Counting Lumps in Word Space: Density as a Measure of Corpus Homogeneity

Magnus Sahlgren and Jussi Karlgren

SICS, Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
{mange, jussi}@sics.se

Abstract. This paper introduces a measure of corpus homogeneity that indicates the amount of topical dispersion in a corpus. The measure is based on the density of neighborhoods in semantic word spaces. We evaluate the measure by comparing the results for five different corpora. Our initial results indicate that the proposed density measure can indeed identify differences in topical dispersion.

1 Introduction

Word space models use co-occurrence statistics to construct high-dimensional semantic vector spaces in which words are represented as *context vectors* that are used to compute semantic similarity between the words. These models are now on the verge of moving from laboratories to practical usage, but while the framework and its algorithms are becoming part of the basic arsenal of language technology, we have yet to gain a deeper understanding of the properties of the high-dimensional spaces.

This study is ment to cast some light on the properties of high-dimensional word spaces; we find that computing a measure for the density of neighborhoods in a word space provides a measure of topical *homogeneity* — i.e. of how topically dispersed the data is. This is a fortunate discovery, since there are no established measures for corpus homogeneity. The hitherto most influential proposal boils down to defining a measure of homogeneity based on the similarity between randomly allocated parts of a corpus: the more similar the parts, the more homogeneous the corpus [3].

As an experimental evaluation of our density measure, we apply it to five different types of text corpus, each of varying degrees of topical homogeneity. The results show that the measure can indeed identify differences in topical dispersion and thus help provide some amount of understanding of what a word space is in relation to the language and the collection of text it models.

2 The density measure

The intuition our measure is based upon is the idea that words in a topically homogeneous data are used in more uniform ways than words in a topically

The final publication is available at Springer via http://dx.doi.org/10.1007/11575832_16

dispersed data. This would imply that the words in a topically homogeneous data have sparser semantic neighborhoods (i.e. fewer semantically related words) than would their topically more promiscuous counterparts. As an example, consider the difference between the semantic neighborhoods of a word with many possible meanings, such as "bark", which has nine meanings in WordNet 2.0, and a word with very few possible meanings, such as "toxin", which has only one meaning in WordNet. Obviously, the semantic neighborhood of "bark" is more populated (in the WordNet space) than the semantic neighborhood of "toxin".

In analogy with such WordNet neighborhoods, we suggest a measure of the number of words that occur within some specified radius around a given word in the word space. A large resulting number means that the word has a dense neighborhood, which indicates that the word occurs in a large number of contexts in the data, while a small resulting number means that it has a sparse neighborhood resulting from occurences in a small number of contexts. We define the density of the neighborhood of a word as the number of unique words that occur within the ten nearest neighbors of its ten nearest neighbors.

3 The word space model

We use the Random Indexing [1, 2] word space methodology, which is an alternative to algorithms such as Latent Semantic Analysis [4] that use factor analytic dimensionality reduction techniques. Rather than first assembling a huge co-occurrence matrix and then transforming it using factor analysis, Random Indexing *incrementally* accumulates context vectors in a two-step operation:

1. First, each word in the text is assigned a unique and randomly generated representation called an *index vector*. These random index vectors have a fixed dimensionality k, and consist of a small number ϵ of randomly distributed $+1$s and -1s.
2. Next, context vectors are produced by scanning through the text, and each time a word occurs, the index vectors of the n surrounding words are added to its context vector.

This methodology has a number of advantages compared to other word space algorithms. First, it is an *incremental* method, which means that the context vectors can be used for similarity computations even after just a few examples have been encountered. Most other algorithms require the entire data to be sampled and represented in a very-high-dimensional space before similarity computations can be performed. Second, it uses fixed dimensionality, which means that new data do not increase the dimensionality of the vectors. Increasing dimensionality can lead to significant scalability problems in other algorithms. Third, it uses implicit dimensionality reduction, since the fixed dimensionality is much lower than the number of contexts in the data. This leads to a significant gain in processing time and memory consumption as compared to algorithms that employ computationally expensive dimensionality reduction techniques. Fourth, it is comparably robust with regards to the choice of parameters. Other algorithms tend to be very sensitive to the choice of dimensionality for the reduced space.

4 Experiment

In order to experimentally validate the proposed measure of corpus homogeneity, we first build a 1,000-dimensional word space for each corpus using Random Indexing, with parameters $n = 4$, $k = 1{,}000$, and $\epsilon = 10$.[1] Then, for each corpus, we randomly select 1,000 words, find their ten nearest neighbors, and then those neighbors' ten nearest neighbors. For each of the 1,000 randomly selected words, we count the number of unique words thus extracted. The maximum number of extracted neighbors for a word is 100, and the minimum number is 10. In order to derive a single measure of the neighborhood sizes of a particular corpus, we average the neighborhood sizes over the 1,000 randomly selected words. The largest possible score for a corpus under these conditions is 100, indicating that it is severly topically dispersed, while the smallest possible score is 10, indicating that the terms in the corpus are extremely homogeneous.

We apply our measure to five different corpora, each with a different degree of topical homogeneity. The most topically homogeneous data in these experiments consist of abstracts of scientifical papers about nanotechnology (NanoTech). Also fairly homogeneous are samples of the proceedings from the European parliament (EuroParl), and newswire texts (ReutersVol1). Topically much more dispersed data are two examples of general balanced corpora, the TASA and the BNC corpora. Since the NanoTech data is very small in comparison with the other corpora (only 384,199 words, whereas the other corpora contain several millions of words), we used samples of comparable sizes from the other data sets. This was done in order to avoid differences resulting from mere sample size. The sampling was done by simply taking the first $\approx 380{,}000$ words from each data set. We did not use random sampling, since that would affect the topical composition of the corpora.

We report results as averages over three runs using different random index vectors. The results are summarized in Table 1.

Table 1. The proposed density measure, as compared with the number of word tokens, the number of word types, and the type-token ratio, for five different English corpora.

Corpus	Word tokens	Word types	Type token ratio	Average density measure	Standard deviation
NanoTech	384,199	13,678	28.09	**49.149**	0.35
EuroParl (sample)	375,220	9,148	41.47	50.9736	0.38
ReutersVol1 (sample)	368,933	14,249	25.89	51.524	0.18
TASA (sample)	387,487	12,153	31.88	52.645	0.45
BNC (sample)	373,621	18,378	20.33	54.488	0.74

[1] These parameters were chosen for efficiency reasons, and the size of the context window n was influenced by [2].

5 Provisional conclusions

The NanoTech data, which is by far the most homogeneous data set used in these experiments, receives the lowest density count, followed by the also fairly homogeneous EuroParl and ReutersVol1 data. The two topically more dispersed corpora receive much higher density counts, with the BNC as the most topically dispersed. This indicates that the density measure does in fact reflect topical dispersion: in more wide-ranging textual collections, words gather more contexts and exhibit more promiscuous usage, thus raising their density score.

Note that the density measure does not correlate with simple type-token ratio. Type-token ratio differentiates between text which tends to recurring terminological usage and text with numerous introduced terms. This can be seen to indicate that terminological variation — in spite of topical homogeneity — is large in the EuroParl data, which might be taken as reasonable in view that individual variation between speakers addressing the same topic can be expected; *text style* and *expression* have less effect on the density measure than the topical homogeneity itself. The ranking according to the density measure:

$$\textbf{Nano} > \text{EuroParl} > \text{ReutersVol1} > \text{TASA} > \text{BNC}$$

and the ranking according to type-token ratio:

$$\text{EuroParl} > \text{TASA} > \textbf{Nano} > \text{ReutersVol1} > \text{BNC}$$

only show a 0.5 rank sum correlation by Spearman's Rho.

This is obviously only a first step in the investigation of the characteristics of the well established word space model. The present experiment has clearly demonstrated that there is more to the word space model than meets the eye: even such a simple measure as the proposed density measure does reveal something about the topical nature of the data. We believe that a stochastic model of the type employed here will give a snapshot of topical dispersal of the text collection at hand. This hypothesis is borne out by the first experimental sample shown above: text of very differing types shows clear differences in the score defined by us. We expect that other measures of more global character will serve well to complement this proposed measure which generalizes from the character of single terms to the character of the entire corpus and the entire word space.

References

1. Kanerva, P., Kristofersson, J., Holst, A.: Random Indexing of text samples for Latent Semantic Analysis. CogSci'00 (2000) 1036.
2. Karlgren, J., Sahlgren, M.: From Words to Understanding. In Uesaka, Y., Kanerva, P., Asoh, H. (eds.): *Foundations of Real-World Intelligence*. CSLI Publications (2001) 294–308.
3. Kilgariff, A.: Comparing Corpora. *Int. Journal of Corpus Linguistics* **6** (2001) 1–37.
4. Landauer, T., Dumais, S.: A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* **104** (1997) 211–240.

This article was processed using the LaTeX macro package with LLNCS style