Attempting to separate inflection and derivation using vector space representations

Rudolf Rosa, Zdeněk Žabokrtský
The problem

- Goal: separate inflection from derivation
The problem

- Goal: separate inflection from derivation (~lemmatization)
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
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...)

Attempting to separate inflection and derivation using vector space representations
Why unsupervised?

- Practical reasons
  - For most languages, there are no or low resources
  - Creation of resources is costly

- 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?

Attempting to separate inflection and derivation using vector space representations
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

Replacing it with an empirical inflectionality score?
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)
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
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
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
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)

Attempting to separate inflection and derivation using vector space representations
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

Attempting to separate inflection and derivation using vector space representations
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, psovity, psův, zepsout...
  • perfective-imperfective, adjective-adverb, possessive, diminuitive, noun gender...
Attempting to separate inflection and derivation using vector space representations
• Goal: unsupervised separation of inflection and derivation
• 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
• 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
• 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)

Additional tweak: average with distance of simplified form

Lowercase, transliterate to ASCII, remove non-initial vowels (a e i o u y)

“Účelový” → “uclv”

Attempting to separate inflection and derivation using vector space representations
Orthographic similarity: string edit distance

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

- Number of single-character edit operations (addition, deletion, substitution)
- ‘prepositions’ $\rightarrow$ ‘postposition’: 4 (r$\rightarrow$o, e$\rightarrow$s, +t, −s)
Orthographic similarity: string edit distance

Levenshtein distance $LD(w_1, w_2)$ (Levenshtein, 1966)
- Number of single-character edit operations (addition, deletion, substitution)
- ‘prepositions’ → ‘postposition’: 4 (r→o, e→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” → “uclv”
Orthographic similarity: string edit distance

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

- Number of single-character edit operations (addition, deletion, substitution)
- ‘prepositions’ → ‘postposition’: 4 (r→o, e→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” → “uclv”

<table>
<thead>
<tr>
<th>Problem</th>
<th>Approach</th>
<th>Evaluation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Meaning similarity: word embeddings

- Word embedding: a vector of many real numbers, e.g. \( vec("king") = [0.12, 5.23, -7.12, \ldots, 2.36] \)
Meaning similarity: word embeddings

- Word embedding: a vector of many real numbers, e.g. \( \text{vec(“king”)} = [0.12, 5.23, -7.12, \ldots, 2.36] \)
- 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

Attempting to separate inflection and derivation using vector space representations
Meaning similarity: word embeddings

- Word embedding: a vector of many real numbers, e.g. $\text{vec}(\text{"king"}) = [0.12, 5.23, -7.12, \ldots, 2.36]$

- 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

Attempting to separate inflection and derivation using vector space representations
Meaning similarity: word embeddings

- Word embedding: a vector of many real numbers, e.g. \( vec(\text{“king”}) = [0.12, 5.23, -7.12, ... , 2.36] \)
- 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

Attempting to separate inflection and derivation using vector space representations

<table>
<thead>
<tr>
<th>Problem</th>
<th>Approach</th>
<th>Evaluation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>7/23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• 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 = cosine similarity of word embeddings
  • Standard way of measuring word embedding similarity

\[
C_{OS}\text{sim}(w_1, w_2) = \frac{\mathbf{vec}(w_1) \cdot \mathbf{vec}(w_2)}{|\mathbf{vec}(w_1)| \cdot |\mathbf{vec}(w_2)|}
\]

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

Attempting to separate inflection and derivation using vector space representations
Meaning similarity: word embeddings cosine similarity

- Meaning similarity = cosine similarity of word embeddings
  - Standard way of measuring word embedding similarity
  - $COS_{sim}(w_1, w_2) = \frac{\text{vec}(w_1) \cdot \text{vec}(w_2)}{|\text{vec}(w_1)| \cdot |\text{vec}(w_2)|}$

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

Attempting to separate inflection and derivation using vector space representations
Meaning similarity: word embeddings cosine similarity

- Meaning similarity = cosine similarity of word embeddings
  - Standard way of measuring word embedding similarity
  - $COS_{sim}(w_1, w_2) = \frac{\text{vec}(w_1) \cdot \text{vec}(w_2)}{|\text{vec}(w_1)| \cdot |\text{vec}(w_2)|}$

- 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

\[ J_W \text{sim}(w_1, w_2) \cdot C_{OS} \text{sim}(w_1, w_2) \]

Distance measure

For technical reasons, we need distances, not similarities

\[ X_{dist} = 1 - X_{sim} \]
Combined measure

- All similarities are scaled to $[0, 1]$ interval
Combined measure

- All similarities are scaled to $[0, 1]$ interval
- Combined similarity measure: multiplication of Jaro-Winkler string similarity and word embedding cosine similarity
Combined measure

• All similarities are scaled to $[0, 1]$ interval
• Combined similarity measure: multiplication of Jaro-Winkler string similarity and word embedding cosine similarity
• $JW_{sim}(w_1, w_2) \cdot COS_{sim}(w_1, w_2)$
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: $X_{dist} = 1 - X_{sim}$
Attempting to separate inflection and derivation using vector space representations
Data

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

Attempting to separate inflection and derivation using vector space representations
• SYN v4 (Hnátková et al., 2014)
  • Tagged corpus
  • Words in sentences annotated by lemmas and morphological features
  • No derivational annotation

Attempting to separate inflection and derivation using vector space representations
• Combine the resources
  • DeriNet derivational trees with lemmas
  • Add inflections from SYN to each lemma
  • Add secondary derivational edges

Attempting to separate inflection and derivation using vector space representations
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

Attempting to separate inflection and derivation using vector space representations
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

Attempting to separate inflection and derivation using vector space representations
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)
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
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

Attempting to separate inflection and derivation using vector space representations
Which pairs of words to evaluate

- Pairs of all words
  - Most realistic
  - Too slow

Attempting to separate inflection and derivation using vector space representations
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

Attempting to separate inflection and derivation using vector space representations
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

Attempting to separate inflection and derivation using vector space representations
Quantitative evaluation: identification of inflection

Pairwise evaluation

- Cosine
- Levenshtein
- Jaro-Winkler
- Cos x JW

Clustering evaluation

- Cosine
- Levenshtein
- Jaro-Winkler
- Cos x JW

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

Attempting to separate inflection and derivation using vector space representations
Further analysis: average count-weighted distances

Orthographic distance

Meaning distance

Combined distance

Attempting to separate inflection and derivation using vector space representations
Further analysis

- Typical inflections have low distance (case, number, gender)
- Typical derivations have high distance (e.g. part of speech change)
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
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 diminutives (hvězda–hvězdička)
  - very regular, very productive
- Inflection-derivation dichotomy: a strictly binary categorization or a continuous scale?

Attempting to separate inflection and derivation using vector space representations
Attempting to separate inflection and derivation using vector space representations
Summary

• Unsupervised separation of inflection from derivation
Summary

- Unsupervised separation of inflection from derivation
- Hypothesis: inflections are more similar than derivations

http://ufal.cz/rudolf-rosa
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
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

http://ufal.cz/rudolf-rosa
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


---

**Attempting to separate inflection and derivation using vector space representations**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Approach</th>
<th>Evaluation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>