Filaments of Meaning in Word Space

Jussi Karlgren, Anders Holst, and Magnus Sahlgren

Swedish Institute of Computer Science

Abstract. Word space models, in the sense of vector space models built on distributional data taken from texts, are used to model semantic relations between words. We argue that the high dimensionality of typical vector space models lead to unintuitive effects on modeling likeness of meaning and that the local structure of word spaces is where interesting semantic relations reside. We show that the local structure of word spaces has substantially different dimensionality and character than the global space and that this structure shows potential to be exploited for further semantic analysis using methods for local analysis of vector space structure rather than globally scoped methods typically in use today such as singular value decomposition or principal component analysis.

Vector space models

Vector space models are frequently used in information access, both for research experiments and as a building block for systems in practical use. There are numerous implementations of methods for modeling topical variation in text using vector spaces. These and related methods are used for information access or knowledge organisation of various levels of abstraction, all more or less based on quasi-geometric interpretations of distributional data of words in documents. Vector space models in various forms have been implicit in information retrieval practice at least since the early 1970's and their origin has usually been attributed to the work of Gerard Salton. His 1975 paper titled "A vector space model for automatic indexing" [1], often cited as the first vector space model, does not in fact make heavy use of vector spaces, but in his later publications the processing model was given more prominence as a convenient tool for topical modeling (see e.g. Dubin for a survey [2]). The vector space model has since become a staple in information retrieval experimentation and implementation.

Distributional data collected from observation of linguistic data can be modeled in many ways, yielding probabilistic language models as well as vector space models. Vector space models have attractive qualities: processing vector spaces is a manageable implementational framework, they are mathematically well-defined and understood, and they are intuitively appealing, conforming to everyday metaphors such as "near in meaning". In this way, vector spaces can be interpreted as a model of meaning, as semantic spaces. In this sense, the term "word space" is first introduced by Hinrich Schütze: "Vector similarity is the only information present in Word Space: semantically related words are close, unrelated words are distant" [3]. While there is some precedent to this definition

in linguistic and philosophical literature, none of the classic claims in fact give license to construct spatial models of meaning: to do so, we first must examine how the model we build in fact preserves and represents the sense of meaning it sets out to capture.

How many dimensions?

Much of the theoretical debate on vector space models has to do with how many dimensions a semantic word space should have. The typical bare vector space of terms by contexts gives a word space of tens or hundreds of thousands or even millions of dimensions, a large number which typically, in most approaches, continues to grow when more data are added. Most every element of a typical word vector will be zero which seems a waste of dimensions, most words appear to be polysemous to some extent, and most concepts – taken on a suitably coarse-grained level of analysis – would seem to be representable by many different lexical items. Without further treatment of the data a bare model fails to generalise between terms with similar but non-identical distribution patterns. This calls for the informed reduction of dimensions to a smaller number, both for ease of processing and to be able to capture similarities.

What then seems to be an appropriate dimensionality? The word space research field frequently searches for a "latent", or intrinsic, dimensionality in the data, lower than the bare dimensionality resulting from the data collection. This intrinsic dimensionality can potentially be found by processing the data set in some informed way [4, e.g.]. Most efforts heretofore have used global measures such as singular value decomposition or principal component analysis to process, examine or reduce dimensions of the observed data. Some claims as to what this intrinsic dimensionality might be are $o(100)$ for data processed by singular value decomposition in the popular latent semantic analysis framework [5] and $o(1000)$ for data processed by us in previous work using the more recent random indexing approach [6]. These figures are obtained by reprocessing the data set with parameter variation based on trial-and-error experimentation, typically evaluated by synonym tests.

In this paper, in contrast with previous dimension reduction approaches, we argue that this appropriate dimensionality is determined locally, not globally, in the space. We base our argument on an inspection of the character of a typical vector space model built from textual data.

Are there large distances in semantic spaces?

There is a huge theoretical leap from the realisation that words are defined by their contexts to furnishing a whole vector space based on the postulated distances between words – however those distances are defined and however the distributional data are collected. What sort of information are the distances supposedly based on? It would seem there is very little purchase in the data to base *any* sort of distance between say "tensor" and "cardamom" or between "chilblain", "child-birth" and "chiliad". There is a limit to what questions one

can expect a word space built by distributional data to answer. The intuitively attractive qualities of a semantic representation where meaning is distributed about a many-dimensional vector space leads us to forget that the *only* interest we ever will show the space is in its local neighbourhoods. Returning to the original discussion on word spaces cited above: "Vector similarity is the only information present in Word Space: semantically related words are close, unrelated words are distant" [3], we claim that "close" is interesting and "distant" is not, and that vector space models are overengineered to handle information that never is relevant in language modeling.

In this paper, the question we address is what sort of dimensionality one might need to model the context of a term – as opposed to how many dimensions would be necessary if one would wish to attempt to model the structure of an entire large sample of language in one contiguous and coherent space. We will investigate the *local character* of word spaces – the structure of the space within which semantically related words are expected to be found.

What is a typical angle between random pairs of words?

Human topological intuitions are based on our experiences in a two-to-three dimensional world. We live our lives more or less on a plane with occasional ventures or glances up or down from it. Many-dimensional spaces are in some important respects very different from two-to-three dimensional spaces. One such unintuitive feature of a high dimensional space is that two randomly picked points on e.g. a unit hypersphere are almost always at near orthogonal angles to each other with respect to the origin. The graphs displayed in Figure 1 show the probability distribution of the resulting angle between two randomly chosen points in 3-, 10-, and 1000-dimensional spaces respectively.

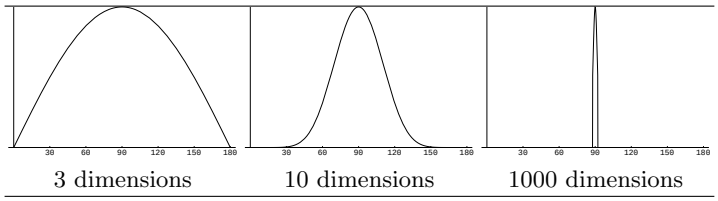

Fig. 1. Probability distribution for angles between directions to randomly chosen points in many-dimensional spaces.

The given distribution of points (and thus angles) has practical consequences for semantic models. In low dimensional spaces, given our experience of the physical world, we find it easy and intuitive to reason about distances. For example, if point A is close to point B, and point B is close to point C, then point A is fairly close to C. This makes it natural to imagine clusters of samples close to each other but separated from samples in other clusters.

However, in high dimensional spaces the transitivity of distance is not necessarily as obvious. It may well be that points A and C are both close to point B but still at considerable distance from each other. The triangle inequality still holds in high dimensional Euclidean spaces; the problem is that whereas an angle of 90° between two words always will mean that the words are completely unrelated, in a thousand-dimensional space an angle of 45° means a remarkable degree of similarity. This makes the notion of a cluster less useful in high dimensional spaces (a similar argument is given by Beyer et al. [7]).

What does a word space look like?

For this experiment we have constructed a vector space, V_{text}, from textual data as provided in the TASA corpus[8], a corpus of short high-school level English texts on various factual subjects such as language, health, business and other general school curriculum related topics. The corpus consists of about 10 million words, 37 600 text samples, and 27 000 distinct terms.

To build the word space, we use the random indexing approach as described by Sahlgren [9]. We choose random indexing for two main reasons. First, its authors make strong claims about appropriate representational dimensionality, and indeed have set $d_{representation}$, the dimension of the representation, as a settable parameter for the algorithm. Second, where in most indexing approaches each distinct term encountered in the text is assigned a binary index vector with all elements zero and one element 1, random indexing assigns each distinct term a ternary index vector with most elements zero and some randomly assigned elements either 1 or -1. Using random indexing thus avoids overloading the positive section of the vector space: by the use of negative vector elements in the representation it has the resulting word space occupy the entire possible vector space. This is desirable for our experiment, since we want to be able to relate the observed distribution of words to the expected distribution over the entire hypersphere, without leaving large swathes of vector space vacant. (The fact that most vector space models only operate in the positive sector of the multi-dimensional space is usually never discussed.)

Random indexing is convenient for our purposes, and the validity of the results is spoken for by Johnson-Lindenstrauss' lemma [10], the basis of random projection approaches, which states that a vector space (in this case, the term-by-context matrix, which is of immense dimensionality) can be projected into a random subspace of appropriate dimensionality (in this case, V_{text} of dimension, $d_{representation}$) without corrupting the relative distances between points in the space.

In this experiment, V_{text} is built from occurrence data collected from a rolling $2+2$ window over the text segments, a setting which has previously been shown by us to provide consistent results in extrinsic evaluation schemes [11]. The dimensionality, $d_{representation}$, is set to 10 000, higher than most published experiments using random indexing, with five randomly positioned 1's and five randomly positioned -1's for the initial index vectors. The relatively high choice

of $d_{representation}$ was chosen to ensure that the global dimensionality is high enough not to distort or constrain any local subspace structure.

A sample of 1 000 000 pairs of words were randomly selected from the material and the angles between them tabulated. The first neighbours in our test material appear at an angle of about 30°. Further neighbours are found at an rapidly increasing rate, and as expected, the distribution peaks around 90°. Figure 2 shows the distribution of angles between words in the sample from V_{text}. By comparing with the expected random distributions given in Figure 1 we find that the shape of the distribution for the observed data yields a global dimensionality of $o(10\ 000)$, around the dimensionality of the representation. But the distribution in Figure 2 does not match the theoretical distributions in every detail. If we zoom in on the base, as shown in the right graph, we find structure in the left tail. This represents a non-homogenous distribution at smaller angles between word vectors than would have been expected if the words were homogeneously distributed in the 10 000-dimensional space. It is in this neighbourhoood we find the non-random, i.e. semantically interesting, word-word relations.

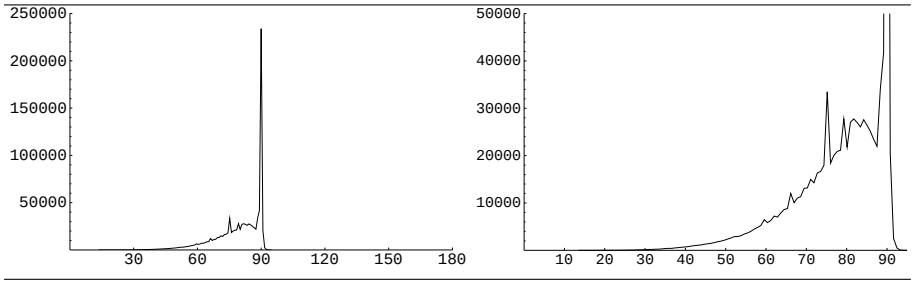

Fig. 2. Observed distance distributions; rightmost graph zoomes in on left tail.

To analyse the intrinsic dimensionality of a local region of word space we use a method from the analysis of fractal dimensions. The method rests on the mathematically trivial observation that if we have samples homogeneously distributed in a $d-$dimensional space, the number of samples within a hypersphere of radius r increases proportionally to r^d. For example, if we double the radius of a circle in a two-dimensional space the number of samples within the circle will increase by a factor of four, and for a sphere in a three dimensional space with a factor of eight. To measure the intrinsic dimensionality of the word space we examine the neighbourhoood of a point within it: we begin by counting the number of samples within a sphere of some radius r, and then we double r and count again. If the number of samples increases at a rapid rate, this means a higher dimensionality. In detail, the dimensionality d in the span between two radii r_1 and r_2 is here computed by

$$d = \frac{log(n_{r_2}/n_{r_1})}{log(r_2/r_1)}$$

where n_{r_1} and n_{r_2} are the observed numbers of samples within those radii.

An advantage with this method to measure intrinsic dimensionality is that it can measure the dimensionality locally on the small scale, as well as medium scale, or large scale, as opposed to e.g. principal component analysis or singular value decomposition which can only be used to compute intrinsic dimensionality on the global scale. This enables us to compare the intrinsic dimensionality of local contexts with the global dimensinality.

Using the above method of analysis of the fractal dimension of the samples in the left tail we find a local intrinsic dimensionality somewhere just below 10, somewhat depending on where the tail is cut off, i.e. exactly how small scale structure we care to investigate. These findings support our claim that there indeed is a local neighbourhood for a typical vector: even allowing for a substantial margin of error due to the amount of noise in the data and the somewhat crude methodology, we find that the local dimensionality is several orders of magnitude less than that of most vector space models today, including those that use dimension reduction techniques.

Can we model the local neighbourhood in word spaces?

In our experimental data and the word space V_{text}, we found, as shown in Figure 2, that the number of neigbours at small distances was greater than what would be expected from a theoretical completely homogeneous case. But we also saw that the small scale dimensionality was much less (< 10) than the large scale dimensionality ($\approx 10\ 000$). So, more neighbours, but lesser dimensionality. This gives us some clues to the local structure of the space. Consider the hand-made example graphs in Figure 3. The first graph shows a fairly homogeneous distribution of observations, both neighbours and dimensionality, in some subspace of a larger-dimensional space. If the observations instead were "clustered" or "lumpy" as in the second graph, then there is a higher number of small distances between samples than in the homogeneous case, but the small scale dimensionality of the subspace will be the same as the large scale dimensionality of the space. However, in the case shown in the third graph where the observations are found to occur in a filamentary structure, the number of small distances are also increased, but the small scale dimensionality (one-dimensional, along the filaments, in the figure) is smaller than the large scale dimensionality (two-dimensional, across the plane, in the figure).

This allows us to suggest that the word space is primarily neither homogeneous nor lumpy, but filamentary in its structure, and that these filaments, which are of much lower dimensionality than usually considered in semantic models is where the key to modelling similarity of meaning may reside.

How much data do we need to train a knowledge representation?

It is often claimed that models of meaning based on distributional data need huge data sets for training. However, it is simple to observe that in actual language use only very few occurrences are in practice necessary to model the approximative

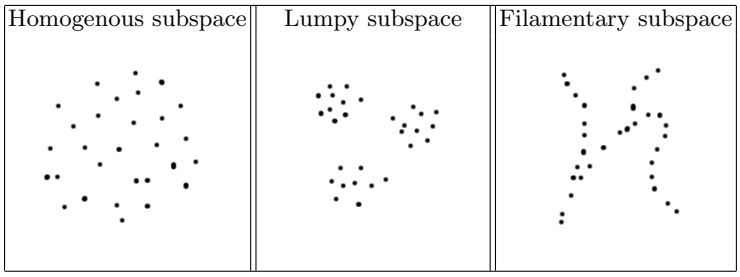

Fig. 3. Neighbourhoods of different character.

meaning of a newly encountered term. How many occurrences of "Jarlsberg" do you need to figure out what manner of beast it is? Training data found by one of the more popular web search engines are given in Figure 4. This would seem to speak to the fact that semantic context in realistic knowledge representations in fact will be saturated rather rapidly even by a small set of sample observations, that semantic similarity in fact is modellable by few dimensions rather than a plentitude. The data collected for any single term can be fairly assumed to tell us a fair amount about the term in question – and of other terms that occur similarly. The distributional data for some term will, however, not tell us much about every other term in the language and relations from the observandum to them. This observation is borne out by the data analysis from our experiment.

The famous Jarlsberg cheese is known for its distinctive sweet and nutty taste ...
The largest producer of Jarlsberg today is the Tine BA factory in ...
Within a few decades Jarlsberg has become one of Norway's greatest export successes ...
Jarlsberg is the most popular Norwegian cheese in the UK. ...

Fig. 4. What does "Jarlsberg" mean?

Ramifications for word spaces

Our conclusions and claims are three-fold. Firstly, that the most interesting qualities of word spaces are found in their local structure rather than their global dimensionality, and that thus much of the discussion of representational dimensionality, latent semantic dimensionality, and of global methods for dimension reduction is of lesser theoretical and practical import. Secondly, since the high dimensionality of the global model saddles the practical system with tractability

bottle-necks in processing, maintenance, and deployment, optimising the global character of the model is likely to provide respectable gains in efficiency. However, any claims of semantic relevance of such optimisations should be viewed with skepticism unless they expressly take local context into account. Thirdly, that elaborating the structure of the local filamentary structure further is likely to lead to less demanding models as regards size of training data sets. We do not claim to yet have a framework for how to realise such models.

We want to stress that these results are not meant to be part of the discussion of appropriate choice of dimensionality in vector space models. However, we believe that studying the local structure of vector space models will cast light on the structure of textual data, give insights into the design of future processing models, and provide new starting points for the informed design of semantic representations based on distributional data, whether in vector space models or not.

References

1. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11) (1975) 613–620
2. Dubin, D.: The most influential paper Gerard Salton never wrote. Library Trends **52**(4) (2004) 748–764
3. Schütze, H.: Word space. In: Proceedings of the 1993 Conference on Advances in Neural Information Processing Systems, NIPS'93, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1993) 895–902
4. Chávez, E., Navarro, G.: Measuring the dimensionality of general metric spaces. Technical Report TR/DCC-2000-1, Department of Computer Science, University of Chile (2000)
5. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the Society for Information Science **41**(6) (1990) 391–407
6. Kanerva, P., Kristofersson, J., Holst, A.: Random indexing of text samples for latent semantic analysis. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, Erlbaum (2000) 1036
7. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is 'nearest neighbor' meaningful? In: Database Theory - Proceedings of ICDT'99: 7th International Conference. Volume 1540/1998 of Lecture Notes in Computer Science., Jerusalem, Israel, Springer (January 1999)
8. Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Processes **25** (1998) 259–284
9. Sahlgren, M.: An introduction to random indexing. In Witschel, H., ed.: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering. Volume 87 of TermNet News: Newsletter of International Cooperation in Terminology. (2005)
10. Johnson, W., Lindenstrauss, J.: Extensions of lipshitz mapping into hilbert space. Contemporary Mathematics **26** (1984) 189–206
11. Sahlgren, M.: The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimens ional vector spaces. PhD thesis, Department of linguistics, Stockholm university (2006)