A Vector Space for Distributional Semantics for Entailment

James Henderson   Diana Nicoleta Popa

Xerox Research Centre Europe

Abstract

We propose a vector-space model which provides a formal foundation for a distributional semantics of entailment. Using a mean-field approximation, we develop approximate inference procedures and entailment operators over vectors of probabilities of features being known (versus unknown). We use this framework to reinterpret the Word2Vec distributional-semantic model as approximating an entailment-based model of words in contexts, thereby predicting lexical entailment. In both unsupervised and semi-supervised experiments on hyponymy detection, we get substantial improvements over previous results.

Motivation

Distributional Semantics:

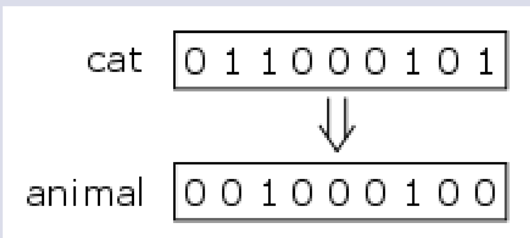
- The distributions of words in contexts reflect semantics
- Lexical semantic similarity can be captured by vector space embeddings trained on these distributions

Entailment:

y entails x $(y \Rightarrow x)$ iff
everything known given x is also known given y

- Entailment is about known versus unknown, not true versus false
- Entailment is asymmetric ($f \Rightarrow unk$, $unk \not\Rightarrow f$), unlike similarity

	unk	f	g	$\neg f$
unk	$\Rightarrow$	$\not\Rightarrow$	$\not\Rightarrow$	$\not\Rightarrow$
f	$\Rightarrow$	$\Rightarrow$	$\not\Rightarrow$	$\not\Rightarrow$
g	$\Rightarrow$	$\not\Rightarrow$	$\Rightarrow$	$\not\Rightarrow$
$\neg f$	$\Rightarrow$	$\not\Rightarrow$	$\not\Rightarrow$	$\Rightarrow$



Proposal:

- A distributional semantics for lexical entailment
- where semantic entailment is captured in a vector space
- of probabilities of known versus unknown features

A Vector Space for Entailment

Framework:

- A framework for the represention and inference of known versus unknown features
- Derived from a mean-field approximation to probabilistic inference for discrete entailment
- Assumes a non-factorised prior, but factorised posterior

Vectors:

- vectors X of log-odds of features x_k being known, $P(x_k{=}1) = \sigma(X_k)$
- a non-factorised prior, $P(x_k) = \theta_k(X_{\bar{k}})$

Operators:

$$\log P(y \Rightarrow x) \approx$$

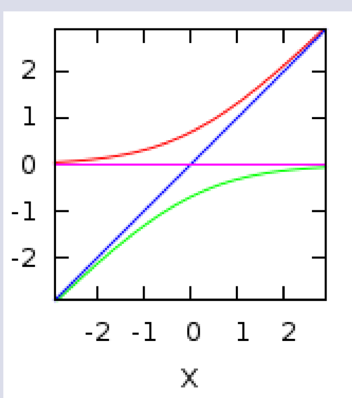
$$⧁ : \quad Y ⧁ X \triangleq \sigma(-Y) \cdot \log \sigma(-X)$$

$$⧀ : \quad X ⧀ Y \triangleq \sigma(X) \cdot \log \sigma(Y)$$

$$\Rrightarrow : \quad Y \Rrightarrow X \triangleq \sum_k \log(1 - \sigma(-Y_k)\sigma(X_k))$$

Inference:

$$X_{ik} = \theta_{ik}(X_{\bar{ik}}) + \sum_{j:x_i \Rightarrow x_j} -\log \sigma(-X_{jk}) + \sum_{j:x_j \Rightarrow x_i} \log \sigma(X_{jk})$$



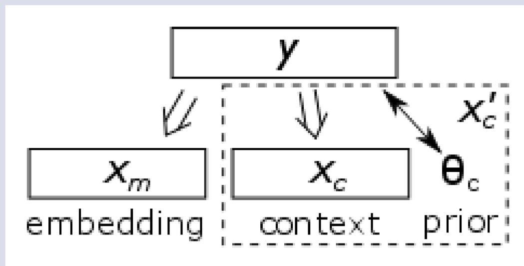
Distributional Semantics for Entailment

Model:

Exists a consistent unification y of the middle word x_m and its context x_c

$$\max_Y (\ \log P(\ y \Rightarrow x_m,\ y \Rightarrow x_c,\ y \mid x_m, x_c\ )\ )$$

$$= \max_Y (\ \log P(y \Rightarrow x_m | x_m) + \log P(y \Rightarrow x_c | x_c) + \log P(y)\ )$$



Embeddings:

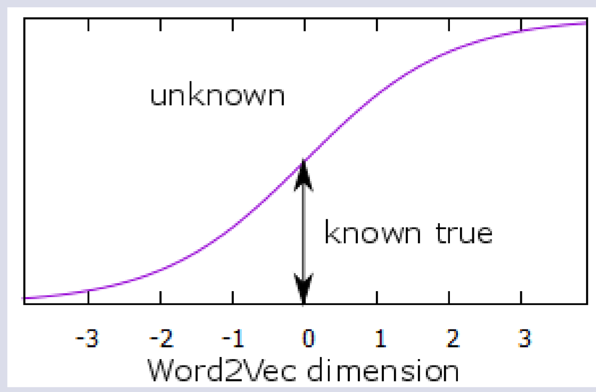
We reinterpret Word2Vec vectors, mapping them into entailment vectors

log-odds: W2V vector interpreted as an entailment vector

$$Y^+ = -\log \sigma(-X_m) - \log \sigma(-X_c) + \theta_c$$

$$\log P(y \Rightarrow x_m, y \Rightarrow x_c, y)$$

$$\approx Y^+ ⧁ X_m + Y^+ ⧁ X_c - \sigma(-Y^+) \cdot \theta_c$$



dup: W2V vector plus its negated duplicate

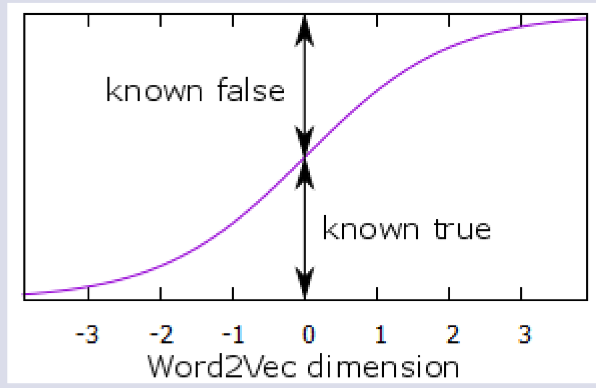
$$Y^+ = -\log \sigma(-X_m) - \log \sigma(-X_c) + \theta_c$$

$$Y^- = -\log \sigma(-(-X_m)) - \log \sigma(-(-X_c)) + -\theta_c$$

$$\log P(y \Rightarrow x_m, y \Rightarrow x_c, y)$$

$$\approx Y^+ ⧁ X_m + Y^+ ⧁ X_c - \sigma(-Y^+) \cdot \theta_c$$

$$+ Y^- ⧁ (-X_m) + Y^- ⧁ (-X_c) - \sigma(-Y^-) \cdot (-\theta_c)$$



unk dup: duplicated W2V vector with unknown around zero

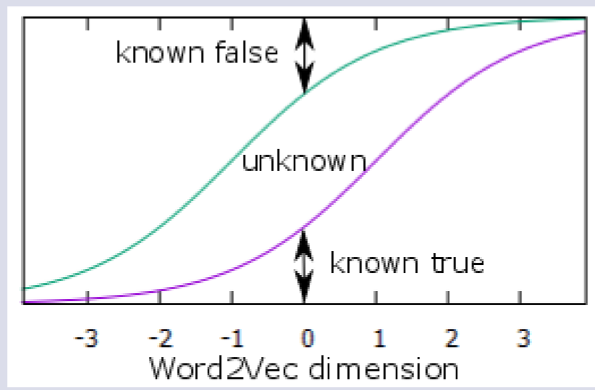
$$Y^+ = -\log \sigma(-(X_m - 1)) - \log \sigma(-X_c) + \theta_c$$

$$Y^- = -\log \sigma(-(-X_m - 1)) - \log \sigma(-(-X_c)) + -\theta_c$$

$$\log P(y \Rightarrow x_m, y \Rightarrow x_c, y)$$

$$\approx Y^+ ⧁ (X_m - 1) + Y^+ ⧁ X_c - \sigma(-Y^+) \cdot \theta_c$$

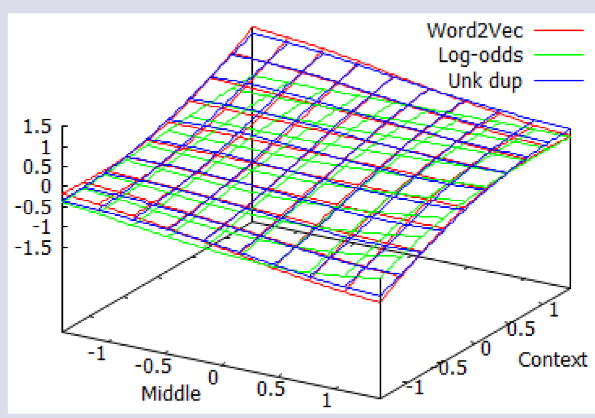
$$+ Y^- ⧁ (-X_m - 1)) + Y^- ⧁ (-X_c) - \sigma(-Y^-) \cdot (-\theta_c)$$



The learning gradients for these three re-interpretations are an increasingly good match with the Word2Vec learning gradient, where

$$\text{word embedding} \approx X_m$$

$$\text{context vector} \approx X_c' \approx -\log \sigma(-X_c) + \theta_c$$



Hyponymy Detection

Unsupervised

operator	50% Acc	Ave Prec	Dir Acc
Weeds et.al.	58%	–	–
log-odds ⧁	54.0%	55.9%	55.9%
weighted cos	55.5%	54.6%	57.9%
dot	56.3%	54.4%	50%
dif	56.9%	56.5%	59.6%
log-odds $\Rrightarrow$	57.0%	58.5%	59.4%
log-odds ⧀	60.1%*	61.3%*	62.2%
dup ⧀	61.7%	61.5%	68.8%
unk dup $\Rrightarrow$	63.4%*	67.3%*	**68.8%**
unk dup ⧀	**64.5%**	**68.8%**	**68.8%**

On BLESS data from Weeds et al. (2014)

Semi-supervised

operator	supervision	50% Acc	Ave Prec	Dir Acc
Weeds et.al.	SVM	75%	–	–
mapped dif	cross ent	64.3%	68.4%	72.3%
mapped ⧁	cross ent	71.2%	73.5%	88.3%
mapped $\Rrightarrow$	cross ent	77.4%	82.4%	**92.6%**
mapped ⧀	cross ent	**80.1%**	**86.3%**	90.0%

- Better accuracies than previous work and several baselines
- More accurate re-interpretations of Word2Vec result in better accuracies
- Better unsupervised accuracies carry over to better semi-supervised accuracies

Conclusions

Contributions:

- a vector-space model for entailment, with entailment operators and vector inference
- a formal foundation for a distributional semantics of entailment
- a reinterpretation of Word2Vec embeddings for lexical entailment
- best unsupervised model of hyponymy detection, and state-of-the-art results with a semi-supervised vector space

Future work:

- compositional models of textual entailment