Compositionality Detection using a Corpus Driven Approach: How to distinguish "couch potato" from "roast potato"

Siva Reddy\textsuperscript{1,2}, Suresh Manandhar\textsuperscript{1}, Diana McCarthy\textsuperscript{2}

\textsuperscript{1} Artificial Intelligence Group, Department of Computer Science, University of York, UK
\textsuperscript{2} Lexical Computing Ltd., UK

2\textsuperscript{nd} International Sketch Engine Workshop
March 16 2011
Outline

1. Compositionality
2. Goal of this work
3. Background
4. Our Approach
5. Evaluation
Multi-word

- A sequence of two or more words describing a meaning together.
- Compound Nouns
  - credit card
  - leather jacket
- Phrasal Verbs
  - look up
  - get over
- Idiomatic expressions
  - kick the bucket
  - spill the beans
Given meanings of

- couch
- roast
- potato
Compositionality

Given meanings of

- couch
- roast
- potato

Can we interpret the meanings of

- couch potato
- roast potato
Couch Potato

Compositionality Detection using Sketch Engine
Compositionality

Compositional Multi-words

- \( m(\text{"A B"}) = m(\text{A}) \oplus m(\text{B}) \)
- e.g. water tank, post man, roast potato
- caution: subjective task
- cracking “\( \oplus \)” is a miracle
Compositionality

Compositional Multi-words

- \( m(\text{"A B"}) = m(\text{A}) \oplus m(\text{B}) \)
- e.g. water tank, post man, roast potato
- caution: subjective task
- cracking "⊕" is a miracle

Non-Compositional multi-words

- think tank, smoking gun, apple polisher, couch potato
Importance of compositionality detection

Dictionary Building: Lexicography and Terminology
- A good dictionary
  - includes non-compositional multi-words
  - does not have redundant information

Machine Translation
- goose egg $\neq$ Gänseei
- goose egg $\rightarrow$ unwichtig
- Compositionality detection in a given context. Much harder.

Word Tokenization
- Search engines
Goal of this work

**Goal:** Identify compositional and non-compositional multi-words for a given language

- My focus is on *compound nouns*
- A sequence of nouns is treated as a *multi-word*
- Vast research on identifying multi-words but not on compositionality detection
Goal of this work

Goal: Identify compositional and non-compositional multi-words for a given language

- My focus is on compound nouns
- A sequence of nouns is treated as a multi-word
- Vast research on identifying multi-words but not on compositionality detection

Unsupervised corpus-based approach

- Huge corpora for many languages available in Sketch Engine
- Sketch Engine provides additional resources like word sketches, distributional thesaurus
Background: Semantics from corpus

Distributional Hypothesis (Harris, 1954)
Words that occur in similar contexts tend to have similar meanings
Background: Semantics from corpus

Distributional Hypothesis (Harris, 1954)
Words that occur in similar contexts tend to have similar meanings

- e.g. Tree and Plant, Tea and Coffee, Bus and Vehicle
Background: Semantics from corpus

Distributional Hypothesis (Harris, 1954)
Words that occur in similar contexts tend to have similar meanings

- e.g. Tree and Plant, Tea and Coffee, Bus and Vehicle

Other variations: (Firth, 1957)
You shall know a word by the company it keeps
Background: Semantics from corpus

Distributional Hypothesis (Harris, 1954)
Words that occur in similar contexts tend to have similar meanings

- e.g. Tree and Plant, Tea and Coffee, Bus and Vehicle

Other variations: (Firth, 1957)
You shall know a word by the company it keeps

Bag of words hypothesis
Two documents tend to be similar if they have similar distribution of similar words
Interpret semantics using VSM

- Backbone: Distributional Hypothesis
Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
- Text entity (we are interested in) as a Vector (point) in dimensional space.
- Context of the entity as dimensions
Vector Space Models (VSMs) of Semantics

- **Interpret semantics using VSM**
  - Backbone: Distributional Hypothesis
  - Text entity (we are interested in) as a Vector (point) in dimensional space.
  - Context of the entity as dimensions
  - Existing methods represent knowledge in VSMs mainly in three types (Turney and Pantel, 2010)
    - term-document
    - term-context
    - pair-pattern
**Term-Document: (Salton et al., 1975)**

Create a word-by-document matrix

<table>
<thead>
<tr>
<th></th>
<th>d1</th>
<th>d2</th>
<th>d3</th>
<th>d4</th>
<th>d5</th>
<th>d6</th>
<th>d7</th>
<th>d8</th>
<th>d9</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>survey</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>trees</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>minors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**d1:** Human machine **interface** for Lab ABC **computer** applications

---

1 Image courtesy: (Landauer et al., 1998)

S. Reddy, S. Manandhar & D. McCarthy

Compositionality Detection using Sketch Engine
Document similarity can be found using Cosine similarity:

\[
\text{sim}(D_1, D_2) = \frac{D_1 \cdot D_2}{\|D_1\| \|D_2\|}
\]

\(^2\)Image courtesy: (Salton et al., 1975)
Term-Document: (Salton et al., 1975)

Document similarity can be found using Cosine similarity

\[ \text{sim}(D1, D2) = \frac{D1 \cdot D2}{\|D1\| \|D2\|} \]

---

2 Image courtesy: (Salton et al., 1975)
**Term-Context: Word Space Model**

<table>
<thead>
<tr>
<th></th>
<th>buy</th>
<th>rent</th>
<th>sell</th>
<th>book</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td>50</td>
<td>60</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>apartment</td>
<td>60</td>
<td>100</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>room</td>
<td>0</td>
<td>40</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>suite</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>80</td>
</tr>
</tbody>
</table>

Words are represented as a vector built from context words:

- I *rent a house*.
- I *bought an apartment*.
- I *booked a room*. 
Semantics of larger entities

*How to interpret semantics of larger entities?*
Semantics of larger entities

How to interpret semantics of larger entities?

The distributional way

<table>
<thead>
<tr>
<th></th>
<th>turn</th>
<th>photon</th>
<th>sign</th>
<th>noise</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Traffic</strong></td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td><strong>Light</strong></td>
<td>2</td>
<td>15</td>
<td>3</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>(TrafficLight_{Dist})</td>
<td>10</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>
Semantics of larger entities

How to interpret semantics of larger entities?

The distributional way

<table>
<thead>
<tr>
<th></th>
<th>turn</th>
<th>photon</th>
<th>sign</th>
<th>noise</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Light</td>
<td>2</td>
<td>15</td>
<td>3</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>TrafficLight_{Dist}</td>
<td>10</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

How to interpret semantics of larger entities from its constituents?
Semantics of larger entities

How to interpret semantics of larger entities?

The distributional way

<table>
<thead>
<tr>
<th></th>
<th>turn</th>
<th>photon</th>
<th>sign</th>
<th>noise</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Light</td>
<td>2</td>
<td>15</td>
<td>3</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>TrafficLight_Dist</td>
<td>10</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>10</td>
</tr>
</tbody>
</table>

How to interpret semantics of larger entities from its constituents?

The Principle of Compositionality:(Partee et al., 1990)

The meaning of a compound expression is a function of, and only of, the meaning of its parts and the way in which the parts are combined.
Idea exploited by many:

Two ways of computing a multi-word’s meaning

- Distributional way: The true meaning vector of the multi-word
- Compositionality function: A way of estimating the meaning vector of the multi-word from the meaning vectors of its parts
Idea exploited by many:

Two ways of computing a multi-word’s meaning

- Distributional way: The true meaning vector of the multi-word
- Compositionality function: A way of estimating the meaning vector of the multi-word from the meaning vectors of its parts

How similar are these two vectors?

- If they are very close, i.e. meaning of the multi-word can be computed from the meaning of the parts, the multi-word is compositional
- else non-compositional
Idea exploited by many:

Two ways of computing a multi-word’s meaning

- Distributional way: The true meaning vector of the multi-word
- Compositionality function: A way of estimating the meaning vector of the multi-word from the meaning vectors of its parts

How similar are these two vectors?

- If they are very close, i.e. meaning of the multi-word can be computed from the meaning of the parts, the multi-word is compositional
- Else non-compositional

Bingo!!
Category 1: Katz and Giesbrecht (2006); Giesbrecht (2009)

- Build
  - $CouchPotato_{Dist}$
  - $CouchPotato_{Comp}$ i.e. Couch $\oplus$ Potato

- $\text{sim} (CouchPotato_{Dist}, CouchPotato_{Comp})$
  - if $\text{sim} > \text{thrsh}$: multi-word is compositional
  - else: multi-word is non-compositional

Pitfalls observed:
- Threshold highly varies
- 48% accuracy
Category 1: Katz and Giesbrecht (2006); Giesbrecht (2009)

- **Build**
  - $CouchPotato_{Dist}$
  - $CouchPotato_{Comp}$ i.e. **Couch ⊕ Potato**
  - sim ($CouchPotato_{Dist}, CouchPotato_{Comp}$)
    - if sim > thrsh: multi-word is compositional
    - else: multi-word is non-compositional

- **Pitfalls observed:**
  - Threshold highly varies
  - 48 % accuracy
Category 2: (Baldwin et al., 2003; Bannard et al., 2003)

Build distributional vectors of

- $CouchPotato_{Dist}$
- $Couch$
- $Potato$

$\text{sim}(CouchPotato_{Dist}, Potato)$

- if $\text{sim} > \text{thrsh}$: multi-word is compositional
- else: multi-word is non-compositional
Background

Compositional Models

Category 2: (Baldwin et al., 2003; Bannard et al., 2003)

Build distributional vectors of

- $CouchPotato_{Dist}$
- $Couch$
- $Potato$

$sim(CouchPotato_{Dist}, Potato)$

- if $sim > thrsh$: multi-word is compositional
- else: multi-word is non-compositional

Pitfalls Observed:

- Was able to capture type-of relations only
- $Threshold$ highly varies
- Moderate results: 51% accuracy
Observations

- Threshold highly varies. Reported by everyone
- What might be the possible reasons?
Observations

- Threshold highly varies. Reported by everyone
- What might be the possible reasons?

Polysemy

- Skewed nature of senses
- \( \text{RiverBank}_{\text{dist}} \) is not similar to \( \text{Bank}_{\text{dist}} \)
- Either of the above approaches fail
- Focus of this presentation is to overcome this using Sketch Engine
Observations

- Threshold highly varies. Reported by everyone
- What might be the possible reasons?

Polysemy

- Skewed nature of senses
- $RiverBank_{dist}$ is not similar to $Bank_{dist}$
- Either of the above approaches fail
- Focus of this presentation is to overcome this using Sketch Engine

Continuum (McCarthy et al., 2003)

- There is no hard boundary to say if a multi-word is compositional
Concordance of Sketch Engine as Term-Context Matrix

the idea of a much worse prison; where No tabernacle-work over the stalls carved in a light and elegant manner. St. John’s, which light; so that, unless when looking at your
He half opened one of them, and as the light poured in, looked round with mournful
beauty heightened by the aid of brilliant lights, of costly jewels, and all the pride
use the cycle paths and have good bright lights, then you should have no problems. Bus
. I think it puts business in a very bad light. Alan Sugar does everyone a great disservice
framework, Tati became an influential guiding light for the generations of comedians and filmmakers
morning - it’s night It’s dark - it’s light It’s raining - it’s sunny life’s serious
or feeling low M Baird 167 The Northern lights and Mackie’s means home sweet home to
This investigation is intended to bring to light some reasons for connecting the notion
24 hours a day and the proprietors keep light security, particularly a local rent a cop
form with the only pleasantness being the light white fluffy foam of the recently sumped
I thought his material had all seen the light of day. TK’s mentor, Henry Stone sent
1973. I am amazed this has n’t seen the legal issues, that it would never see the
light of day. It is wonderful, and definitely
quite rightly so. Bright and breezy and light of day. Frank has taken the reins, as requested. â€œ Ideal as a clip-on book
cover at the end of the arm. The Flexi light up a few dancefloors as well as receiving
and clips in for compact storage. Flexi Light â€œ Reaches places other torches can
Light requires two AAA batteries (not included
Light FAQ’s: Q) Hi, what bulb should I use

Much powerful than conventional Term-Context matrix
Ignored fact: Words are polysemous

Current trend: Prototype Vectors
Currently most methods represent each word as a single vector i.e. a prototype vector for each word.

S. Reddy, S. Manandhar & D. McCarthy Compositionality Detection using Sketch Engine
Our Approach

Ignored fact: Words are polysemous

Current trend: Prototype Vectors
Currently most methods represent each word as a single vector i.e. a prototype vector for each word.

Light occur in many contexts
- Quantum theory, Optics, Bulbs and Traffic
- Not all contexts are relevant for building compositional vectors.
- Light is noisy $\rightarrow \text{TrafficLight}_{\text{Comp}}$ is noisy

S. Reddy, S. Manandhar & D. McCarthy Compositionality Detection using Sketch Engine
Our Approach

Ignored fact: Words are polysemous

Current trend: Prototype Vectors

Currently most methods represent each word as a single vector i.e. a prototype vector for each word.

Light occur in many contexts

- Quantum theory, Optics, Bulbs and Traffic
- Not all contexts are relevant for building compositional vectors.
- Light is noisy $\Rightarrow$ TrafficLight\textsubscript{Comp} is noisy

Exemplars of Light

'-interest-n': 1.0, 'round-n': 1.0, 'open-v': 1.0
'business-n': 1.0, 'bad-j': 1.0, 'put-v': 1.0
'framework-n': 1.0, 'generation-n': 1.0, 'technique-n': 1.0, 'follow-v': 1.0
'material-n': 1.0, 'day-n': 1.0, 'complete-j': 1.0
Proposed Solution

Prototype vectors are more noisy
A need for refined vectors
Proposed Solution

Prototype vectors are more noisy
A need for refined vectors

Exemplar-based Vector Space Model

- Select (examples) exemplars of *Light* which have similar context of *Traffic*
- Prunes out irrelevant exemplars
- Use selected exemplars to build *Light*_{Traffic}
- Motivated from the work of Erk and Pado (2010)
Proposed Solution

Prototype vectors are more noisy
A need for refined vectors

Exemplar-based Vector Space Model

- Select (examples) exemplars of Light which have similar context of Traffic
- Prunes out irrelevant exemplars
- Use selected exemplars to build $Light_{Traffic}$
- Motivated from the work of Erk and Pado (2010)

*How to select (examples) exemplars of Light which have similar context of Traffic??*
First order co-occurrences of *traffic* from Sketch Engine

<table>
<thead>
<tr>
<th><strong>object of</strong></th>
<th><strong>18950</strong></th>
<th><strong>1.4</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>divert</td>
<td>198</td>
<td>7.78</td>
</tr>
<tr>
<td>slow</td>
<td>212</td>
<td>7.58</td>
</tr>
<tr>
<td>block</td>
<td>319</td>
<td>7.5</td>
</tr>
<tr>
<td>generate</td>
<td>711</td>
<td>7.34</td>
</tr>
<tr>
<td>speed</td>
<td>149</td>
<td>7.25</td>
</tr>
<tr>
<td>motorise</td>
<td>93</td>
<td>7.17</td>
</tr>
<tr>
<td>encrypt</td>
<td>83</td>
<td>6.79</td>
</tr>
<tr>
<td>rout</td>
<td>78</td>
<td>6.76</td>
</tr>
<tr>
<td>direct</td>
<td>218</td>
<td>6.05</td>
</tr>
<tr>
<td>calm</td>
<td>75</td>
<td>6.5</td>
</tr>
<tr>
<td>congest</td>
<td>55</td>
<td>6.43</td>
</tr>
<tr>
<td>wheel</td>
<td>55</td>
<td>6.28</td>
</tr>
<tr>
<td>redirect</td>
<td>52</td>
<td>6.24</td>
</tr>
<tr>
<td>monitor</td>
<td>260</td>
<td>6.21</td>
</tr>
<tr>
<td>increase</td>
<td>985</td>
<td>6.18</td>
</tr>
<tr>
<td>drive</td>
<td>407</td>
<td>6.12</td>
</tr>
<tr>
<td>stop</td>
<td>330</td>
<td>6.1</td>
</tr>
<tr>
<td>reduce</td>
<td>698</td>
<td>6.08</td>
</tr>
<tr>
<td>induce</td>
<td>83</td>
<td>6.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>and/or</strong></th>
<th><strong>10005</strong></th>
<th><strong>0.8</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>pedestrian</td>
<td>197</td>
<td>7.56</td>
</tr>
<tr>
<td>congestion</td>
<td>134</td>
<td>6.92</td>
</tr>
<tr>
<td>pollution</td>
<td>168</td>
<td>6.42</td>
</tr>
<tr>
<td>noise</td>
<td>188</td>
<td>5.8</td>
</tr>
<tr>
<td>parking</td>
<td>173</td>
<td>5.57</td>
</tr>
<tr>
<td>freight</td>
<td>39</td>
<td>5.45</td>
</tr>
<tr>
<td>roadwork</td>
<td>14</td>
<td>5.26</td>
</tr>
<tr>
<td>traffic</td>
<td>190</td>
<td>5.17</td>
</tr>
<tr>
<td>transportation</td>
<td>32</td>
<td>5.09</td>
</tr>
<tr>
<td>highway</td>
<td>38</td>
<td>5.05</td>
</tr>
<tr>
<td>passenger</td>
<td>99</td>
<td>4.87</td>
</tr>
<tr>
<td>fume</td>
<td>14</td>
<td>4.73</td>
</tr>
<tr>
<td>pedestrianisation</td>
<td>8</td>
<td>4.63</td>
</tr>
<tr>
<td>lorry</td>
<td>20</td>
<td>4.53</td>
</tr>
<tr>
<td>commuter</td>
<td>13</td>
<td>4.49</td>
</tr>
<tr>
<td>motorway</td>
<td>20</td>
<td>4.38</td>
</tr>
<tr>
<td>transport</td>
<td>130</td>
<td>4.33</td>
</tr>
<tr>
<td>TCP</td>
<td>11</td>
<td>4.33</td>
</tr>
<tr>
<td>road</td>
<td>240</td>
<td>4.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>pp_along-i</strong></th>
<th><strong>95</strong></th>
<th><strong>5.2</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>road</td>
<td>25</td>
<td>0.92</td>
</tr>
<tr>
<td>route</td>
<td>14</td>
<td>0.83</td>
</tr>
<tr>
<td>street</td>
<td>9</td>
<td>0.62</td>
</tr>
<tr>
<td>乒乓</td>
<td>64</td>
<td>4.6</td>
</tr>
<tr>
<td>road</td>
<td>25</td>
<td>0.92</td>
</tr>
<tr>
<td>route</td>
<td>9</td>
<td>0.2</td>
</tr>
<tr>
<td>乒乓</td>
<td>54</td>
<td>4.4</td>
</tr>
<tr>
<td>road</td>
<td>31</td>
<td>1.23</td>
</tr>
<tr>
<td>乒乓</td>
<td>305</td>
<td>2.8</td>
</tr>
<tr>
<td>firewall</td>
<td>10</td>
<td>4.35</td>
</tr>
<tr>
<td>village</td>
<td>46</td>
<td>2.74</td>
</tr>
<tr>
<td>port</td>
<td>18</td>
<td>2.7</td>
</tr>
<tr>
<td>tunnel</td>
<td>8</td>
<td>2.69</td>
</tr>
<tr>
<td>Internet</td>
<td>482</td>
<td>5.83</td>
</tr>
<tr>
<td>coal</td>
<td>117</td>
<td>5.75</td>
</tr>
<tr>
<td>multicast</td>
<td>37</td>
<td>5.63</td>
</tr>
<tr>
<td>tourist</td>
<td>106</td>
<td>5.48</td>
</tr>
<tr>
<td>barge</td>
<td>40</td>
<td>5.27</td>
</tr>
<tr>
<td>data</td>
<td>143</td>
<td>5.25</td>
</tr>
<tr>
<td>container</td>
<td>72</td>
<td>5.23</td>
</tr>
</tbody>
</table>

*Word Sketch of traffic to select exemplars of light*
Our Approach

Similar Words to *Traffic* from Sketch Engine

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Score</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>transport</td>
<td>0.362</td>
<td>134717</td>
</tr>
<tr>
<td>road</td>
<td>0.339</td>
<td>324641</td>
</tr>
<tr>
<td>train</td>
<td>0.336</td>
<td>114514</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.331</td>
<td>160671</td>
</tr>
<tr>
<td>bus</td>
<td>0.322</td>
<td>131884</td>
</tr>
<tr>
<td>route</td>
<td>0.312</td>
<td>168121</td>
</tr>
<tr>
<td>network</td>
<td>0.311</td>
<td>262162</td>
</tr>
<tr>
<td>trade</td>
<td>0.308</td>
<td>165216</td>
</tr>
<tr>
<td>market</td>
<td>0.307</td>
<td>379175</td>
</tr>
<tr>
<td>travel</td>
<td>0.306</td>
<td>115459</td>
</tr>
<tr>
<td>communication</td>
<td>0.3</td>
<td>171501</td>
</tr>
<tr>
<td>flow</td>
<td>0.299</td>
<td>77846</td>
</tr>
<tr>
<td>station</td>
<td>0.295</td>
<td>175788</td>
</tr>
<tr>
<td>operation</td>
<td>0.295</td>
<td>198053</td>
</tr>
<tr>
<td>car</td>
<td>0.293</td>
<td>419404</td>
</tr>
<tr>
<td>access</td>
<td>0.289</td>
<td>376109</td>
</tr>
<tr>
<td>business</td>
<td>0.287</td>
<td>700710</td>
</tr>
<tr>
<td>speed</td>
<td>0.286</td>
<td>146544</td>
</tr>
<tr>
<td>sale</td>
<td>0.285</td>
<td>239489</td>
</tr>
</tbody>
</table>

*Not only context words of Traffic but also similar words to Traffic are useful [courtesy: Diana]*
Exemplar-based Composition

Exemplars of $Light_{Traffic}$

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0
'bright-j': 15.0, 'day-n': 15.0
Exemplar-based Composition

Exemplars of $Light_{Traffic}$

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0  
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0  
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0  
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0  
'bright-j': 15.0, 'day-n': 15.0

*Build vector of Light using the above exemplars: $Light_{Traffic}$*
Our Approach

Exemplar-based Composition

Exemplars of $\text{Light}_\text{Traffic}$

'speed-n': 4.0, 'create-v': 1.0, 'mass-n': 1.0
'road-n': 2.0, 'good-j': 1.0, 'white-j': 3.0
'street-n': 1.0, 'road-n': 2.0, 'limit-n': 1.0, 'sign-n': 1.0
'road-n': 2.0, 'side-n': 1.0, 'wrong-j': 1.0, 'drive-v': 1.0
'bright-j': 15.0, 'day-n': 15.0

Build vector of Light using the above exemplars: $\text{Light}_\text{Traffic}$

Exemplar-based Composition

$\text{TrafficLight}_{\text{Comp}} = \text{TrafficLight} \oplus \text{Light}_\text{Traffic}$
Traffic Light: Evaluation

- ukWaC in Sketch Engine
- Traffic: 97492 examples
- Light: 316133 examples
- $Traffic_{Light}^{Dist}$: 6730 examples
Traffic Light: Evaluation

- ukWaC in Sketch Engine
- Traffic: 97492 examples
- Light: 316133 examples
- TrafficLight$_{Dist}$: 6730 examples

Prototype-based Model

\[ \text{sim}(\text{TrafficLight}_{Dist}, \text{TrafficLight}_{Comp}^{Prt}) = 0.434 \]
Traffic Light: Evaluation

- ukWaC in Sketch Engine
- Traffic: 97492 examples
- Light: 316133 examples
- TrafficLight\textsubscript{Dist}: 6730 examples

Prototype-based Model

\[ \text{sim}(\text{TraficLight}_{\text{Dist}}, \text{TraficLight}_{\text{Comp}}^{\text{Prt}}) = 0.434 \]

Exemplar-based Model

- Traffic\textsubscript{Light}: 1949 exemplars
- Light\textsubscript{Trafic}: 6322 exemplars
- Just 1% of the examples

\[ \text{sim}(\text{TraficLight}_{\text{Dist}}, \text{TraficLight}_{\text{Comp}}^{\text{Exm}}) = 0.509 \]
Couch Potato: Evaluation

- Couch: 5103 examples
- Potato: 23277 examples
- *CouchPotato*\textsubscript{Dist}: 407 examples
Couch Potato: Evaluation

- Couch: 5103 examples
- Potato: 23277 examples
- \(\text{CouchPotato}_{\text{Dist}}\): 407 examples

Prototype-based Model

- \(\text{sim}(\text{CouchPotato}_{\text{Dist}}, \text{CouchPotato}_{\text{Comp}}^{\text{Prt}}) = 0.134\)
Couch Potato: Evaluation

- Couch: 5103 examples
- Potato: 23277 examples
- $CouchPotato_{Dist}$: 407 examples

Prototype-based Model

$$sim(CouchPotato_{Dist}, CouchPotato_{Comp}^{Prt}) = 0.134$$

Exemplar-based Model

- $Couch_{Potato}$: 41 exemplars
- $Potato_{Couch}$: 5 exemplars
- $$sim(CouchPotato_{Dist}, CouchPotato_{Comp}^{Exm}) = 0.005$$
Observations and Evaluation

Observations
Exemplar-based composition using Sketch Engine
- Rewards compositional multi-words
- Penalizes non-compositional multi-words
- This is what we want!!

Evaluation [along with Diana]
- 200 Mechanical turkers annotated 90 words for multi-word compositionality
- To be evaluated on this data
- More findings about the Principle of Compositionality
Other aspects

- Compositionality function and its parameters
- Parameter estimation
- Principle of Compositionality: Its pitfalls
- Anatomy aware model of compositionality detection
Summary

- Compositionality and its importance
- Vector Space Models for Compositionality Detection
- Polysemy is a major problem
- Exemplar-based VSM using Sketch Engine
Bibliography I


