On the relevance of documents for semantic representation

Laurianne Sitbon
National ICT Australia
Queensland University of Technology
Brisbane, Australia
laurianne.sitbon@nicta.com.au

Peter Bruza
Queensland University of Technology
Brisbane, Australia
p.bruza@qut.edu.au

Abstract *The subject of this paper is the quality of semantic vector representation with random projection under various conditions. The main effect we are watching is the size of the context in which words are observed. We are also interested in the stability of such representations since they rely on random initialisation. In particular we investigate the possibility of stabilising terms representations through documents representations. The quality of semantic representation was tested by means of synonym finding task using the TOEFL test on the TASA corpus. It was found that small context windows produces the best semantic vectors with 59.4 % of the questions correctly answered. Processing the projection between terms and documents representations several times was found not to improve the stability of the representation. It was also found not to improve the average quality of representations.*

Keywords Natural Language Techniques and Documents, Semantic spaces, Random projection.

1 Introduction

In computational linguistics, information retrieval and applied cognition, words are often represented as vectors in a high dimensional space computed from a corpus of text. In a variety of studies from cognitive science there have been encouraging results using such representations to replicate human word association norms, for example, semantic association (see, for example, [11], [10], [14]). Therefore, there is some evidence such vector representations do capture semantics of words in a way which accords with those we carry around "in our heads". We will call such representations "semantic vectors". The aim of this paper is to evaluate the effect of granularity on the quality of semantic vectors. Both documents (low granularity) and windows (high granularity) will be used to compute semantic vectors.

Dimensionally reducing the term-documents matrix has often been shown to improve the quality of semantic vectors, for example, latent semantic analysis. However, singular value decomposition, the means for dimension reduction is computationally expensive. Random Indexing [12] offers a computationally inexpensive alternative to dimension reduction [15]. However, the semantic vectors computed by RP are not stable due to the final semantic vector representations depending on initial random seeds. We aim at investigating if the use of an iterative repetition of the projection process between terms and documents representation will lead to more stable representations. We will also investigate if it is more efficient.

The structure of this paper is as follows. In the next section the semantic vector model of random projection will be described. Thereafter the TOEFL test will be used as a means of evaluating semantic vector representations in relation to the questions just raised.

2 Semantic space models

The idea behind semantic spaces is that the meaning of a word is carried by the words that co-occurs with it, and that two words are semantically related if they tend to co-occur with the same words. Co-occurrence is defined with respect to a context, for example, a window of fixed length, or even a document. Co-occurring words can be stored into matrices where the rows can represent the terms and the columns can represent contexts. Each row corresponds to a vector representation of a word. The strength of semantic association between words can be computed by using cosine - the smaller the angle between words representations, the more semantically related they are assumed to be.

Proceedings of the 13th Australasian Document Computing Symposium, Hobart, Australia, 8 December 2008.

2.1 Random Projection

Random Projection (RP) is based on the fact that a term-document matrix computed from a corpus is sparse. The sparsity is large enough that the vector representations can be projected onto a basis comprising a smaller number of randomly allocated vectors. Due to sparseness condition, the basis of random vectors has, in general, a high probability of being orthonomal [2]. The algorithm proceeds in 4 steps after the creation of a document- (or term-) term matrix : (1) create an empty matrix where rows are documents and the columns new random vectors of dimension t, (2) randomly insert in each document vector $t/6$ of positive seeds and $t/6$ of negative seeds, (3) generate a matrix where the rows are terms and the columns new dimensions by adding the corresponding random vector to a term each time it appears in a document, (4) generate the new matrix of documents in new dimensions by adding the corresponding term vector each time a document contains a term.

This can be seen mathematically as the new representation M^{random} of an initial term-document matrix M spanning N terms in d documents and then reduced to t dimensions through a random matrix as in Equation 1.

$$M^{random}_{t\times N} = Random_{k\times d} M_{d\times N} \quad (1)$$

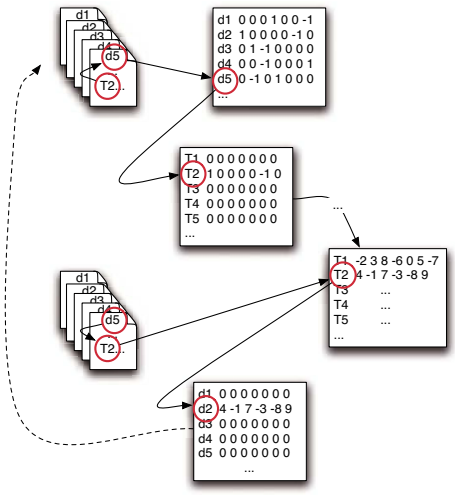

Figure 1: Process of random projection to compute term and documents matrices.

The process is illustrated on Figure 1 where it is suggested that the last two steps could be repeated, using the previously computed matrix for documents instead of the initial random one.

The number of positive and negative random seeds initially followed a Gaussian distribution but it has been shown [1] that a probabilistic distribution with 1/6 is equivalent. This method can be applied to retrieve documents and is referred to as Random Indexing [12]. The initial representation can also be based on contexts [8].

3 Experiments

3.1 Experimental setup

As a means to compare the semantic vector models above, the TOEFL synonym task on the the TASA corpus was used. The basic hypothesis is the higher the TOEFL score, the better the quality of the underlying semantic vector. This choice follows many similar evaluations in the literature and allows our results to be placed in the perspective of other published results. The TOEFL synonym test comprises 80 questions. Each question is multiple choice made of a question word and four potential answers. A question is "incomplete" if the question term is unknown to the model in question, for example, because the question words were not present in the model. In the main experiment both the number of correct answers and the number of answerable questions will be reported. In the best results section the scores will be calculated according to the measure introduced in [9] where non-answerable questions will be scored 0.25 each thereby simulating guessing. The TASA[1] contains 44,486 documents of "General Reading up to 1st year college" . It is assumed American students can learn relevant vocabulary and language usage from these readings. These documents contain 148,221 different non-stop terms for a total of 8,605,497 words. We have performed the experiments using a java implementation of Random Projection provided by the semantic vectors package[2][15]. Both corpus and questions were stemmed with a Porter Stemmer implementation[3] and the corpus is indexed with Lucene[4] to generate the initial matrix. Both term-document and term-context matrices were investigated. The minimum frequency of terms in the initial representation is set to 2 and the values of the initial seeds are either -1 or +1. Over the 80 questions of the TOEFL test, two are incomplete within all models constructed using Random Projection with stemming and 6 are incomplete without stemming. As mentioned previously, the semantic vectors produced by Random projections are somewhat unstable due to the use of

[1]We are grateful to Tom Landauer for providing the TASA corpus
[2]http://code.google.com/p/semanticvectors/
[3]http://tartarus.org/ martin/PorterStemmer/
[4]http://apache.lucene.org

random seeds during initialization. Therefore, the experiments are reported on the basis of 5 runs.

3.2 The effect of dimension reduction

The effect of varying dimension size is evaluated on word-document matrices constructed from the corpus. The notion of projection aims at some generalisation over the initial content of documents. Stemming is a first dimension reduction in that sense since it projects words onto word stems. In the context of random projection the number of random vectors used as a basis for representing the terms is another means of dimension reduction.The average results of testing RP with various dimensions are reported on Figure 2.The average results for non-stemmed data represented on the dashed line are well under the accuracy of stemmed data. Interestingly the highest average value of 37.4

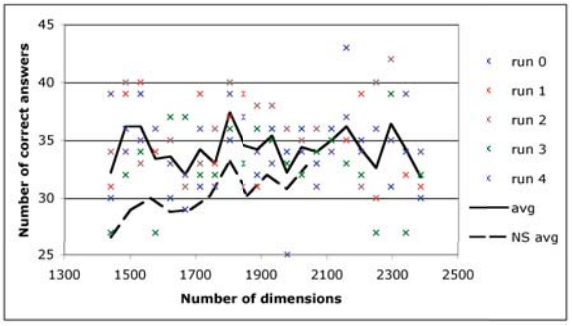

Figure 2: Accuracy of Random Projection for various numbers of dimensions.

correct answers out of 78 (33.2 without stemming) is obtained with 1800 dimensions which is the number of dimensions recommended in [2]. The five individual results for each run on stemmed data consistently exhibit a quite large variation suggesting the underlying vector representations are not that stable.

3.3 Stabilising representations with cycles

The idea suggested in Figure 1 of repeating the last two steps could lead to more stable representations. We have experimented with this idea by using different numbers of cycles (iterations) for the best set of parameters according to previous experiment : stemmed documents projected on 1800 random dimensions. The iterative reallocation of values on random vectors from terms matrix to document matrix and vice-versa doesn't improve neither the quality of representations according to figure 3 nor the stability between two representations.

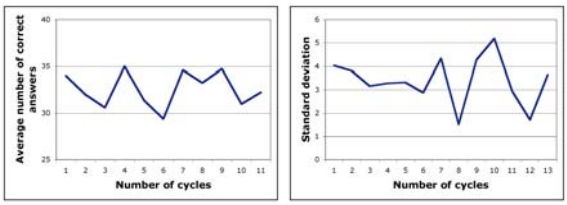

Figure 3: Average (left) and standard deviation (right) of the number of correct answers on 5 RP models for various numbers of cycles.

3.4 Reducing the granularity : context windows

As a mean of evaluating the semantic impact of full documents on semantic representation we have also built models based on an initial term-term matrix computed with a sliding window. This leads to the optimal size of the context in which words are considered to be co-ocurring. Figure 4 shows the average results for various context window sizes with random vectors. The random vector have been computed with 1800 dimensions since this size lead to the best results in previous experiments. The window sizes refer to the total number of words taken into account including the target word. The results show

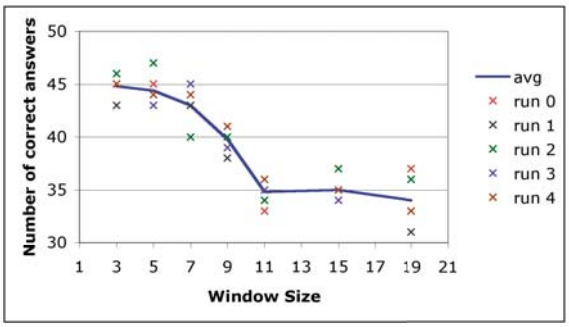

Figure 4: Accuracy of Random Projection using a word-word matrix with different context window sizes.

that the smallest context window (3 words) provides the most accurate results on the TOEFL test with 45 correct answers out of 78 in average. This implies that the model constructed based on the co-ocurrencies of the words only with the previous and the next word (these not being stop-words) performs the best for the synonym test. With a context window size of up to 9, the results are higher (40 correct answers out of 78) than when using whole a document as context. It is however important to note that context based models are more computationally expensive.

4 Conclusion

The best average obtained with a minimal window of size 3 words leads to a score of 59.4% of accuracy according to TOEFL evaluation. Comparison with previous published work should be viewed in light of doubt regarding the size of the underlying corpus. In this paper, the corpus comprises 44,486 documents whereas in other studies reported in the literature, the size is either 37,600 or 30,473 articles. We are unable to explain this discrepancy. Several results have been reported on the use of LSA. [9] had 64.5% of correct answers, [6] report results of 55.31% correctly answered questions for LSA and [4] found 63.6 % of correct responses using the cosine similarity and 61.5% using an inner product instead. Random Indexing [7] using word contexts gave 35-44% with unnormalised 1800 dimensional vectors and 48-51% with normalised vectors .

The models developed for information retrieval purposes tend to show that representing semantic spaces at the document level might benefit from a context window representation of words. One of the reason of the failure of document sized contexts could be the variety of topics present in single documents resulting in noisy representations. A intermediate solution still to be tested is to create topically coherent sub-documents using a linear segmentation algorithm.

In the future, it will also be worth investigating how stable semantic vectors are with slight corpus changes, or on larger corpora. Potential other tasks for examining semantic vectors are replications of free association [13] and priming [5].

Acknowledgements NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] D. Achlioptas. Database-friendly random projections. In *Proc. of the Symposium on Principles of Database Systems*, pages 274–281, 2001.

[2] E. Bingham and H. Mannila. Random projection in dimensionality reduction : applications to image and text data. In *Proc. of the 7th KDDM*, pages 245–250, New York, NY, USA, 2001.

[3] D. M. Blei, A. Y. Ng and M. I. Jordan. Latent dirichlet allocation. *Journal of machine learning research*, Volume 3, pages 993–1022, 2003.

[4] T. L. Griffiths and M. Steyvers. Topics in semantic representation. *Psychological review*, Volume 114, Number 2, pages 211–244, 2007.

[5] M. N. Jones, W. Kintsch and D. J. K. Mewhort. High-dimensional semantic space accounts of priming. *Journal of memory and language*, Volume 55, pages 534–552, 2006.

[6] M. N. Jones and D. J. K. Mewhort. Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, Volume 114, Number 1, pages 1–37, 2007.

[7] P. Kanerva, J. Kristoferson and A. Holst. Random indexing of text samples for latent semantic analysis. In Erlbaum (editor), *Proc. of the 22nd annual conference of the cognitive science society*, New Jersey, USA, 2000.

[8] J. Karlgren and M. Sahlgren. *Foundations of real-world intelligence*, Chapter From Words to Understanding, pages 294–308. Uesaka, Y., Kanerva, P. & Asoh, H., 2001.

[9] T. Landauer and S. T. Dumais. A solution to Plato's problem : the latent semantic analysis theory of acquisitionm induction and representation of knowledge. *Psychological review*, Volume 104, Number 2, pages 211–240, 1997.

[10] W. Lowe. Towards a theory of semantic space. In J. D. Moore and K. Stenning (editors), *Proc. of the 23rd Annual Conference of the Cognitive Science Society*, pages 576–581. Lawrence Erlbaum Associates, 2001.

[11] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour research methods, instruments and computers*, Volume 28, Number 2, pages 203–208, 1996.

[12] M. Sahlgren. An introduction to random indexing. In *Proc. of Methods and Applications of Semantic Indexing Workshop*, Copenhagen, Denmark, 2005.

[13] L. Sitbon, P. Bellot and P. Blache. Evaluation of lexical resources and semantic networks on a corpus of mental associations. In *Proc. of the 6th LREC*, Marrakech, Morocco, 2008.

[14] D. Widdows. *Geometry and Meaning*. CSLI Publications, 2004.

[15] D. Widdows and K. Ferraro. Semantic vectors : a scalable open source package and online technology management application. In *Proc. of the 6th LREC*, Marrakech, Morocco, 2008.