Measuring the compositionality of collocations via word co-occurrence vectors

Shared task system description

Alfredo Maldonado-Guerra    Martin Emms

School of Computer Science and Statistics
Trinity College Dublin
Ireland

DISCo 2011: Distributional Semantics and Compositionality Workshop at ACL/HLT 2011
Outline

1. Introduction
2. System description
3. Results and discussion
4. Future work
Introduction

- A shared task system that measures the compositionality of bigrams

Basic intuition

A highly compositional bigram would tend to have a considerable semantic overlap with its constituents whereas a bigram with low compositionality would share little semantic content with its constituents.

- Intuition operationalised via three configurations that exploit cosine similarity measures to detect the semantic overlap between the bigram and its constituents

- Fully unsupervised system that could be employed for any language, including under-resourced languages
This work uses vectors as defined by Schütze (1998):

- **Word (co-occurrence) vector** $W(w)$:
  - Counts words that co-occur with target word $w$ in corpus
  - 20 word window centred at target word

- **Second-order context vector** $C^2(p)$:
  - Sum of word vectors of words co-occurring with target word at position $p$ in corpus.
  - 20 word window centred at target word

In these vectors the simplest approach possible was used: no normalisation, no weighting, etc. → just counts
We assume each bigram is made up of a headword and a modifier.

<table>
<thead>
<tr>
<th>Type</th>
<th>Headword</th>
<th>Modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-N</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>S-V</td>
<td>V</td>
<td>S</td>
</tr>
<tr>
<td>V-O</td>
<td>V</td>
<td>O</td>
</tr>
</tbody>
</table>
System description

Three configurations:

- Two configurations that use cosine similarity measures in two different ways (configurations 1 and 2)

- One configuration that attempts to address the issue of polysemy (configuration 3)
Conf 1: Average of cosine similarity measures

Build word vectors for:

- Modifier $W(x)$
- Headword $W(y)$
- Bigram $W(xy)$

Compositionality score for Configuration 1

$$c_1 = \frac{1}{2} \left[ \cos(W(xy), W(x)) + \cos(W(xy), W(y)) \right]$$ (1)
Conf 2: Headword in bigram vs not in bigram

In this configuration we want to look at:

- Contexts where the modifier and the headword form a bigram
- Contexts where the headword occurs but does not form a bigram with the modifier

<table>
<thead>
<tr>
<th>red herring</th>
<th>Indeed, reflexive practice in the arts is a red herring, not because it doesn’t exist, but because all practice is inherently reflexive.</th>
</tr>
</thead>
<tbody>
<tr>
<td>red herring</td>
<td>Peterhead enjoys an increasingly important role in the trade of pelagic species of herring and mackerel, particularly with the processing plant at Albert Quay.</td>
</tr>
</tbody>
</table>

Sentences taken from the UK WaC corpus.
Conf 2: Headword in bigram vs not in bigram

Build word vectors for:

- Headword $y$ when forming a bigram with modifier $x$: $W^x(y)$

- Headword $y$ when not forming a bigram with modifier $x$: $W^{\bar{x}}(y)$

Compositionality score for Configuration 2

$$c_2 = \cos(W^x(y), W^{\bar{x}}(y))$$

(2)
Conf 3: Cluster potential bigram senses

**Intuition**

Different senses of a bigram might have different degrees of compositionality. E.g.:

1. Two cans of soup for the price of one is such a great deal!
2. The tsunami caused a great deal of damage to the country’s infrastructure.
Conf 3: Cluster potential bigram senses

- Cluster occurrences of headword $y$, modifier $x$ and bigram $xy$ via second-order context vectors

- Each cluster could represent a different sense of the bigram

- If we knew what cluster represents the bigram sense seen by human annotators, we could compute compositionality score from the sub-corpus represented by that cluster only.
Conf 3: Cluster potential bigram senses

But since we do not know what sense is used, we choose to compute the compositionality score as a weighted average from each cluster → a polysemy-enhanced version of Conf 1

For each cluster $k$ build the word vectors:
- $W_k(x, y)$ for the bigram
- $W_k(x)$ for the modifier
- $W_k(y)$ for the headword

Compositionality score for Configuration 3

$$c_3 = \sum_{k=1}^{K} \frac{\|k\|}{N} \frac{1}{2} \left[ \cos(W_k(x, y), W_k(x)) + \cos(W_k(x, y), W_k(y)) \right]$$

where $\|k\|$ is the number of contexts in cluster $k$ and $N$ is the total number of contexts across all clusters.
## Results and discussion

<table>
<thead>
<tr>
<th>C</th>
<th>Average diffs (numeric)</th>
<th>Precision (coarse)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ALL</td>
<td>A-N</td>
</tr>
<tr>
<td>1</td>
<td>17.95</td>
<td>18.56</td>
</tr>
<tr>
<td>2</td>
<td>18.35</td>
<td>19.62</td>
</tr>
<tr>
<td>3</td>
<td>25.59</td>
<td>24.16</td>
</tr>
<tr>
<td>R</td>
<td>32.82</td>
<td>34.57</td>
</tr>
</tbody>
</table>

- Conf1 and Conf 2 show very similar performance
- Disappointingly, Conf 3 — the polysemy enhanced version of conf 1 — did much **worse**
- S-V came out worse than A-N and V-O
Results and discussion

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>A-N</th>
<th>S-V</th>
<th>V-O</th>
<th>ALL</th>
<th>A-N</th>
<th>S-V</th>
<th>V-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold Standard</td>
<td>16.86</td>
<td>17.73</td>
<td>15.54</td>
<td>16.52</td>
<td>58.5</td>
<td>65.4</td>
<td>34.6</td>
<td>65.0</td>
</tr>
</tbody>
</table>

Alfredo Maldonado-Guerra, Martin Emms

Compositionality via word vectors
Future work

- Investigate effects of weighting schemes (IDF and others)
- Similarity measures other than cosine
- Further research into the role played by context in determining the compositionality of a bigram
  - In configuration 2, involve modifier in computation of compositionality score
  - In configuration 3, create separate clustering spaces for bigram, headword and modifier
- Explore other ways of clustering
Thank you for your attention! Questions?

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College Dublin.
References I

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta.
2009.
The WaCky wide web: a collection of very large linguistically processed web-crawled corpora.

Chris Biemann and Eugenie Giesbrecht.
2011.
Distributional Semantics and Compositionality 2011: Shared Task Description and Results.
References II


References


Amruta Purandare and Ted Pedersen.
2004.
Word sense discrimination by clustering contexts in vector and similarity spaces.

Hinrich Schütze.
1998.
Automatic word sense discrimination.
Appendix A: Preliminary definitions

Definitions

First-order context vector

\[ C^1(p)(w) = \sum_{p' \neq p \atop p-10 \leq p' \leq p+10} (1 \text{ if } w = \text{doc}(p'), \text{ else } 0) \]  \hspace{1cm} (4)

Word (co-occurrence) vector

\[ W(w) = \sum_p (1 \text{ if } w = \text{doc}(p), \text{ else } 0) \cdot C^1(p) \]  \hspace{1cm} (5)
Appendix A: Preliminary definitions

Definitions

Second-order context vector

\[ C^2(p) = \sum_{\substack{p' \neq p \\ p-10 \leq p' \\ p' \leq p+10}} W(doc(p)) \]  

Vectors based on work by Schütze (1998)
Appendix A: Preliminary definitions

Generalisation to MWEs:

Single token: *make*

<table>
<thead>
<tr>
<th>They</th>
<th>will</th>
<th>make</th>
<th>a</th>
<th>decision</th>
<th>based</th>
<th>on</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p-2$</td>
<td>$p-1$</td>
<td>$p$</td>
<td>$p+1$</td>
<td>$p+2$</td>
<td>$p+3$</td>
<td>$p+4$</td>
<td>…</td>
</tr>
</tbody>
</table>

MWE: *make decision*

<table>
<thead>
<tr>
<th>They</th>
<th>will</th>
<th>make a decision</th>
<th>based</th>
<th>on</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p-2$</td>
<td>$p-1$</td>
<td>$p$</td>
<td>$p+1$</td>
<td>$p+2$</td>
<td>…</td>
</tr>
</tbody>
</table>

• Up to 3 intervening words allowed.

Alfredo Maldonado-Guerra, Martin Emms  Compositionality via word vectors
Appendix A: Preliminary definitions

- Similarity measure between vectors done via cosine, defined in the standard way:

$$\cos(v, w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$ (7)
Appendix A: Conf 2 Word vector definitions

Definitions

Headword vector forming a bigram with $x$:

$$W^x(y) = \sum_p (1 \text{ if } \text{doc}(p) = y, \text{coll}(p, x), \text{else } 0) \cdot C^1(p)$$ (8)

Headword vector not forming a bigram with $x$:

$$W^\bar{x}(y) = \sum_p (1 \text{ if } \text{doc}(p) = y, \text{coll}(p, x), \text{else } 0) \cdot C^1(p)$$ (9)

where $y$ is the headword and $\text{coll}(p)$ is a Boolean function that determines whether the word at position $p$ forms a bigram with modifier $x$. 
Appendix B: Results and conclusion

Table: Some corpus statistics: the number of matched bigrams per subtype (Instances) and the average number of intervening words per subtype (Avg intervening).

<table>
<thead>
<tr>
<th></th>
<th>A-N</th>
<th>S-V</th>
<th>V-O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>177,254</td>
<td>11,092</td>
<td>121,317</td>
</tr>
<tr>
<td>Avg intervening</td>
<td>0.0684</td>
<td>0.3867</td>
<td>0.4612</td>
</tr>
</tbody>
</table>

Table: A few bigram examples provided by organisers.

- Digital radio future lie add value
- Small island government intend address issue
- Hard copy business need help children
- Black hole event occur raise bar