Measuring Distributional Similarity in Context

Georgiana Dinu and Mirella Lapata

Matija Hanževački, Saarland University

Recent Advances in Computational Semantics
Seminar

November 2013

Introduction

Vector-based Models

- Wide-spread usage for computing meaning similarity.
- Various NLP tasks (synonym and paraphrase acquisition, word sense disambiguation, textual entailment, IR, etc.).
- Popular because of unsupervised nature and ease of computation.
- Usually one word is a point in a high-dimensional vector space.

This Paper

- Probabilistic framework for representing word meaning and measuring similarity in context.
- Meaning of words represented as a probability distribution over latent "senses".
- This distribution represents the a priori out-of-context likelihood of each sense.

Related Work

- Mitchell and Lapata's (2008) point-wise multiplication approach we've already heard about.
- Mitchell and Lapata (2009) show that their models yield improvement in language modeling.
- Erk and Pado's (2008) selectional preference technique we've also talked about.
- Plus some other clustering and exemplar-based approaches.

The Probabilistic Approach

The Probabilistic Approach

- The model assumes input data as a co-occurrence matrix of target words to context features.
- Given a set of latent senses:

$$Z = \{z_k | k : 1...K\}$$

- We represent the target word *i* with:

$$\mathbf{v(t_i) = (P(z_1|t_i), ..., P(z_K|t_i))}$$

- And the meaning of a target word given a context feature:

$$\mathbf{v(t_i, c_j) = (P(z_1|t_i, c_j), ..., P(z_K|t_i, c_j))}$$

The Probabilistic Approach Cont.

- The conditional probability of a latent sense given a target word and a context feature:

$$P(z_k|t_i, c_j) = \frac{P(t_i, z_k)P(c_j|z_k, t_i)}{\sum_k P(t_i, z_k)P(c_j|z_k, t_i)}$$

- The conditional probability of a context feature given a latent sense and a target word is difficult to estimate.
- We make a simplifying assumption of conditional independence of target words and context features given a latent sense:

$$P(z_k|t_i, c_j) \approx \frac{P(z_k|t_i)P(c_j|z_k)}{\sum_k P(z_k|t_i)P(c_j|z_k)}$$

Non-negative Matrix Factorization

- Approximate a non-negative input matrix V by two non-negative factors W and H, under a given loss function.

$$V_{I,J} \approx W_{I,K} H_{K,J}$$

- Using Kullback-Leibler divergence as a loss function:

$$\min \sum_{i,j} (V_{i,j} \log \frac{V_{i,j}}{WH_{i,j}} - V_{i,j} + WH_{i,j})$$

- If we interpret V as a matrix of joint probabilities of target words and context features, W as joint probabilities of target words and senses, and H as conditional probabilities of context features given senses, we get:

$$P(t_i, c_j) = \sum_k P(t_i) P(z_k|t_i) P(c_j|z_k).$$

Non-negative Matrix Factorization Cont.

- Given diagonal matrices A and B:

$$A_{ii} = \sum_k (WB)_{ik}.$$
$$B_{kk} = \sum_j H_{kj}.$$

- We can rewrite the factorization of WH as:

$$WH = AA^{-1}WBB^{-1}H = A(A^{-1}WB)(B^{-1}H)$$

Latent Dirichlet Allocation

- Parametrized probabilistic method of text generation.
- Each document *d* modeled as a distribution over *K* topics.
- Topics characterized by their distributions over words.
- Individual words in a document generated by repeatedly sampling a topic according to the topic distribution and then sampling a single word from the chosen topic.

Experiments

Tasks

- Word similarity task:
  - Involves judging similarity of two words out-of-context.
  - Using dataset of Finkelstein et al. (2002.) which contains 353 pairs of words.
- Lexical substitution task:
  - Requires finding appropriate substitutions of words in-context.
  - System typically required to rank a set of candidate substitutions.
  - Using SemEval 2007 Lexical Substitution Task benchmark dataset which contains 200 words (N,Adj, V, Adv) in 10 distinct contexts.

Baselines

- Latent Semantic Analysis:
  - Singular Value Decomposition of the original matrix for rank k = 1000.
  - Experimented with tf-idf and line normalization.
- Simple Vector Space Model:
  - No dimensionality reduction.
  - Measuring similarity with cosine, scalar product, Lin, and inverse JSD.
- Vector addition and multiplication for contextualized.

Word similarity

Model	Spearman ρ
SVS	38.35
LSA	49.43
NMF	**52.99**
LDA	**53.39**
LSA_{MIX}	49.76
NMF_{MIX}	51.62
LDA_{MIX}	51.97

Senses	Word Distributions
TRAFFIC (0.18)	*road, traffic, highway, route, bridge*
MUSIC (0.04)	*music, song, rock, band, dance, play*
FAN (0.04)	*crowd, fan, people, wave, cheer, street*
VEHICLE (0.04)	*car, truck, bus, train, driver, vehicle*

Lexical Substitution

Model	Kendall's τ_b
SVS	11.05
Add-SVS	12.74
Add-NMF	12.85
Add-LDA	12.33
Mult-SVS	14.41
Mult-NMF	13.20
Mult-LDA	12.90
Cont-NMF	14.95
Cont-LDA	13.71
Cont-NMF$_{\text{MIX}}$	**16.01**
Cont-LDA$_{\text{MIX}}$	**15.53**

Model	Adv	Adj	Noun	Verb
SVS	22.47	14.38	09.52	7.98
Add-SVS	22.79	14.56	11.59	10.00
Mult-SVS	22.85	16.37	13.59	11.60
Cont-NMF$_{\text{MIX}}$	**26.13**	**17.10**	15.16	**14.18**
Cont-LDA$_{\text{MIX}}$	21.21	16.00	**16.31**	13.67

Conclusion

Conclusion

- Presented a general framework of computing similarity in context.
- Meaning is a distribution over a set of global senses.
- Contextualized meaning modeled as a change in this distribution.
- Future work should expand on types and quantities of contextual features.

Thank you! Questions?