Rudolf Rosa
rosa@ufal.mff.cuni.cz

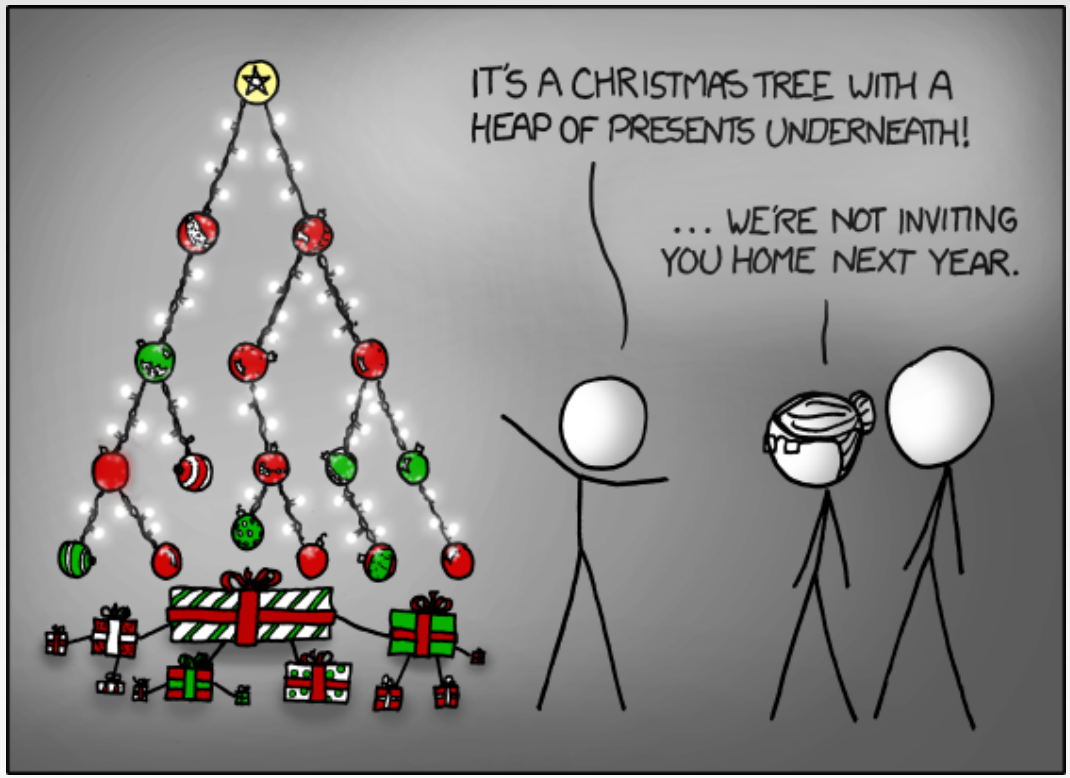
Using a Collection of Many Treebanks for Exploring the Structure of Natural Language Sentences

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

ÚFAL Week of Doctoral Students, Prague, 19 May 2014

The questions

- Are trees good?
  - as a representation of sentence syntactic structure
- If no...
  - ...what is?
- If yes...
  - ...which ones?
  - ...how to get them?



The answers?

- Are trees good?
  - as a representation of sentence syntactic structure
- If no...
  - ...what is?
- If yes...
  - ...which ones?
  - ...how to get them? *semi-supervised parsing?*

1

The answers?

- Are trees good?
  - as a representation of sentence syntactic structure
- If no...
  - ...what is?
- If yes...
  - 2 ...which ones? *as in treebanks? harmonized?*
  - 1 ...how to get them? *semi-supervised parsing?*

The answers?

- Are trees good?
  - as a representation of sentence syntactic structure
- If no...
3. ...what is? *some fuzzier structures?*
- If yes...
2. ...which ones? *as in treebanks? harmonized?*
1. ...how to get them? *semi-supervised parsing?*

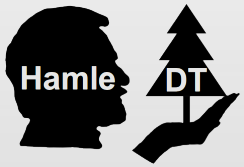
Supervised parsing is limited

- take a treebank (TB) and train a parser on it
- HamleDT
  - collection of **30** TBs
    - ar, eu, bn, bg, ca, cs, da, nl, en, et, fi, de, el, grc, hi, hu, it, ja, la, fa, pt, ro, ru, sk, sl, es, sv, ta, te, tr
  - some TBs small or low quality...
- there are about **7000** languages!
  - plus various types (e.g. speech)
  - annotating TBs is costly

Semi-supervised techniques

- e.g. we have an English TB, but no Punjabi TB
- delexicalized parsing (no words, just tags)
  - train a delexicalized parser on EN TB
  - apply to PA text (there is a PA tagger)
- parse tree projection
  - train a full parser on EN TB
  - take parallel PA-EN text (can use SMT)
  - parse the EN side
  - project the tree through word-alignment to PA side

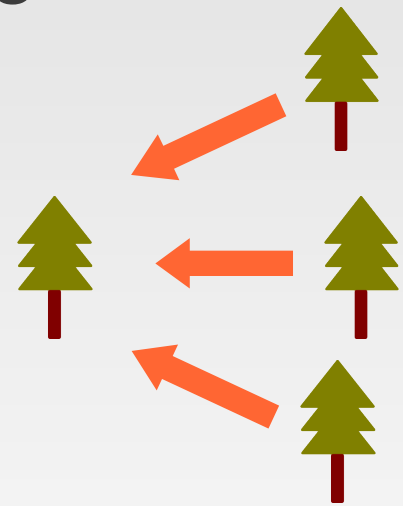
Harmonization of trees and tags

- HamleDT (ÚFAL) 

  - PDT-like trees, tags converted to Interset
- Universal Dependency Treebanks (Google)
  - Universal Stanford Dependencies, Universal POS
- 1 style fits all? 1 style for a language family?
- how to evaluate the style fitness? (**unsolved**)
  - parser-friendliness (achievable parsing accuracy)?
  - extrinsic evaluation (e.g. in Depfix post-editing)?

My plan: Multisource projection

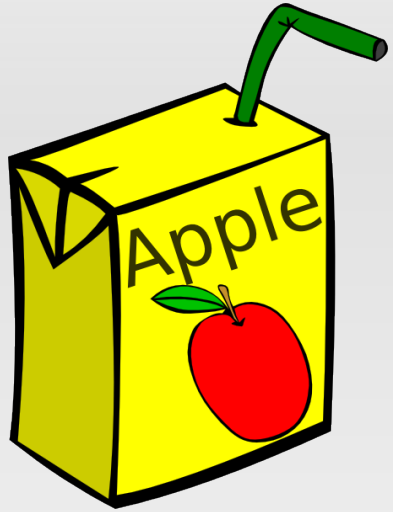
- translate the (Punjabi) text into the 30 languages
- parse each of them by a supervised parser
- each potential edge in (PA) tree gets a score
  - number of trees that confirm it?
  - or sum of scores assigned to it by the parsers?
    - normalized, weighted by language similarity...?



- find the parse tree = the maximum spanning tree
  - McDonald's parser (Chu-Liu-Edmonds algorithm)

Making it more fuzzy?

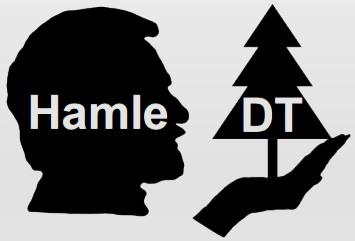
- suggested using fuzzy input features
  - edge score instead of edge existence
  - similar approach for tags?
    - e.g. **apple** juice:
      60% noun, 40% adjective



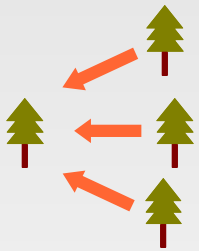
- edge scores more useful than the tree?
  - applications use individual edges, not whole tree
    - e.g. Depfix post-editing (edge-local)

Conclusion

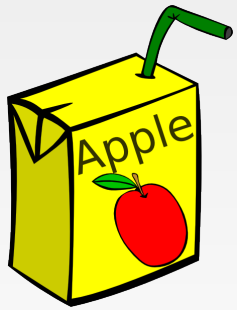
- HamleDT: 30 treebanks
  - Prague style, Interset tags
  - extrinsic evaluation of harmonization?



- semi-supervised parsing
  - multisource projection?



- adding more fuzziness
  - fuzzy input features?
  - and/or even fuzzy output?



Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Using a Collection of Many Treebanks for Exploring the Structure of Natural Language Sentences

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

For this presentation and other information, please visit:

http://ufal.mff.cuni.cz/rudolf-rosa