Entailment above the word level in distributional semantics

Marco Baroni
Raffaella Bernardi

Free University of Bozen-Bolzano

Ngoc-Quynh Do

Cornell University

Chung-chieh Shan

University of Trento

Outline

- Abstract
- Background
- Problem: Entailment
- Datasets
- Experiments
- Conclusion

Introduction

- Goal: detecting (phrase level) entailment using Distributional Semantics
- Corpus-harvested Datasets
- Experimental evidence

Background

Formal Semantics

- Sentence level
- Complex models of meaning
- Lack of resources

Background

Distributional semantics

- Large-scale
- Successful in many tasks
- Word level

Previous work

Two strands of research

- Model compositionality within DS

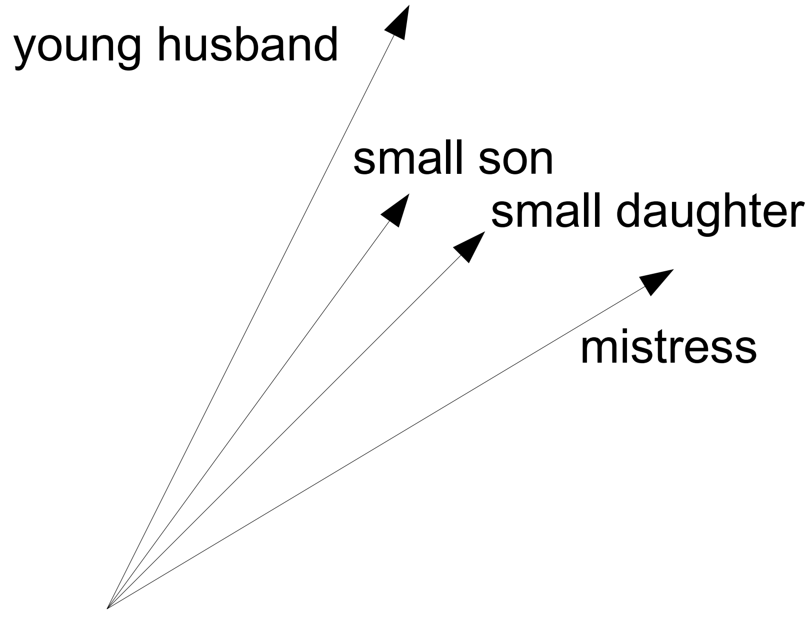
  (Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011; Guevara, 2010: Mitchell and Lapata, 2012)

- Reformulate logical inference (FS) in DS

  (Erk, 2009; Geffet and Dagan, 2005; Kotlerman et al., 2010)

DS above the word level

Harvesting AN is meaningful



Entailment

Entailment relation ($\models$) is a core notion of logic

$$A \models B$$

it cannot be that A is true and B is false

Entailment in FS

Sentences in FS denote a truth value

$$A \models_S B$$

it cannot be that A is true and B is false

Entailment in FS

Nouns in FS denote sets of entities

$$A \models_N B$$

inclusion relation between sets

Entailment in FS

Quantifiers in FS denote sets of sets of entities

$$A \models_{QP} B$$

inclusion relation between of sets of sets

Entailment in DS

Ability of one term to substitute for another

- *baseball* $\models$ *sport*
- Inclusion of contexts (or features)
- Asymmetric!
- Possible measure: balAPinc (Kotlerman et al. 2010)

Semantic space

- British National Corpus, WackyPedia and ukWaC
- Tokenized, lemmatized, POS-tagged
- Pointwise Mutual Information matrix
  - 48K rows (phrases of interest)

    N, AN, QP
  - 27K columns (content words)

AN $\models$ N dataset

big cat $\models$ *cat*

- Restrictive adjectives
- Nouns from BLESS dataset (not too polysemous)
- 1246 AN sequences for which AN $\models$ N holds
- Negative examples: $AN_1 \not\models N_2$

$N_1 \models N_2$ dataset

pope $\models$ *leader*

- WordNet nouns
- Hyponym-hypernym chains
- 1385 positive instances
- Negative examples by inversion

  leader $\not\models$ *pope*

$Q_1N \vDash Q_2N$ dataset

many dogs ⊨ several dogs

- 12 quantifiers:

  all, both, each, either, every, few, many, most, much, no, several, some

- 13 clear cases where $Q_1 \vDash Q_2$
- 17 clear cases where $Q_1 \nvDash Q_2$
- Cartesian product with WordNet nouns
- 7537 positive examples, 8455 negative examples

Classification methods

- balAPinc

  average two terms:

  APinc (Capture feature (context) inclusion)

  LIN (Capture relative relevance of features)

- SVM classifier (plus SVD)
- Two baselines (frequency and cosine distance)

First experiment

Generalizing from $AN \models N$ to $N_1 \models N_2$

- Training set $AN \models N$, test set $N_1 \models N_2$
- Tune balAPinc threshold
  - $\text{balAPinc}_{AN \models N}$
  - $\text{balAPinc}_{\text{upper}}$
- Train SVM classifier
  - $\text{SVM}_{AN \models N}$
  - $\text{SVM}_{\text{upper}}$

First experiment

Generalizing from $AN \models N$ to $N_1 \models N_2$

	P	R	F	Accuracy (95% C.I.)
$\text{SVM}_{\text{upper}}$	88.6	88.6	88.5	88.6 (87.3–89.7)
$\text{balAPinc}_{\text{AN} \models \text{N}}$	65.2	87.5	74.7	70.4 (68.7–72.1)
$\text{balAPinc}_{\text{upper}}$	64.4	90.0	75.1	70.1 (68.4–71.8)
$\text{SVM}_{\text{AN} \models \text{N}}$	69.3	69.3	69.3	69.3 (67.6–71.0)
$\cos(\text{N}_1, \text{N}_2)$	57.7	57.6	57.5	57.6 (55.8–59.5)
$\text{fq}(\text{N}_1) < \text{fq}(\text{N}_2)$	52.1	52.1	51.8	53.3 (51.4–55.2)

Second experiment

Generalizing QN $\models$ N entailment

- Hold out one quantifier pair as testing data
  - $\text{SVM}_{\text{pair-out}}$
- Hold out one of the 12 quantifiers
  - $\text{SVM}_{\text{quantifier-out}}$
- Ignore the nouns altogether
  - $\text{SVM}^{Q}_{\text{pair-out}}$
  - $\text{SVM}^{Q}_{\text{quantifier-out}}$
- Include classifiers from previous experiment

Second experiment

Generalizing QN $\models$ N entailment

	P	R	F	Accuracy (95% C.I.)
$\text{SVM}_{\text{pair-out}}$	76.7	77.0	76.8	78.1 (77.5–78.8)
$\text{SVM}_{\text{quantifier-out}}$	70.1	65.3	68.0	71.0 (70.3–71.7)
$\text{SVM}^{\text{Q}}_{\text{pair-out}}$	67.9	69.8	68.9	70.2 (69.5–70.9)
$\text{SVM}^{\text{Q}}_{\text{quantifier-out}}$	53.3	52.9	53.1	56.0 (55.2–56.8)
$\cos(\text{QN}_1, \text{QN}_2)$	52.9	52.3	52.3	53.1 (52.3–53.9)
$\text{balAPinc}_{\text{AN} \models \text{N}}$	46.7	5.6	10.0	52.5 (51.7–53.3)
$\text{SVM}_{\text{AN} \models \text{N}}$	2.8	42.9	5.2	52.4 (51.7–53.2)
$\text{fq}(\text{QN}_1) < \text{fq}(\text{QN}_2)$	51.0	47.4	49.1	50.2 (49.4–51.0)
$\text{balAPinc}_{\text{upper}}$	47.1	100	64.1	47.2 (46.4–47.9)

Conclusions

- Semantic vector representations of AN constructions encode entailment
- Semantic vectors of quantifiers also encode entailment
- QN entailment is different from feature inclusion