Improving Verb Metaphor Detection by Propagating Abstractness to Words, Phrases and Individual Senses
Overview
In this talk ...

- Abstractness and Concreteness
- Metaphor detection using abstractness
- Learning abstractness norms for:
  - Single words (comparison of approaches)
  - Phrases
  - Individual senses
In this talk ...

- Abstractness and Concreteness
  - Metaphor detection using abstractness
- Learning abstractness norms for:
  - Single words (comparison of approaches)
  - Phrases
  - Individual senses
Abstractness & Concreteness

- Concrete words refer to things, events, and properties that we can perceive directly with our senses 🌨️。
- **Abstract**: anxiety, happiness, democracy, opportunity
- **Concrete**: flower, stone, cinnamon, wall ...
Obtaining Ratings

- Dataset, created by manual annotation

<table>
<thead>
<tr>
<th>word</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>spaghetti</td>
<td>5.0</td>
</tr>
<tr>
<td>green</td>
<td>4.0</td>
</tr>
<tr>
<td>fitness</td>
<td>3.1</td>
</tr>
<tr>
<td>scheme</td>
<td>2.4</td>
</tr>
<tr>
<td>logic</td>
<td>1.7</td>
</tr>
<tr>
<td>belief</td>
<td>1.1</td>
</tr>
</tbody>
</table>

- Brysbaert et al. (2014): 40 000 words
- MRC Psycholinguistic Database: ≈8 200 words
- Extension of manually collected ratings with supervised learning techniques.
- Turney et al. (2011) : 114 501 words
Metaphor Detection

• Example 1:

The detective **digs up** enough evidence about the crime.

• Example 2:

The dog **digs up** a small bone in our garden.

• Features

  • Rating of the verbs **subject**
  • Rating of the verbs **object**
  • Average rating of **all nouns**
  • Average rating of **all proper names**
  • Average rating of **all verbs**, excluding the target verb
  • Average rating of **all adjectives**
  • Average rating of **all adverbs**
Metaphor Detection

• Example 1:

The **detective** **digs up** enough **evidence** about the **crime**.

• Example 2:

The **dog** **digs up** a small **bone** in our **garden**.

• Features
  - Rating of the verbs subject
  - Rating of the verbs object
  - **Average rating of all nouns**
  - Average rating of all proper names
  - Average rating of all verbs, excluding the target verb
  - Average rating of all adjectives
  - Average rating of all adverbs
Metaphor Detection

- **Example 1 = metaphorical**
  
  The detective digs up enough evidence about the crime.

- **Example 2 = literal**
  
  The dog digs up a small bone in our garden.
Propagate abstractness to all words for which we have vector space representation

- **Propagate abstractness**

- **Compare supervised methods:**
  - Linear regression (L-Reg), used by Tsvetkov et al. (2014)
  - Regression forest (Reg-F)
  - Feed forward neural network (NN)

- **Experiment: comparison of methods**
Learning Abstractness Results

<table>
<thead>
<tr>
<th></th>
<th>T&amp;L 03</th>
<th>L-Reg.</th>
<th>Reg-F.</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove50</td>
<td>.76</td>
<td>.76</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td>Glove100</td>
<td>.80</td>
<td>.79</td>
<td>.79</td>
<td>.85</td>
</tr>
<tr>
<td>Glove200</td>
<td>.78</td>
<td>.78</td>
<td>.76</td>
<td>.84</td>
</tr>
<tr>
<td>Glove300</td>
<td>.76</td>
<td>.78</td>
<td>.74</td>
<td>.85</td>
</tr>
<tr>
<td>W2V300</td>
<td>.83</td>
<td>.84</td>
<td>.79</td>
<td>.90</td>
</tr>
</tbody>
</table>

- Split Brysbaert ratings into 20% test, 80% train
  (small part of train used for tuning)
- Results show Spearman’s $\rho$ against human ratings
## Learning Abstractness Results

<table>
<thead>
<tr>
<th></th>
<th>T&amp;L 03</th>
<th>L-Reg.</th>
<th>Reg-F.</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glove50</td>
<td>.76</td>
<td>.76</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td>Glove100</td>
<td>.80</td>
<td>.79</td>
<td>.79</td>
<td>.85</td>
</tr>
<tr>
<td>Glove200</td>
<td>.78</td>
<td>.78</td>
<td>.76</td>
<td>.84</td>
</tr>
<tr>
<td>Glove300</td>
<td>.76</td>
<td>.78</td>
<td>.74</td>
<td>.85</td>
</tr>
<tr>
<td>W2V300</td>
<td>.83</td>
<td>.84</td>
<td>.79</td>
<td>.90</td>
</tr>
</tbody>
</table>

- Split Brysbaert ratings into 20% test, 80% train (small part of train used for tuning)
- Results show Spearman’s $\rho$ against human ratings
- Applied best setup to create a resource of automatically created ratings for 3,000,000 English words and multi-words (available!)
Abstractness for Individual Senses
“Today we’re tackling sloth”, said the student.
Abstractness & Metaphor Detection

“Today we’re tackling sloth”, said the student.
"Today we’re tackling sloth", said the student.
"Today we’re tackling sloth", said the student.

- Words are ambiguous
  - fan – supporter/ventilator
  - outlet – socket/escape
  - latex – rubber latex/LATEX

- Assumption: Sense specific ratings are more accurate
  - → improve token-based verb metaphor classification
Learning Multi-Sense Embeddings

- Many approaches for learning sense embeddings
- Pelevina et al. (2016) “Making Sense of Word Embeddings”
  - Learn single sense embeddings \((\text{word2vec})\)
  - Cluster based on WordSimilarity Graph
  - Pooling of Word Vectors = multi sense embeddings
Learning Multi-Sense Embeddings

- Many approaches for learning sense embeddings
- Pelevina et al. (2016) “Making Sense of Word Embeddings”
  - Learn single sense embeddings (word2vec)
  - Cluster based on WordSimilarity Graph
  - Pooling of Word Vectors = multi sense embeddings
- Apply previous setting: W2V300 vectors → Multi-Sense + Brysbaert ratings + NN method
- Sense embeddings “live” in the same space as single sense embeddings
Multi-Sense Embeddings & Sense Specific Abstractness Ratings

- Compare each sense embedding with each context word and pick the one with the highest cosine similarity
- Example

“Today we’re tackling sloth”, said the student.
Multi-Sense Embeddings & Sense Specific Abstractness Ratings

- Compare each sense embedding with each context word and pick the one with the highest cosine similarity

- Example sloth sense-1, rating: 6.7 (C)

"Today we’re tackling sloth", said the student.
Multi-Sense Embeddings & Sense Specific Abstractness Ratings

- Compare each sense embedding with each context word and pick the one with the highest cosine similarity
- Example sloth sense-2, rating: 0.4 (A)

"Today we’re tackling sloth", said the student.
Multi-Sense Embeddings & Sense Specific Abstractness Ratings

- Compare each sense embedding with each context word and pick the one with the highest cosine similarity
- VU Amsterdam Metaphor Corpus (VUA) 2,174 verbs and 23,113 sentences
- TroFi metaphor dataset 50 verbs 3,737 labeled sentences
- 10-fold Cross Validation
Results

<table>
<thead>
<tr>
<th>Feat.</th>
<th>TroFi(_{10F})</th>
<th>VUA(_{10F})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1S</td>
<td>.72</td>
<td>.42</td>
</tr>
<tr>
<td>MS</td>
<td><strong>.74</strong></td>
<td><strong>.44</strong>*</td>
</tr>
</tbody>
</table>

- F1 Metaphor Class
- Classifier: (balanced) Logistic Regression
- Ratings obtain state of the art results
  - (VUA\(_{test}\)) Klebanov et al. (2016)
- Effect vanishes when adding target verb itself as feature