Latent Semantic Analysis Models on Wikipedia and TASA

Dan Ştefănescu, Rajendra Banjade, Vasile Rus
Department of Computer Science, Institute for Intelligent Systems
The University of Memphis, USA
E-mail: {dstfnscu, rbanjade, vrus}@memphis.edu

Abstract

This paper introduces a collection of freely available Latent Semantic Analysis models built on the entire English Wikipedia and the TASA corpus. The models differ not only on their source, Wikipedia versus TASA, but also on the linguistic items they focus on: all words, content-words, nouns-verbs, and main concepts. Generating such models from large datasets (e.g. Wikipedia), that can provide a large coverage for the actual vocabulary in use, is computationally challenging, which is the reason why large LSA models are rarely available. Our experiments show that for the task of word-to-word similarity, the scores assigned by these models are strongly correlated with human judgment, outperforming many other frequently used measures, and comparable to the state of the art.

Keywords: Latent Semantic Analysis, semantic models, word-to-word similarity, Wikipedia, TASA

1. Introduction

This paper introduces a collection of freely available Latent Semantic Analysis (LSA) semantic models constructed on two well-known corpora: Wikipedia[1] and the Touchstone Applied Science Associates (TASA) corpus (Ivens & Koslin, 1991; Landauer et al., 1998). The quality of these models is assessed through a series of experiments aiming at finding a practical LSA-based solution for quantifying word similarity that would best match human judgments. Moreover, we compare our LSA models to other well-known measures of semantic similarity such as WordNet based similarity measures, Explicit Semantic Analysis (ESA; Gabrilovich & Markovitch, 2007) and Latent Dirichlet Allocation (LDA; Blei et al. 2003).

This research is motivated by the fundamental need for better semantic similarity measures in general and in particular by our efforts to further develop the semantic processing capabilities of our DeepTutor system, a dialogue-based intelligent tutoring system (Rus et al., 2013a; Rus et al., 2013b). Semantic similarity is a core component in our DeepTutor system because it plays an important role in assessing the correctness of student answers, which in turn impacts the quality of the feedback the system provides to students. Word-to-word similarity, the focus of this paper, is a main ingredient of the semantic similarity solution used in DeepTutor and therefore a better understanding of these word-level similarity measures is needed.

The proposed LSA models are freely available for research, non-commercial purposes at http://www.semanticsimilarity.org/WikiLSA/.

2. Related Work

There are several publicly available tools for constructing LSA models. The most known ones are *Infomap NLP Software*[2] (2004), *SemanticVectors Package* (Widdows & Ferraro, 2008), *SenseClusters* (Pedersen, 2008), *S-Space Package* (Jurgens & Stevens, 2010) and *Gensim* (Řehůřek & Sojka, 2010). The *irlba*[3] and *lsa*[4] packages for *R*[5] (the software environment for statistical computing) can also be used for generating LSA spaces. The earliest implementations date back to the late 80's and the 90's, the most known one being developed at the University of Colorado Boulder (Landauer et al., 1998). The Colorado-Boulder group is still providing [6] "several pre-computed semantic spaces and tools to manipulate those spaces in a number of ways". Yet, the semantic spaces themselves are not available for download. We should also mention here *Text to Matrix Generator* (Zeimpekis & Gallopoulos, 2006), a MATLAB Toolbox that can generate term-document matrices from texts, with support for LSA.

Given the amount of research conducted on LSA and that assessing text-to-text semantic similarity is a crucial component for many Natural Language Processing (NLP) tasks, it is quite surprising that freely downloadable LSA models are extremely rare. In what regards the English Wikipedia, the only known work in this direction we are aware of is that of Řehůřek (2010)[7], but no model is available for download. Because such resources are virtually non-existent at this moment, our aim is to offer the scientific community a number of LSA models constructed over Wikipedia and TASA that can be effortlessly used, and even combined with other methods, for estimating word-to-word similarity.

3. Wikipedia and TASA

We decided to build our models on Wikipedia, the free encyclopedia, and on TASA, a corpus well suited for

[1] http://en.wikipedia.org/wiki/Main_Page

[2] http://infomap-nlp.sourceforge.net/
[3] http://cran.r-project.org/web/packages/irlba/index.html
[4] http://cran.r-project.org/web/packages/lsa/index.html
[5] http://www.r-project.org/
[6] http://lsa.colorado.edu/
[7] http://radimrehurek.com/gensim/wiki.html

educational applications. These are two natural choices given the fact that DeepTutor is a virtual tutor whose aim is to help high-school and freshmen college students master Newtonian physics.

The English Wikipedia is the largest publicly available descriptive corpus which encapsulates significant general and domain specific knowledge. We used an early Spring 2013 Wikipedia version, containing 4,208,450 articles. TASA comprises 60,527 samples from 6,333 textbooks, works of literature, and popular works of fiction and nonfiction. It "contains approximately the quantity and quality of text that the average college-level student has experienced across his lifetime" (cf. Jones et al., 2006). Our version has 37,652 documents. Both corpora were pre-processed using the Stanford NLP tools[8].

4. The LSA Models

Latent Semantic Analysis (LSA) is "a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text" (Landauer at al., 1998). LSA is often employed in NLP for knowledge representation and to assess semantic similarities between words or documents. An LSA semantic model is generated starting with a term-document matrix in which each row represents a term and each column represents a document in the given collection. The cell at row *t* and column *d* represents the frequency of term *t* in document *d* or a more complex weight representing the relation between *t* and *d*. The derivation of the LSA space means applying a Singular Value Decomposition (SVD) on the term-document matrix, obtaining an approximation of it using only the largest *k* singular values of the decomposition. Usually, *k* ranges from 300-500 latent dimensions.

We created 7 different LSA models for Wikipedia and 7 models for TASA, depending on different logical views of the documents in the two text collections, e.g. using all words documents or only the content words. All our LSA spaces have 300 dimensions, as it has been shown that in practice, this number appears to be optimal (Landauer & Dumais, 1997). This means that each word/term is represented in the reduced LSA space by a dense vector of 300 dimensions.

All our models were generated using *R* and the *irlba* package, which contains an implementation of the implicitly-restarted Lanczos bi-diagonalization algorithm, an efficient solution for applying SVD and performing dimensionality reduction on huge matrices.

For Wikipedia, we created the following models:

1. **Wiki_NVAR_f7** (**W1** in Table 1**)**: only the lemmas of content words were considered, occurring at least 7 times (in this way we filtered out spelling errors or exceptionally rare words, but still keeping a vocabulary of over 1 Million). Stop-words were excluded, and so were the light verbs (e.g. take, have, make, etc.) as they are often used in expressions that completely change their original sense.
2. **Wiki_NVAR_WN** (**W2**): only the lemmas of content words were considered, as long as they could be found as literals in Princeton WordNet (PWN; Fellbaum, 1998). Light verbs were excluded;
3. **Wiki_NVAR_pos_WN** (**W3**): same as W2, but the grammatical meta-categories are part of the terms. Thus, we distinguish between different grammatical functions (and senses) of the same word, e.g. *project* as a verb (becomes *project_v*) and a noun (*project_n*);
4. **Wiki_NV_WN (W4)**: same as W2, but the adjectives and the adverbs were filtered out; the reason for this elimination is that, for instance, adjectives may act too often as link/bridge among nouns leading to an over-estimate of the similarity of nouns. For instance, *red car* and *red apple* may lead to a similarity between *car* and *apple* due to the "common context" of *red*.
5. **Wiki_NV_pos_WN** (**W5**): same as W3, but the adjectives and the adverbs were filtered out;
6. **Wiki_NV_pos_11_150_f5_WN** (**W6**): same as W5, but documents having less than 11 and more than 150 words, were filtered out. Only lemmas occurring at least 5 times in the entire corpus were considered. By keeping only medium-length documents, the initial frequency matrix is reduced and so are the computational requirements for the LSA derivation. We were interested to see whether this would have any impact on the results of the model;
7. **Wiki_CONCEPT (W7)**: this LSA space was created using only the highlighted words and expressions in the Wikipedia articles (usually pointing to other articles) as terms.

Model	Initial matrix size	Number of non-zero values
W1	1,096,192 x 3,837,895	399,696,922
W2	72,930 x 3,577,017	346,934,093
W3	78,237 x 3,572,727	350,327,948
W4	68,187 x 3,550,591	284,093,540
W5	60,013 x 3,534,787	284,330,023
W6	51,641 x 2,318,925	92,876,186
W7	16,832,889 x 4,208,487	136,772,250

Table 1: The size of the initial term- document sparse matrices used for LSA on Wikipedia

For TASA, we constructed the following models:

1. **TASA_NVAR** (**T1** in Table 2): similar to W1, but with no frequency threshold
2. **TASA_NVAR_WN (T2)**: similar to W2;
3. **TASA_NVAR_pos_WN (T3)**: similar to W3;
4. **TASA_NV_WN (T4)**: similar to W4
5. **TASA_NV_pos_WN (T5)**: similar to W5
6. **TASA_OF_swr (T6)**: all the words as they appear in documents (**o**ccurrence **f**orms) were considered, but stop words were removed. We could afford keeping all non-stop words because the collection is smaller.
7. **TASA_OF (T7)**: all the words as they appear in documents were considered.

[8] http://nlp.stanford.edu/software/

To assess the quality of these models, we chose the most used test set in literature for word-to-word similarity: the WordSimilarity-353 Test Collection (Finkelstein et al., 2001). It contains 353 word pairs whose similarity was judged by at least 13 different subjects. The average scores can be considered an accurate human evaluation of the similarity between word pairs and provide a means for evaluating different methods against human judgment. Traditionally, this is done by computing Pearson (*r*) or Spearman rank-order (*rho*) correlations (Agirre et al., 2009; Gabrilovich & Markovitch, 2007).

Model	Original matrix size	Number of non-zero values
T1	69,824 x 37,652	2,462,727
T2	33,223 x 37,652	2,362,072
T3	36,110 x 37,652	2,379,187
T4	25,509 x 37,652	1,838,152
T5	26,938 x 37,652	1,854,306
T6	91,112 x 37,652	2,798,212
T7	91,365 x 37,652	4,418,938

Table 2: The size of the initial term-document sparse matrices used for LSA on TASA

We evaluated all our models against this dataset, to see which one is closer to human judgments. Moreover, we were interested to find out if the models are similar to each other. Tables 3 and 4 show the correlations with human judgments and among the models themselves.

	W1	W2	W3	W4	W5	W6
H	0.588 0.599	0.589 0.600	0.587 0.599	**0.589** **0.603**	0.584 0.600	*0.520* *0.542*
W1		0.999 0.998	0.987 0.986	0.996 0.996	0.984 0.983	0.893 0.896
W2			0.986 0.985	0.995 0.994	0.982 0.980	0.889 0.892
W3				0.986 0.985	0.998 0.997	0.911 0.913
W4					0.987 0.986	0.899 0.903
W5						0.917 0.919

Table 3: Pearson (top) & Spearman (bottom) correlations between Wiki models and human judgment (H)

	T1	T2	T3	T4	T5	T6	T7
H	0.566 0.583	0.564 0.583	0.554 0.568	**0.576** **0.591**	0.563 0.571	0.542 0.540	*0.517* *0.503*
T1		1 1	0.980 0.984	0.979 0.983	0.973 0.971	0.934 0.931	0.882 0.876
T2			0.981 0.984	0.979 0.983	0.973 0.971	0.935 0.931	0.883 0.875
T3				0.974 0.979	0.984 0.986	0.935 0.932	0.884 0.882
T4					0.987 0.988	0.924 0.916	0.861 0.852
T5						0.919 0.913	0.855 0.850
T6							0.941 0.946

Table 4: Pearson and Spearman correlations between TASA models and human judgment (H)

All TASA models yield results whose distributions have a highly positive skew. Since Pearson correlation is a more accurate estimator on bivariate normal distributions, we applied a transformation (i.e. sqrt; Newton & Rudestam, 1999) to bring the distributions closer to normality.

The results obtained by the LSA models constructed using lemmas of the content words (T1 to T5) have a (slightly) higher correlation with human judgments than the models built on occurrence forms (T6 and T7) (Table 4). This difference between the two types of models is statistically significant for T1, T2 and T4 when compared to T7 ($p < 0.05$). This comes as a confirmation of previous research on constructing LSA models from lemmatized text which were shown to be even more effective for highly inflected languages (Alumäe & Kirt, 2007).

Models	T6	T7
T1	Z= 1.446; p= 1.1482	**Z= 2.197; p= 0.0280**
T2	Z= 1.334; p= 0.1822	**Z= 2.115; p= 0.0344**
T3	Z= 0.725; p= 0.4688	Z= 1.665; p= 0.0960
T4	Z= 1.919; p= 0.0550	**Z= 1.989; p= 0.0466**
T5	Z= 1.223; p= 0.2214	Z= 1.504; p= 0.1326

Table 5: Steiger's Z (Meng et al., 1992) significance test on the differences between Pearson correlations

Overall, on WordSimilarity-353, the correlations between any of W1 to W5 models and T1 to T5 models are close to 0.6 (for both *r* and *rho*), while for the others, the correlations are lower, but still higher than 0.4.

We also assessed the inter-model agreement on a different collection of word-pairs than WordSimilarity-353. Considering the vocabulary for the model with the most restrictive constraints (i.e. T4) and the scores given by this model, we randomly selected 1,000 word pairs in a uniform manner so that we have 100 pairs having relatedness scores in the interval [0,0.1), 100 pairs in [0.1,0.2) and so forth.

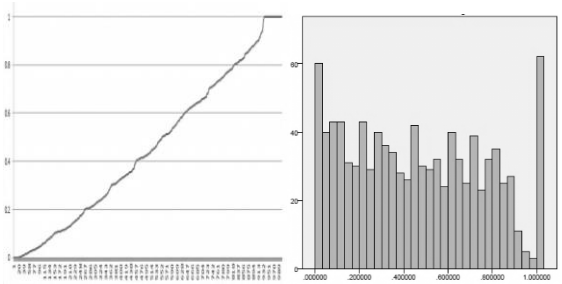

Figure 1: The distribution of the scores for the selected word pairs according to the T4 model

The inter-model agreement evaluations are presented in Tables 6 and 7. On this test set, the correlations between any of W1 to W5 models and T1 to T5 models are close to 0.5 (for both *r* and *rho*).

Interestingly enough, we noticed a very high correlation (> 0.95) between the models constructed considering all the grammatical categories and those in which the adjectives and adverbs were filtered out. One reason that could explain this finding is the fact that the adjectives and adverbs are not self-explanatory, but they always

determine nouns and verbs that were already taken into account when generating the LSA space(s). Also, another reason that requires further investigation may be that adjectives and adverbs have a tendency to modify content words that are actually semantically close.

	W2	W3	W4	W5	W6
W1	0.999 0.999	0.970 0.970	0.990 0.990	0.968 0.967	0.870 0.873
W2		0.970 0.970	0.990 0.991	0.969 0.968	0.871 0.875
W3			0.977 0.977	0.998 0.998	0.899 0.901
W4				0.979 0.978	0.883 0.885
W5					0.902 0.904

Table 6: Pearson and Spearman correlations among Wiki models on the 1,000 word pairs

	T2	T3	T4	T5	T6	T7
T1	1 1	0.954 0.951	0.985 0.984	0.947 0.944	0.893 0.885	0.870 0.861
T2		0.954 0.951	0.985 0.984	0.947 0.943	0.893 0.885	0.870 0.861
T3			0.957 0.954	0.994 0.993	0.868 0.861	0.846 0.836
T4				0.960 0.957	0.879 0.872	0.853 0.845
T5					0.860 0.852	0.835 0.823
T6						0.979 0.967

Table 7: Pearson and Spearman correlations among TASA models on the 1,000 word pairs

Method	Pearson	Spearman
Our experiments		
WordNet::Similarity (WN)	0.187-0.380	0.196-0.381
Our ESA Wiki	0.542	0.568
Best LSA Wiki	**0.589**	**0.603**
Best LSA TASA	0.576	0.591
Best LDA TASA	0.345	0.326
Linear Regression Models		
LSA Wiki + LSA TASA	0.632	0.646
LSA TASA + ESA + WN	0.673	0.712
LSA Wiki + LSA TASA + ESA + WN	**0.676**	**0.723**
Other experiments		
PWN (Jarmasz, 2003)		0.33-0.35
Roget's Thesaurus (Jarmasz, 2003)		0.55
LSA (Finkelstein et al., 2002)		0.56
Wikipedia (Strube & Ponzetto, 2006)	0.19-0.48	
ESA Wiki (Gabrilovich & Markovitch, 2007)		**0.75**
ESA ODP (Gabrilovich & Markovitch, 2007)		0.65
PWN 3.0 (Agirre et al., 2009)		0.56
PWN 3.0 + glosses (Agirre et al., 2009)		0.66
Context Windows (CW) (Agirre et al., 2009)		0.57-0.63
Bag of Words (Agirre et al., 2009)		0.64-0.65
Syntactic Vectors (Syn) (Agirre et al., 2009)		0.55-0.62
CW + Syn (Agirre et al., 2009)		0.48-0.66
SVM (Agirre et al., 2009)		**0.78**

Table 8: Our results compared to other or similar methods

Next, we obtained word-to-word results using the measures in the WordNet::Similarity software package (WN; Pedersen et al., 2004), ESA and LDA against human judgments. In what regards ESA, we tried to reproduce the approach of Gabrilovich & Markovitch (2007), but our results were lower than those they reported (0.57 vs. 0.75). This might be due to the different version of Wikipedia we have used (Gabrilovich and Markovitch worked on a 2006 version) and also due to some other implementation details we were not aware of. Yet, this large difference in results should be further investigated.

Observing the fact that some of the models we evaluated yielded similar results, we were interested in finding whether they can be combined in order to obtain models that can better correlate with human judgments. Dieterich (1998) showed that combining two or more methods is effective only when two conditions are met: (i) they should have comparable performance scores and (ii) they should make different errors. Some of the models do in fact perform comparably (the LSA ones and ESA; see Table 8), but in order to tell whether they are different enough (make different errors), we need to look at their correlations, one versus the other (see Table 9). Here, we have to mention that given the fact that WordNet (WN) measures have a relatively low correlation with human judgments, in these particular experiments we used only one WN measure, which yielded the best correlation with humans (i.e. a Lesk type measure [Banerjee and Pedersen 2002]).

	LSA TASA	LDA	ESA	WN
LSA Wiki	**0.594** **0.600**	0.354 0.249	**0.599** **0.633**	0.304 0.224
LSA TASA		0.635 0.374	**0.585** **0.584**	0.282 0.248
LDA			0.407 0.249	0.278 0.168
ESA				0.314 0.292

Table 9: Pearson (top) and Spearman (bottom) correlations among the selected measures on the 1,000 word pairs

The values in Table 9 show that the models are indeed different enough to also meet Dieterich's second condition. This should indicate that combining them would result in better models, but on the other hand, if the correlation is high, for example 0.60 *rho* between Wiki and TASA LSA models, it might be the case that the improvement would not be a spectacular one.

The most straightforward way to combine the available measures is by fitting linear regression models. The results obtained versus human judgments using two by two combinations, are presented in Table 10 and indeed, they confirm the intuition stated in the previous paragraph: the correlation of the combined LSA models with human judgment increases only to 0.65 (*rho*).

On the same line of thinking, the low correlation values between our selected WN method and the other methods indicate that they are making mostly different mistakes. However, in this case Dieterich's first condition is not fully met since its correlation score against humans is too low compared to the others.

Table 10 shows correlation scores versus human judgments for all possible two-by-two linear combinations of the selected measures. These values are obtained using a ten-fold cross-validation evaluation method: 10 consecutive chunks of 31 word-pairs were considered as test sets, while the rest were used for training. Furthermore, we looked at all the possible linear combinations between subsets of these selected measures. The best combinations are shown in Table 8, under *Linear Regression Models*. We obtained the best results using a combined model that included LSA Wiki, LSA TASA, ESA and WN, with correlation values of **0.676** (*r*) and **0.723** (*rho*) and it is interesting to notice that the different measures involved are capturing different similarity aspects. For instance, in the case of WN-based measures, similarity is usually computed based on distances between concepts in an ontological hierarchical structure. Conversely, in LSA and ESA similarity is based on word co-occurrences. We consider this to be a clear indication of the potential of combining Knowledge and Corpus-based measures for word-to-word similarity.

	LSA TASA	LDA	ESA	WN
LSA Wiki	**0.632** 0.646	0.591 0.623	0.643 0.658	0.621 0.647
LSA TASA		0.559 0.576	0.640 0.661	0.618 0.642
LDA			0.594 0.637	0.479 0.501
ESA				0.627 **0.672**

Table 10: Correlations with humans obtained by linearly combining the selected measures, two by two (Pearson – top; Spearman - bottom).

5. Conclusions

This paper introduced a freely available collection of Latent Semantic Analysis models for word-to-word similarity built on the entire English Wikipedia and the TASA corpus. The models differ not only on their source (Wikipedia versus TASA), but also on the linguistic items they considered: all words, content-words, nouns-verbs, and main concepts. The quality of the proposed LSA models was assessed through a series of experiments against human judgments and other well-known measures of semantic similarity, such as WN-based measures, ESA or LDA. Among the conclusions of these experiments we mention:

- The newly created LSA models clearly outperform the WN methods, W4 being the best;
- ESA model correlates somewhat better with the WN methods than the LSA models;
- It is slightly better to build an LSA model on lemmatized text than on raw text;
- Interestingly, filtering out adjectives and adverbs does not have a significant impact on the final LSA space;
- When constructing an LSA model, in case the collection is very big, filtering it based on document size must be done with care as the results might differ significantly.

Moreover, observing the fact that some of the models we evaluated yielded similar results while making different mistakes, we were interested in finding whether their combination would produce better results. Consequently, we constructed linear regressions models using all the available measures, and again, we investigated how much they correlate to human judgment.

The best results obtained by our models (0.723 *rho*) are comparable to state of the art, being higher than any others described in the literature, except for ESA Wiki (0.75 *rho*) and SVM (0.78 *rho*) reported by Gabrilovich and Markovitch (2007) and Agirre and colleagues (2009), respectively.

6. Acknowledgments

This research was supported in part by Institute for Education Sciences under awards R305A100875. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors'.

7. References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. In Proceedings of *HLT: The 2009 Annual Conference of the NAACL* (pp. 19-27). ACL.

Alumäe, T., & Kirt, T. (2007). Lemmatized latent semantic model for language model adaptation of highly inflected languages.

Banerjee, S., & Pedersen, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. In *IJCAI*, 3: 805–810.

Blei, D., M., Ng, A., Y., & Jordan, M., I. (2003). Latent Dirichlet Allocation. *The Journal of ML research* 3: 993–1022.

Christiane, F. (1998). WordNet: an electronic lexical database. *Cambrige, MIT Press, Language, Speech, and Communication*.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. In *Neural Computation*, 10(7): 1895–1924.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web* (pp. 406-414). ACM.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In*Proceedings of the 10th international conference on World Wide*

Web (pp. 406-414). ACM.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI* (Vol. 7, pp. 1606-1611).

Ivens, S. H., & Koslin, B. L. (1991). *Demands for Reading Literacy Require New Accountability Methods*. Touchstone Applied Science Associates.

Jarmasz, M. (2012). Roget's thesaurus as a lexical resource for natural language processing. *arXiv preprint arXiv:1204.0140*.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, *55*(4), 534-552.

Jurgens, D., & Stevens, K. (2010). The S-Space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 30-35).Association for Computational Linguistics.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological bulletin*, *111*(1), 172.

Newton, R. R., & Rudestam, K. E. (1999). Your Statistical Consultant: Answer To Your Research & Data Analysis Questions.

Pedersen, T. (2008). Computational approaches to measuring the similarity of short contexts: A review of applications and methods. *arXiv preprint arXiv:0806.3787*.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). WordNet::Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004* (pp. 38-41). ACL.

Řehůřek, R. (2010). Fast and Faster: A Comparison of Two Streamed Matrix Decomposition Algorithms. In *NIPS 2010 Workshop on Low-rank Methods for Large-scale Machine Learning*.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks* (pp. 46-50).

Rus, V., D'Mello, S., Hu, X., & Graesser, A. (2013a). Recent Advances in Conversational Intelligent Tutoring Systems. *AI Magazine*, *34*(3), 42-54.

Rus, V., Niraula, N., Lintean, M., Banjade, R., Stefanescu, D., & Baggett, W. (2013b). Recommendations For The Generalized Intelligent Framework for Tutoring Based On The Development Of The DeepTutor Tutoring Service. In *AIED 2013 Workshops Proceedings Volume 7* (p. 116).

Strube, M., & Ponzetto, S. P. (2006, July). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI* (Vol. 6, pp. 1419-1424).

Widdows, D., & Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. *Proceedings of LREC 2008*.

Zeimpekis, D., & Gallopoulos, E. (2006). TMG: A MATLAB toolbox for generating term-document matrices from text collections. In *Grouping multidimensional data* (pp. 187-210). Springer Berlin Heidelberg.