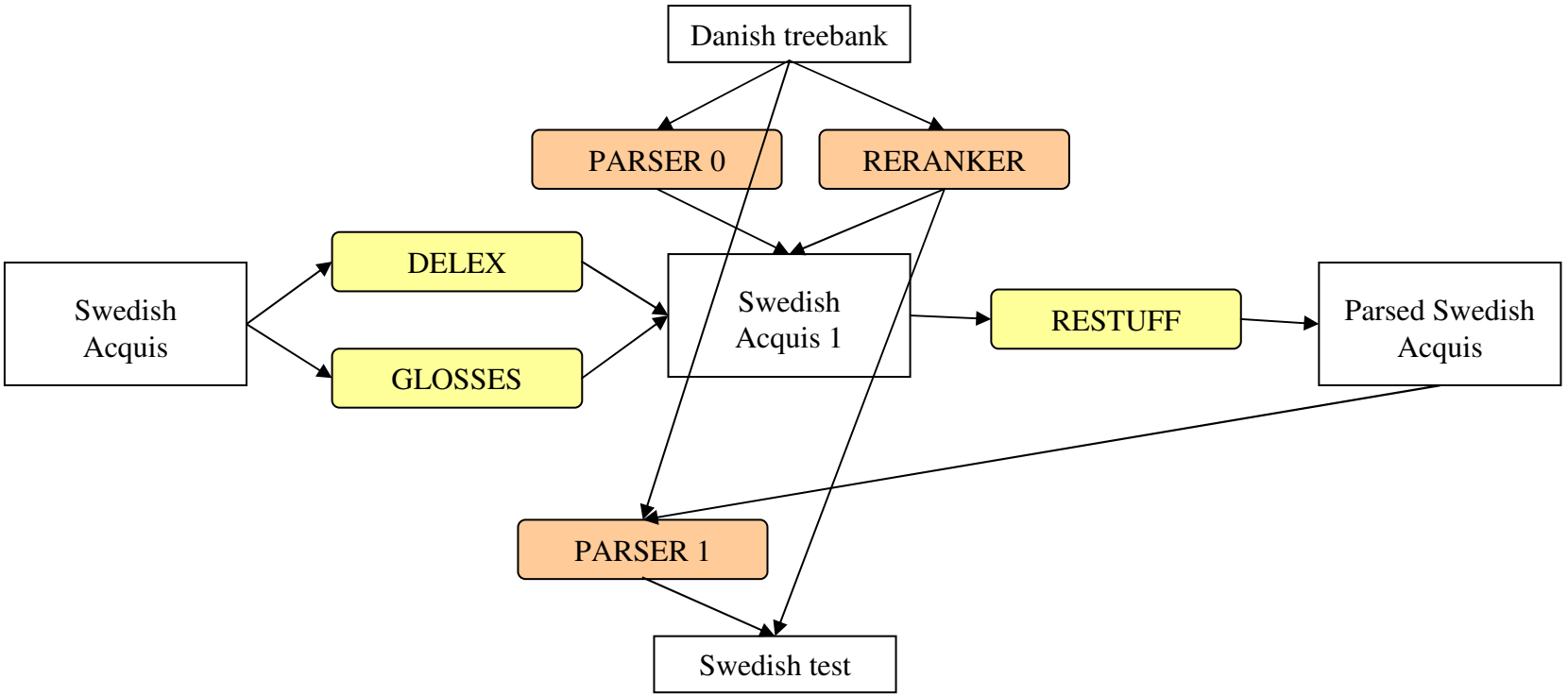
- behandlingsaktörer
- behandlingsanläggning
- behandlingsanläggningar
- behandlingsanläggningen
- behandlingsdatum
- behandlingsformer
- behandlingsfrister
- behandlingsförfaranden
- behandlingsförsök
- behandlingsindikation
- behäftad
- behåll
- behandlingsvirksomheder
- behandlingsanlæg
- behandlingsvirksomheders
- behandlingsanlægget
- datøn
- behandlingsmuligheder
- frister
- behandlingsprocedurer
- befolkningsforsøg
- indikation
- behæftet
- behold

Glossed parsing

- Trained on Danish training data
- Translate Swedish test data to Danish
- Parse it using Danish-trained model
- Restuff trees with Swedish and evaluate
- Following are results on CoNLL test

Parser	F = 2PR/(P+R)
Charniak	63.40
Brown	61.50

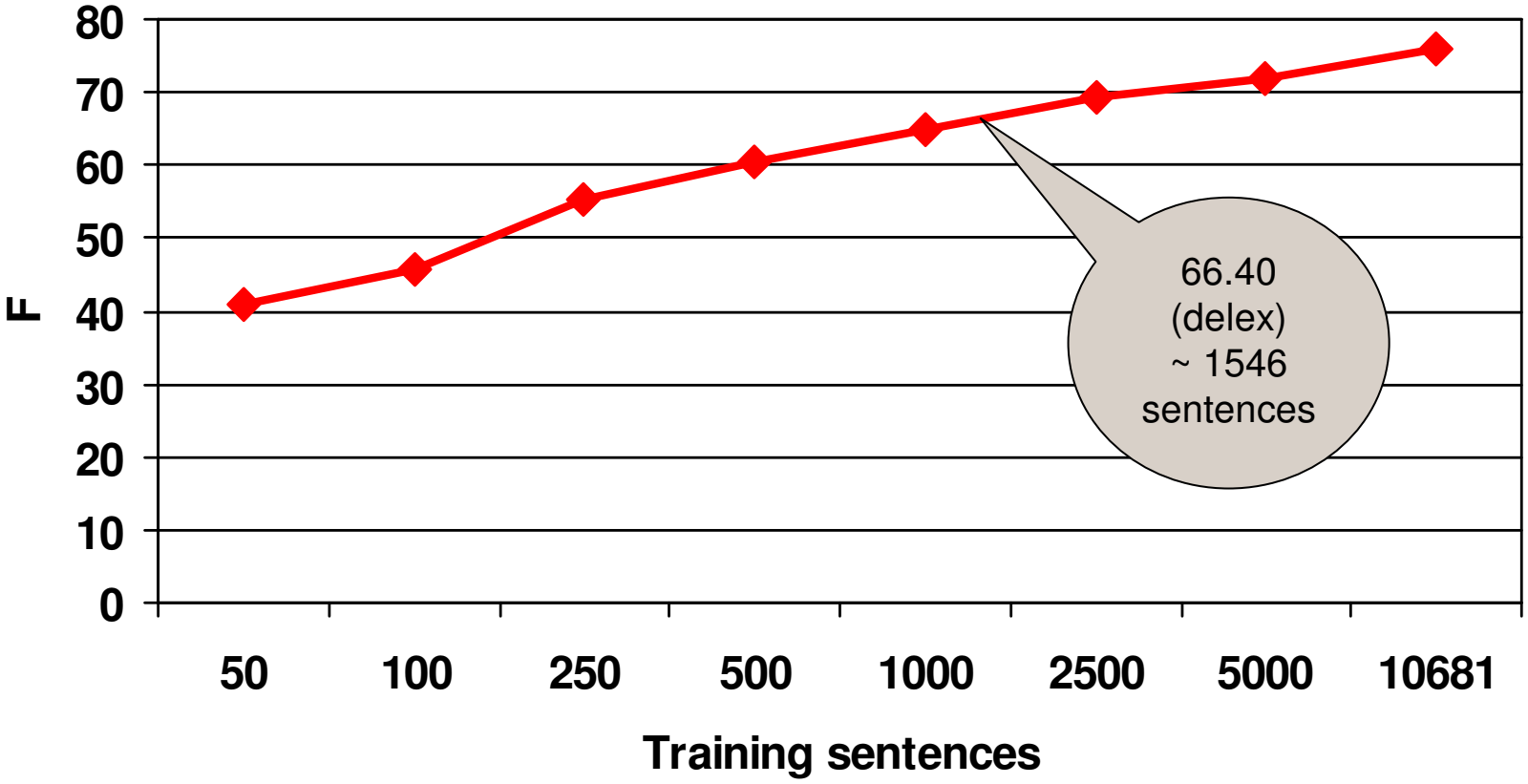
Finally back to Acquis



Train Danish, parse Swedish

	No Acq Charniak	No Acq Brown	Acquis Charniak	Acquis Brown
Plain	43.28	41.84	44.54	42.67
Delex	65.50	66.40	59.60	57.41
Gloss	63.40	61.50	64.48	63.32

How big a Swedish treebank can produce the same results?



Tagset mapping

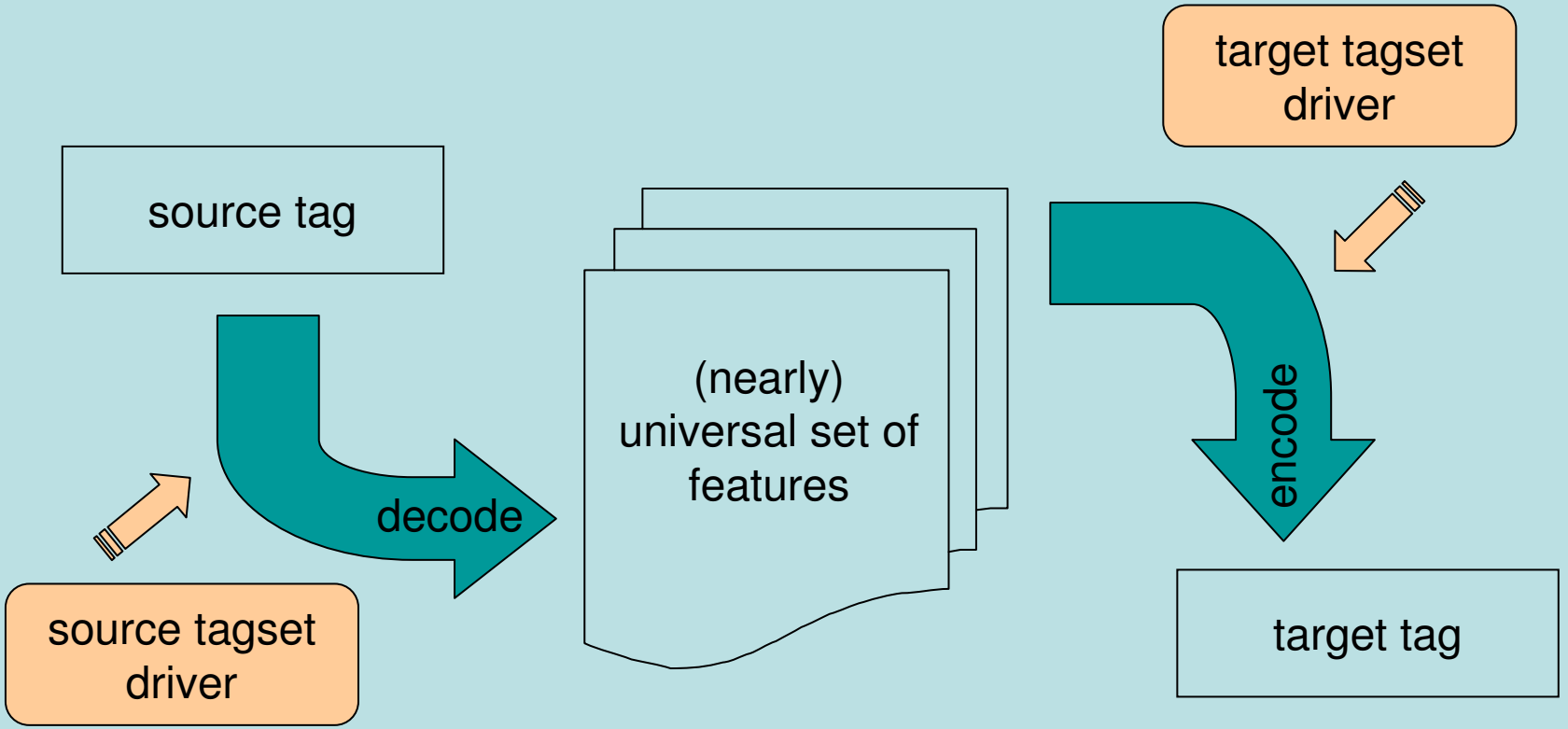
- Already mentioned: da/sv → Penn
- We want to preserve features that
  - are present in both da and sv
  - are not present in Penn
- This is CRUCIAL:
  - unmapped tags are unknown words again
- Mapping tags is always hard even for the same language
- Languages can be similar, approaches way different!

Tagset discrepancy examples

- No determiners in DA, pronouns instead.
- Subject/object pronoun forms in SV *(he/him in EN)*, nominative vs. “unmarked” case in DA
- Masculine gender in SV (pronouns)
- Numbers as adjectives in DA

- Supine in SV — probably the only really language-caused difference

Tagset mapping



reusable!

Tagset mapping

- Key function is **`encode()`**
- Example:
  - pos = noun, gender = masc
  - target set allows:
    - noun + com | neut
    - pronoun + masc | fem | com | neut
    - **but not** noun + masc!
  - priority(pos) > priority(gender)
    $\Rightarrow$ pos = noun
  - then gender is forced to common

Future work

- Delex + glosses for 100 most frequent words
- Iterative bootstrapping (P0 + Acquis $\Rightarrow$ P1 + Acquis $\Rightarrow$ P2 + Acquis ...)
- N-best Acquis trees, classifier combination (delex + gloss)...
- Other language pairs (less related etc.)
- Same approach with dependency parsing (McDonald's parser)

ధన్యవాదములు

THANK YOU!