Identification of root morphs in morphologically segmented data

Vojtěch John, Magda Ševčíková, Zdeněk Žabokrtský

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

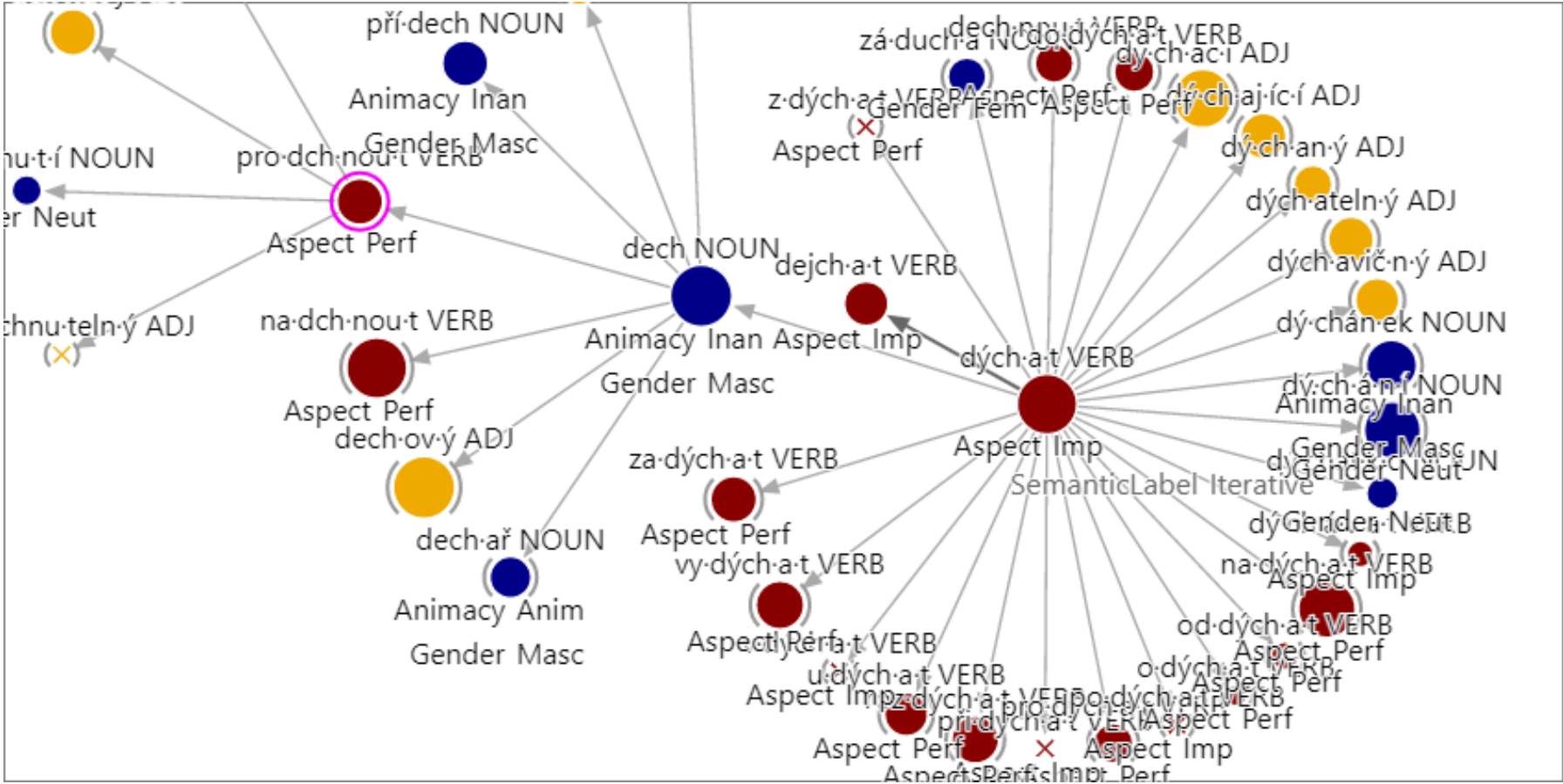
Introduction

Terminology

- **Morpheme** = the smallest meaning-bearing unit of language
- **Morph** = the concrete form of morpheme
    - Need to distinguish between several forms of the same morpheme (modified by phonological changes): **dech** x **dých**at x pro**dch**nout
    - Simplification: Words are strings of morphs
        - Cf. Arabic.
    - Simplification: **Root** morph conveys lexical meaning
        - Cf. In Czech, the same morph (with the same lexical meaning) can be used as both root and non-root: **Před**poklad x **před**nosta; cf. **Over** x **over**bearing x **over**all
- **Morphological segmentation**: Given a word, divide the word to morphs

Motivation

- Multiple resources contain morphological segmentation
  - Without morphological classification
  - With low-quality morphological classification
- State-of-the-art morphological segmentation (Sigmorphon 2022) often does not include morphological classification
  - There will probably be more segmentation-only resources in the future
- Identification of the root could help in building derivational networks


pří·dech NOUN
Animacy Inan
Gender Masc
pro·dch·nou·t VERB
Aspect Perf
dech NOUN
Animacy Inan
Gender Masc
dejch·a·t VERB
Aspect Imp
dých·a·t VERB
Aspect Imp
SemanticLabel Iterative
na·dch·nou·t VERB
Aspect Perf
dech·ov·ý ADJ
dech·ař NOUN
Animacy Anim
Gender Masc
za·dých·a·t VERB
Aspect Perf
vy·dých·a·t VERB
Aspect Perf
z·dých·a·t VERB
Aspect Perf
Gender Fem
Aspect Perf
dý·ch·ac·í ADJ
dý·ch·aj·íc·í ADJ
dý·ch·an·ý ADJ
dých·ateln·ý ADJ
dých·avič·n·ý ADJ
dý·chán·ek NOUN
dý·ch·á·n·í NOUN
Animacy Inan
Gender Masc
Gender Neut
na·dých·a·t VERB
Aspect Imp
od·dých·a·t VERB
Aspect Perf
o·dých·a·t VERB
Aspect Perf
u·dých·a·t VERB
Aspect Imp
Aspect Perf
Aspect Imp

Root identification

Data

- Gold data
  - 7 Indo-European languages with manually annotated data
  - Data for 6 of the languages (not for Czech) taken from UniSegments
  - French, English, German, Croatian, Italian, Russian, Czech
  - For each language, 5000 words train data, 5000 words test data.
- Universal Derivations
  - Treebanks for all the 7 languages, not necessarily manually annotated

Methods - simple statistics

- **MaxLen**: The longest morph
- **MinFreq**: Frequency of morph in dictionary
- **MinNeighborEntropy**: min(max(H(wi-1/i+1|wi)))
- **UnweighedMix**: Unweighted combination of the above
- **ProbabMix**: Run UnweighedMix on all the data and pick the most common tag (root x non-root) for every morph.

- ***Limitations***: *MinFreq*, *MinNeighborEntropy* and *UnweighedMix* only pick the best candidate, which significantly decreases accuracy for languages with common compounding (German – oracle picking only one root: 57.5 %).

Methods – derivational trees, CRF

- **DerivTree**: Shortest edit distance from the ***derivational*** root
- **DerivTree + UnweightedMix**: add DerivTree as one of the factors in UnweightedMix
- **LongestInDerivTree**: Apply the previous on common substring (with simple wildcards) of all the derivationally related words
- **CRF classifier**, trained on the training data (5000 words).

- ***Limitations***: the *DerivTree* methods also pick only the best candidate.

Results – word-level accuracy

Language	ProbabMix	UnweightedMix + DerivTree	CRF tagger
Czech	95.4 %	98.6 %	97.6 %
Croatian	91.9 %	97.1 %	98.3 %
English	91.0 %	85.5 %	94.0 %
French	92.9 %	94.8 %	94.4 %
German	83.4 %	55.9 %	92.2 %
Italian	90.8 %	96.1 %	96.2 %
Russian	80.1 %	78.1 %	90.2 %

Error analysis - compounding

- Most of the unsupervised methods cannot deal with multiple roots
- On data without compounds:
    - For Croatian and Italian, the best word-level accuracy is achieved by MinFreq (98.7 % and 97.5 %)
    - UnweightedMix achieves 93.1 % to 98.1 %, is best for English and in 4 out of 7 cases achieves better results than the CRF tagger.
    - DerivTree + UnweightedMix is the best solution for all the remaining languages and in 6 out of 7 cases is better than CRF tagger; in the remaining case (Italian), the difference is 0.1%

Error analysis - homomorphy

- Homomorphy = two morphemes are expressed by the same morph
- What is seen as root in the training data may not always be root (or even the same morpheme).
- Bad also for the unsupervised methods - the statistics gets mixed up
- Root-Affix homomorphy for all the languages in less than 1.6 % of words
- Errors disproportionately common in words with root-affix homomorphy

Method	Czech	German	English	French	Croatian	Italian	Russian
Unweighted Mix	6 %	8 %	8 %	6 %	14 %	17 %	22 %
ProbabMix	16 %	23 %	18 %	15 %	49 %	24 %	39 %
DerivTree + UM	4 %	8 %	8 %	6 %	12 %	16 %	22 %
CRF tagger	12 %	40 %	19 %	11 %	32 %	23 %	45 %

Problems and future work

- Homomorphy (pod l ý x pod klad)
- Allomorphy (dých a t x pro dch nou t)
- Gold data hard to get (Biggest collection to date: USeg)
- Multiple roots recognition (iterative?)
- Resource-light inflection/derivation desambiguation

Summary

- Root identification given segmented words is fairly easy
- Simple statistical methods can be relatively strong
- Biggest problems are compounding and homomorphy