Attempting to separate inflection and derivation using vector space representations

Rudolf Rosa, Zdeněk Žabokrtský

September 19, 2019

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

The problem

- Goal: separate inflection from derivation (~lemmatization)
- Classical approach: supervised methods
  - Manually annotate a corpus
  - Train a tagger and lemmatizer on the corpus
  - (Or: manually create a rule-based tool)
  - Apply to text
- Our focus: unsupervised methods
  - Use no annotated data
  - Discover lemmasets solely based on unannotated plain-text corpora
  - (Also interesting: semi-supervised methods, using a handful of annotated data, and/or data for another language...)

Why unsupervised?

- Practical reasons
  - For most languages, there are no or low resources
  - Creation of resources is costly
  - (Also: resources are not consistent across languages)
  - Plain text data available for hundreds of languages
    - Bible (or part of it): 1,400 languages (Mayer and Cysouw, 2014)
    - JW300: Watchtower texts (~100k sentences) for 300 languages (Agić and Vulić, 2019)
- Research reasons
  - It is an interesting challenge
  - We can learn something about language
    - Empirical research independent of linguistic traditions and annotations
    - Whatever we discover is true about the language itself, not only about a particular annotation
  - Question the traditional strictly binary inflection-derivation dichotomy
    - Replace it with an empirical inflectionality score?

This work

- A modest beginning of a probably long journey
- Currently, we only present experiments for Czech language
- For evaluation, we rely on existing annotated resources
  - lemmas and their inflections: PDT (Böhmová et al., 2003), SYN (Hnátková et al., 2014)
  - derivational relations between lemmas: DeriNet (Žabokrtský et al., 2016)
- Inflection: lemma → word form
  - take → take, takes, taking, took, taken
  - pes (dog) → pes, psa, psu, psovi, pse, psem, psi, psů, psům, psy, psech
  - case, number, gender, person, tense, degree, negation, voice
- Derivation: parent lemma → child lemma
  - take → overtake, taker, intake, takeout, mistake...
  - pes → pejsek, psí, psisko, psoun, psovitý, psův, zepsout...
  - perfective-imperfective, adjective-adverb, possessive, diminuitive, noun gender...

Outline

Approach

- Goal: unsupervised separation of inflection and derivation
- Hypothesis: inflections are closer than derivations
  - Word forms that are inflections of one lemma are *more similar* than word forms belonging to different lemmas
  - We explore two kinds of similarity:
    - Orthographic similarity, via string edit distance
    - Meaning similarity, via word embeddings similarity
- Note: there are other potentially testable criteria (Stump, 1998)
  - inflection is semantically more regular than derivation (Bonami and Paperno, 2018)
  - syntax may determine inflection
  - inflection is more productive
  - ...

Orthographic similarity: string edit distance

Levenshtein distance $LD(w_1, w_2)$ (Levenshtein, 1966)

- Number of single-character edit operations (addition, deletion, substitution)
- 'p<u>re</u>position<u>s</u>' $\rightarrow$ 'p<u>ost</u>position': 4 (r$\rightarrow$o, e$\rightarrow$s, +t, –s)

Jaro-Winkler distance $JW(w_1, w_2)$ (Winkler, 1990)

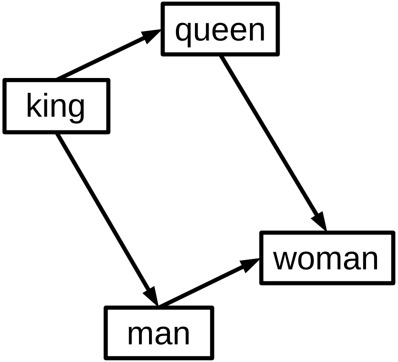
- Similar idea to Levenshtein distance
- The JW distance is a number between 0 and 1
- Imbalanced: matching at the beginning of the string is more important
  - Useful for predominantly suffixing languages (typical for languages we usually encounter)

Additional tweak: average with distance of simplified form

- Lowercase, transliterate to ASCII, remove non-initial vowels (a e i o u y)
- "Účelový" $\rightarrow$ "uclv"

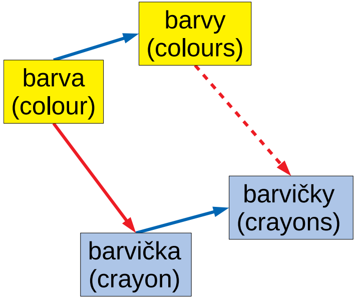
Meaning similarity: word embeddings

- Word embedding: a vector of many real numbers, e.g. $vec(``king") = [0.12, 5.23, -7.12, \ldots, 2.36]$
- Computed unsupervisedly from large text corpora
    - Tools to compute word embeddings from text corpora are easy to download and use
    - Pre-computed embedding dictionaries available for download for hundreds of languages
- Based on the distributional hypothesis
    - Embedding of a word determined by contexts in which it appears in the corpus
    - Words appearing in similar contexts have similar embeddings
    - Embedding similarity can serve as a proxy to meaning similarity
    - Also, some interesting regularities can be observed



Meaning similarity: word embeddings

- Inflection tends to correspond to a vector shift (Mikolov et al., 2013)
- Derivation tends to correspond to a vector shift (Musil et al., 2019)
- Our hypothesis: an inflectional shift should be smaller than a derivational shift



Meaning similarity: word embeddings cosine similarity

- Meaning similarity = cosine similarity of word embeddings
    - Standard way of measuring word embedding similarity
    - $COS_{sim}(w_1, w_2) = \frac{vec(w_1) \cdot vec(w_2)}{|vec(w_1)| \cdot |vec(w_2)|}$
- FastText word embeddings, downloaded from FastText website (Grave et al., 2018)
    - Combine embeddings of full words and of character n-grams
    - Provides a vector even for out-of-vocabulary words

Combination, conversion to distance

Combined measure

- All similarities are scaled to $[0, 1]$ interval
- Combined similarity measure: multiplication of Jaro-Winkler string similarity and word embedding cosine similarity
- $JW_{sim}(w_1, w_2) \cdot COS_{sim}(w_1, w_2)$

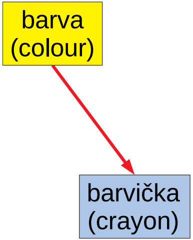
Distance measure

- For technical reasons, we need distances, not similarities
- Distance: $X_{dist} = 1 - X_{sim}$

Outline

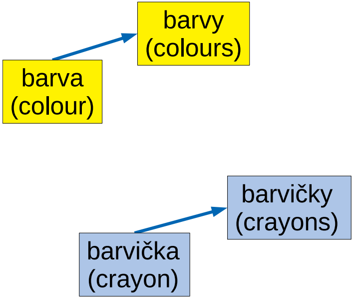
Data

- DeriNet v1.7 (Žabokrtský et al., 2016)
  - Derivational dictionary
  - Lemmas in one derivational family linked by derivational edges
  - No inflections



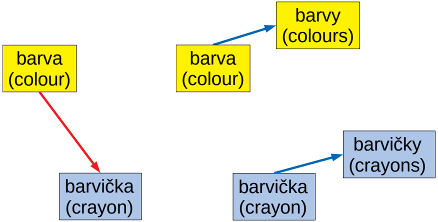
Data

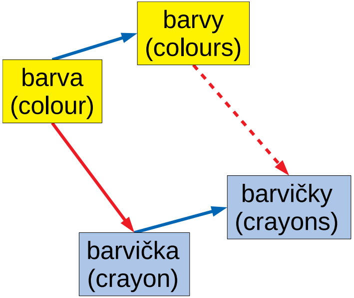
- SYN v4 (Hnátková et al., 2014)
  - Tagged corpus
  - Words in sentences annotated by lemmas and morphological features
  - No derivational annotation



Data

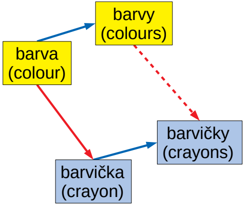
- Combine the resources
  - DeriNet derivational trees with lemmas
  - Add inflections from SYN to each lemma
  - Add secondary derivational edges





Task

- For a pair of words, decide if they are inflections of the same lemma
  - barva (colour), barvy (colours) → yes
  - barvička (crayon), barvičky (crayons) → yes
  - barva (colour), barvička (crayon) → no
  - barvy (colours), barvičky (crayons) → no
  - barva (colour), barvičky (crayons) → no

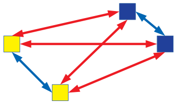


- We use only several of the largest derivational families from DeriNet
  - Small derivational families are uninteresting (not many derivational relations)
  - 561 derivational families with at least 50 lemmas
    → sample 42 families
    → 4,514 lemmas
    → 69,743 word forms

Evaluation types

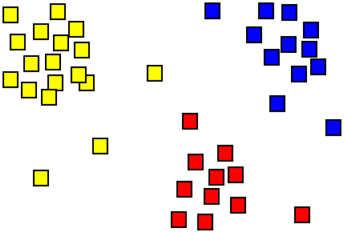
Pairwise evaluation

- Is the distance of the two words higher than a threshold?
- Inflections should be below the threshold, derivations above
- Oracle threshold



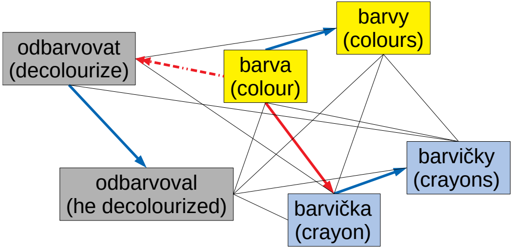
Clustering-based evaluation

- Use the word distances to find clusters of nearby words
- Agglomerative clustering algorithm
- Inflections of one lemma should fall into one cluster, derivations into separate clusters
- Oracle number of clusters

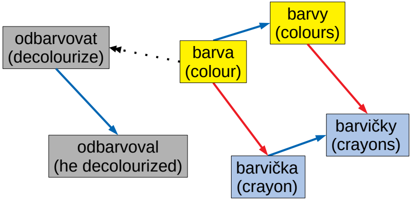


Which pairs of words to evaluate

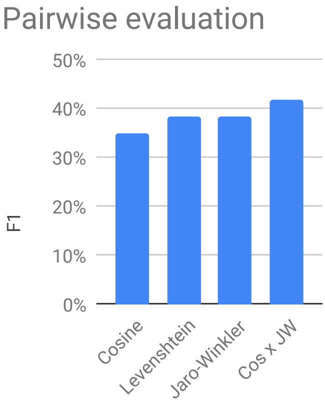
- Pairs of all words
  - Most realistic
  - Too slow
- Pairs of all words in one derivational family
  - Reasonably realistic
  - Most pairs are very distant words – boring
  - Use this for quantitative evaluation

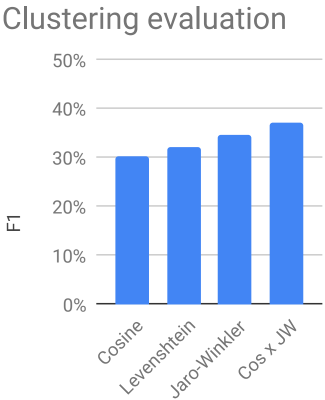


- Pairs of words linked by a single derivational or inflectional operation
  - Not realistic, many close pairs omitted
  - Focuses on the hard cases – interesting
  - Use this for further manual analysis



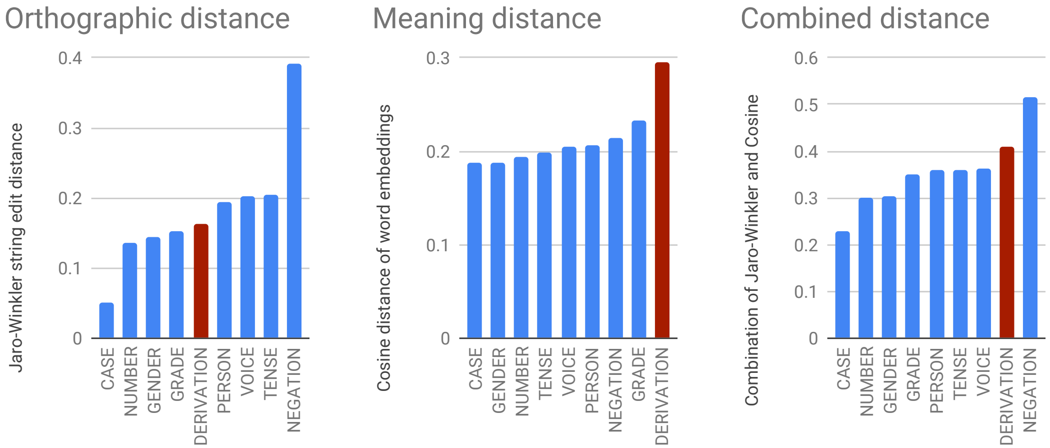
Quantitative evaluation: identification of inflection





- Inflection and derivation separable to some extent
- Combination better than individual measures

Further analysis: average count-weighted distances



Further analysis

- Typical inflections have low distance (case, number, gender)
- Typical derivations have high distance (e.g. part of speech change)
- Some inflections have high distance: negation, grade, voice
  - limited productivity, larger meaning shift
- Some derivations have low distance:
  adjective→adverb (barevný–barevně), noun→possessive (hvězdář–hvězdářův), perfective→imperfective (bloknout–blokovat), noun diminuitives (hvězda–hvězdička)
  - very regular, very productive
- Inflection-derivation dichotomy: a strictly binary categorization or a continuous scale?

Outline

Summary

- Unsupervised separation of inflection from derivation
- Hypothesis: inflections are more similar than derivations
  - Orthographic similarity: Jaro-Winkler edit distance
  - Meaning similarity: cosine similarity of FastText word embeddings
- Combined similarity measure achieves respectable accuracy
- Inflection-derivation boundary is vague

http://ufal.cz/rudolf-rosa

References I

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1310.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer, 2003.

Olivier Bonami and Denis Paperno. Inflection vs. derivation in a distributional vector space. *Lingue e linguaggio*, 17(2):173–196, 2018.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Milena Hnátková, Michal Kren, Pavel Procházka, and Hana Skoumalová. The SYN-series corpora of written Czech. In *LREC*, pages 160–164, 2014.

Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

Thomas Mayer and Michael Cysouw. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3158–3163, Reykjavik, Iceland, May 2014. European Languages Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Tomáš Musil, Jonáš Vidra, and David Mareček. Derivational morphological relations in word embeddings. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 173–180, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4818.

Gregory T Stump. Inflection. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 13–43. London: Blackwell, 1998.

William E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pages 354–359, 1990. URL http://www.amstat.org/sections/srms/Proceedings/papers/1990_056.pdf.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. Merging data resources for inflectional and derivational morphology in Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1307–1314, 2016.