Semantic Topology

Jussi Karlgren Martin Bohman Ariel Ekgren
Gabriel Isheden Emelie Kullmann David Nilsson

Gavagai
Stockholm, Sweden

KTH, Royal Institute of
Technology
Stockholm, Sweden

1. REQUIREMENTS FOR A PRACTICAL MODEL OF MEANING

A reasonable requirement (among many others) for a lexical or semantic component in an information system is that it should be able to learn incrementally from the linguistic data it is exposed to, that it can distinguish between the topical impact of various terms, and that it knows if it knows stuff or not.

We work with a specific representation framework – *semantic spaces* – which well accommodates the first requirement; we study the global qualities of semantic spaces by a topological procedure – *mapper* – which gives an indication of topical density of the space; we examine the local context of terms of interest in the semantic space using another topologically inspired approach which gives an indication of the neighbourhood of the terms of interest. Our aim is to be able to establish the qualities of the semantic space under consideration without resorting to inspection of the data used to build it.

2. DISTRIBUTIONAL MODELS

Distributional models, such as collocational analyses or probabilistic language models, are based on the analysis of observed item distribution and collocation in linguistic data and have a long history in linguistics. [3] Today, they provide a theoretical base and profitable results for tasks such as speech recognition, language modelling and information retrieval.

In general, distributional semantic models use the notion of *distance* between two words to describe relation in meaning. This combination of distributional data with a geometric interpretation is what defines *semantic spaces*. [15, 14] The geometric model is appealing: the notion of *closeness in meaning* speaks to our intuitions about how semantics work. This, however, would seem to be a somewhat false friend. Our geometric intuitions do not hold water for several thousand-dimensional spaces. Also, the metaphor of closeness does not deliver useful help if more complex semantic relations are considered or larger distances in the space are queried: "What is the relation between *bell pepper* and *one-pass compiler*"?; "Is *cow* closer to *horse* than *coffee* is to *tea*"?; "Is a *bullfinch* closer to *bird* than LaTeX is to *language*"? Arguably those questions are meaningless for human semantics, but are handily and uselessly answered with great exactitude by geometric semantic spaces. [7]

3. GEOMETRY AND TOPOLOGY

The insight that geometric models are overly specific and unwieldy, especially if built on realistic scale, is the motivation for e.g. dimensionality reduction approaches, various latent variable models [2, 12], graph-based models [11], and e.g. Laplacian transforms such as in self-taught hashing [18, 17, 5]. We propose here to use generalise some of those insights, and move from a semantic geometry to a semantic topology.

Semantic space models have no natural scale and no given base vectors. Topological models are resilient with respect to scale, rotational transformations, deformations, and coordinate choice, and can be constructed to focus on local structure and similarity in near relations. [1] A topological perspective of the data affords us an effective view of the structure of models, and is useful for the diagnosis and practical quality assessment of models which already have proven to be of value in real-world applications.

The basis for our experimentation are semantic spaces created using *random indexing*, [12] trained on various corpora of relevance for information processing tasks which require lexical semantics, e.g. ontology mapping, media monitoring, or topic tracking. We currently use such semantic spaces in practical large-scale industrial applications to find synonyms or near-synonyms of terms of interest, and to track associative concepts over time, as an up-to-date lexical resource. We frequently find we need to examine the models we have trained to ascertain their qualities with respect to some topic of interest. In these following experiments we will use models which are trained on traditional research corpora using the same procedure we would use on internet data for commercial purposes.

4. THE MAPPER PROCEDURE

Mapper, first introduced by Singh, Mémoli, and Carlsson [16], is an algorithm based on topological principles to visualize high dimensional data. The intuition behind Mapper is to analyze the structure of the data as a whole instead of analyzing the entire dataset in detail. Mapper is intended to capture such regularities of massive data sets which are

CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ACM 978-1-4503-2598-1/14/11 $15.00.
http://dx.doi.org/10.1145/2661829.2662028 .

obscured by focus on geometric coordinates, by transforming the data set to a *simplicial complex*, a combinatorial and discrete data structure. If the steps of this transformation is done well, the resulting structure can be inspected to understand the characteristics of the data set.

Given a dataset $D = d_i : d_i \in X$, the Mapper procedure can be given in 4 steps:

Filtering We analyze the data using a *filter function* $f : X \rightarrow \mathbb{R}$ which creates an image of the datapoints in $\mathbb{R}$. The filter function should capture some interesting or relevant aspect of the data. In this case, in an analysis of a semantic space, we can set f to be the distance in that space of each point from a target notion of interest.

Cover Given our dataset D, we structure it by appliying a *covering* to its image $f(D)$ by a set C of subsets of $\mathbb{R}$: where $C = c_i : c_i \subset \mathbb{R}$. This covering can be given by an expert or through other means. In our case we use overlapping intervals of the filter function itself, essentially grouping data into overlapping bins of a histogram. The datapoints in each bin are given by the datapoints with their image in the specified interval of $\mathbb{R}$.

Clustering The points in each bin, meaning each subset of C (the elements in C are themselves subsets of $\mathbb{R}$) are then clustered individually. The clustering algorithm can be chosen to be whatever clustering algorithm is required. We use single-linkage clustering.

Graph A graph G is created with every individual cluster from the clustering of the points in the subsets of C as a node of G. When two clusters share a common datapoint in D, an edge is drawn between the two nodes that represent them.

We illustrate the procedure first using artificially generated point cloud data, as shown in Figure 1. The data consists of 5000 points randomly generated from a Gaussian distribution surrounding three centroids at $[x, y]$ coordinates: $[10, 20]$, $[-10, -10]$, $[17, -10]$ with a standard deviation of 9. The filter function f was chosen to be Gaussian kernel density estimation. The coloring of the points in the graph follow the density estimation. The covering was set to 7 intervals with an overlap of 10 percent. After Mapper processing those same data can be visualized as shown in Figure 2 using a similar colouring scheme. Here, the graph shows that if the points at high density are clustered, there are three clusters; the points at low density cluster into one. Overlapping density ranges show the expected correspondences from high to low.

This procedure, in our application of it to semantic spaces, serves to illuminate shared structure across different distance scales of the semantic space by showing if the cluster structure in one distance range correspond or differ from another distance range.

5. GLOBAL TOPOLOGICAL CHARACTER

One of the specific questions we wish to investigate is that of *expertise*. Given two semantic spaces, what is the extent of training in some topical domain? We will assume that expertise, in the sense of being trained on a set of texts, should have effects on the topological makeup of the semantic space. We trained two semantic spaces on general English-language text[1] and then added some selected topics to each of the spaces. One semantic space was trained by including entire Wikipedia articles related to the topics; another semantic space was only given the introductory paragraphs of those same articles. Thus both semantic spaces are familiar with the foundational vocabulary of the topics, but one of them would have a passing knowledge while the other would have a more in-depth understanding of the topic.

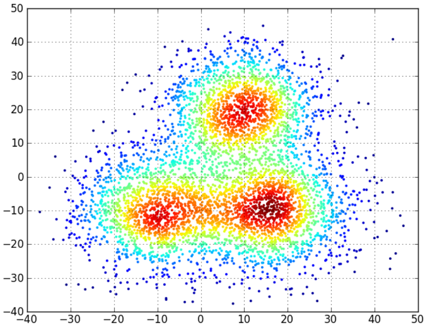

Figure 1: Artificially generated geometric data

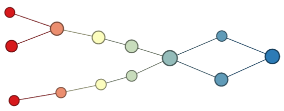

Figure 2: Geometric data transformed by Mapper

If we now apply a filter function to the points of the semantic space based on relation to T, the target concept of interest, our expectation would be that in the one semantic space, terms for probe concepts t_i known to be related to T should cluster relatively close to T; in the other they would be more or less randomly distributed over the scale intervals. This is borne out in experiments. Figure 3 shows the difference, as measured by a filter function defined by ten probe words relative to the target topic “Finland”. The graph shows how the probe words cluster both better with respect to each other, and closer to the target.

6. LOCAL TOPOLOGICAL STRUCTURE

The second question we wish to address is that of differential qualities of terms we have observed. Some words are more topical than others, which has been observed in numerous different research traditions, but most notably in practically oriented text analysis. [9, 10, 6, 4, 13] We wish to examine

[1] Settings: 2000 random indexing dimensions, 2 + 2 context, trained on the The British National Corpus. (Distributed by Oxford University Computing Services at url http://www.natcorp.ox.ac.uk/).

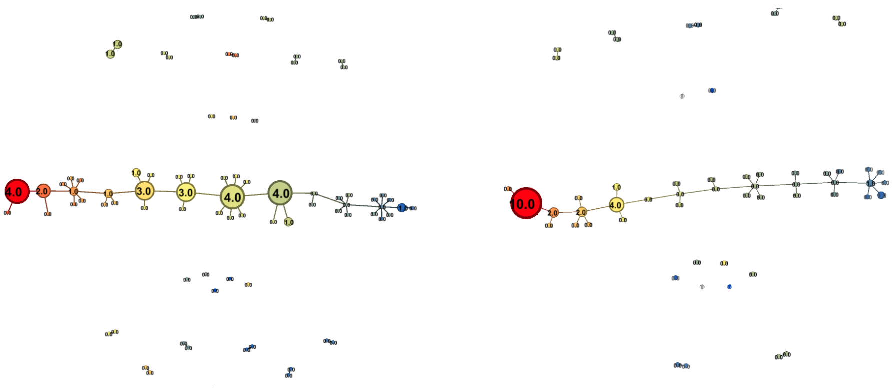

Figure 3: Data for passing knowledge vs expertise for "Finland"

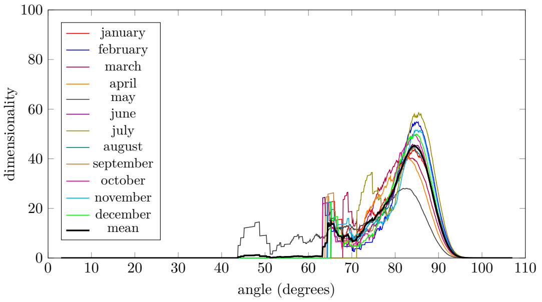

Figure 4: Local dimensionality at various angular separation for names of months

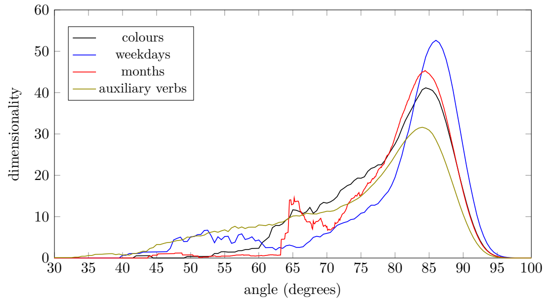

Figure 5: Local dimensionality at various angular separation for names of months, weekdays, colours and auxiliary verbs

the local structure of the semantic space around a term of interest.

We recently experimented using a topologically related approach to establish the density of a neighbourhood for terms in a semantic space and to thus infer the *intrinsic dimensionality* of the local space around the term. While we expand the radius around the spatial coordinates of a term of interest we record the rate of increase in number of neighbours within that radius. We begin by establishing how rapidly the number of neighbours of a term grows in relation to the growth of the radius of the neighbourhood. We define the rate of growth in the interval $r \in I = [r_1, r_2]$ to be

$$d = \frac{log(n_2/n_1)}{log(r_2/r_1)}, \tag{1}$$

with n_i being the number of observed term neighbours within the radius r_i. We use d as an estimate of the local dimensionality around the probe term in r. Averaging the results of those computations over an entire semantic space we find that the local dimensionality was considerably lower than that of the representation itself. [8]

Here, we follow a similar approach, but instead study the particularities of individual terms, or specific categories of term.[2] In Figures 4, 5, 6, 7, we plot d at various radius ranges on the surface of a unit hypersphere, with the radius here graphed as the angle of separation between the probe term vector and the neighbouring vectors.

As a first illustration, Figure 4 shows the rate of neighbourhood growth curves for names of months, with the angle as computed from the origin on the x-axis. Note that *May* behaves differently from the other months, due to the polysemy of the word. Figure 5 shows a comparison of results between names of months, weekdays, colours, and auxiliary verbs. The latter have a much more flattened distribution; the former three all have higher neighbourhood density at lower angular distances, and then the majority of neighbours at around ninety degrees, i.e. at maximum distance from the term, indicating no semantic relation. This is to be expected since months, weekdays, and colours all have fairly well delimited semantics and thus contexts of use, whereas auxiliaries can be expected to cooccur with numerous subjects and verbs and thus have a much more promiscuous context. This comparison would lead us to expect that content-heavy words are likely to have neighbours accrue earlier, at smaller angular distances.

A comparison between the figures in Figures 6 and 7 confirms this. One shows the neighbourhood growth curves for the 150 most frequent words found in the corpus: *the, be, to, of, and, ...*. The other shows the same curves for some 300 terms which are found in Wikipedia: topic headers *england,*

[2] Settings: 1000 random indexing dimensions, $2+2$ context, trained on the TASA corpus.

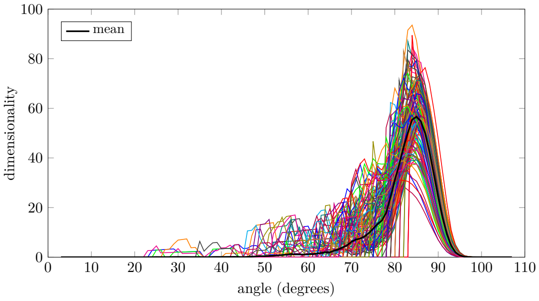

Figure 6: Local dimensionality at various angular separation for the most frequent words in the corpus

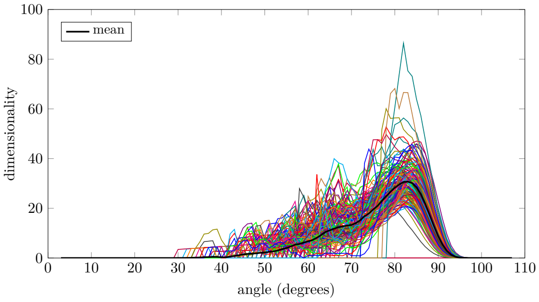

Figure 7: Local dimensionality at various angular separation for the most frequent words in the corpus

mississippi, instagram, socrates ... . The form of the curves are clearly different even at a cursory inspection.

To verify this observation, we performed several simple categorisation experiments, based on minimising square error to the dimensionality graprh, to distinguish parts of speech and term lists of various classes of word. Table 1 shows the result of categorising the classes given in Figure 5. Similarly useful results were found between other categories of term such as various semantic categories of verbs.

7. CONCLUSIONS

Semantic spaces, a useful learning framework for lexical resources, are typically treated as black boxes and applied using geometric and linear algebraic processing tools. We have found that topological methods are useful for exploring the makeup of a semantic space.

8. REFERENCES

[1] Gunnar Carlsson. Topology and data. *American Mathematical Society*, 46, 2009.

[2] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), 1990.

[3] Zellig Harris. *Mathematical Structures of Language.* Interscience publishers, 1968.

[4] Stephen Harter. A probabilistic approach to automated keyword indexing. *Journal of the American Society for Information Science*, 26, 1975.

[5] Xiaofei He, Deng Cai, Haifeng Liu, , and Wei-Ying Ma. Locality preserving indexing for document representation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[6] Aurélie Herbelot and Mohan Ganesalingam. Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 2013.

[7] Jussi Karlgren. Meaningful models for information access systems. In Arppe, Carlson, Heinämäki, Lindén, Miestamo, Piitulainen, Tupakka, Westerlund, and Yli-Jyrä, editors, *A Finnish Computer Linguist: Kimmo Koskenniemi Festschrift on the 60th birthday.* CSLI Publications, 2005.

[8] Jussi Karlgren, Anders Holst, and Magnus Sahlgren. Filaments of meaning in word space. In *Proceedings of the 30th European Conference on Information Retrieval*, 2008.

[9] Slava Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1), 1996.

[10] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[11] Irina Matveeva, Gael Dìaz, and Ahmed Hassan, editors. *TextGraphs-7 '12: Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[12] Gabriel Recchia, Michael Jones, Magnus Sahlgren, and Pentti Kanerva. Encoding sequential information in vector space models of semantics: Comparing holographic reduced representation and random permutation. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.

[13] Stephen Robertson, Cornelis J. van Rijsbergen, and Michael Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3d Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1980.

[14] Magnus Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.* PhD Dissertation, Department of Linguistics, Stockholm University, 2006.

[15] Hinrich Schütze. Word space. In *Advances in Neural Information Processing Systems 5*, San Francisco, CA, USA, 1993. Morgan Kaufmann.

[16] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In M. Botsch and R. Pajarola, editors, *Eurographics Symposium on Point-Based Graphics.* The Eurographics Association, 2007.

[17] D. Zhang, J. Wang, D. Cai, and J. Lu. Laplacian co-hashing of terms and documents. In *Proceedings of the 32nd European Conference on Information Retrieval*, 2010.

[18] Dell Zhang, Jun Wang, Deng Cai, and Jinsong Lu. Self-taught hashing for fast similarity search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

Inferred class → Actual class ↓	Colour	Month	Auxiliary verb	Weekday
Colour	56%	33%	11%	0%
Month	17%	67%	8%	8%
Auxiliary verbs	3%	13%	84%	0%
Weekday	0%	0%	0%	100%

Table 1: Confusion matrix for categorisation