From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions

Recent Developments in Computational Semantics, WS 2013/14

Susanne Fertmann

February 25th 2014
Motivation

• Drawing inferences is crucial for natural language understanding

  *people shopping in a supermarket*

• Distributional models are good at finding related words
• Difficult to capture entailment between complex expressions

  → New similarity measure: *Denotational similarity*
  → Use images as denotations
Outline

• Denotation

• Denotation graph
  – Graph generation
  – Reduction rules

• Experiments
  – Denotational similarity
  – Approximate entailment
  – Semantic textual similarity

• Summary
• Truth-conditional semantic theories:
  – Denotation = the set of all situations or possible worlds in which a sentence is true

• New Idea: **Images as Visual denotations:**
  – Denotation = the set of images a linguistic expression describes

\[ [s] = \{ \text{set of images } i \in U \mid \text{s is a truthful description of } i \} \]
• Truth-conditional semantic theories:
  – Denotation = the set of all situations or possible worlds in which a sentence is true

• New Idea: Images as Visual denotations:
  – Denotation = the set of images a linguistic expression describes

$$\left| [s] \right| = \{ \text{set of images } i \in \text{of } U \mid \text{s is a truthful description of } i \}$$

$$\left| [\text{A dog runs.}] \right| = \{ \text{images} \}$$
Use visual denotations to define new linguistic similarity measures

→ Denotation graph = subsumption hierarchy, where each node is a string and its visual denotation
Use visual denotations to define new linguistic similarity measures.

Denotation graph = subsumption hierarchy, where each node is a string and its visual denotation.

- 'animal running'
  - 'dog running'
    - 'dog running with ball'
    - 'dog running on beach'
  - 'animal running on beach'
    - 'dog running on beach'
    - ...
Data and Preprocessing

- Data: ~ 30,000 photographs of everyday situations, 5 captions each

- Spell checker, tokenizer, POS tagger, chunker and parser

- Own heuristics to adapt them to the specific domain
  - Systematic errors (climbs is never a noun)
  - Lexicon of common entity types (people, clothing, food)
  - Normalize spelling variations (barbecue, barbeque, BBQ)
  - Identify boundaries of complex NPs

- Hypernym Lexicon based on WordNet

- Normalization: drop punctuation, singular determiners, lemmatizing

- ...

Reduction Rules

- Drop Pre-Nominal Modifiers
  - *big red shirt* vs. *ice hockey player*

- Drop Other Modifiers
  - *run quickly* → *run*
  - *on sky* → *sky*

- Replace Nouns by Hypernyms
  - *poodle* → *dog*

- Handle Partitive NPs
  - *cup of tea* → *cup, tea*

- Handle VP-to-VP Cases
  - *jump to catch* vs. *wait to jump* vs. *seem to jump*

- Extract Simpler Constituents
  - *man laughs while drinking* → *man laugh* and *man drink*
Graph Generation

- Generation is top down
- Start at general root nodes, stop at nodes with one single denotation

1. Reduce each caption as far as possible to obtain a generic string
2. Use the generic strings as root nodes
3. As long as the string a 'describes' more than one caption/image: generate more specific strings
Graph Generation

1. Reduce each caption as far as possible to obtain a generic string:

   • 'A dog running on an empty beach.'
   • 'dog running on empty beach'
   • 'dog running on beach'
   • 'dog running'
   • 'animal running'
Graph Generation

1. Reduce each caption as far as possible to obtain a generic string:

- 'A dog running on an empty beach.'
- 'dog running on empty beach'
- 'dog running on beach'
- 'dog running'
- 'animal running'

'A dog running on an empty beach.'

'animal running'
Graph Generation

1. Reduce each caption as far as possible to obtain a generic string:
   'animal running'
2. Use the generic strings as root nodes
Graph Generation

1. Reduce each caption as far as possible to obtain a generic string
2. Use the generic strings as root nodes
3. As long as the string a 'describes' more than one caption/image:
   generate more specific strings

'animal running'
1. Reduce each caption as far as possible to obtain a generic string
2. Use the generic strings as root nodes
3. As long as the string a 'describes' more than one caption/image:
   generate more specific strings

• Generate new captions:
• 'A dog running on an empty beach.'
  → 'animal running on beach'
  → 'dog running'
The Denotation Graph

'animal running'

'dog running'  

'animal running on beach'

'animal running'
The Denotation Graph

- 'animal running'
  - 'dog running'
  - 'dog running with ball'
  - 'dog running on beach'
  - 'animal running on beach'

- 'animal running on beach'
  - ...
The Denotation Graph

'animal running'
'dog running'
'dog running with ball'
'Dog runs with a colored ball.'

'dog running'
'dog running on beach'
'A dog running on an empty beach.'

'animal running on beach'
'animal running on beach'
...
### Denotational Similarity

- Denotational pointwise mutual information: $nPMI_{[]}$

#### Denotational PMI similarity finds actions that are part of playing football

<table>
<thead>
<tr>
<th>$nPMI_{[]}$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.623</td>
<td>tackle person</td>
</tr>
<tr>
<td>0.597</td>
<td>hold football</td>
</tr>
<tr>
<td>0.545</td>
<td>run down field</td>
</tr>
<tr>
<td>0.519</td>
<td>wear white jersey</td>
</tr>
<tr>
<td>0.487</td>
<td>avoid</td>
</tr>
<tr>
<td></td>
<td>0.826</td>
</tr>
<tr>
<td></td>
<td>play game</td>
</tr>
<tr>
<td></td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>play rugby</td>
</tr>
<tr>
<td></td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>play soccer</td>
</tr>
<tr>
<td></td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>play on field</td>
</tr>
<tr>
<td></td>
<td>0.773</td>
</tr>
<tr>
<td></td>
<td>play ball</td>
</tr>
</tbody>
</table>

- The **compositional** $\Sigma$ similarity find events similar to playing football
- The **denotational** PMI similarity finds actions that are part of playing football
Approximate Entailment

- Similar to RTE
- Decision: Does $h$ describe the same image as the set of captions $P$?

<table>
<thead>
<tr>
<th>Premises</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man editing a black and white photo on a computer.</td>
<td>man sit</td>
</tr>
<tr>
<td>A man in a white shirt is working at a computer.</td>
<td></td>
</tr>
<tr>
<td>A guy in a white t-shirt on a mac computer.</td>
<td></td>
</tr>
<tr>
<td>A young man is using an apple computer.</td>
<td></td>
</tr>
</tbody>
</table>

- Data generation based on the denotation graph (~ 700,000 items)
- Hypotheses: short, represented by nodes S, SBJ, VP, V, OBJ
### Approximate Entailment

<table>
<thead>
<tr>
<th></th>
<th>VP</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>58.7</td>
<td>71.2</td>
</tr>
<tr>
<td>Best distributional (cosine)</td>
<td>71.9</td>
<td>78.9</td>
</tr>
<tr>
<td>Best compositional ((\prod, \Sigma))</td>
<td>72.7</td>
<td>79.6</td>
</tr>
<tr>
<td>Denotational nPMI_{[[]]}</td>
<td>74.9</td>
<td>80.2</td>
</tr>
<tr>
<td>Denotational P</td>
<td>73.8</td>
<td>79.5</td>
</tr>
<tr>
<td>Denotational (combined)</td>
<td><strong>75.5</strong></td>
<td><strong>81.2</strong></td>
</tr>
</tbody>
</table>

- The best denotational model outperforms distributional and compositional models
- For different varying hypothesis length
Semantic Textual Similarity

- 1500 sentence pairs from MSR Video Description Corpus
- Scores between 0 and 5 (equivalent to unrelated)
- DKPro = state of the art system (Bär et al, 2013)
- Add compositional and denotational similarity features

<table>
<thead>
<tr>
<th></th>
<th>$DKPro$</th>
<th>$+\sum, \Pi \text{ (img)}$</th>
<th>$+nPMI$</th>
<th>$+both$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson $r$</td>
<td>0.868</td>
<td>0.880</td>
<td>0.888</td>
<td>0.890</td>
</tr>
</tbody>
</table>
• Images used as visual denotations
• Denotation graph combined captions, generalised captions and images
• Define denotational measures of linguistic similarity
• These showed to be competitive with/slightly better than distributional similarities (for approximate entailment and semantic textual similarity)
References

• P. Young, A. Lai, M. Hodosh, J. Hockenmaier (2014): From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. To appear in TACL.

• Images:
  • http://www.horsenation.com/wp-content/uploads/2012/01/V3mcp.jpg
  • http://www.kimballstock.com/pix/DOG/02/DOG_02_KH0038_01_P.JPG
  • http://www.installitdirect.com/blog/is-artificial-grass-pet-friendly/
  • http://cache.desktopnexus.com/thumbnails/34599-bigthumbnail.jpg
Generic node 'animal running' describes more than one image:

- 'A dog running on an empty beach.'
- 'There is a horse running freely on the street.'
- 'Dog runs with a colored ball.'