How Important Is Size? An Investigation of Corpus Size and Meaning in both Latent Semantic Analysis and Latent Dirichlet Allocation

Scott A. Crossley,[a] Mihai Dascalu,[b] and Danielle S. McNamara[c]

[a]Department of Applied Linguistics, Georgia State University, USA

[b]Department of Computer Science, University Politehnica of Bucharest, Romania

[c]Department of Psychology, Arizona State University, USA

Abstract

This study examines how differences in corpus size influence the accuracy of Latent Semantic Analysis (LSA) spaces and Latent Dirichlet Allocation (LDA) spaces in two tasks: a word association task and a vocabulary definition test. Specific optimizations were considered in building each semantic model. Initial results indicate that larger corpora lead to greater accuracy and that LDA probabilistic models, similar to LSA vector spaces, can provide insights into cognitive processing at semantic levels.

Introduction

The statistical redundancies found in language afford the opportunity to model higher order representations of word meaning using unsupervised learning techniques (Kintsch, 2001). The majority of this work was completed in the 1990s and centered around Latent Semantic Analysis (LSA; Landauer and Dumais, 1997) models, which derives semantic representations of words from large corpora of texts. The central notion of LSA semantic models is that the combined contexts in which a word occurs provide a set of mutual constraints that can be used to estimate a word's meaning based on context (Jones and Mewhort, 2007). LSA models have replicated human judgments of semantic similarity (Landauer and Dumais, 1997), judgments of word synonyms, and judgments of essay quality (Landauer, Laham, Rehder, and Schreiner, 1997).

One consideration for deriving accurate semantic representations using LSA is the size of the corpus. Deerwester et al. (1990) suggested a reasonably sized corpus should comprise about 1,000-2,000 documents and contain about 5,000-7,000 words. Such a corpus would be representative of natural language and contain sufficient redundancies (i.e., enough conceptually related terms appearing together). Landauer and Dumais (2008) suggested that the overall minimum size of the initial term-document matrix should be at least 20,000 terms with 20,000 passages. Following this trend, the most commonly used corpus to derive LSA spaces is the Touchstone Applied Science Association (TASA) corpus (http://lsa.colorado.edu/spaces.html), which comprises about 38,000 documents and 92,000 terms.

Corpus size is a concern because large, heterogeneous corpora may provide more noise or too much specific information from a single domain, reducing the accuracy of the derived models. However, there is little agreement on the expected size of the corpus or what comprises a large or small corpus (Giesbers, Rusman, and van Bruggen, 2006; Villalon and Calvo, 2009). Some research reports that LSA performs best on an entire corpus (Wiemer-Hastings, 1999), while other research reports that the impact of corpus size asymptotes at about 80% of a corpus (Terra and Clark, 2003). However, these findings are corpus specific and often specific to the number of vectors selected.

Recently, researchers have expanded beyond LSA models of semantic representations and have begun to explore topic-based models of semanticity, of which the most common is Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan, 2003). LDA, like LSA, depends on large corpora to infer topics based on a combination of words and documents. However, LDA have mostly been trained on specific, information retrieval oriented, corpora that needed to be semantically annotated. Most LDA models focus on topic extraction (Chang et al., 2009) or dynamic modeling (Blei and Lafferty, 2006) of topic trends, but have not analyzed the cognitive and psychological implications of LDA in a similar degree to which LSA has been analyzed.

Cognitively, LSA can be seen as an expression of meaning because each word can be represented as a context-free vector in the semantic vector-space model (Kintsch, 2001). The actual dimensions of concepts do not bear a specific individual meaning, but the overall

representation generated by LSA can be considered a map of meanings (Landauer and Dumais, 2008). In addition, positive correlations between LSA similarity scores and human recall using word association lists support a semantic proximity effect (Howard and Kahana, 1999) in which LSA bears resemblance to human memory (e.g., memory search, free recall; Zaromb et al., 2006; Landauer and Dumais, 2008). In contrast to LSA, LDA does not support theories of cognition. This is chiefly because LDA is a probabilistic topic model in which the connotations of the latent space behind the model are ignored because only the distributions of words within documents are observable (Chang et al., 2009). Although LDA topics are not equi-probable and semantic significances cannot be automatically deduced (Arora and Ravindran, 2008), LDA has proven to be reliable in extracting topics from texts (Blei et al., 2003).

The purpose of this study is to examine differences in LSA and LDA models derived from corpora of two different sizes: TASA and the Corpus Contemporary American English (COCA; Davies, 2010). Specifically, we examine the accuracy of LSA and LDA spaces derived from these corpora in terms of simulating word association norms and selecting vocabulary test answers. Our primary goal is to assess the degree to which corpus size increases, decreases, or has no effect on replications of the human semantic knowledge. Our approach also allows us to test differences between a large corpus comprised of a single domain written at a similar level of complexity (COCA) and a smaller corpus comprised of multiple domains written at different complexity levels (TASA). Lastly, this approach provides the opportunity to assess the cognitive and psychological implications of LDA.

Method

Training Corpora

LSA and LDA models used in this study were trained for the English language using the TASA corpus and COCA. We selected TASA because it is a common corpus used in developing LSA spaces. TASA consists of educational texts spanning the 1st through the 12th grade and contains a number of domains chief among them language arts, sciences, and social studies. COCA has five different subgenres: spoken, fiction, popular magazine, newspaper, and academic. For our semantic spaces, we selected the newspaper genre because it is the most general.

A specific NLP pre-processing cleaning was applied to each corpus, in which non-dictionary word forms and stop-words were disregarded, and all inflected word forms were reduced to their corresponding lemmas. While building the sparse term-document matrix stored in Hadoop, we relied on log-entropy for our LSA space. Due to the highly computational SVD decomposition, our training relied on a distributed version of stochastic SVD from the Mahout framework (Owen, Anil, Dunning, and Friedman, 2011). We ensured that LSA spaces for both TASA and COCA had the same number of vectors ($N = 300$). Similarly, we ensured that the optimal number of LDA topics inferred from both corpora were similar. Descriptions of both corpora are found in Table 1.

Descriptors	TASA	COCA
N# lemmas	5,864,529	41,732,161
N# paragraphs	44,486	57,037
Dictionary size	43,012	55,449
LSA *k*	300	300
LDA / HDP-inferred *k*	230	175

Table 1. Corpus descriptions.

USF Word Association Norms

We use the University of South Florida (USF) association norms (Nelson, McEvoy, and Schreiber, 1998) to examine similarities between the LSA and LDA spaces with standardized word association norms. The USF norms report the number of stimuli words (N = 5,019) that resulted in production of a target word (N = 10,470) as an associate in a free association task. We examined the maximum similarity between the response words and the average cosines derived from our semantic models along with the average similarity between the top three responses and the derived cosines.

Vocabulary Levels Test

We use Vocabulary Levels Tests (VLT; Nation, 1990; Schmitt, Schmitt and Clapham, 2001) to investigate the potential for our LSA and LDA spaces to predict word definitions in a standardized test. VLTs are designed to assess knowledge of common and uncommon English words. The tests assess word knowledge at levels based on 1,000 word frequency bands. The original VLT (Nation, 1990) and the second VLT (Schmitt et al., 2001) tested knowledge at the second, third, fifth, tenth word frequency bands along with words taken from the Academic Word List (Coxhead, 2000). Test-takers see six words on the left side and three definitions on the right side. The learner must match the three definitions to three of the words on the left side. We used a bipartite graph in which we initially made all possible association between words and their potential definition followed by a maximization flow within the resulting network to select the correct answer from our semantic models.

Results

Student *t*-tests were conducted to compare differences in the strength of associations between the LSA and LDA word vectors and the Nelson Word Association Norms for

both the TASA and COCA corpora. *t*-tests were calculated for the average vector similarity for the top 3 associated words in the Nelson norms along with the similarity to the most strongly associated word (i.e., maximum similarity). There were significant advantages for the LDA space using COCA compared to TASA in terms of both average similarity; $t(9044) = 25.68$, $p < .001$, and maximum similarity; $t(9044) = 12.22$, $p < .001$ (see Table 2). Likewise, there were significant advantages for the LSA space using COCA for both average similarity; $t(9084) = 19.92$, $p < .001$, and maximum similarity; $t(9084) = 13.98$, $p < .001$ (see Table 3). These results indicate that LDA and LSA spaces derived from a larger corpus (i.e., COCA) showed stronger links to word association norms than spaces derived from a smaller corpus (i.e., TASA).

Variable	LDA COCA	LDA TASA
Average similarity	0.414 (0.141)	0.332 (0.162)
Maximum similarity	0.553 (0.167)	0.504 (0.208)

Table 2: Comparisons between LDA TASA and COCA spaces.

Variable	LSA COCA	LSA TASA
Average similarity	0.265 (0.130)	0.211 (0.127)
Maximum similarity	0.395 (0.183)	0.340 (0.190)

Table 3: Comparisons between LSA TASA and COCA spaces.

A factorial ANOVA was conducted to examine differences in VLT accuracy scores reported by LSA and LDA spaces for both the TASA and the COCA corpus across the entire test and among the vocabulary levels. The ANOVA showed a significant main effect for semantic space/corpus, $F(3, 12) = 3.68$, $p < .050$ and a significant main effect for vocabulary level, $F(4, 12) = 9.89$, $p < .001$. No significant interaction between semantic space/corpus and level was reported, $F(12, 920) = 1.19$, $p > .05$. Pairwise comparisons demonstrated that the LSA space derived from COCA outperformed both LDA spaces (COCA and TASA) and that the LSA space derived from TASA outperformed the LDA space derived from COCA ($p < .05$). Pairwise comparisons indicated that lower level words (i.e., 2000 and 3000 level words) were defined most accurately by the LSA and LDA spaces. Overall, this analysis indicates that LSA spaces based on COCA outperform other spaces and that LDA spaces (especially those based on COCA) were the worst performing.

Discussion

Consensus on the effect of corpus size on the accuracy of LSA spaces is still lacking. At the same time there seems to be little research about the effects of corpus size on LDA spaces. This study helps to address corpus size differences by developing similar LSA and LDA spaces on two different corpora of different sizes and assessing the resulting spaces on two cognitive assessments of word knowledge: a word association task and a vocabulary test. The results indicate that a larger corpus leads to accuracy gains for LSA models in terms of word association similarities and vocabulary test scores (although the latter was not significant). The results for the LDA models are more nuanced with LDA models based on larger corpora performing better on word association tasks, but worse on vocabulary tests. We discuss these findings below.

In terms of matching human judgments of word association tasks, both LSA and LDA spaces derived from the larger COCA performed significantly better than models derived from TASA indicating that the larger coverage of words found in COCA allowed the models to develop both similarity and topic matrices that were better aligned with human judgments of word associations. This finding provides some evidence that larger corpora may lead to the development of more accurate semantic spaces. Also of interest is the difference in similarity strengths reported by the LDA and LSA spaces with the LDA spaces reporting stronger similarities with human judgments of word associations than LSA spaces. This may be a result of differences in the similarity functions used. The cosine is bounded within the [-1; 1] interval, with extremely few cases of negative values for high dimensional LSA spaces, whereas Jensen Shannon Dissimilarity ranges from 0 to 1. This means that overall association strengths for all words pairs are generally higher for LDA spaces than for LSA spaces.

In terms of vocabulary test results, the LSA spaces based on COCA reported higher mean scores (75%) on the VLT than LSA spaces based on TASA (72%) although this difference was not significant. The LSA spaces based on COCA reported stronger accuracies than the LSA spaces based on TASA for the more difficult word levels (i.e., the 10,000 and academic word levels) indicating that gains were made with more difficult words. This is possibly the result of the corpus design (i.e., TASA has many lower level texts while COCA does not) indicating a strength of using more complex corpora in developing semantic spaces. Our LDA results were mixed with spaces based on TASA outperforming spaces based on COCA, except in the higher word levels (the 5,000 and 10,000 word levels) in a manner similar to the LSA spaces. TASA may better represent vocabulary knowledge than COCA in topic models because the corpus contains more domains of knowledge than COCA, thus allowing for more topics to be induced by HDP, to support stronger representations of knowledge.

Of secondary interest is the assessment of LDA spaces in light of cognitive assessments. Previous research (e.g., McNamara, 2010) has emphasized the importance of LSA models in understanding cognition by enabling large-scale representations of human knowledge. This study provides

some evidence that LDA models, like LSA models, can also provide insights into cognitive processing at the semantic level.

Conclusion

In conclusion, this study demonstrates benefits for developing LSA and LDA models using larger corpora but also opens avenues for future research. Primary among these are the need to investigate LSA and LDA models developed on large and small corpora that are counterbalanced in terms of domains covered and linguistic complexity. In addition, future studies should continue to examine the potential for LDA models to contribute to our understanding of cognitive processing.

Acknowledgments

This research was partially supported by NSF grants 1417997, 144 PRJ88JH and 1418378.

References

Arora, R., and Ravindran, B. 2008. Latent dirichlet allocation based multi-document summarization. In *2nd Workshop on Analytics for Noisy Unstructured Text Data* (pp. 91–97). Singapore: ACM.

Blei, D.M., and Lafferty, J. 2006. Dynamic topic models. In *23rd Int. Conf. on Machine Learning (ICML '06)*. Pittsburgh, PA: ACM.

Blei, D.M., Ng, A.Y., and Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(4-5), 993–1022.

Cha, S.H. 2007. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences, 1*(4), 300–307.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D.M. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (Eds.), *23rd Annual Conference on Neural Information Processing Systems (NISP 2009)* (pp. 288–296). Vancouver, Canada.

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly, 34* (2), 213-238.

Davies, M. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing, 25*(4), 447–465.

Deerwester, S., Dumais, S.T., Furnas, G.W., Harshman, R., Landauer, T.K., Lochbaum, K., and Streeter, L. 1989. USA Patent No. 4,839,853. 4,839,853: USPTO.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dumais, S.T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1), 188–230.

Giesbers, B., Rusman, E. and van Bruggen, J. 2006. State of the art report in knowledge sharing, recom- mendation and latent semantic analysis, Technical report, Cooper Consortium.

Golub, G.H., and Reinsch, C. 1970. Singular value decomposition and least squares solutions. *Numerische Mathematik, 14*(5), 403–420.

Howard, M.W., and Kahana, M.J. 1999. Temporal Associations and Prior-List Intrusions in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 923–941.

Jones, M. N., and Mewhort,D. J. K. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1-37.

Kintsch, W. 2001. Predication. *Cognitive Science, 25*(2), 173–202.

Kullback, S., and Leibler, R.A. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics, 22*(1), 79–86.

Landauer, T.K., and Dumais, S. 2008. Latent semantic analysis. *Scholarpedia, 3*(11), 4356.

Landauer, T.K., and 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), 211–240.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E. 1997. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In M. G. Shafto and P. Langley (Eds.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society (pp. 412–417).

Manning, C.D., and Schütze, H. 1999. *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McCallum, A.K. 2002. MALLET: A Machine Learning for Language Toolkit. Amherst, MA: University of Massachusetts Amherst. Retrieved from http://mallet.cs.umass.edu/

Nation, P. 1990. *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.

Nelson, D. L., McEvoy, C., and Schreiber, T. 1998. The University of South Florida word association, rhyme, and word fragment norms. 1998 http://www. usf. edu.

Owen, S., Anil, R., Dunning, T., and Friedman, E. 2011. *Mahout in Action*: Manning Publications Co.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Level Test. Language Testing 18, 55- 88.

Terra, E., and Clarke, C. L. A. 2003. Frequency estimates for statistical word similarity measures. In *Naacl '03: Proceedings of the 2003 conference of the north amer- ican chapter of the association for computational linguistics on human language technology* (pp. 165–172). Morristown, NJ, USA: Association for Computational Linguistics.

Villalon, J. and Calvo, R. A. 2009. Single Document Semantic Spaces. *The Australasian Data Mining conference*, Melbourne,

Wiemer-Hastings, 1999. How Latent is Latent Semantic Analysis?. In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence*, pp. 932–937 San Francisco. Morgan Kaufmann.

Zaromb, F.M., Howard, M.W., Dolan, E.D., Sirotin, Y.B., Tully, M., Wingfield, A., and Kahana, M.J. 2006. Temporal Associations and Prior-List Intrusions in Free Recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 792–804.