New Measures to Investigate Term Typology by Distributional Data

Jussi Karlgren
from
Kungliga Tekniska Högskolan
and
Gavagai, Stockholm
jussi@kth.se

ABSTRACT
This report describes a series of exploratory experiments to establish whether terms of different semantic type can be distinguished in useful ways in a semantic space constructed from distributional data. The hypotheses explored in this paper are that some words are more variant in their distribution than others; that the varying semantic character of words will be reflected in their distribution; and this distributional difference is encoded in current distributional models, but that the information is not accessible through the methods typically used in application of them. This paper proposes some new measures to explore variation encoded in distributional models but not usually put to use in understanding the character of words represented in them. These exploratory findings show that some proposed measures show a wide range of variation across words of various types.

KEYWORDS: Term typology, distributional semantics.

1 Words and terms in use — general requirements for a language model

For text analysis tasks such as information retrieval, terminology mining, or conceptual modelling, words and terms naturally lend themselves as surrogates for documents or representations of concepts. A necessary component in any system for text analysis is a representation, explicit or implicit in the text analysis process, for the concepts expressed in text by computation over the terms that express them. Since the lexical variation is great, such a representation must select or weight or consider the words, terms, and constructions judiciously, typically based on observable characteristics of the items in the text collection under consideration, sometimes consulting language resources compiled elsewhere such as lexica or ontologies.

Target notions for the usefulness of a word or term for some typical text analysis tasks are e.g. *representativeness* for some topic, *specificity* in discriminating between documents of different topics, and *topicality* in general, meaning how likely its appearance in a text is evidence of it being a carrier of text topic. (Spärck Jones, 1972; Hisamitsu et al., 2000; Katz, 1996) A representation or language model, whether probabilistic, geometric, or symbolic, should obviously be designed to capture the target notions most relevant for the task at hand. The three characteristics given above are not immediately observable in themselves — they are derived from observed distributional behaviour of words and terms in text and discourse and from an understanding of what topic is, extratextually.

But, aside from a word's potential usefulness for a task, it will have semantic characteristics which usually are acknowledged by the community of language users: terminological *vagueness*, *abstraction*, *indefiniteness*, and *change* over time are obvious, ubiquitous, and salient characteristics of words and terms in human language, but seldom afforded any place in computation. Any language model, whether it addresses hands-on tasks in information access application, machine translation, dialogue systems or other application area for human alnguage processing or if it built to elucidate the workings of human communicative behaviour in the abstract should be expected to address those semantic characteristics which are most obvious to human users of language.

What character can words and terms then have, observably? An observable difference between words and terms is their distribution over semantic neighbourhoods. Some words and terms are focussed and specific; others are inspecific and spread over several usage patterns. In this initial paper we propose some measures to explore this distributional variation.

2 Experiments on word distributions

These first experiments concern the behaviour of *words*. Terms are frequently composed of several words in conventional fixed configurations (Smadja, 1993; Justeson and Katz, 1995) and the arguments and experiments given here can be generalised to multi-word terms and even constructions, but for practical reasons these initial experiments are performed on single-token words. The distribution of words is here studied in two text collections — a collection of newprint from Reuters comprising 200 million words in short news telegrams, and one month of collected English-language blog text comprising 189 million words in short blog posts.

Our hypotheses are (1) that some words are more variant in their distribution than others — an assumption that is not difficult to defend; (2) that the varying semantic character of words will be reflected in their distribution; and (3) this distributional difference is encoded in current distributional models, but that the information is not accessible through the methods typically used in application of them.

abstract terms

ability adventure amazed anger anxious awe bad beauty belief brave brutal calm chaos charity childhood clarity comfort communication compassion confidence content courage crime curious customer death deceit dedication defeat delight democracy despair determined dictatorship disappointment disbelief disquiet disturbance education ego elegance energy enhancement enthusiasm envy evil excited failure faith faithful fascination fear forgive fragile frail free freedom friend friendship generous glitter good grace gracious grief happiness happy hate hatred hearsay helpful helpless home honest hope hurt idea imagination impression improvement infatuation inflation insanity intelligence jealousy joy justice kindness knowledge laughter law liberty life loss love loyal loyalty luck lucky luxury mature maturity memory mercy moral motivation move movement music need opinion opportunity pain patience peace peculiar peculiarity pleasure poor poverty power pride principle real reality refreshment relief restoration rich rumour sacrifice sad sadness sanity satisfaction self-control sensitive service shock silly skill sparkle speculation speed strength stupid success surprise sympathy talent thrill tired tolerance trust unemployment upset warm weak weakness wealth wisdom worth

Swadesh terms

i you we this that who what not all many one two big long small woman man person fish bird dog louse tree seed leaf root bark skin flesh blood bone grease egg horn tail feather hair head ear eye nose mouth tooth tongue nail foot knee hand belly neck breast heart liver drink eat bite see hear know sleep die kill swim fly walk come lie sit stand give say sun moon star water rain stone sand soil cloud smoke fire ashes burn path mountain red green yellow white black night hot cold full new good round dry name

Figure 1: Abstract and Swadesh terms used in the experiment

To test the hypotheses under consideration we established two word lists to provide a basis for differentiating behaviour of different types of word. We initially took the 100 word list by Swadesh with cross-linguistically translatable words as an example of concrete and fairly invariant concept references. (Swadesh, 1971) We added a list of some 160 abstract terms taken from various author guides.[1] The terms used are given in Figure 1.

2.1 Simple distributional measures

For practical semantic tasks such as search engines, topic modelling, or text categorisation, term usage is computed from their occurrence statistics. The typical search engine categorises terms according to their potential information content inasmuch can be determined from its distribution over a document collection, the target task being to separate documents from each other using search terms as a criterion. This sort of computation serves well to distinguish topical from non-topical words. Table 2.1 shows how the Okapi BM25 formula (Robertson and Zaragoza, 2009), the base for many, or even most, practical search engines today weights terms differentially depending on their topical qualities based on a combination of term frequency within documents, the number of documents the term appears in often with a document length factor added in to compensate for the effect of long documents on term frequency counts.

In general, formulæ such as BM25 do a good job of taking out words with a broad and uninteresting distribution, as well as infrequent words. Words with a low BM25 tend to be function words, misspellings, and unusual compounds. Words with a high BM25 score will be trade marks, names, and technical terms. This serves the needs of a topical search engine well.

A more principled approach to modelling word distribution is the three-parameter burstiness model formulated by Slava Katz (Katz, 1996). His model uses occurrence statistics to estimate α, the likelihood of encountering a term, γ the likelihood of its being topical if encountered, and B, the burstiness of the term, if established as topical. These estimates are handily calculated with α being the observed relative frequency of the term, γ the observed number of documents where the term occurs more than once, and B the average frequency of a term in those documents

[1] Author guidelines tend to issue blanket warnings to aspiring writers, discouraging abstract and vague vocabulary, irrespective of if it would be warranted or not.

	Newsprint		
	average tf	idf	$BM25$
and	4.29	2.92e-7	3.20
it	2.74	6.45e-7	5.61
	...		
food	1.72	2.09e-5	13.2
hunger	1.50	0.000453	14.8
eat	1.21	0.000653	12.3
beef	2.40	8.50e-5	10.56
seafood	1.43	0.00239	10.56
barbecue	1.19	0.0149	2.97
jerky	1.00	0.0454	0.841

Table 1: Search engine measures for some sample words for news text.

	Newsprint		
	α $\hat{p}(occurrence)$	γ $\hat{p}(topicality)$	B $\hat{p}(burstiness)$
and	0.270	0.819	6.20
it	0.177	0.589	3.52
		...	
food	0.0123	0.347	3.09
hunger	0.000616	0.319	2.59
eat	0.0000545	0.159	2.3
beef	0.00236	0.613	3.29
seafood	0.000131	0.242	2.79
barbecue	0.0000239	0.121	2.57
jerky	0.00000907	0	0

Table 2: Katz measures for some sample words for news text.

where its frequency is over 1. Katz' α is high for frequent and low for infrequent words; Katz' γ is high for words that are likely to be repeated and thus likely to be topical in the contexts where it is used; for Katz' burstiness we find very high scores for function words, high scores indicative of technical terms and trade marks and low burstiness scores for hapax legomena and misspellings, as exemplifed in Table 2.1. A word such as *music* will has in our present data set an observed likelihood of 0.38 of being repeated once seen in text, and thus, following Katz' estimate, a relatively high probability of being a topical term in discourse; a word such as *stupid* an observed repeat likelihood of 0.09 and a lower attendant estimated probability of being topical.

2.2 Variation in local context size

As a first test we collected the variation in the immediate neighbourhood of the words under consideration. For each word, we tabulated its occurrences and its immediately adjacent words

	#different words/#observations	
	blog text	newsprint
eat	0.09	0.35
sun	0.18	0.58
		...
talent	0.91	0.95
gossip	1.10	1.56

Table 3: Number of neighbours in local context

in a window of two words to the left and to the right. We find that function words such as *and it*, vague terms such as *good*, and very general verbs such *see* have a more wide range of neighbours than do more referentially specific terms such as *frail soil* or *defeat*.

Table 3 gives some examples of terms ranging from 0.09 to 1.56. The theoretical maximum of this score is 4 for a completely dispersed neighbourhood: with two words collected to the left and two to the right, if every observation has all those four words different the score will be 4. The table of results is sorted by score, and subsequent rank sum tests show significant differences between the Swadesh terms and the abstract and vague term list, with the former having a more focussed and presumably semantically more constrained neighbourhood. (Mann Whitney U, $p > 0.95$) This gives us reason to experiment further, to see how their neighbourhood evaloves, related to the two different types of word and to their respective occurrence statistics over a growing number of observations.

2.3 Variation in a semantic space

In the following experiments a distributional semantic space representation is used. Distributional semantic models are based on the assumption that similarity in meaning between entails similarity in usage and thus in the distribution of the words under consideration over a text collection. This assumption holds, as shown in numerous word semantic evaluation experiments. The model in these experiments is built from the two data sets given above using a standard implementation of the random indexing framework (Kanerva et al., 2000) using 1000-dimensional ternary vectors of density 3. Each word encountered in the text is at time of first observation accorded one randomly generated index vector. The distributional context of a word is then represented with another 1000-dimensional *context vector*, into which, for each occurrence of the word, are added the index vectors of adjacent words in the immediate left and right contexts (Sahlgren et al., 2008). This context vector will over time, after a number of observations, capture the close distributional neighbourhood of the word in all occurrences encountered thus far. Words with similar contextual distributions will converge towards similar context vectors. The context in these experiments are modelled by a $2 + 2$ window which has been shown by previous work to best capture tight semantic relations such as synonymy (Sahlgren, 2006). This setting is naturally an obvious one to vary in future experiments — is the semantic variation more noticeable in the tight semantic relations or in the broader more associative relations given by a broad window.

This context vector constitutes the basic data on which the following experiments are performed. This model is a fairly general distributional model and is used not for its specific characteristics but for processing convenience. It is reasonable to assume that most results would in the main

honest	25
person	74
education	85
law	96

Table 4: Number of second-order neighbours for some selected words

translate to any more sophisticated semantic space or probabilistic language model.

2.3.1 Semantic tightness: Overlap

In a semantic space, the meaning of a word is determined by its neighbours. This first measure computes the number of second order-neighbours by taking the ten nearest neigbours to a word and then the ten nearest neighbours for each of those neighbours. If the ten closest neighbours of the ten closest neighbours overlap well, the neighbourhood is semantically well connected. If they do not, the neighbourhood is semantically diverse. The minimum score is 10, the maximum is 110, if all terms have ten different nearest neighbours. A graph of how the number of neighbours evolves as text is being processed is shown in Figure 2 and some example words are given in Table 4. The range of variation for words with a score starting to flatten out is between about 20 and 90 and do so at about a hundred occurrences. Rank sum tests show significant differences between the Swadesh terms and the abstract and vague term list, with the former having a larger average number of second-order neighbours, in partial contradiction to the above result of local neighbourhood size given in Table 3. (Mann Whitney U, $p > 0.95$) Presumably this shows that the while the immediate distributional neighbourhood of the concrete Swadesh words is close, the words used with it are of a general character.

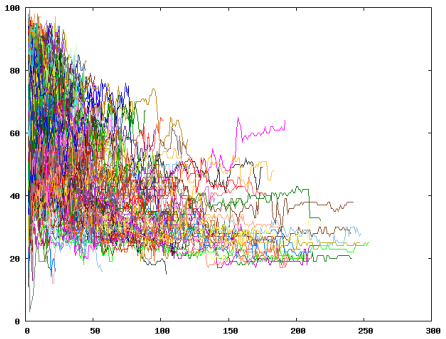

Figure 2: Semantic tightness measured by the number of second order neighbours among the ten nearest neighbours.

2.3.2 Semantic tightness: Angle-at-10

This measure is computed by taking cosine between the target word and its tenth nearest neighbour in the word space. If the word has few semantically similar words, this measure should remain close to 0; if this measure is closer to one, there is a tighter semantic context around this word. A graph of how this measure evolves as text is processed is shown in Figure 3 and some example words are given in Table 5. The range of variation for words with a score

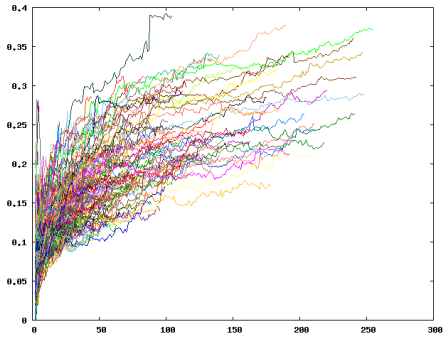

Figure 3: Semantic tightness measured by the cosine to the tenth neighbour.

$know \rightarrow feel$	0.37
$bad \rightarrow substandard$	0.24
$peace \rightarrow harmony$	0.20

Table 5: Cosine to the tenth neighbour for some selected words

starting to flatten out is between about 0.15 and 0.5 — the latter is already a noteworthy score of closeness in this type of model. Rank sum tests show significant differences between the Swadesh terms and the abstract and vague term list, with the former having a lower cosine score to the tenth neighbour, again in partial contradiction to the above result of local neighbourhood size given in Table 3. (Mann Whitney U, $p > 0.95$) Presumably this shows that the constrained neighbourhood may not extend to as many as ten neighbouring words.

2.3.3 Semantic wobble

The third measure we propose is how much the position of the word changes with additional information as new observations of the word are encountered. This measure is computed by taking the cosine between the position of the word in word space before and after each observation. This measure thus captures the semantic impact of the latest observation. A graph of how this measure evolves as text is processed is shown in Figure 4. The great majority of words converge relatively rapidly towards values of over 0.9, indicating that their meaning remains stable in this implementation. Some vary more: *group*, *position*, and *break* with rather less specific semantics have lower scores than words such as *election* and *university*.

2.3.4 Converging to known synonym

For words with very obviously identifiable synonyms — a small minority of terms in human language — a measure of *convergence* can be used to test the quality of the implementation at hand. Semantically very similar words should be assumed to relatively rapidly converge towards each other — meaning that their representations in this implementation should find each other after not too many occurrences. As an example, words such as *he* and *she* as well as *man* and *woman* should be expected to end up with similar representations. This assumption is borne out by data from this experiment as shown in Table 6 with the notable exception of the delay for *man* to match with *woman* as its closest synonym in blog text, where various names

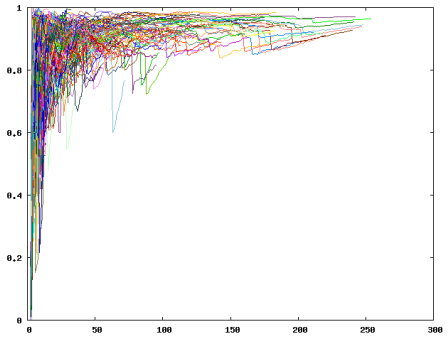

Figure 4: Semantic wobble measured by divergence from centroid.

	blog text		newsprint	
	first occurrence	stable occurrence	first occurrence	stable occurrence
he → *she*	12	37	216	384
she → *he*	21	34	4	5
woman → *man*	214	284	4	4
man → *woman*	228	562	40	74

Table 6: Number of occurrences until most reasonable synonym is established

and synonyms such as *guy* confound the process.

3 Conclusion

This first — quite explorative — study is intended to provide a first basis for measures beyond first-order term and document frequency as a measurement of terminological specificity. Some of the proposed measures appear to give purchases to separate terminology quite broadly: the range of variation shows promise for semantically useful distinctions to be made. A notable observation is that by most measures, words appear to find their semantic position in but a few mentions. Most words are fairly consistent in their meaning, even in the more fluid material given by blog authors.

Acknowledgements

This work is supported by Vetenskapsrådet, the Swedish Research Council through a grant for the project “Distributionally derived grammatical analysis models”.

References

Hisamitsu, T., Niwa, Y., and Tsujii, J.-i. (2000). A method of measuring term representativeness: baseline method using co-occurrence distribution. In *Proceedings of the 18th conference on Computational linguistics*, pages 320–326, Morristown, NJ, USA. Association for Computational Linguistics.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Kanerva, P., Kristofersson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society, CogSci'00*, page 1036. Erlbaum.

Katz, S. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–60.

Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389.

Sahlgren, M. (2006). *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.

Sahlgren, M., Holst, A., and Kanerva, P. (2008). Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society, CogSci'08*, pages 1300–1305, Washington D.C., USA.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Swadesh, M. (1971). *The origin and diversification of language*. Aldine, Chicago. Edited by Joel Sherzer post mortem.