Semantic Compositionality/decomposability and Idiomaticity

Timothy Baldwin and Dominic Widdows
INTRODUCTION
Semantic Decomposability and Compositionality

• Decomposability = degree to which the semantics of an MWE can be ascribed to those of its parts

  kick the bucket → die'

  spill the beans → reveal'(secret')

• Compositionality = degree to which the semantics of the parts of an MWE contribute towards those of the whole
Syntactic Compositionality

- **Degree to which the syntactic properties of the parts of an MWE combine to make up the syntax of the whole**
  - Fixed expressions: *by and large*, *San Francisco*
  - Verb particles: *eat up*, *look up*

- **Syntactic compositionality** binary effect; non-compositional MWEs lexicalised

- **Semantic decomposability** continuum of regularity with more subtle effects and syntactic corollaries
Decomposability and Syntactic Flexibility

• Consider:

*the bucket was kicked by Kim
Strings were pulled to get Sandy the job.
The FBI kept closer tabs on Kim than they kept on Sandy.
... the considerable advantage that was taken of the situation

• The syntactic flexibility of an idiom can generally be explained in terms of its decomposability
Ideal Research Objective

- Automatically decompose a given MWE/align component words with semantic primitives
- Classification of MWEs into 3 classes:
  1. non-decomposable MWEs (e.g. kick the bucket, shoot the breeze, hot dog)
  2. idiosyncratically decomposable MWEs (e.g. spill the beans, let the cat out of the bag, radar footprint)
  3. simple decomposable MWEs (e.g. kindle excitement, traffic light)
Realistic Short-term Objective

- Demarcate simple decomposable MWEs from idiosyncratically decomposable and non-decomposable MWEs (roughly equivalent to endocentric vs. exocentric distinction)

- Binary distinction vs. mapping onto continuum of relative decomposability
Procedural Definition of Decomposability

- In simple decomposable MWEs, the parts are similar to the whole

  \( \text{e.g. } \text{sim}(\text{house boat}, \text{house}) \) and \( \text{sim}(\text{house boat}, \text{boat}) \) both high
A Statistical Approach to the Semantics of Verb-particles

Dekang Lin
ACL 1999
Basic Method

• Use substitution as a test of compositionality:

  \[\text{red tape} \rightarrow \text{yellow tape, red cassette}\]
  \[\text{economic impact} \rightarrow \text{political impact, economic effect}\]

• Evaluate based on a dictionary of idioms
System Resources

- POS-conditioned thesaurus (nouns, verbs, adjectives/adverbs)
  - derived from dependency data (Minipar):
- Collocation data
  - dependency tuples \((H, R, M)\) with high log-likelihood ratio \((H = \text{head}, R = \text{relation}, M = \text{modifier})\)
(Point-wise) Mutual Information

- Measure of the level of association between two events $A$ and $B$:

$$MI(A, B) = \log_2 \frac{P(A, B)}{P(A)P(B)}$$

- Commonly used in collocation extraction
- Not appropriate for low-frequency events
Mutual Information and Compositionality

- Scaling up to 3 events $A$, $B$ and $C$, where $B$ and $C$ are conditionally independent given $A$:

$$MI(A, B, C) = \log_2 \frac{P(A, B, C)}{P(B|A)P(C|A)P(A)}$$

$$MI(H, R, M) = \log_2 \frac{|H \ R \ M|}{|* \ * \ *|}$$

$$= \log_2 \frac{|H \ R \ *| \ |* \ R \ M| \ |* \ R \ *|}{|H \ R \ *| \ |* \ R \ M|}$$

$$= \log_2 \frac{|H \ R \ M| \ |* \ R \ *|}{|H \ R \ *| \ |* \ R \ M|}$$

21 November, 2003
Definition of Compositionality

• A phrase $\alpha$ is non-compositional iff there is no $\beta$ s.t.:
  
  (a) $\beta$ can be produced by substitution of the components of $\alpha$ for any of 10 most-similar words, and
  
  (b) there is an overlap between the 95% confidence interval of the MI values of $\alpha$ and $\beta$

• 10 most-similar words tested for each of $H$ and $M$ ($R$ fixed)

\[
sim(w_1, w_2) = \frac{\sum (r,w) \in T(w_1) \cap T(w_2) MI(w_1, r, w) + MI(w_2, r, w)}{\sum (r,w) \in T(w_1) MI(w_1, r, w) + \sum (r,w) \in T(w_2) MI(w_2, r, w)}
\]
Example 1: *spill (one’s) guts*

- **(spill, V:comp1:N, gut):**
  - *spill*: leak, pour, spew, ..., spray
  - *gut*: intestine, instinct, foresight, ..., charisma

- Check for each of (leak, V:comp1:N, gut), (spill, V:comp1:N, gut), ...
  in the collocation database

- None found, so *spill (one’s) guts* is non-compositional
Example 2: *red tape*

- \((tape,N:adj:N,red)\):
  - *tape*: videotape, cassette, videocassette, ..., audio
  - *red*: yellow, purple, pink, ..., shade

- Find \((tape,N:adj:N,yellow)\), \((tape,N:adj:N,orange)\), \((tape,N:adj:N,black)\) in the collocation database but with very different MI values

- *red tape* is non-compositional
MI Confidence Interval: the Z-test

- Possible to calculate the “true” MI of (H,R,M) according to the Z-test:

\[
\bar{p} \pm z_N \sqrt{\frac{p(1-p)}{n}} = \frac{k}{n} \pm z_N \sqrt{\frac{k(1-k)}{n}} \approx \frac{k \pm z_N \sqrt{k}}{n}
\]

where \(\bar{p}\) is the MLE of \(p\), \(n\) is \(|* * *|\), \(k\) is \(|H R M|\), and \(z_N\) is a constant determined by the confidence level \(N\), e.g. \(z_{0.95} = 1.96\)
Applying the Z-test

- Determine the “fit” between two MI values by calculating the Z-score interval for the putative non-compositional MWE and determining whether the MI of the second falls into that interval.
Evaluation

• Evaluate the method relative to an idiom dictionary

• OK precision, and significant numbers of the extracted MWEs not contained in the dictionary appear to be non-compositional based on manual inspection
Reflections

• Is substitution really a good test for non-compositionality?
  ✴ institutionalised phrases: *frying pan, salt and pepper, many thanks*
  ✴ productive MWEs: *call/phone/ring up*

• Look to alternative methods
A Statistical Approach to the Semantics of Verb-particles

Colin Bannard, Timothy Baldwin and Alex Lascarides

ACL 2003 Workshop on MWE
Basic Method

- Define similarity in terms of distributional similarity, i.e. assume that if an MWE is compositional, it will occur in the same lexical context as its parts

- Divide up compositionality to look at verb and particle similarity independently

- Evaluate against human judgements
Verb-particle constructions (VPCs)

- **VPC** = A verb plus one or more obligatory (prepositional) particles

  *Peter put the picture up*
  *Susan finished up her paper*
  *Philip gunned down the intruder*
  *Barbara and Simon made out*

- **Assumption:** VPCs are not always fully compositional or fully non-compositional, but rather populate a continuum between the two extremes
Compositionality by Entailment

- Peter put the picture up
  \[ \Rightarrow \text{Peter put the picture somewhere} \]
  \[ \Rightarrow \text{the picture was up} \]

- Susan finished up her paper
  \[ \Rightarrow \text{Susan finished her paper} \]

- Philip gunned down the intruder
  \[ \Rightarrow \text{the intruder was down} \]

- Barbara and Simon made out
Obtaining Human Judgements

1. Extract VPCs from British National Corpus (Baldwin and Villavicencio, 2002)

2. Randomly select 5 sentence tokens for each of 40 randomly selected VPC types

3. Present native English speakers with tokens and asked whether verb and/or particle is implied by the VPC

4. Response: Yes, No or Don’t Know
Example Sentence Tokens: *round up*

A dog started to *round up* sheep.

In three years they had *rounded up* fifty captive orangs.

*Owned by Jo Rutherford, Trigo rounded up the milking herd and brought it back to the milking parlour in Devon.*

*On 9 August, 349 Arrests were made as the military swooped to *round up* serving and former IRA activists.*

*Ten days later, when the agents moved in to *round up* their targets, El-Jorr checked out and returned to Cyprus, charging the hotel bill to his American Express card as instructed.*
Human Judgements

- Does *round up* imply *round*?
- Does *round up* imply *up*?
- Obtain gold-standard analysis by taking majority judgement (ignore *Don’t Know* responses)
## Sample Judgements

<table>
<thead>
<tr>
<th>VPC</th>
<th>Component word</th>
<th>Yes</th>
<th>No</th>
<th>Don’t Know</th>
</tr>
</thead>
<tbody>
<tr>
<td>dig up</td>
<td>dig</td>
<td>21</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>up</td>
<td>18</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>stay up</td>
<td>stay</td>
<td>20</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>up</td>
<td>21</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>brighten up</td>
<td>brighten</td>
<td>9</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>up</td>
<td>16</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>add up</td>
<td>add</td>
<td>12</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>up</td>
<td>19</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

* [http://www.language-experiments.org/](http://www.language-experiments.org/)

---

21 November, 2003
Binary Classification Tasks ($\times 4$)

1. Is the item completely compositional?

2. Does the item include at least one item that is compositional?

3. Does the verb contribute its simplex meaning?

4. Does the particle contribute its simplex meaning?
Classification Methods

- Four different classifiers implemented
- Method 1 based on Lin (1999), Methods 2-4 address theoretical concerns with this model
- All methods based on co-occurrence vector representation of VPC and component words
Method 1

- Reimplemented Lin (1999) over VPCs.
- Tested over all four tasks - assuming that the substitutability of each item will give us some insight into its semantic contribution.
- Reconstruct Lin’s thesaurus to include all of verbs, nouns, adjectives/adverbs and prepositions.
Method 2

- Similar to Lin (1999) except for use of knowledge-free approach to obtaining thesaurus
- Very similar to Schütze (1998) “context space” method
- Similarities from pairwise comparison of all verbs, particles and VPCs
- Obtain thesaurus by taking the $N$ most similar words of a given POS
Method 3

• Use same method of substitution

• A component is said to be contributing simplex meaning if expression formed by substitution occurs among the nearest 100 verb-particle constructions
Method 4

- Hypothesis is that if a verb or particle is contributing simplex meaning to a VPC then it will be semantically similar to the VPC according to cosine measure
  - a verb is judged to be contributing simplex meaning if it occurs within the 20 most similar items to the VPC.
  - a particle is judged to be contributing simplex meaning if it is in top 10 most similar items to the VPC.
## Task 4 Results

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>.750</td>
<td>.750</td>
<td>1.000</td>
<td>.857</td>
</tr>
<tr>
<td>Method 1</td>
<td>.425</td>
<td>.818</td>
<td>.300</td>
<td>.442</td>
</tr>
<tr>
<td>Method 2</td>
<td>.425</td>
<td>.818</td>
<td>.300</td>
<td>.442</td>
</tr>
<tr>
<td>Method 3</td>
<td>.425</td>
<td>.769</td>
<td>.333</td>
<td>.480</td>
</tr>
<tr>
<td>Method 4</td>
<td>.725</td>
<td>.758</td>
<td>.833</td>
<td>.793</td>
</tr>
</tbody>
</table>

**Baseline:** classify according to the most frequent class
Task 4: Does the particle in the item contribute its simplex meaning?
# Task 3 Results

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>.525</td>
<td>.525</td>
<td>1.000</td>
<td>.690</td>
</tr>
<tr>
<td><strong>Method 1</strong></td>
<td>.450</td>
<td>.474</td>
<td>.429</td>
<td>.450</td>
</tr>
<tr>
<td><strong>Method 2</strong></td>
<td>.600</td>
<td>.608</td>
<td>.666</td>
<td>.639</td>
</tr>
<tr>
<td><strong>Method 3</strong></td>
<td>.525</td>
<td>.531</td>
<td>.810</td>
<td>.641</td>
</tr>
<tr>
<td><strong>Method 4</strong></td>
<td>.550</td>
<td>.666</td>
<td>.286</td>
<td>.400</td>
</tr>
</tbody>
</table>
Task 3: Does the verb in the item contribute its simplex meaning?

21 November, 2003
## Task 2 Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.750</td>
<td>.750</td>
<td>1.000</td>
<td>.860</td>
</tr>
<tr>
<td>Method 1</td>
<td>.550</td>
<td>.731</td>
<td>.633</td>
<td>.678</td>
</tr>
<tr>
<td>Method 2</td>
<td>.600</td>
<td>.769</td>
<td>.667</td>
<td>.717</td>
</tr>
<tr>
<td>Method 3</td>
<td>.700</td>
<td>.765</td>
<td>.866</td>
<td>.810</td>
</tr>
<tr>
<td>Method 4</td>
<td>.725</td>
<td>.771</td>
<td>.900</td>
<td>.830</td>
</tr>
</tbody>
</table>
Task 2: Does the item include at least one item that is compositional?
### Task 1 Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.525</td>
<td>0.525</td>
<td>1.000</td>
<td>0.680</td>
</tr>
<tr>
<td>Method 1</td>
<td>0.575</td>
<td>0.577</td>
<td>0.714</td>
<td>0.638</td>
</tr>
<tr>
<td>Method 2</td>
<td>0.575</td>
<td>0.575</td>
<td>0.714</td>
<td>0.638</td>
</tr>
<tr>
<td>Method 3</td>
<td>0.575</td>
<td>0.558</td>
<td>0.905</td>
<td>0.690</td>
</tr>
<tr>
<td>Method 4</td>
<td>0.500</td>
<td>0.514</td>
<td>0.857</td>
<td>0.642</td>
</tr>
</tbody>
</table>
Task 1: Is the item completely compositional?
Reflections

• Difficult task!

• Methods 3 and 4 generally perform the best

• Question mark over appropriateness of substitution test for VPCs

• Possibility of mapping human judgements onto continuous scale of compositionality?
Detecting a Continuum of Compositionality in Phrasal Verbs

Diana McCarthy, Bill Keller and John Carroll

ACL 2003 Workshop on MWE
Basic Method

• Compare different methods for modelling compositionality based on distributional similarity and statistical tests traditionally used in collocation extraction

• Map 111 VPCs onto a ranked list, based on human judgements over an 11-point compositionality scale

• Evaluate according to the rank order correlation with the gold-standard ranked list
System Resources

- Build thesaurus a la Lin (1998), based on dependency tuple output of RASP
Similarity-based Methods

- **overlap**: relative overlap between the top $N$ neighbours of the VPC and its simplex verb

- **sameparticle**: the number of VPCs which select for the same particle as the given VPC amongst the top $N$ neighbours of that VPC

- **sameparticle – simplex**: the value for **sameparticle** minus the number of top $N$ neighbours of the simplex verb which select for that same particle
• **simplexasneighbour**: does the simplex verb occur in the top 50 neighbours of the VPC?

• **rankofsimplex**: what is the rank of the simplex verb in the neighbours of the VPC?

• **overlaps**: the overlap of neighbours in the top $N$ neighbours of the VPC and simplex verb, where VPC neighbours are converted to simplex verbs in the VPC case
Statistical Methods

- $\chi^2$
- Log-likelihood ratio
- (Point-wise) mutual information
- Simple frequency of the VPC
- Simple frequency of the simplex verb
Resource-based Method

- Binary test for the occurrence of the VPC in:
  - WordNet
  - Alvey Tools (ANLT) VPC data
  - Alvey Tools (ANLT) prepositional verb data
Correlation with the Gold-standard Data

- For binary tests (simplexasneighbour, WordNet, ANLT), use the Mann-Whitney U test (rank sum test)
- For other methods, map each output value onto a rank and apply the Spearman Rank Correlation test (rank test)
- In each case, calculate the Z score and the probability of the null hypothesis (i.e. no correlation between the output of the method and the gold-standard data)
Results

• **same particle** — **simplex** best performer of similarity-based methods

• MI best performer of statistical tests

• Question of how to apply the results to the proposed task of lexical acquisition?
Overall Reflections

- Promising results observed for detecting compositionality/decomposability, but less so for determining the semantic contribution of individual words in an overall MWE.

- What about MWEs where the simplex words don’t occur with that same POS (e.g., *chicken out*).

- Effects of polysemy (e.g., *run down, run over*).

- How to move on to actually semantically decompose an MWE?