An Ordering of Terms Based on Semantic Relatedness

Peter Wittek
Department of Computer Science
National University of Singapore

Sándor Daránya
Swedish School of Library and Information Science
Göteborg University
Sandor.Daranyi@hb.se

Chew Lim Tan
Department of Computer Science
National University of Singapore
tancl@comp.nus.edu.sg

Abstract

Term selection methods typically employ a statistical measure to filter or weight terms. Term expansion for IR may also depend on statistics, or use some other, non-metric method based on a lexical resource. At the same time, a wide range of semantic similarity measures have been developed to support natural language processing tasks such as word sense disambiguation. This paper combines the two approaches and proposes an algorithm that provides a semantic order of terms based on a semantic relatedness measure. This semantic order can be exploited by term weighting and term expansion methods.

1 Introduction

Since the early days of the vector space model, it has been debated whether it is a proper carrier of meaning of texts [23], arguing if distributional similarity is an adequate proxy for lexical semantic relatedness [3]. With the statistical, i.e. devoid of word semantics approaches there is generally no way to improve both precision and recall at the same time, increasing one is done at the expense of the other. For example, casting a wider net of

Proceedings of the 8th International Conference on Computational Semantics, pages 235–247,
Tilburg, January 2009.

search terms to improve recall of relevant items will also bring in an even greater proportion of irrelevant items, lowering precision. In the meantime, practical applications in information retrieval and text classification have been proliferating, especially with developments in kernel methods in the last decade [9, 4].

Ordering of terms based on semantic relatedness seeks an answer to the simple question, can statistical term weighting be eclipsed? Namely, variants of weighting schemes based on term occurrences and co-occurrences dominate the information retrieval and text classification scenes. However, they also have a number of limitations. The connection between statistics and word semantics is in general not understood very well. In other words, a systematic discussion of mappings between theories of word meaning and modeling them by mathematical objects is missing for the time being. Further, enriching weighting schemes by importing their sense content from lexical resources such as WordNet lacks a theoretical interpretation in terms of lexical semantics. Combining co-occurrence and lexical resource-based approaches for term weighting and term expansion may offer further theoretical insights, as well as performance benefits.

Using vectors in the vector space model as such mathematical objects for the representation of term, document or query meaning necessarily expresses content mapped on form as a set of coordinates. These coordinates, at least in the case of the tfidf scheme, are corpus-specific, i.e. term weights are neither constant over time nor database independent. Introducing a semantic ordering of terms, hence loading a coordinate with semantic content, reduces the dependence on a specific corpus.

In what follows, we will argue that:

- By assigning specific scalar values to terms in an ontology, terms represented by sets of geometric coordinates can be outdone;
- Such values result from a one-dimensional ordering based on the idea of a sense-preserving distance between terms in a conceptual hierarchy;
- Sense-preserving distances mapped onto a line condense lexical relations and express them as a kind of within-language referential meaning pertinent to individual terms, quasi charging their occurrences independent of their occurrence rates, i.e. *from the outside*;
- This linear order can be used to assist term expansion and term weighting.

This paper is organized as follows. Section 2 discusses the most important measures for semantic relatedness with regard to the major linguistic theories. Section 3 introduces an algorithm that creates a linear semantic order of terms of a corpus, and Section 4 both offers first results in text classification and discusses some implications. Finally, Section 5 concludes the paper.

2 Measuring Semantic Relatedness

Several methods have been proposed for measuring similarity. One of such early proposals was the semantic differential which analyzes the affective meaning of terms into a range of different dimensions with the opposed adjectives at both ends, and locates the terms within semantic space [20].

Semantic similarity as proposed by Miller and Charles is a continuous variable that describes the degree of synonymy between two terms [16]. They argue that native speakers can order pairs of terms by semantic similarity, for example *ship-vessel, ship-watercraft, ship-riverboat, ship-sail, ship-house, ship-dog, ship-sun*. This concept may be extended to quantify relations between non-synonymous but closely related terms, for example *airplane-wing*. Semantic distance is the inverse of semantic similarity [17].

Semantic relatedness is defined between senses of terms. Given a relatedness formula $\mathrm{rel}(s_1, s_2)$ between two senses s_1 and s_2, term relatedness between two terms t_1 and t_2 can be calculated as

$$\mathrm{rel}(t_1, t_2) = \max_{s_1 \in \mathrm{sen}(t_1), s_2 \in \mathrm{sen}(t_2)} \mathrm{rel}(s_1, s_2),$$

where $\mathrm{sen}(t)$ is a set of senses of term t [3].

Automated systems assign a score of semantic relatedness to a given pair of terms calculated from a relatedness measure. The absolute score itself is typically irrelevant on its own, what is important is that the measure assigns a higher score to term pairs which humans think are more related and comparatively lower scores to term pairs that are less related [17].

The best known theories of word semantics fall in three major groups:

1. "Meaning is use" [30]: habitual usage provides indirect contextual interpretation of any term. In accord with Carnap, frequency of use expresses aspects of a conceptual hierarchy. In terms of logical semantics, one regards document groups as value extensions (classes) and index terms as value intensions (properties) of a (semantic) function

'f'. Extensions and intensions are inverse proportional: the more properties defined, the less entities they apply to - there are more flowers in general than tulips in particular, for instance.

2. "Meaning is change": the stimulus-response theory by Bloomfield and the biological theory of meaning by von Uexküll both stress that the meaning of any action is its consequences.

3. "Meaning is equivalence": referential or ostensional theories of meaning suggest that 'X = Y for/as long as Z' [22].

Point 2 refers to theories which assign a temporal structure to word meaning, they are not discussed here. Measures that rely on distributional measures (Point 1) and those that use knowledge-rich resources (Point 3) both exist, and they have been individually shown to good quantifiers of term similarity each [17], These theories have been individually shown to be good, therefore their combination must be a valid research alternative.

A lexical resource in computer science is a structure that captures semantic relations among terms. Such a resource necessarily entails some sort of world view with respect to a given domain. This is often conceived as a set of concepts, their definitions and their inter-relationships; this is referred to as a conceptualization. The following types of resources are commonly used in measuring semantic similarity between terms: dictionary [12], semantic networks, such as WordNet [5], thesauri modeled on Roget's Thesaurus [19].

All approaches to measuring semantic relatedness that use a lexical resource regard the resource as a network or a directed graph, making use of the structural information embedded in the graph [8, 3].

Distributional similarity, as studied by language technology, covers an important kind of theories of word meaning and can be hence seen as contributing to semantic document indexing and retrieval. Its predecessors go back a long way, building on the notion of term dependence and structures derived therefrom [2, 18]. Also called the contextual theory of meaning (see [15] for the historical development of the concept), the underlying distributional hypothesis is often cited for explaining how word meaning enters information processing [10], and basically equals the claim "meaning is use" in language philosophy. Before attempts to utilize lexical resources for the same purpose, this used to be the sole source of word semantics in information retrieval, inherent in the exploitation of term occurrences (tfidf) and term co-occurrences [7, 21, 27], including multiple-level term co-occurrences [11].

Statistical techniques typically suffer from the sparse data problem: they perform poorly when the terms are relatively rare, due to the scarcity of data. Hybrid approaches attempt to address this problem by supplementing sparse data with information from a lexical database [24, 8]. In a semantic network, to differentiate between the weights of edges connecting a node and all its child nodes, one needs to consider the link strength of each specific child link. This is a situation in which corpus statistics can contribute. Ideally the method chosen should be both theoretically sound and computationally efficient [8].

Following the notation in information theory, the information content (IC) of a concept c can be quantified as follows.

$$\mathrm{IC}(c) = \frac{1}{\log P(c)}.$$

where $P(c)$ is the probability of encountering an instance of concept c. In the case of the hierarchical structure, where a concept in the hierarchy subsumes those ones below it, this implies that $P(c)$ is monotonic as one moves up in the hierarchy. As the node's probability increases, its information content or its informativeness decreases. If there is a unique top node in the hierarchy, then its probability is 1, hence its information content is 0. Given the monotonic feature of the information content value, the similarity of two concepts can be formally defined as follows.

$$\mathrm{sim}(\mathrm{c}_1, \mathrm{c}_2) = \max_{c \in \mathrm{Sup}(c_1, c_2)} \mathrm{IC}(c) = \max_{c \in \mathrm{Sup}(c_1, c_2)} -\log p(c)$$

where $\mathrm{Sup}(c_1, c_2)$ is the set of concepts that subsume both c_1 and c_2. To maximize the representativeness, the similarity value is the information content value of the node whose IC value is the largest among those higher order classes.

The information content method requires less information on the detailed structure of a lexical resource and it is insensitive to varying link types [24]. On the other hand, it does not differentiate between the similarity values of any pair of concepts in a sub-hierarchy as long as their lowest super-ordinate class is the same. Moreover, in the calculation of information content, a polysemous term will have a large content value if only term frequency data are used.

The distance function between two terms can be written as follows:

$$d(t_1, t_2) = \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2\mathrm{IC}(\mathrm{LSuper}(c_1, c_2)),$$

where LSuper$(c1, c2)$ denotes the lowest super-ordinate of c_1 and c_2 in a lexical resource. This distance measure also satisfies the properties of a metric [8].

3 A Semantic Ordering of Terms

Traditional distributional term clustering methods do not provide significantly improved text representation [13]. Distributional clustering has also been employed to compress the feature space while compromising document classification accuracy [1]. Applying the information bottleneck method to find term clusters that preserve the information about document categories has been shown to increase text classification accuracy in certain cases [28].

On the other hand, term expansion has been widely researched, with varying results [21]. These methods generate new features for each document in the data set. These new features can be synonyms or homonyms of document terms [26], or expanded features for terms, sentences and documents as in [6]. Several distributional criteria have been used to select terms related to the query. For instance, [25] proposed the principle that the selected terms should have a higher probability in the relevant documents than in the irrelevant documents. Others examined the impact of determining expansion terms using a minimum spanning tree and some simple linguistic analysis [29].

This section proposes an algorithm that connects term clustering and term expansion. It employs a pairwise comparison between the terms to find a linear order, instead of finding clusters. In this order, the transition from a term to an adjacent one is "smooth" if the semantic distance between two neighboring terms is small. The dimension of the feature space is not compressed, yet, groups of adjacent terms can be regarded as semantic clusters. Hence, following the idea of term expansion, adjacent terms can help to improve the effectiveness of any vector space-based language technology.

Let V denote a set of terms $\{t_1, t_2, \ldots, t_n\}$ and let $d(t_i, t_j)$ denote the semantic distance between the terms t_i and t_j.

Let $G = (V, E)$ denote a weighted undirected graph, where the weights on the set E are defined by the distances between the terms.

Finding a semantic ordering of terms can be translated to a graph problem: a minimum-weight Hamiltonian path S of G gives the ordering by reading the nodes from one end of the path to the other. G is a complete graph, therefore such a path always exists, but finding it is an NP-complete problem. The following greedy algorithm is similar to the nearest neighbor

heuristic for the solution of the traveling salesman problem. It creates a graph $G' = (S, T)$, where $S = V$ and $T \subset E$. This G' graph is a spanning tree of G in which the maximum degree of a node is two, that is, the minimum spanning tree is a path between two nodes.

Step 1 Find the term at the highest stage of the hierarchy in a lexical resource.

$$t_s = \text{argmin}_{t_i \in V} \text{depth}(t_i).$$

This seed term is the first element of V', $V' = \{t_s\}$. Remove it from the set V:

$$V := V \backslash \{t_s\}.$$

Step 2 Let t_l denote the leftmost term of the ordering and t_r the rightmost one. Find the next two elements of the ordering:

$$t'_l = \text{argmin}_{t_i \in V} d(t_i, t_l),$$

$$t'_r = \text{argmin}_{t_i \in V \backslash \{t'_l\}} d(t_i, t_r).$$

Step 3 If $d(t_l, t'_l) < d(t_r, t'_r)$ then add t'_l to V', $E' := E' \cup \{e(t_l, t'_l)\}$, and $V := V \backslash \{t'_l\}$. Else add t'_r to V', $E' := E' \cup \{e(t_r, t'_r)\}$ and $V := V \backslash \{t'_r\}$.

Step 4 Repeat from *Step 2* until $V = \emptyset$.

The computational cost of the algorithm is $O(n^2)$. The above algorithm can be thought of as a modified Prim's algorithm, but it does not find the optimal minimum-weight spanning tree.

The validity of the ordering algorithm is discussed as follows.

1. The ordering is possible. Starting from the seed term, the candidate sets will always contain elements, which either share the same hypernym or are hypernyms of each other.

2. The ordering is good enough. The quality will also depend on the lexical resource in question. Further, the complexity of human languages makes the creation of even a near perfect semantic network of its concepts impossible. Thus in many ways the lexical resource-based measures are as good as the networks on which they are based.

3. The distance between adjacent terms is uniform. By the construction of the ordering, it is obvious that the distances will not be uniform.

4 Discussion

We were interested in how the distances of consecutive index terms change if we apply the semantic ordering. We indexed the ModApte split of Reuters-21578 benchmark corpus with a WordNet-based stemmer. The indexing found 12643 individual terms. Prior to the semantic ordering, terms were assumed to be in an arbitrary order. Measuring the Jiang-Conrath distance between the arbitrarily ordered terms, the average distance was 1.68. Note that the Jiang-Conrath distance was normalized to the interval $[0, 2]$. Figure 1 shows the distribution of distances. The histogram has a high peak at the maximum distance, indicating that the original arrangment had little to do with semantic distance. However, there were few terms with zero or little distance between them. This is due to terms which are related and start with the same word or stem. For example, *account, account executive, account for, accountable, accountant, accounting principle, accounting standard, accounting system, accounts payable, accounts receivable.*

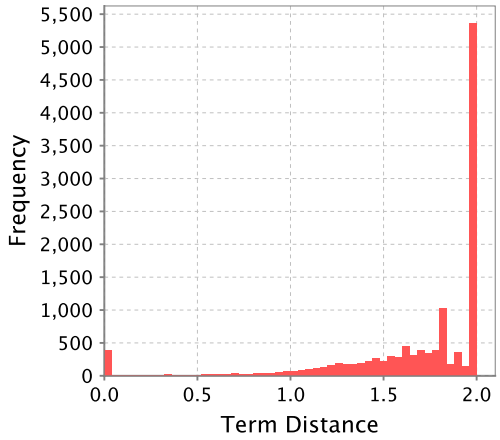

Figure 1: Distribution of Distances Between Adjacent Terms in an Arbitrary Order

After the semantic ordering of the term by the proposed algorithm, both the average distance and the Jiang-Conrath distance were 0.56. About one third of the terms had very little distance between each other (see Figure 2). Nevertheless, over 10 % of the total terms still had the maximum distance. This is due to the non-optimal nature of the proposed term-ordering algorithm. These terms add noise to the classification. The noisy terms occur

typically at the two sides of the scale, being the leftmost and the rightmost ones. While it is easy to find terms close to each other in the beginning, as the algorithm proceeds, fewer terms remain in the pool to be chosen. For instance, *brand, brand name, trade name, label* are in the 33rd, 34th, 35th and 36th position on the left side counting from the seed respectively, while *windy, widespread, willingly, whatsoever, worried, worthwhile* close the left side, apparently sharing little in common.

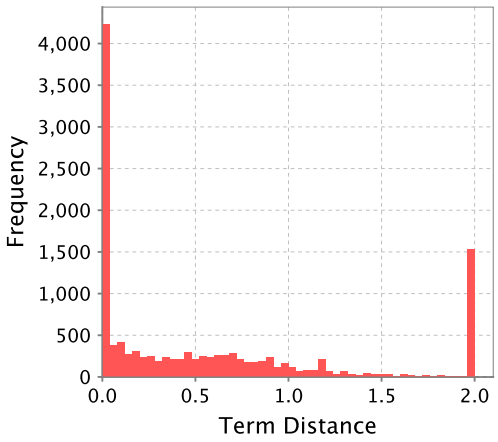

Figure 2: Distribution of Distances Between Adjacent Terms in a Semantic Order Based on Jiang-Conrath Distance

We conducted experiments on the ten most common categories of the ModApte split of Reuters-21578. We trained support vector machines with a linear kernel to compare the micro- and macro-average F_1 measures for different methods. Table 1 summarizes the results. The baseline vector space model has zero expansion terms. Neighboring terms of the semantic order were chosen as expansion terms. We found that increasing the number of expansion terms also increases the effectiveness of classification, however, effectiveness decreases after 4 expansions for micro-F1 and after 6 expansions for macro-F1.

5 Conclusions

Terms can be corpus- or genre-specific. Manually constructed general-purpose lexical resources include many usages that are infrequent in a particular cor-

Number of Expansion Terms	Micro-F_1	Macro-F_1
0	0.900	0.826
2	0.901	0.826
4	**0.905**	0.828
6	0.898	**0.830**
8	0.896	0.827

Table 1: Micro-Average and Macro F_1-measure, Reuters-21578

pus or genre of documents, and therefore of little use. For example, one of the 8 senses of *company* in WordNet is a *visitor/visitant*, which is a hyponym of *person* [14]. This usage of the term is practically never used in newspaper articles, hence distributional attributes should be taken into consideration when creating a linear ordering of terms.

Integrating lexical resources into an upgraded semantic weighting scheme that could augment statistical term weighting is a prospect that cannot be overlooked in information retrieval and text categorization. Our first results with such a scheme in text categorization. At the same time, the results also raise the question, does assigning specific scalar values to terms in an ontology, this far represented by their geometric coordinates only, turn them metaphorically into band lines of elements in a conceptual spectrum. We anticipate that applying other types of kernels to the task may bring a new set of challenging results.

References

[1] L.D. Baker and A.K. McCallum. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 96–103, Melbourne, Australia, August 1998. ACM Press, New York, NY, USA.

[2] M.A. Bärtschi. Term dependence in information retrieval models. Master's thesis, Swiss Federal Institute of Technology, 1984.

[3] A. Budanitsky and G. Hirst. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[4] N. Cristianini, J. Shawe-Taylor, and H. Lodhi. Latent semantic kernels. *Journal of Intelligent Information Systems*, 18(2):127–152, 2002.

[5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA, 1998.

[6] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *Proceedings of IJCAI-05, 19th International Joint Conference on Artificial Intelligence*, volume 19, Edinburgh, UK, 2005. Lawrence Erlbaum Associates Ltd.

[7] S. I. Gallant. A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation*, 3:293–309, 1991.

[8] J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, pages 19–33, Taipei, Taiwan, 1997.

[9] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142, Chemnitz, Germany, April 1998. Springer-Verlag, London, UK.

[10] J. Karlgren and M. Sahlgren. From words to understanding. *Foundations of Real-World Intelligence*, pages 294–308, 2001.

[11] A. Kontostathis and W.M. Pottenger. A framework for understanding latent semantic indexing (LSI) performance. *Information Processing and Management*, 42(1):56–73, 2006.

[12] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone? In *Proceedings of SIGDOC-86, 5th Annual International Conference on Systems Documentation*, pages 24–26, New York, NY, USA, 1986. ACM Press.

[13] D.D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of SIGIR-92, 15th ACM*

International Conference on Research and Development in Information Retrieval, pages 37–50, Copenhagen, Denmark, June 1992. ACM Press, New York, NY, USA.

[14] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, volume 98, pages 768–773, Montréal, Québec, Canada, August 1998. ACL, Morristown, NJ, USA.

[15] J. Lyons. *Semantics*. Cambridge University Press, New York, NY, USA, 1977.

[16] G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[17] S. Mohammad and G. Hirst. Distributional measures as proxies for semantic relatedness. *Submitted for publication*, 2005.

[18] J. Morris, C. Beghtol, and G. Hirst. Term relationships and their contribution to text semantics and information literacy through lexical cohesion. In *Proceedings of the 31st Annual Conference of the Canadian Association for Information Science*, Halifax, Nova Scotia, Canada, May 2003.

[19] J. Morris and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991.

[20] C.E. Osgood. The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197–237, 1952.

[21] H.J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[22] C.S. Peirce. Logic as semiotic: The theory of signs. *Philosophical Writings of Peirce*, pages 98–119, 1955.

[23] V.V. Raghavan and S.K.M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, 1986.

[24] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI-95, 14th International Joint Conference on Artificial Intelligence*, volume 1, pages 448–453, Montréal, Québec, Canada, August 1995.

[25] S.E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[26] M.D.E.B. Rodriguez and J.M.G. Hidalgo. Using WordNet to complement training information in text categorisation. In *Procedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing*. John Benjamins Publishing, Amsterdam, The Netherlands, 1997.

[27] H. Schutze and T. Pedersen. A co-occurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 3(33):307–318, 1997.

[28] N. Slonim and N. Tishby. The power of word clusters for text classification. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, Darmstadt, Germany, 2001.

[29] A.F. Smeaton and C.J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[30] L. Wittgenstein. *Philosophical Investigations*. Blackwell Publishing, Oxford, UK, 1967.