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

Denotations

- 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

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

Denotations

- 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

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

|[A dog runs.]| = {

}

Denotation Graph

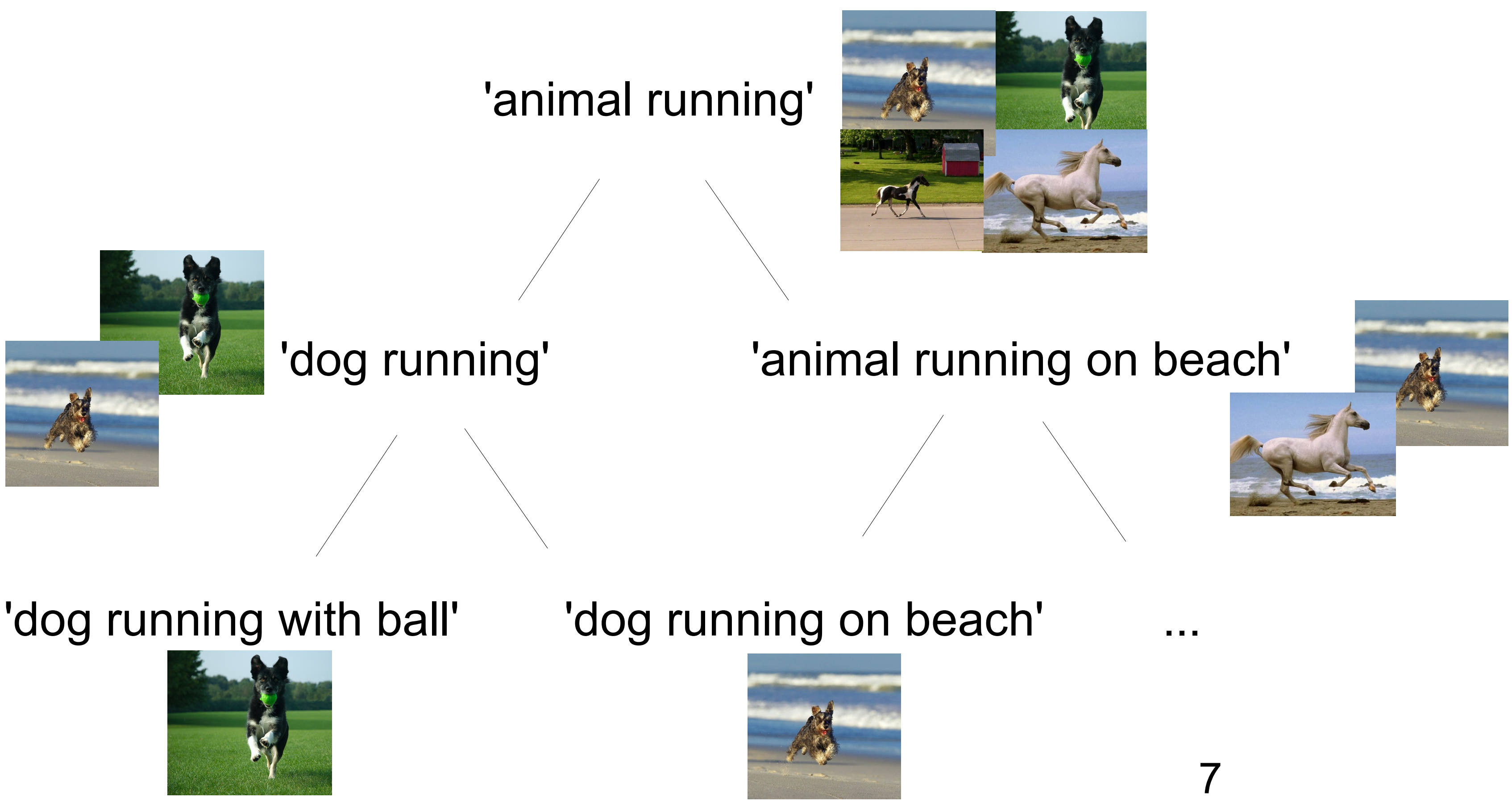
- Use visual denotations to define new linguistic similarity measures

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

Denotation Graph

- Use visual denotations to define new linguistic similarity measures

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



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 errorrs (*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:

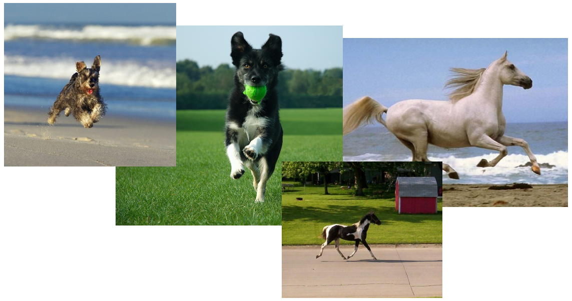
- **'A dog running on an empty beach.'**
- 'dog running on empty beach'
- 'dog running on beach'
- 'dog running'
- **'animal running'**

⟨ 'A dog running on an empty beach.'  , 'animal running' ⟩

Graph Generation

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

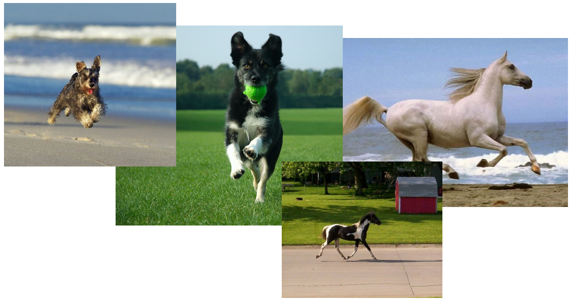
'animal running'



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'



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

- 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'

The Denotation Graph

'animal running'

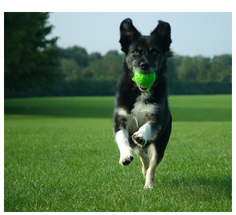
'dog running'

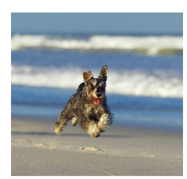
'animal running on beach'

'dog running with ball'

'dog running on beach'

...





The Denotation Graph

'animal running'

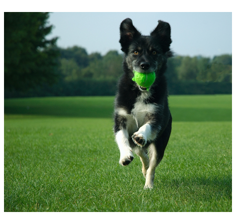
'dog running'

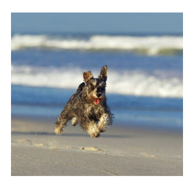
'animal running on beach'

'dog running with ball'

'dog running on beach'

...





'Dog runs with a colored ball.'

'A dog running on an empty beach.'

Denotational Similarity

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

play football

$nPMI_{[\![\,]\!]}$		Σ	
0.623	*tackle person*	0.826	*play game*
0.597	*hold football*	0.817	*play rugby*
0.545	*run down field*	0.811	*play soccer*
0.519	*wear white jersey*	0.796	*play on field*
0.487	*avoid*	0.773	*play ball*

- The **compositional** Σ 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**?

Premises:	A man editing a black and white photo on a computer.
	A man in a white shirt is working at a computer.
	A guy in a white t-shirt on a mac computer.
	A young man is using an apple computer.
Hypothesis:	man sit

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

Approximate Entailment

	VP	S
Bag of Words	58.7	71.2
Best distributional (cosine)	71.9	78.9
Best compositional (∏, ∑)	72.7	79.6
Denotational $nPMI_{[[\ ]]}$	74.9	80.2
Denotational P	73.8	79.5
Denotational (combined)	**75.5**	**81.2**

- 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 composistional and denotational similarity features

	$DKPro$	$+\Sigma, \Pi$ (img)	$+nPMI_{[\![]\!]}$	$+both$
Pearson r	0.868	0.880	0.888	0.890

Summary

- 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

Graph Generation

- Generic node 'animal running' describes more than one image:

⟨ 'A dog running on an empty beach.'  , 'animal running' ⟩

⟨ 'There is a horse running freely on the street.'  , 'animal running' ⟩

⟨ 'Dog runs with a colored ball.'  , 'animal running' ⟩