Word Semantic Representations using Bayesian Probabilistic Tensor Factorization

Jingwei Zhang, Jeremy Salwen, Michael Glass and Alfio Gliozzo

Department of Computer Science Columbia University
IBM T.J. Watson Research Center

Tuesday 21st October, 2014

Outline

Outline

Objectives

Combining word relatedness measures

Many approaches to word relatedness

- Manually constructed lexical resources
- Distributional vector space approaches
- Topic-based vector spaces

Continuous word representation

Word embedding method capable of distinguishing synonyms and antonyms.

Motivating Idea

Resources for word relatedness can be complementary

Manual resources get at interesting relationships
Automatic methods provide high coverage without extensive human effort.

Outline

Collaborative Filtering

Bayesian Probabilistic Matrix Factorization (BPMF) introduced for collaborative filtering (Salakhutdinov and Minh 2008 [10])
Bayesian Probabilistic Tensor Factorization (BPTF) incorporated temporal factors (Xiong et al 2010 [13])
Competitive results on real-world recommendation data sets.

Hypothesis

There is some latent set of word vectors

The word relatedness measures are constructed through these latent vectors.

Each word relatedness measure has some associated perspective vector

Combining the perspective with the dot product of the word vectors gives the word relatedness measure. There is also some Gaussian noise.

Basics

Bayesian Probabilistic

We determine the probability for a parameterization of our model by considering the probability of the data given the model, and the prior for the model.

Tensor Factorization

We will find vectors that when combined, give high probability to the observed tensor.

BPTF Model - Tensor

Relatedness tensor $\mathbf{R} \in \mathbb{R}^{N \times N \times K}$.

	joy	gladden	sorrow	sadden	anger
joyfulness	1	1	-1		
gladden	1	1		-1	
sad	-1		1	1	

$R^{(1)}$: Lexical similarity

	joy	gladden	sorrow	sadden	anger
joyfulness	.3	.1	-.1	.1	.3
gladden	.2	1	.2	.7	-.1
sad	.6	0	.4	.5	.1

$R^{(2)}$: Distributional similarity

BPTF Model[10][13]

$$R_{ij}^{k} | V_i, V_j, P_k \sim \mathcal{N}(< V_i, V_j, P_k >, \alpha^{-1}),$$

where $< \cdot, \cdot, \cdot >$ is a generalization of dot product:

$$< V_i, V_j, P_k > \equiv \sum_{d=1}^{D} V_i^{(d)} V_j^{(d)} P_k^{(d)},$$

α is the precision, the reciprocal of the variance.
V_i and V_j are the latent vectors of word i and word j
P_k is the latent vector for perspective k

Vectors and Perspectives

$$V_i \sim \mathcal{N}(\mu_V, \Lambda_V^{-1}),$$
$$P_i \sim \mathcal{N}(\mu_P, \Lambda_P^{-1}),$$

μ_V and μ_P are D dimensional vectors

Λ_V and Λ_P are D-by-D precision matrices.

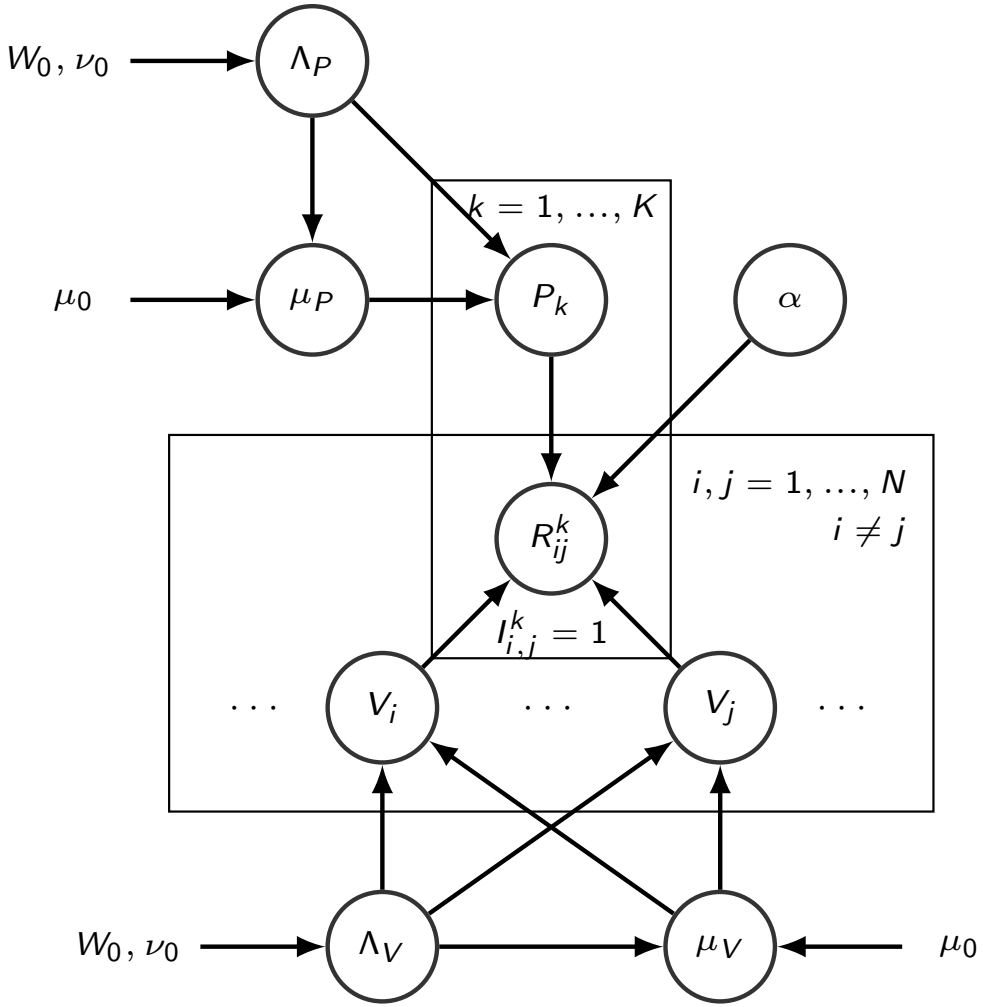
Hyper parameters

Conjugate Priors

$$p(\alpha) = \mathcal{W}(\alpha|\hat{W}_0, \nu_0),$$

$$p(\mu_V, \Lambda_V) = \mathcal{N}(\mu_V|\mu_0, (\beta_0\Lambda_V)^{-1})\mathcal{W}(\Lambda_V|W_0, \nu_0),$$

$$p(\mu_P, \Lambda_P) = \mathcal{N}(\mu_P|\mu_0, (\beta_0\Lambda_P)^{-1})\mathcal{W}(\Lambda_P|W_0, \nu_0),$$


W_0, ν_0
Λ_P
$k = 1, \ldots, K$
μ_0
μ_P
P_k
α
$i, j = 1, \ldots, N$
$i \neq j$
R_{ij}^k
$I_{i,j}^k = 1$
V_i
V_j
W_0, ν_0
Λ_V
μ_V
μ_0

Gibbs sampling

Algorithm 1 Gibbs Sampling for BPTF

Initialize the parameters.
repeat
  Sample the hyper-parameters α, μ_V, Λ_V, μ_P, Λ_P
  for $i = 1$ **to** N **do**
    Sample V_i
  end for
  for $k = 1$ **to** 2 **do**
    Sample P_k
  end for
until convergence

Out-of-vocabulary embedding

Generalize to words not present in a perspective

Can include all words in the BPTF procedure.
More efficient: compute the $R_{i,j}$ for the perspective of interest using only the V_i Gibbs sampling and the perspective dot product.

Predictions

Generalize and regularize the relatedness tensor by averaging over samples

$$p(\hat{R}_{ij}^k|\mathbf{R}) \approx \frac{1}{M}\sum_{m=1}^{M} p(\hat{R}_{ij}^k|V_i^m, V_j^m, P_k^m, \alpha^m),$$

Tuning

Number of dimensions for latent word and perspective vectors: $D = 40$

Untuned hyper-priors

- $\mu_0 = 0$
- $\nu_0 = \hat{\nu}_0 = D$
- $\beta_0 = 1$
- $W_0 = \hat{W}_0 = \mathbf{I}$

Outline

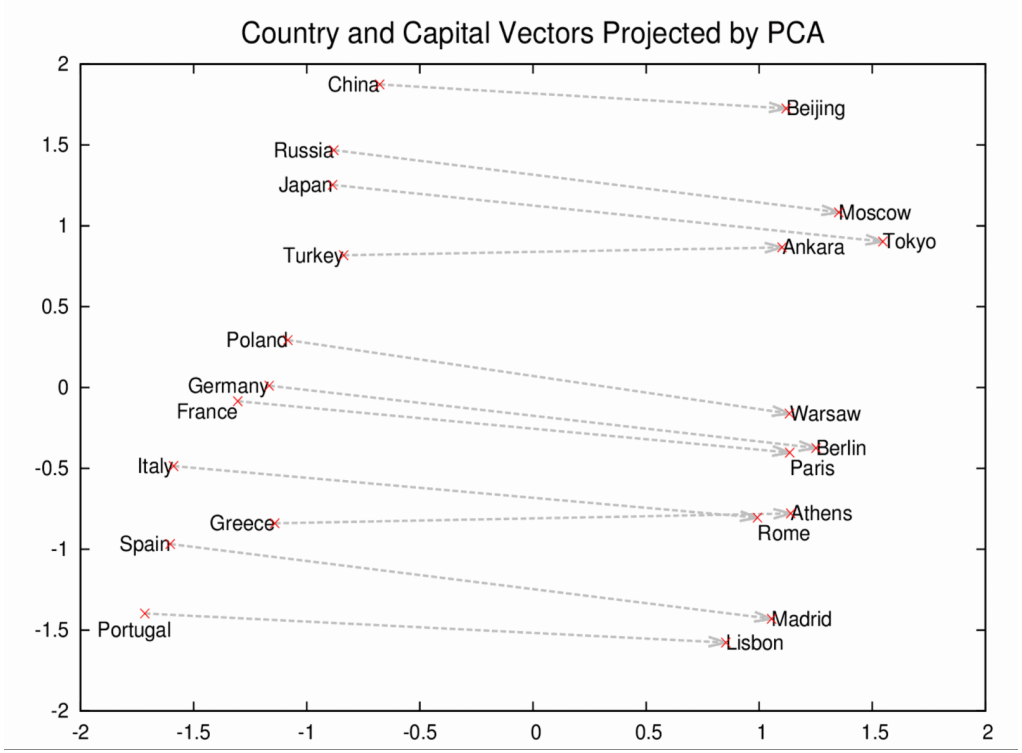
Thesaurus

1. WordNet
2. Roget's Thesaurus
3. Encarta Thesaurus[1]
4. Macquarie Thesaurus[2]

[1] Not available.

Neural word embeddings

- Linguistic regularities [7] (e.g. King−Man+Woman ≈Queen).
- Better for rare word: morphologically-trained word vectors [5].



Source: T. Minkolov

Evaluation

The GRE test dataset by Mohammad

- Development set: 162 questions
- Test set: 950 questions

Example GRE Antonym Question

desultory

1. phobic
2. entrenched
3. fabulous
4. systematic
5. inconsequential

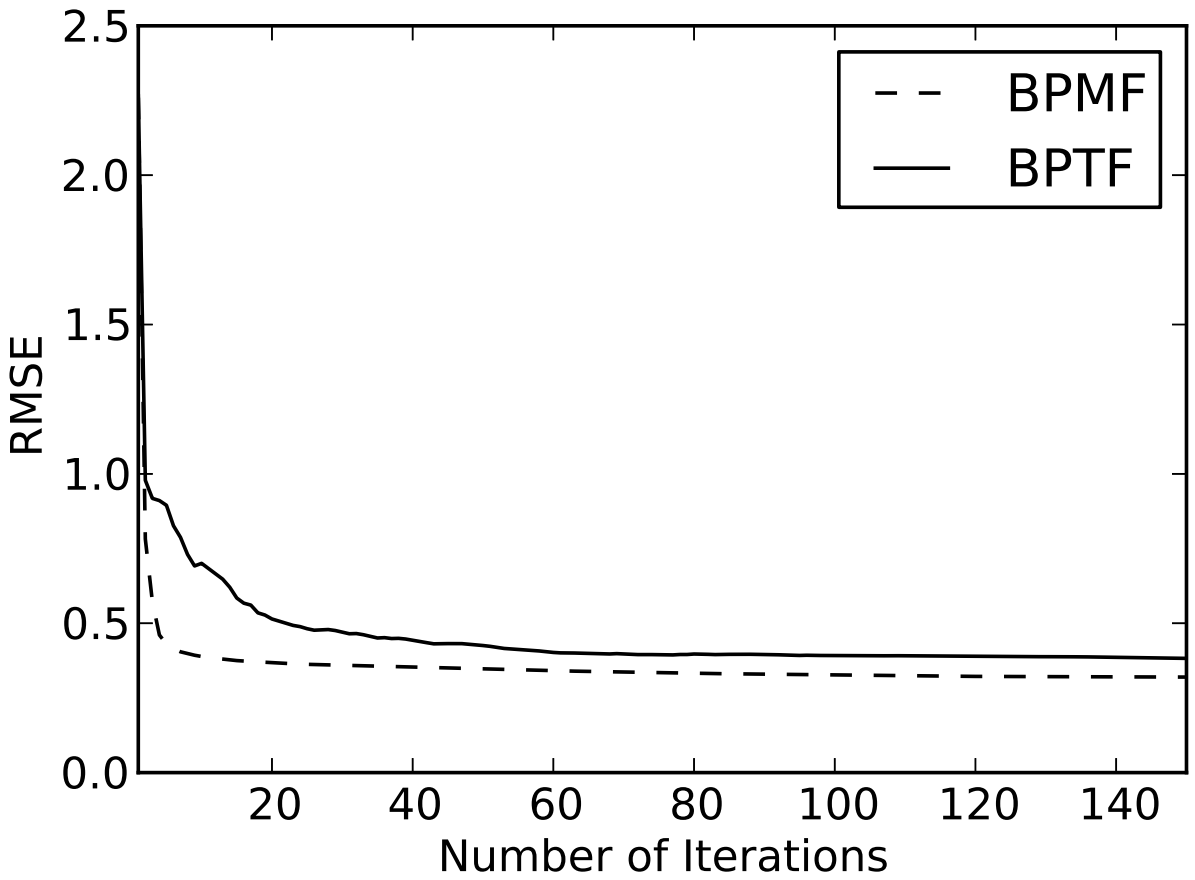
Previous Work

Lin [4] identifies antonyms by looking for pre-identified phrases in corpus datasets
Turney [12] uses supervised classification for analogies, transforming antonym pairs into analogy relations.
Mohammad et al. [8, 9] uses corpus co-occurrence statistics and the structure of a published thesaurus.
PILSA from Yih et al. [14] achieves the state-of-the-art performance in answering GRE antonym questions.

Evaluation

	Dev. Set			Test Set		
	Prec.	Rec.	F_1	Prec.	Rec.	F_1
WordNet lookup	0.40	0.40	0.40	0.42	0.41	0.42
WordNet PILSA	0.63	0.62	0.62	0.60	0.60	0.60
WordNet MRLSA	0.66	0.65	0.65	0.61	0.59	0.60
Encarta lookup	0.65	0.61	0.63	0.61	0.56	0.59
Encarta PILSA	0.86	0.81	0.84	0.81	0.74	0.77
Encarta MRLSA	0.87	0.82	0.84	**0.82**	0.74	0.78
Encarta PILSA + S2Net + Emebed	**0.88**	**0.87**	**0.87**	0.81	**0.80**	**0.81**
W&E MRLSA	0.88	0.85	0.87	0.81	0.77	0.79
WordNet lookup	0.48	0.44	0.46	0.46	0.43	0.44
WordNet&Morpho BPTF	0.63	0.63	0.63	0.63	0.62	0.62
Roget lookup	0.61	0.44	0.51	0.55	0.39	0.45
Roget&Morpho BPTF	0.80	0.80	0.80	0.76	0.75	0.76
W&R lookup	0.62	0.54	0.58	0.59	0.51	0.55
W&R BPMF	0.59	0.59	0.59	0.52	0.52	0.52
W&R&Morpho BPTF	**0.88**	**0.88**	**0.88**	**0.82**	**0.82**	**0.82**

Convergence Curve



Outline

Word Vector Representations

Core Methods

- Latent Semantic Analysis LSA (Deerwester et al 1990 [2])
- Polarity Inducing LSA (PILSA): LSA on a thesaurus (Yih et al 2012 [14])
- Distributional Similarity (Harris 1954 [3])
- Neural language models (Mikolov 2012 [6]), (Socher 2011 [11]), (Luong et al 2013 [5])

Multi-Source

Multi-Relational LSA does Tucker decomposition over tensor (Chang et al 2013 [1]).

Outline

Conclusion

Combining word relatedness measures

BPTF can combine matrices expressing word relatedness as a number

Word embedding to distinguish antonyms

Key limitation of distributional approaches can be improved with lexicon slice

`https://github.com/antonyms/AntonymPipeline`

References I

K.-W. Chang, W.-t. Yih, and C. Meek.

Multi-relational latent semantic analysis.

In *EMNLP*, 2013.

S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman.

Indexing by latent semantic analysis.

JASIS, 41(6):391–407, 1990.

Z. Harris.

Distributional structure.

Word, 10(23):146–162, 1954.

D. Lin and S. Zhao.

Identifying synonyms among distributionally similar words.

In *In Proceedings of IJCAI-03*, page 14921493, 2003.

M.-T. Luong, R. Socher, and C. D. Manning.

Better word representations with recursive neural networks for morphology.

In *CoNLL*, Sofia, Bulgaria, 2013.

References II

T. Mikolov.

Statistical language models based on neural networks.

PhD thesis, Ph. D. thesis, Brno University of Technology, 2012.

T. Mikolov, W.-t. Yih, and G. Zweig.

Linguistic regularities in continuous space word representations.

In *Proceedings of NAACL-HLT*, page 746751, 2013.

S. Mohammad, B. Dorr, and G. Hirst.

Computing word-pair antonymy.

In *EMNLP*, pages 982–991. Association for Computational Linguistics, 2008.

S. M. Mohammad, B. J. Dorr, G. Hirst, and P. D. Turney.

Computing lexical contrast.

Computational Linguistics, 39(3):555–590, 2013.

R. Salakhutdinov and A. Mnih.

Bayesian probabilistic matrix factorization using markov chain monte carlo.

In *ICML*, pages 880–887. ACM, 2008.

References III

R. Socher, C. C. Lin, C. Manning, and A. Y. Ng.

Parsing natural scenes and natural language with recursive neural networks.

In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129–136, 2011.

P. D. Turney.

A uniform approach to analogies, synonyms, antonyms, and associations.

Coling, pages 905–912, Aug. 2008.

L. Xiong, X. Chen, T.-K. Huang, J. G. Schneider, and J. G. Carbonell.

Temporal collaborative filtering with bayesian probabilistic tensor factorization.

In *SDM*, volume 10, pages 211–222. SIAM, 2010.

W.-t. Yih, G. Zweig, and J. C. Platt.

Polarity inducing latent semantic analysis.

In *EMNLP-CoNLL*, pages 1212–1222. Association for Computational Linguistics, 2012.

Word Semantic Representations using Bayesian Probabilistic Tensor Factorization

Jingwei Zhang, Jeremy Salwen, Michael Glass and Alfio Gliozzo

Department of Computer Science Columbia University
IBM T.J. Watson Research Center

Tuesday 21st October, 2014