Rudolf Rosa, Zdeněk Žabokrtský
{rosa,zabokrtsky}@ufal.mff.cuni.cz

Error Analysis of Cross-lingual (Tagging and) Parsing

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

TLT 16, Praha, 24th Jan 2018
Key points of the talk on 1 slide!

- **cross-lingual** parser transfer
  - train on *source* treebank, eval on *target* treebank
  - 1 source (English), 32 target languages – case study
- most frequent errors: incorrectly parsed *nouns*
  - average LAS: 24% on nouns  x 33% on all tags
  - only 3% of predicted *compound* edges correct
- source-target **grammatical similarity** important
  - word order (e.g. ADJ ↔ NOUN, ADP ↔ NOUN)
  - function words (e.g. AUX, DET, PRON, ADP)
Cross-lingual parsing

How to parse a target-language text

- if we **have** a target treebank
  - train a (tagger and) parser on the target treebank
  - apply it to the target text, obtain a parse tree

- if we **don't have** a target treebank
  - take a treebank for a **source** language
  - transfer it to the **target** language (e.g. machine transl.)
    - conversion to the previous case
  - train a (tagger and) parser on the resulting **pseudo-target** treebank
  - apply it to the target text, obtain a parse tree

~70 languages, news/books/wiki

~7000 languages

Other good approaches also exist
Our setup

- 32 targets (UD 1.4), 1 source (English)
- translate source treebank into target language
  - OpenSubtitles2016, MGiza intersection alignment
  - word-based Moses (1:1), KenLM, no reordering
  - translate word forms, keep annotation
- train UDPipe tagger and parser on it
  - tagger: form → UPOS
  - parser: form & UPOS → head & basic deprel
    - form: word2vec on target side of parallel data
Based on our SFNW setup

- Slavic Forest, Norwegian Wood (Rosa+, 2017)
  - winner of VarDial Cross-lingual parsing shared task

- in VarDial, source languages were pre-defined
  - this work: source \equiv English; to do: source selection
Motivation for error analysis

1. initial setup
2. identify common problems
3. think up possible remedies
4. try them out in experiments (preliminary)
5. final improved setup (future work)
Target languages (UD 1.4)

- grouped by cross-lingual tagging accuracy
  - source always English
  - **High** (pt, no, it, fr, da, de, sv)
    - Germanic and Romance languages with large parallel data
  - **Med** (bg, ca, gl, nl, sk, cs, ru, id, el, hr, ro, pl, et, lv, sl)
    - mostly European languages more distant from English and/or with smaller parallel data
  - **Low** (fi, he, hi, uk, tr, ar, fa, vi, eu, hi)
    - non-European or non-Indo-European languages
LAS factored by (gold) UPOS

![Graph showing LAS by (gold) UPOS categories]

- NOUN
- PROPN
- ADJ
- ADV
- VERB
- AUX
- PRON
- DET
- ADP

Legend:
- Low
- Med
- High
- EN mono

Rudolf Rosa and Zdeněk Žabokrtský – Error Analysis of Cross-lingual Tagging and Parsing

8/18
LAS factored by (gold) deprel

- nmod
- nsubj
- dobj
- compound
- advmod
- amod
- det
- case
- aux
- cop

Legend:
- Low
- Med
- High
- EN mono
Nouns

- problematic tagging & parsing of named entities
  - many OOVs, already in translation → non-target words
  - many capitalized NOUNs mistagged as PROPNs
  - *name* annotation seems inconsistent in UD 1.4
  - simplify names? truecase? casing feature?
- *nmod*, *compound*, *nsubj*, *dobj*...
  - different languages mark the relations differently
    - word order, adpositions, determiners, morphology...
  - most frequent error: *nmod* → *compound*
    - compound very specific for English – remove?
Verbs

- auxiliary verbs (AUX tag, aux & cop deprels)
  - good only in High langs – grammar similarity crucial
  - VERB/AUX mistagging, unreliable parsing
- clausal relations (ccomp, xcomp, advcl, acl...)
  - very hard to get right (even for monolingual parser)
  - head assignment: long-distance relations
  - deprel assignment: confused for each other
Regular phenomena

- case, nummod, punct, det, advmod, amod, cc
- usually easy to parse if tagging correct
- head attachment usually rather easy, except for:
  - amod in NOUN ADJ languages (Romance)
  - case in post-positional Low languages
- deprel assignment mostly trivial
  - ADP → case, NUM → nummod, PUNCT → punct, DET → det, ADV → advmod, ADJ → amod
Adjectives

- confused for NOUN compounds
  - *en*: “fruit salad” \(\times\) *cs*: “ovocný salát”
    - NOUN compound ADJ amod
  - remove such confusing words from training data?
- ADJ NOUN / NOUN ADJ (Romance) word order
  - reorder in MT? pre-reorder? shuffle words locally?
- otherwise parsing easy
Pronouns, determiners, adpositions

- PRONs hard & often cannot align 1:1
  - extra PRONs (reflexive), missing PRONs (pro-drop)
- DET/PRON mistagging, esp. if form ambiguous
  - e.g. “le”, “la” in French – quite common ambiguity
  - leave decision to parser?
- DETs rare in target → much confusion
  - remove some from source?
- ADP tagging good (sometimes aligned to DETs)
  - parsing good unless post-positional target
How to address the issues

1. select source language similar to target
   - especially in the problematic phenomena
     - word order, function words usage, noun phrases

2. try a workaround
   - diverge from 1:1 monotonic MT (but M:N hard)
     - allow 2:1? remove words? (pre-)reordering?
   - relabel some phenomena to get a closer match?
   - remove some phenomena from source data?
   - mix multiple sources (in a clever way?)
     - different mix for different phenomena?
Simple preliminary experiments

- relabel `PROPN` → `NOUN`
  - deterioration for most targets (PROPN signal useful)
- relabel `AUX` → `VERB`
  - helps for Med and Low targets (different grammar)
- relabel `DET` → `PRON`
  - helps for half of targets (across all groups)
- relabel `compound` → `nmod` (not in test!)
  - helps by +0.6% LAS (compound too specific for `en`)
- word reordering in Moses
  - large deterioration (translation literalness?)
Key points of the talk on 1 slide!

- **cross-lingual** parser transfer
  - train on *source* treebank, eval on *target* treebank
  - 1 source (English), 32 target languages – case study
- most frequent errors: incorrectly parsed **nouns**
  - average LAS: **24%** on nouns \(\times\) **33%** on all tags
  - only **3%** of predicted **compound** edges correct
- **source-target** **grammatical similarity** important
  - word order (e.g. ADJ ↔ NOUN, ADP ↔ NOUN)
  - function words (e.g. AUX, DET, PRON, ADP)
Thank you for your attention

Rudolf Rosa, Zdeněk Žabokrtský
{rosa,zabokrtsky}@ufal.mff.cuni.cz

Error Analysis of Cross-lingual Tagging and Parsing

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ufal.cz/rudolf-rosa