Improving Lexical Semantics for Sentential Semantics: Modeling Selectional Preference and Similar Words in a Latent Variable Model

Weiwei Guo
Department of Computer Science
Columbia University
weiwei@cs.columbia.edu

Mona Diab
Department of Computer Science
George Washington University
mtdiab@gwu.edu

Abstract

Sentence Similarity [SS] computes a similarity score between two sentences. The SS task differs from document level semantics tasks in that it features the sparsity of words in a data unit, i.e. a sentence. Accordingly it is crucial to robustly model each word in a sentence to capture the complete semantic picture of the sentence. In this paper, we hypothesize that by better modeling lexical semantics we can obtain better sentential semantics. We incorporate both corpus-based (selectional preference information) and knowledge-based (similar words extracted in a dictionary) lexical semantics into a latent variable model. The experiments show state-of-the-art performance among unsupervised systems on two SS datasets.

1 Introduction

Sentence Similarity [SS] is emerging as a crucial step in many NLP tasks that focus on sentence level semantics such as word sense disambiguation (Guo and Diab, 2010; Guo and Diab, 2012a), summarization (Zhou et al., 2006), text coherence (Lapata and Barzilay, 2005), tweet clustering (Sankaranarayanan et al., 2009; Jin et al., 2011), etc. SS operates in a very small context, on average 11 words per sentence in Semeval-2012 dataset (Agirre et al., 2012), resulting in inadequate evidence to generalize to robust sentential semantics.

Weighted Textual Matrix Factorization [WTMF] (Guo and Diab, 2012b) is a latent variable model that outperforms Latent Semantic Analysis [LSA] (Deerwester et al., 1990) and Latent Dirichelet Allocation [LDA] (Blei et al., 2003) models by a large margin in the SS task, yielding state-of-the-art performance on the LI06 (Li et al., 2006) SS dataset. However, all of these models make harsh simplifying assumptions on how a token is generated: (1) in LSA/WTMF, a token is generated by the inner product of the word latent vector and the document latent vector; (2) in LDA, all the tokens in a document are sampled from the same document level topic distribution. Under this framework, they ignore rich linguistic phenomena such as inter-word dependency, semantic scope of words, etc. This is a result of simply using document IDs as features to represent a word.

Modeling quality lexical semantics in latent variable models does not draw enough attention in the community, since people usually apply dimension reduction techniques for documents, which have abundant words for extracting the document level semantics. However, in the SS setting, it is crucial to make good use of each word, given the limited number of words in a sentence. We believe a reasonable word generation story will avoid introducing noise in sentential semantics, encouraging robust lexical semantics which can further boost the sentential semantics. In this paper, we explicitly encode lexical semantics, both corpus-based and knowledge-based information, in the WTMF model, by which we are able to achieve even better results in SS task.

The additional corpus-based information we exploit is selectional preference semantics (Resnik, 1997), a feature already existing in the data yet ignored by most latent variable models. Selectional preference focuses on the admissible arguments for a word, thus capturing more nuanced semantics than the sentence IDs (when applied to a corpus of sentences as opposed to documents). Consider the following example:

Proceedings of NAACL-HLT 2013, pages 739–745,
Atlanta, Georgia, 9–14 June 2013.

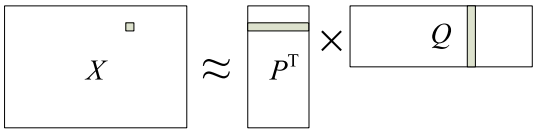


Figure 1: matrix factorization

Many analysts say the global Brent crude oil benchmark price, currently around $111 a barrel ...

In WTMF/LSA/LDA, a word will receive semantics from all the other words in a sentence, hence, the word *oil*, in the above example, will be assigned the incorrect *finance* topic that reflects the sentence level semantics. Moreover, the problem worsens for adjectives, adverbs and verbs, which have a much narrower semantic scope than the whole sentence. For example, the verb *say* should only be associated with *analyst* (only receiving semantics from *analyst*), as it is not related to other words in the sentence. In contrast, *oil*, according to its selectional preference, should be associated with *crude* indicating the *resource* topic. We believe modeling selectional preference capturing local evidence completes the semantic picture for words, hence further rendering better sentential semantics. To our best knowledge, this is the first work to model selectional preference for sentence/document semantics.

We also integrate knowledge-based semantics in the WTMF framework. Knowledge-based semantics, a human-annotated clean resource, is an important complement to corpus-based noisy co-occurrence information. We extract similar word pairs from Wordnet (Fellbaum, 1998). Leveraging these pairs, an infrequent word such as *purchase* can exploit robust latent vectors from its synonyms such as *buy*. Similar words pairs can be seamlessly modeled in WTMF, since in the matrix factorization framework a latent vector profile is explicitly created for each word, while in LDA all the data structures are designed for documents/sentences. We construct a graph to connect words according to the extracted similar word pairs, to encourage similar words to share similar latent vector profiles. We will refer to our proposed novel model as WTMF+PK.

2 Weighted Textual Matrix Factorization

Our previous work (Guo and Diab, 2012b) models the sentences in the weighted matrix factorization framework (Figure 1). The corpus is stored in an $M \times N$ matrix X, with each cell containing the TF-IDF values of words. The rows of X are M distinct words and columns are N sentences. As in Figure 1, X is approximated by the product of a $K \times M$ matrix P and a $K \times N$ matrix Q. Accordingly, each sentence s_j is represented by a K dimensional latent vector $Q_{\cdot,j}$. Similarly a word w_i is generalized by $P_{\cdot,i}$. P and Q is optimized by minimize the objective function:

$$\sum_i \sum_j W_{ij} \left(P_{\cdot,i} \cdot Q_{\cdot,j} - X_{ij}\right)^2 + \lambda ||P||_2^2 + \lambda ||Q||_2^2$$
$$W_{i,j} = \begin{cases} 1, & \text{if } X_{ij} \neq 0 \\ w_m, & \text{if } X_{ij} = 0 \end{cases} \quad (1)$$

where λ is a regularization term. Missing tokens are modeled by assigning a different weight w_m for each 0 cell in the matrix X. We can see the inner product of a word vector $P_{\cdot,i}$ and a sentence vector $Q_{\cdot,j}$ is used to approximate the cell X_{ij}.

The graphical model of WTMF is illustrated in Figure 2a. A w_i/s_j node is a latent vector $P_{\cdot,i}/Q_{\cdot,j}$, corresponding to a word/sentence, respectively. A shaded node is a non-zero cell in X, representing an observed token in a sentence. For simplicity, the missing tokens and weights are not shown in the graph.

3 Corpus-based Semantics: Selectional Preference

In this paper, we focus on selectional preference that reflects the association of two words: if two words form a bigram, then the two words should share similar latent dimensions. In the previous example, *crude* and *oil* form a bigram, and they share the *resource* topic. In our framework, this is implemented by adding extra columns in X, so that each additional column corresponds to a bigram, treating each bigram as a *pseudo-sentence* for the two words. The graphical model is illustrated in Figure 2b. Therefore, *oil* will receive more *resource* topic from *crude* through the bigram *crude oil*, instead of only *finance* topic from the sentence as a whole.

Each non-zero cell in the new columns of X, i.e. an observed token in a bigram (pseudo-sentence), is given a different weight:

$$W_{i,j} = \begin{cases} 1, & \text{if } X_{ij} \neq 0 \text{ and j is a sentence index} \\ \gamma \cdot freq(j), & \text{if } X_{ij} \neq 0 \text{ and j is a bigram index} \\ w_m, & \text{if } X_{ij} = 0 \end{cases}$$

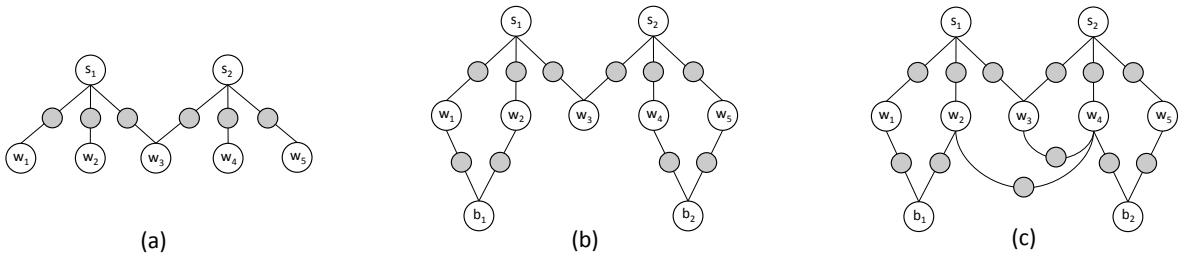


Figure 2: WTMF+PK model (WTMF + corpus-based Selectional [P]references semantics + [K]nowledge-based semantics): a w/s/b node represents a word/sentence/bigram, respectively

$freq(j)$ denotes the frequency of bigram j appearing in the corpus, hence the strength of association is differentiated such that higher weights are assigned on the more probable bigrams. The coefficient γ is the importance of selectional preference. A larger γ indicates that we trust the selectional preference over the global sentential semantics.

4 Knowledge-based Semantics: Similar Word Pairs

We first extract synonym pairs from WordNet, which are words associated with the same sense, synset. We further expand the set by exploiting the relations defined in WordNet. For the extracted words, we consider the first sense of each word, and if it is connected to other senses by any of the WordNet defined relations (*hypernym, similar words*, etc.), then we treat the words associated with the other senses as similar words. In total, we are able to discover $80K$ pairs of similar words for the $46K$ distinct words in our corpus.

Given a pair of similar words w_{i_1}/w_{i_2}, we want the two corresponding latent vectors $P_{\cdot,i_1}/P_{\cdot,i_2}$ to be as close as possible, namely the cosine similarity to be close to 1. Accordingly, a term is added in equation 1 for each similar word pair w_{i_1}/w_{i_2}:

$$\delta \cdot (P_{\cdot,i_1} \cdot P_{\cdot,i_2} - |P_{\cdot,i_1}||P_{\cdot,i_2}|)^2 \qquad (2)$$

$|P_{\cdot,i}|$ denotes the length of the vector $P_{\cdot,i}$. The coefficient δ, analogous to γ, denotes the importance of the knowledge-based evidence. The Figure 2c shows the final WTMF+PK model.

5 Inference

In (Guo and Diab, 2012b) we use Alternating Least Square [ALS] for inference, which is to set the derivative of equation 1 for P/Q to 0 and iteratively compute P/Q by fixing the other matrix (Srebro and Jaakkola, 2003). However, it is no longer applicable with the new term (equation 2) involving the length of word vectors $|P_{\cdot,i}|$. Therefore we approximate the objective function by treating the vector length $|P_{\cdot,i}|$ as fixed values during the ALS iterations:

$$\begin{aligned} Q_{\cdot,j} &= \left(P\tilde{W}^{(j)}P^\top + \lambda I\right)^{-1} P\tilde{W}^{(j)}X_{\cdot,j} \\ P_{\cdot,i} &= \left(Q\tilde{W}^{(i)}Q^\top + \lambda I + \delta P_{\cdot,s(i)}P_{\cdot,s(i)}^\top\right)^{-1} \\ &\quad \left(Q\tilde{W}^{(i)}X_{i,\cdot}^\top + \delta L_i P_{\cdot,s(i)} L_{s(i)}\right) \end{aligned} \qquad (3)$$

where $P_{\cdot,s(i)}$ are the latent vectors of similar words of word i; the length of these vectors in the current iteration are stored in $L_{s(i)}$ (similarly L_i is the current length of $P_{\cdot,i}$) (*cf.* (Steck, 2010; Guo and Diab, 2012b) for optimization details).

6 Experimental Setting

We build the model WTMF+PK on the same **corpora** as used in our previous work (Guo and Diab, 2012b), comprising the following: Brown corpus (each sentence is treated as a document), sense definitions from Wiktionary and Wordnet (only definitions without target words and usage examples). We follow the preprocessing steps in (Guo and Diab, 2012c): tokenization, pos-tagging, lemmatization and further merge lemmas. The corpus is used for building matrix X.

The **evaluation datasets** are LI06 dataset and Semeval-2012 STS [STS12] (Agirre et al., 2012) dataset. LI06 consists of 30 sentence pairs (dictionary definitions). For STS12,[1] the training data (2000 pairs) are used as the tuning set for setting the

[1] A detailed description of the data sets is provided in (Agirre et al., 2012).

parameters of our models. This data comprises msr-par, msr-vid, smt-eur. Once the models are tuned, we evaluate them on the STS12 test data that comprises 3150 sentence pairs from msr-par, msr-vid, smt-eur, smt-news, On-WN. It is worth noting that smt-news and On-WN are not part of the tuning data. We use cosine similarity to measure the similarity scores between two sentences. Pearson correlation between the system's answer and gold standard similarity scores is used as the evaluation metric.

We include three **baselines** LSA, LDA and WTMF using the setting described in (Guo and Diab, 2012b). We run Gibbs Sampling based LDA for 2000 iterations and average the model over the last 10 iterations. For WTMF, we run 20 iterations and fix the missing words weight at $w_m = 0.01$ with a regularization coefficient set at $\lambda = 20$, which is the best condition found in (Guo and Diab, 2012b).

7 Experiments

Table 1 summarizes the results at dimension $K = 100$ (the dimension of latent vectors). To remove randomness, each reported number is the averaged results of 10 runs. Based on the STS tuning set, we experiment with different values for the selectional preference weight ($\gamma = \{0, 1, 2\}$), and likewise for the similar word pairs weight varying the δ value as follows $\delta = \{0, 0.1, 0.3, 0.5, 0.7\}$. The performance on STS12 tuning and test dataset as well as on the LI06 dataset are illustrated in Figures 3a, 3b and 3d. The parameters of model 6 in Table 1 ($\gamma = 2, \delta = 0.3$) are the chosen values based on tuning set performance.

7.1 Evaluation on the STS12 datasets

Table 1 shows WTMF is already a very strong baseline: it outperforms LSA and LDA by a large margin. Same as in (Guo and Diab, 2012b), LSA performance degrades dramatically when trained on a corpus of sentence sized documents, yielding results worse than the surface words baseline 31% (Agirre et al., 2012). Using corpus-based selectional preference semantics **alone** (model 4 WTMF+P in Table 1) boosts the performance of WTMF by $+1.17\%$ on the test set, while using knowledge-based semantics **alone** (model 5 WTMF+K) improves the over the WTMF results by an absolute $+2.31\%$. Combining them (model 6 WTMF+PK) yields the best results, with an absolute increase of $+3.39\%$, which suggests that the two sources of semantic evidence are useful, but more importantly, they are complementary for each other.

Table 1 also presents the performance on each individual dataset. The gain on each individual source is not as much as the overall gain, which suggests part of the overall gain comes from the correct ranking of intra-source pairs. Note that WTMF+PK improves all individual datasets except smt-eur. This may be caused by too many overlapping words in the sentence pairs in smt-eur, while our approach focuses on extracting similarity between different words.

Observing the performance using different values of weights in figure 3a and 3b, we can conclude that the selectional preference and similar word pairs yield very promising results. The trends hold in different parameter conditions with a consistent improvement. Figure 3c illustrates the impact of dimension $K = \{50, 75, 100, 125, 150\}$ on WTMF and WTMF+PK. Generally a larger K leads to a higher Pearson correlation, but the improvement is tiny when $K \geq 100$ (0.1% increase).

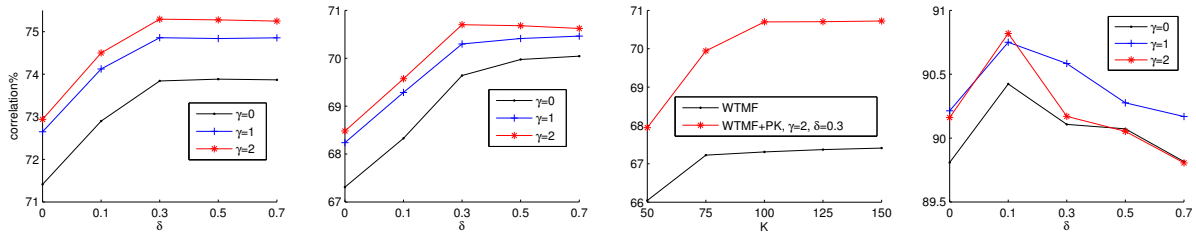
Compared to all the unsupervised systems that participated in Semeval STS 2012 task, WTMF+PK yields state-of-the-art performance (70.70%).[2] In (Guo and Diab, 2012c) we also apply WTMF ($K = 100$) on STS12, achieving a correlation of 69.5%. However, additional data is incorporated in the training corpora: (1) STS12 tuning set; (2) for WordNet and Wiktionary data, the target words are also included in the definitions (hence synonym pairs were used); (3) the usage examples of target words were also appended to the definitions.[3] While trained with this experimental setting, our model WTMF+PK ($\gamma = 2, \delta = 0.3, K = 100$) is able to reach an even higher correlation of 72.0%.

[2] WTMF+PK is an unsupervised system, since the gold standard similarly scores are never used in the objective function. Moreover, even without a tuning set, a non-zero value of γ or δ will always improve the baseline WTMF according to figure 3a and 3b.

[3] We do not adopt this corpora schema, since some definitions are test set sentences in On-WN, thereby adding target words and usage examples introduces additional information for some of the test set sentences

Models	Parameters	STS12 tune	STS12 test	msr-par	msr-vid	On-WN	smt-eur	smt-news	LI06
1. LSA	-	21.67%	24.41%	27.18%	9.91%	50.93%	27.86%	19.73%	63.77%
2. LDA	$\alpha = 0.05, \beta = 0.05$	71.10%	63.18%	29.15%	76.73%	62.81%	47.81%	27.2%	83.71%
3. WTMF	-	71.41%	67.31%	44.00%	82.59%	70.78%	**50.89%**	37.77%	89.81%
4. WTMF+P	$\gamma = 2, \delta = 0$	72.94%	68.48%	46.21%	83.29%	70.61%	49.54%	39.50%	90.16%
5. WTMF+K	$\gamma = 0, \delta = 0.3$	73.84%	69.64%	45.04%	83.04%	70.40%	49.88%	**41.66%**	90.11%
6. WTMF+PK	$\gamma = 2, \delta = 0.3$	**75.29%**	**70.70%**	**46.77%**	**83.90%**	**71.03%**	49.77%	40.48%	**90.17%**

Table 1: Evaluation Results using Pearson Correlation on STS12 and LI06



(a) STS12 tuning set ($K = 100$) (b) STS12 test set ($K = 100$) (c) STS12 test set (d) LI06 ($K = 100$)

Figure 3: Pearson correlation at different parameter settings

7.2 Evaluation on the LI06 dataset

Figure 3d presents the results obtained on the LI06 data set at different weight values for the corpus-based selectional preference semantics γ and for the knowledge-based semantics δ. Our previous experiments (Guo and Diab, 2012b) show that WTMF is the state-of-the-art model on LI06. With lexical semantics explicitly modeled, WTMF+PK yields better results than WTMF (see Table 1). It should be noted that LI06 prefers a smaller similar word pair weight (a $\delta = 0.1$ yields the best performance around of 90.75%), yet in almost all conditions WTMF+PK outperforms WTMF as shown in Figure 3d.

8 Related Work

SS has progressed immensely in recent years, especially with the establishment of the Semantic Textual Similarity task in SEMEVAL 2012. Early work in SS focused on word pair similarity in the high dimensional space (Li et al., 2006; Liu et al., 2007; Islam and Inkpen, 2008; Tsatsaronis et al., 2010; Ho et al., 2010), where co-occurrence information was not efficiently exploited. Researchers (O'Shea et al., 2008) find LSA does not yield good performance. In (Guo and Diab, 2012b; Guo and Diab, 2012c), we show the superiority of the latent space approach in WTMF. In this paper, we improve the WTMF model and achieve state-of-the-art Pearson correlation on two standard SS datasets.

There are latent variable models designed for lexical semantics, such as word senses (Boyd-Graber et al., 2007; Guo and Diab, 2011), function words (Griffiths et al., 2005), selectional preference (Ritter et al., 2010), synonyms and antonyms (Yih et al., 2012), etc. However little improvement is shown on document/sentence level semantics: (Ritter et al., 2010) and (Yih et al., 2012) focus on selectional preference and antonym identification, respectively; in (Griffiths et al., 2005) the LDA performance degrades in the text categorization task including the modeling of function words. Rather, we concentrate on nuanced lexical semantics phenomena that could benefit sentential semantics.

9 Conclusion

We incorporate corpus-based (selectional preference) and knowledge-based (similar word pairs) lexical semantics into a latent variable model. Our system yields state-of-the-art unsupervised performance on two most popular and standard SS datasets.

10 Acknowledgment

This work is supported by the IARPA SCIL program.

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *Advances in Neural Information Processing Systems*.

Weiwei Guo and Mona Diab. 2010. Combining orthogonal monolingual and multilingual sources of evidence for all words wsd. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Weiwei Guo and Mona Diab. 2011. Semantic topic models: Combining word distributional statistics and dictionary definitions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Weiwei Guo and Mona Diab. 2012a. Learning the latent semantics of a concept by its definition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Weiwei Guo and Mona Diab. 2012b. Modeling sentences in the latent space. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Weiwei Guo and Mona Diab. 2012c. Weiwei: A simple unsupervised latent semantics based approach for sentence similarity. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Chukfong Ho, Masrah Azrifah Azmi Murad, Rabiah Abdul Kadir, and Shyamala C. Doraisamy. 2010. Word sense disambiguation-based sentence similarity. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Aminul Islam and Diana Inkpen. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*, 2.

Ou Jin, Nathan N. Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*.

Yuhua Li, Davi d McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Transaction on Knowledge and Data Engineering*, 18.

Xiao-Ying Liu, Yi-Ming Zhou, and Ruo-Shi Zheng. 2007. Sentence similarity based on dynamic time warping. In *The International Conference on Semantic Computing*.

James O'Shea, Zuhair Bandar, Keeley Crockett, and David McLean. 2008. A comparative study of two short text semantic similarity measures. In *Proceedings of the Agent and Multi-Agent Systems: Technologies and Applications, Second KES International Symposium (KES-AMSTA)*.

Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*

Alan Ritter, Mausam, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Nathan Srebro and Tommi Jaakkola. 2003. Weighted low-rank approximations. In *Proceedings of the Twentieth International Conference on Machine Learning*.

Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

George Tsatsaronis, Iraklis Varlamis, and Michalis Vazirgiannis. 2010. Text relatedness based on a word thesaurus. *Journal of Articial Intelligence Research*, 37.

Wentau Yih, Geoffrey Zweig, and John C. Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Proceedings of Human Language Technology Conference of the North American Chapter of the ACL,*.