Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics

Angeliki Lazaridou  Marco Marelli  Roberto Zamparelli  Marco Baroni

University of Trento

ACL 2013
Making sense of the title

Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics
Distributional Semantic Models (DSMs)

- A representation of meaning based on the *Distributional Hypothesis* ...

  *curtains open and the moon shining in on the barely ars and the cold, close moon*”. And neither of the w rough the night with the moon shining so brightly, it made in the light of the moon. It all boils down, wr surely under a crescent moon, thrilled by ice-white sun, the seasons of the moon? None, alone, Jay plam is dazzling snow, the moon has risen full and cold un and the temple of the moon, driving out of the hug in the dark and now the moon rises, full and amber a bird on the shape of the moon over the trees in front. But I could n’t see the moon or the stars, only the rning, with a sliver of moon hanging among the stars they love the sun, the moon and the stars. None of

- ...useful for...

  ![Graph showing nearest neighbors](image)

  **Nearest Neighbors (NNs)**
  - star
  - sun
  - Armstrong
  - space
Introduction

Distributional Semantic Models (DSMs)

- A representation of meaning based on the *Distributional Hypothesis* ...

he curtains open and the moon shining in on the barely ars and the *cold*, close moon " And neither of the w rough the night with the moon shining so brightly, it made in the light of the moon. It all boils down, we surely under a crescent moon, thrilled by ice-white sun, the *seasons* of the moon? Home, alone, Jay pla m is dazzling snow, the moon has risen full and cold un and the temple of the moon, driving out of the hug in the dark and now the moon rises, full and amber a bird on the shape of the moon over the trees in front. But I could n’t see the moon or the *stars*, only the rning, with a sliver of moon hanging among the *stars* they love the sun, the moon and the *stars*. None of

- ...useful for...

<table>
<thead>
<tr>
<th></th>
<th>planet</th>
<th>night</th>
<th>full</th>
<th>shadow</th>
<th>shine</th>
<th>crescent</th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>10</td>
<td>22</td>
<td>43</td>
<td>16</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>sun</td>
<td>14</td>
<td>10</td>
<td>4</td>
<td>15</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

- But *word frequency* affects *vector quality* [Bullinaria and Levy, 2007]
Introduction

Making sense of the title

Compositional-ly Derived Representations of
Morphologically Complex Words
in Distributional Semantics
Introduction

Morphologically Complex Words

- high productivity of word derivations
  - complex word = stem and affix
  - e.g. rebuild, colorful, frequently ...
- 55% of the words in CELEX Lexical Database are morphologically complex
- BAD NEWS: **Sparsity**
  - lexicalizable and affixless do not occur at all in our large corpus!
Morphologically Complex Words

- high productivity of word derivations
  - complex word = stem and affix
  - e.g. rebuild, colorful, frequently ...
- 55% of the words in CELEX Lexical Database are morphologically complex
- BAD NEWS: Sparsity
  - lexicalizable and affixless do not occur at all in our large corpus!
- GOOD NEWS: 80% of times in our large corpus, the stem appears more frequently than the complex word!
Making sense of the title

Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics
Compositional DSMs (cDSMs)

- Increasing need to represent meaning **beyond words** (e.g. paraphrase detection, etc)
- Productivity of syntax causes **sparsity**, which does not allow to collect co-occurrence statistics for whole sentences
- Solution à la Frege...

...operationalized in DSM with different **composition functions** on vectors for **phrases**

cDSMs to the rescue for morphology

- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?

The meaning of *complex word* is a function of the meaning of its *morphemes* and their composition rules. .
cDSMs to the rescue for morphology

- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?

The meaning of **complex word** is a function of the meaning of its **morphemes** and their composition rules....

- **affixless** means **without affix**
cDSMs to the rescue for morphology

- **Sparsity** is not a new issue for DSMs...
  - ..remember when scaling up to phrases!
- cDSMs can address productivity-caused sparsity in morphology just like they do in syntax...but how?

Our proposal: \( \text{affixless} = f_{\text{composition function}}(\text{affix,less}) \)

The meaning of complex word is a function of the meaning of its morphemes and their composition rules ....

affixless means without affix
Morphological Data

- CELEX English Lexical Database
  - Rich morphological annotations
- Extract stem/derived pairs
  - e.g. valid/invalid, frequent/frequently, run/rerun etc
- 18 derivational affixes
  - 3 prefixes: in, re, un
  - 15 suffixes: al, er, ful, ic, ion, ist, ity, ize, less, ly, ment, ness, ous, y
- Report results on 50 stem/derived pairs (900 in total)
Representing corpus-extracted vectors

- **stems** and derived words:
  - collect co-occurrence statistics from 2-word windows around each target item from a large *large*\textsuperscript{1} corpus

- **affixes** as vectors
  - accumulate context vectors of derived words with the same affix
  - e.g. \( \vec{less} = \vec{effortless} + \vec{colorless} + \ldots \)

Compositional Models for Derivational Morphology

Derive representation of complex word by composing its stem and affix

<table>
<thead>
<tr>
<th>Model</th>
<th>Composition function</th>
<th>By Who?</th>
</tr>
</thead>
<tbody>
<tr>
<td>weighted additive</td>
<td>$w_1 \vec{stem} + w_2 \vec{affix}$</td>
<td>[Mitchell and Lapata, 2010]</td>
</tr>
<tr>
<td>multiplicative</td>
<td>$\vec{stem} \odot \vec{affix}$</td>
<td>[Mitchell and Lapata, 2010]</td>
</tr>
<tr>
<td>dilation</td>
<td>$</td>
<td></td>
</tr>
<tr>
<td>full additive</td>
<td>$W_1 \vec{stem} + W_2 \vec{affix}$</td>
<td>[Guevara, 2010]</td>
</tr>
<tr>
<td>lexical function</td>
<td>$A_{affix} \vec{stem}$</td>
<td>[Baroni and Zamparelli, 2010]</td>
</tr>
</tbody>
</table>

Lazaridou et al. (University of Trento)
NON-Compositional models for Derivational Morphology

- **corpus**: How problem is currently been handled in DSMs i.e. it isn’t!!
  - ignore morphology and extract representation of complex word from the corpus
    - **frequently**: 😊
    - **affixless**: 😞
- **stem**: back off to the meaning of the stem
  - **build** instead of **rebuild**: 😊
  - **cycle** instead of **recycle**: 😞
NON-Compositional models for Derivational Morphology

- **corpus**: How problem is currently been handled in DSMs i.e. it isn’t!! ignore morphology and extract representation of complex word from the corpus
  - *frequently* 😊
  - *affixless* 😞

- **stem**: back off to the meaning of the *stem*
  - *build* instead of *rebuild* 😊
  - *cycle* instead of *recycle* 😞

- Introduce 2 annotations to dichotomize dataset and test our intuitions
Dichotomy of dataset based on high OR low vector quality

- Approximate quality of a corpus-extracted vector by the quality of its nearest neighbors (NNs)
  - **car**: automobile, truck, drive 😊
  - **river**: potato, school, t-shirt 😞
- Relatedness judgments\(^2\) to classify items as high or low quality

<table>
<thead>
<tr>
<th>Affix</th>
<th>Type</th>
<th>Derived form</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ist</td>
<td>High</td>
<td>transcendentalist</td>
<td>mythologist, futurist, theosophist Harrod, wholesaler, stockist</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>florist</td>
<td></td>
</tr>
<tr>
<td>in-</td>
<td>High</td>
<td>inaccurate</td>
<td>misleading, incorrect, erroneous metastasis, colorectal, biopsy</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>inoperable</td>
<td></td>
</tr>
<tr>
<td>re-</td>
<td>High</td>
<td>recapture</td>
<td>retake, besiege, capture</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>rename</td>
<td>defunct, officially, merge</td>
</tr>
</tbody>
</table>

\(^2\)http://crowdflower.com/
Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. `in-` (valid-invalid)
  - ...simply change the POS-tag of the stem, e.g. `-ly` (sad-sadly)

- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems

- Examples:
  - **High**: compulsory/compulsorily, chaos/chaotic
  - **Low**: cycle/recycle, believable/unbelievable
Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. in- (valid-invalid)
  - ...simply change the POS-tag of the stem, e.g. -ly (sad-sadly)
- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems
- Examples:
  - **High**: compulsory/compulsorily, chaos/chaotic
  - **Low**: cycle/recycle, believable/unbelievable
- What is causing the low judgements?:
  - opacity
Dichotomization of dataset based on high OR low stem/derived relatedness

- Morphological processes can...
  - ...completely change the meaning of a stem, e.g. in- (valid-invalid)
  - ...simply change the POS-tag of the stem, e.g. -ly (sad-sadly)
- Stem/Derived Relatedness judgments to classify derived words as having high or low relatedness to their stems
- Examples:
  - **High**: compulsory/compulsorily, chaos/chaotic
  - **Low**: cycle/recycle, believable/unbelievable
- What is causing the low judgements?
  - opacity
  - affixes triggering antonymy
Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the model-generated ones. Focus only on **high quality** complex words.

<table>
<thead>
<tr>
<th></th>
<th>stem</th>
<th>mult</th>
<th>dil.</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.47</td>
<td>0.39</td>
<td>0.48</td>
<td>0.50</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>High Relatedness</td>
<td>0.52</td>
<td>0.43</td>
<td>0.53</td>
<td>0.55</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>Low Relatedness</td>
<td>0.32</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.41</td>
<td>0.42</td>
</tr>
</tbody>
</table>

- Overall composition models apart from **mult** outperform **stem**
  - **fulladd** and **lexfunc** significantly outperform **stem**
Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the model-generated ones. Focus only on high quality complex words.

<table>
<thead>
<tr>
<th></th>
<th>stem</th>
<th>mult</th>
<th>dil.</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.47</td>
<td>0.39</td>
<td>0.48</td>
<td>0.50</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>High Relatedness</td>
<td>0.52</td>
<td>0.43</td>
<td>0.53</td>
<td>0.55</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>Low Relatedness</td>
<td>0.32</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.41</td>
<td>0.42</td>
</tr>
</tbody>
</table>

- Overall composition models apart from mult outperform stem
- fulladd and lexfunc significantly outperform stem
- Surprisingly, even when meanings of stem/derived are very close!!
Experiment 1: Is composition better than backing off to stem meaning?

- Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the model-generated ones. Focus only on high quality complex words.

<table>
<thead>
<tr>
<th></th>
<th>stem</th>
<th>mult</th>
<th>dil.</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.47</td>
<td>0.39</td>
<td>0.48</td>
<td>0.50</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>High Relatedness</td>
<td>0.52</td>
<td>0.43</td>
<td>0.53</td>
<td>0.55</td>
<td>0.61</td>
<td>0.58</td>
</tr>
<tr>
<td>Low Relatedness</td>
<td>0.32</td>
<td>0.28</td>
<td>0.33</td>
<td>0.38</td>
<td>0.41</td>
<td>0.42</td>
</tr>
</tbody>
</table>

- Overall composition models apart from mult outperform stem
- fulladd and lexfunc significantly outperform stem
- Surprisingly, even when meanings of stem/derived are very close!!
- stem is a particularly bad surrogate when meanings of stem/derived are not close
- fulladd, lexfunc, and to a lesser extent wadd, can capture affix-triggered semantics
Experiment 2: Focus on high-quality antonym complex words

Need to use corpus-extracted representations as a gold standard to compute cosine similarity between those and the model-generated ones. Focus only on **high quality** complex words.

<table>
<thead>
<tr>
<th></th>
<th>stem</th>
<th>mult</th>
<th>dil</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>-less</td>
<td>0.22</td>
<td>0.23</td>
<td>0.24</td>
<td>0.30</td>
<td>0.38</td>
<td><strong>0.44</strong></td>
</tr>
<tr>
<td>in-</td>
<td>0.39</td>
<td>0.34</td>
<td>0.40</td>
<td>0.45</td>
<td><strong>0.47</strong></td>
<td>0.45</td>
</tr>
<tr>
<td>un-</td>
<td>0.33</td>
<td>0.33</td>
<td>0.34</td>
<td>0.41</td>
<td>0.44</td>
<td><strong>0.46</strong></td>
</tr>
</tbody>
</table>

Encouraging results for modeling negation in DSM and cDSMs [Mohammad et al., 2013, Preller and Sadrzadeh, 2011]

- modeling inoperable is close to modeling syntactic negation compositionally, i.e. **not** operable
Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on **low quality** complex words
- Evaluate quality of **model**-generated vector representations.
  - Higher scores = better quality

<table>
<thead>
<tr>
<th></th>
<th>corpus</th>
<th>stem</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2.28</td>
<td>3.26</td>
<td>4.12</td>
<td>3.99</td>
<td>3.09</td>
</tr>
<tr>
<td>HR</td>
<td>2.29</td>
<td>3.56</td>
<td>4.48</td>
<td>4.31</td>
<td>3.31</td>
</tr>
<tr>
<td>LR</td>
<td>2.22</td>
<td>2.48</td>
<td>3.14</td>
<td>3.12</td>
<td>2.52</td>
</tr>
</tbody>
</table>

- **lexfunc** is significantly outperformed by **wadd** and **fulladd**.
  - **fulladd** and **wadd** feature almost always the stem as **NN**
- Stemploitation stategy for **fulladd** and **wadd**?
Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on **low quality** complex words
- Evaluate quality of **model**-generated vector representations.
  - Higher scores = better quality

<table>
<thead>
<tr>
<th></th>
<th>corpus</th>
<th>stem</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2.28</td>
<td>3.26</td>
<td>4.12</td>
<td>3.99</td>
<td>3.09</td>
</tr>
<tr>
<td>HR</td>
<td>2.29</td>
<td>3.56</td>
<td>4.48</td>
<td>4.31</td>
<td>3.31</td>
</tr>
<tr>
<td>LR</td>
<td>2.22</td>
<td>2.48</td>
<td>3.14</td>
<td>3.12</td>
<td>2.52</td>
</tr>
</tbody>
</table>

- **lexfunc** is significantly outperformed by **wadd** and **fulladd**.
  - **fulladd** and **wadd** feature almost always the stem as **NN**
- Stemploitation stategy for **fulladd** and **wadd**? **NO**
  - Both models over **All** significantly outperform **stem**
Experiment 3: Is compositionality able to overcome the sparsity problem?

- Scenario: Complex words appear in the corpus but with low frequency → low frequency might result in low quality
  - Focus only on low quality complex words
- Evaluate quality of model-generated vector representations.
  - Higher scores = better quality

<table>
<thead>
<tr>
<th></th>
<th>corpus</th>
<th>stem</th>
<th>wadd</th>
<th>fulladd</th>
<th>lexfunc</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>2.28</td>
<td>3.26</td>
<td>4.12</td>
<td>3.99</td>
<td>3.09</td>
</tr>
<tr>
<td>HR</td>
<td>2.29</td>
<td>3.56</td>
<td>4.48</td>
<td>4.31</td>
<td>3.31</td>
</tr>
<tr>
<td>LR</td>
<td>2.22</td>
<td>2.48</td>
<td>3.14</td>
<td>3.12</td>
<td>2.52</td>
</tr>
</tbody>
</table>

- lexfunc is significantly outperformed by wadd and fulladd.
  - fulladd and wadd feature almost always the stem as NN
- Stemploitation stategy for fulladd and wadd? NO
  - Both models overAll significantly outperform stem
  - In LR, corpus is significantly outperformed by both models, but not stem
## Experiment 3: Qualitative analysis of NNs

<table>
<thead>
<tr>
<th>Target</th>
<th>Model</th>
<th>Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>florist</td>
<td><strong>wadd</strong></td>
<td><strong>flora</strong>, fauna, ecosystem</td>
</tr>
<tr>
<td></td>
<td><strong>fulladd</strong></td>
<td><strong>flora</strong>, fauna, ecologist</td>
</tr>
<tr>
<td></td>
<td><strong>lexfunc</strong></td>
<td>ornithologist, naturalist, botanist</td>
</tr>
<tr>
<td></td>
<td><strong>corpus</strong></td>
<td>Harrod, wholesaler, stockist</td>
</tr>
<tr>
<td>inoperable</td>
<td><strong>wadd</strong></td>
<td><strong>operable</strong>, palliation, biopsy</td>
</tr>
<tr>
<td></td>
<td><strong>fulladd</strong></td>
<td><strong>operable</strong>, inoperative, ventilator</td>
</tr>
<tr>
<td></td>
<td><strong>lexfunc</strong></td>
<td>inoperative, unavoidably, flaw</td>
</tr>
<tr>
<td></td>
<td><strong>corpus</strong></td>
<td>metastasis, colorectal, biopsy</td>
</tr>
<tr>
<td>rename</td>
<td><strong>wadd</strong></td>
<td><strong>name</strong>, later, namesake</td>
</tr>
<tr>
<td></td>
<td><strong>fulladd</strong></td>
<td><strong>name</strong>, namesake, later</td>
</tr>
<tr>
<td></td>
<td><strong>lexfunc</strong></td>
<td>temporarily, reinstate, thereafter</td>
</tr>
<tr>
<td></td>
<td><strong>corpus</strong></td>
<td>defunct, officially, merge</td>
</tr>
</tbody>
</table>
Conclusions

- First work to test validity of compositionality principle for generating representations of complex words.
- Morpheme composition can improve on corpus-based representations when the latter are low-quality (e.g., because of sparsity problems).
- Morpheme composition outperforms stem baseline, even when similarity between stem and derived is high.
- Finally, new evaluation task for compositional approaches, in which fulladd emerged as the best model.
Future Work

- Model jointly morphological induction and composition
- Try to “decompose” frequent complex words to derive the representations of their less frequent stems.
- Test compositionality principle for inflectional morphology
- Test recursive morpheme composition (e.g. unoperable), especially useful for agglutinative languages
Thank you for your attention!

The implementation is based on the DISSECT toolkit, demo at the ACL demo session this year!
http://clic.cimec.unitn.it/composes/toolkit
Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space.
In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Mathematical foundations for a compositional distributional model of meaning.

A regression model of adjective-noun compositionality in distributional semantics.

Vector-based models of semantic composition.
In *Proceedings of ACL*, pages 236–244, Columbus, OH.

Composition in distributional models of semantics.

Computing lexical contrast.
*Computational Linguistics.*
In press.
Bell states and negative sentences in the distributed model of meaning.

Semantic compositionality through recursive matrix-vector spaces.