Improving Text Classification by a Sense Spectrum Approach to Term Expansion

Peter Wittek
Department of Computer Science
National University of Singapore
Computing 1, Law Link
Singapore 117590
wittek@comp.nus.edu.sg

Sándor Darányi
Swedish School of Library
and Information Science
Göteborg University &
University of Borås
Allégatan 1
50190 Borås, Sweden
sandor.daranyi@hb.se

Chew Lim Tan
Department of Computer Science
National University of Singapore
Computing 1, Law Link
Singapore 117590
tancl@comp.nus.edu.sg

Abstract

Experimenting with different mathematical objects for text representation is an important step of building text classification models. In order to be efficient, such objects of a formal model, like vectors, have to reasonably reproduce language-related phenomena such as word meaning inherent in index terms. We introduce an algorithm for sense-based semantic ordering of index terms which approximates Cruse's description of a sense spectrum. Following semantic ordering, text classification by support vector machines can benefit from semantic smoothing kernels that regard semantic relations among index terms while computing document similarity. Adding expansion terms to the vector representation can also improve effectiveness. This paper proposes a new kernel which discounts less important expansion terms based on lexical relatedness.

1 Introduction

Generally, building an automated text classification system consists of two key subtasks. The first task is text representation which converts the content of documents into compact format so that they can be further processed by the text classifiers. Another task is to learn the model of a text classifier which is used to classify the unlabeled documents. This paper proposes a substantially new model for text representation to improve effectiveness of text classification by semantic ordering.

Our motivation for the research presented here came from (Dorrer et al., 2001) who demonstrated the viability of database searching by visible light using a quantum algorithm, albeit on meaningless items. The question was, what kind of document representation would be necessary to extend their in-principle results to include semantics, one that has been leading us to test both periodic and non-periodic functions for this purpose. Since representation and retrieval by colors was implied in their method, we speculated that the following components could be useful in a rephrased model: (a) a metaphorically presented spectral expression of lexical semantic phenomena, (b) a ranked one-dimensional condensate of multidimensional sense structure, and (c) representation of documents and queries by functions in L_2 space with a similarity measure. Our anticipation was that by matching these components, a new model could demonstrate new capacities in general, and contribute to computing meaning by waves in particular.

Semantic ordering (component b) is an approximation of what (Cruse, 1986) referred to as a sense spectrum, i.e. a series of points - called local senses and constituting lexical units -, in a one-dimensional semantic continuum (component a). Apart from differentiating between the conceptual content of the same word in terms of its senses in word pairs, i.e. their semantic relatedness, it also compresses the result in spectral form. The scalar values of this spectrum have the double potential of being a condensed measure for semantic weighting, and, tentatively, they can play the role of mass in experiments where gravity is called in as a metaphor for text categorization and information retrieval (Paijmans, 1997; Shi et al., 2005; Wittek et al., 2009).

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 183–191,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

This paper addresses text categorization by means of non-periodical functions only.

In support of Cruse's point, recently it has been demonstrated by measurements that sense classification errors made by their maximum entropy based word sense disambiguation system were partly remedied once instead of a fine-grained view, a more coarse-grained view of senses was adopted (Palmer et al., 2006). Improvement of sense classification accuracy linked with "zooming out" in terms of observation granularity indicates, in our eyes, the "fluid", perhaps spectral nature of sense inasmuch as it is impossible to precisely distinguish between the borderlines and some fuzziness is implied both in the phenomenon and its perception. This "fluidity of language", as Palmer et al. call it, is in accord with the theory of shared semantic representations in psycholinguistics (Rodd et al., 2002), according to which related senses share a portion of their meaning representation in the mental lexicon; it also supports an earlier observation of two of the present authors based on the same methodology as outlined in this paper, namely that using continuous functions for information retrieval leads to content representation without exact term or document locations, one which is regional in its nature and subject to a mathematical uncertainty principle (Wittek and Darányi, 2007).

We approach our problem in three steps: (1) whether distributional semantics alone is enough for the representation of word meaning, (2) whether semantic relatedness between word pairs can be expressed in an ordered form while preserving lexical field structure, and if (3) the uniqueness of entries in such an order can be expressed by functions rather than scalars such as distance. As we will show, this line of thought leads to performance improvement in text classification by using kernel-based feature weighting.

Since the early days of the vector space model, it has been debated whether it is a proper carrier of meaning of texts (Raghavan and Wong, 1986), arguing if distributional similarity is an adequate proxy for lexical semantic relatedness (Budanitsky and Hirst, 2006). We argue for the need to enrich distributional semantics-based text representation by other components because 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 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 approaches have been proliferating, especially with developments in kernel methods in the last decade (Joachims, 1998; Cristianini et al., 2002). Some researchers suggested a more general mathematical framework to accommodate the needs that the vector space model cannot satisfy (van Rijsbergen, 2004). This paper explores the opportunities of this representation in the domain of text classification by introducing it as a new nonlinear semantic kernel.

Another aspect of the same problem is term expansion for document classification and retrieval. By automatically selecting expansion terms for a text classification system to expand a document vector by adding terms that are related to the terms already in the document, performance can be improved (Hu et al., 2008). Such new terms can either be statistically related to the original terms or chosen from lexical resources such as thesauri, controlled vocabularies, ontologies and the like.

However, in doing so the fundamental question often overlooked is whether the expansion terms extracted are equally related to the document and are useful for text classification. In what follows we propose a form of term expansion with decreasing importance of those terms that are less related, as contrasted with rigid term expansion. This can be carried out by a combination of semantic ordering and using function space for classification.

This paper is organized as follows. Section 2 overviews text classification by support vector machines, expanding on traditional text similarity measures (Section 2.1), semantic smoothing kernels (Section 2.2), term expansion strategies (Section 2.3), and finally introduces our semantic kernels in the L_2 space (Section 2.4). Section 3 discusses experimental results and Section 4 concludes the paper.

2 Text Classification with Support Vector Machines

Text categorization is the task of assigning unlabeled documents into predefined categories. Given a collection of $\{d_1, d_2, \ldots, d_N\}$ documents, and a $C = \{c_1, c_2, ..., c_{|C|}\}$ set of predefined categories, the task is, for each document d_j ($j \in \{1, 2, \ldots, N\}$), to assign a decision to file d_j under c_i or a decision not to file d_j under c_i ($c_i \in C$) by virtue of a function Φ, where the function Φ is also referred to as the classifier, or model, or hypothesis, or rule. Supervised text classification is a machine learning technique for creating the function Φ from training data. The training data consist of pairs of input documents, and desired outputs (i.e., classes).

Support vector machines have been found the most effective by several authors (Joachims, 1998). The proposed semantic text classification method is grounded in the kernel methods underlying support vector machines.

A support vector machine is a kind of supervised learning algorithm. In its simplest, linear form, a support vector machine is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin (Shawe-Taylor and Cristianini, 2004). The strength of kernel methods is that they allow a mapping $\phi(.)$ of $\mathbf{x}$ to a higher dimensional space. In the dual formulation of the mathematical programming problem, only the kernel matrix $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_i)$ is needed in the calculations.

2.1 Traditional Text Similarity Measure

Intuitively, if a text fragment of two documents address similar topics, it is highly possible that they share lots of substantive terms. After having removed the stopwords and stemmed the rest, the stemmed terms construct a vector representation for each text document. Let $\mathbf{a}_j$ be a document vector in the vector space model, that is, $\mathbf{a}_j = \sum_{k=1}^{M} a_{kj}\mathbf{e}_k$, where M is the number of index terms, a_{kj} is some weighting (e.g., term frequency), and $\mathbf{e}_k$ is a basis vector of the M-dimensional Euclidean space. This representation is also referred to as the bag-of-words (BOW) model.

Given this representation, semantic relatedness of a pair of text fragments is computed as the cosine similarity of their corresponding term vectors which is defined as:

$$S(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i\mathbf{a}_j}{|\mathbf{a}_i||\mathbf{a}|_j}. \qquad (1)$$

2.2 Linear Semantic Kernels

One enrichment strategy is to use a semantic smoothing kernel while calculating the similarity between two documents. Any linear kernel for texts is characterized by $K(\mathbf{a}_i, \mathbf{a}_j) = \mathbf{a}_i'S'S\mathbf{a}_j$, where S is an appropriately shaped matrix commonly referred to as semantic smoothing matrix (Siolas and d'Alché Buc, 2000; Shawe-Taylor and Cristianini, 2004; Basili et al., 2005; Mavroeidis et al., 2005; Bloehdorn et al., 2006). The presence of S changes the orthogonality of the vector space model, as this mapping should introduce term dependence. A recent attempt tried to manually construct S with the help of a lexical resource (Siolas and d'Alché Buc, 2000). The entries in the symmetric matrix S express the semantic similarity between the terms i and j. Entries in this matrix are inversely proportional to the length of the WordNet hierarchy path linking the two terms. The performance, measured over the 20NewsGroups corpus, showed an improvement of 2 % over the the basic vector space method. Moreover, the semantic matrix S is almost fully dense, hence computational issues arise. In a similar construction, (Bloehdorn et al., 2006) defined the matrix entries as weights of superconcepts of the two terms in the WordNet hierarchy. Focusing on special subcategories of Reuters-21578 and on the TREC Question Answering Dataset, they showed consistent improvement over the baseline. As (Mavroeidis et al., 2005) pointed out, polysemy will remain a problem in semantic smoothing kernels. A more complex way of calculating the semantic similarity as the matrix entries was also proposed (Basili et al., 2005). For a more general discussion on semantic similarity see Section 2.4.1.

An early attempt to overcome the untenable orthogonality assumption of the vector space model was proposed under the name of generalized vector space model (Wong et al., 1985). The article which proposed the model did not provide empirical results, and since then the model has been regarded of large theoretical importance with less impact on actual applications. The model takes a distributional approach, focusing on term co-occurrences. The underlying assumption is that term correlations are captured by the co-occurrence information. That is, two terms are semantically related if they co-occur often in the same documents. By eliminating orthogonality, documents can be seen as similar even if they do not share any terms. The term co-occurrence matrix is AA', hence the model takes A' as the semantic similarity matrix S. A major drawback of the generalized vector space model is that it replaces the orthogonality assumption with another questionable assumption. The computational needs are tremendous too, if the dimensions of A are considered. Moreover, the co-occurrence matrix is not sparse anymore.

Latent semantic indexing (or latent semantic analysis) was another attempt to bring more linguistic and psychological aspects to language processing via a kernel. Conceptually, latent semantic indexing is similar to the generalized vector space model, it measures semantic information through co-occurrence analysis in the corpus. From the algorithmic perspective it is an enormous problem that textual data have a large number of relevant features. This results in huge computational needs and the classification models may overfit the data. The number of features can be reduced by multivariate feature extraction methods. In latent semantic indexing, the dimension of the vector space is reduced by singular value decomposition (Deerwester et al., 1990).

Using rank reduction, terms that occur together very often in the same documents are merged into a single dimension of the feature space. The dimensions of the reduced space correspond to the axes of greatest variance. For latent semantic indexing, by dual representation the kernel matrix is $K = V\Sigma_k^2 V'$, where Σ_k is a diagonal matrix containing the k largest singular values of the singular value decomposition of the vector space, and V holds the right singular vectors of the decomposition. The new kernel matrix can be obtained directly from K by applying an eigenvalue decomposition of K (Cristianini et al., 2002). The computational complexity of performing an eigenvalue decomposition on the kernel matrix is a major drawback of latent semantic indexing.

2.3 Text Representation Enrichment Strategies by Term Expansion

In order to eliminate the bottleneck of the traditional BOW representation, previous approaches in term expansion enriched this convention by external lexical resources such as WordNet.

As a first step, these methods generate new features for each document in the dataset. These new features can be synonyms or homonyms of document terms as in (Hotho et al., 2003; Rodriguez and Hidalgo, 1997), or expanded features for terms, sentences and documents as in (Gabrilovich and Markovitch, 2005), or term context information for word sense disambiguation such as topic signatures (Agirre and De Lacalle, 2003; Agirre et al., 2004).

Then, the generated new features replace the old ones or are appended to the document representation, and construct a new vector representation $\hat{\mathbf{a}}_i$ for each text document. The similarity measure of document pairs is defined as:

$$S(\hat{\mathbf{a}}_i, \hat{\mathbf{a}_j}) = \frac{\hat{\mathbf{a}}_i \hat{\mathbf{a}_j}}{|\hat{\mathbf{a}}_i||\hat{\mathbf{a}}_j|}. \qquad (2)$$

2.4 Our Framework

The basic assumption of our framework is that terms can be arranged in an order such that consecutive terms are semantically related. Hence each term acquires a unique position, and this position ties the term to its semantically related neighbors. However, given a BOW representation with a cosine similarity measure, this position would not improve classification performance. Therefore we suggest to associate a mathematical function with each term, thus mapping terms and documents to the L_2 space, and using the inner product of this space to express similarity. The choice of function will determine to which extent neighboring terms, i.e., the enriching terms, are considered in calculating the similarity between two documents. This section first introduces an algorithm that produces the aforementioned semantic order, then the semantic kernels in the L_2 space are discussed.

2.4.1 An Algorithm for a Semantic Ordering of Terms

The proposed kernels assume that there is a semantic order between terms. 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. The initial order of the terms is not relevant, though it is assumed to be alphabetic. Let $G = (V, E)$ denote a weighted undirected graph, where the weights in the set E are defined by the distances between the terms.

Various lexical resource-based (Budanitsky and Hirst, 2006) and distributional measures (Mohammad and Hirst, 2005) have been proposed to measure semantic relatedness and distance between terms. Terms can be corpus- or genre-specific. Manually constructed general-purpose lexical resources include many usages that are infrequent in a particular corpus or genre of documents. For example, one of the 8 senses of *company* in WordNet is a *visitor/visitant*, which is a hyponym of *person* (Lin, 1998). This sense of the term is practically never used in newspaper articles, hence distributional attributes should be taken into consideration. Composite measures that combine the advantages of both approaches have also been developed (Resnik, 1995; Jiang and Conrath, 1997). This paper relies on the Jiang-Conrath composite measure (Jiang and Conrath, 1997), which has been shown to be superior to other measures (Budanitsky and Hirst, 2006), and we also found that this measure works the best for the purpose. The Jiang-Conrath metric measures the distance between two senses by using the hierarchy of WordNet. By denoting the lowest superordinate of two senses s_1 and s_2 in the hierarchy with LSuper(s_1,s_2), the metric is calculated as follows:

$$d(s_1, s_2) = \mathrm{IC}(s_1) + \mathrm{IC}(s_2) - 2\mathrm{IC}(\mathrm{LSuper}(s_1, s_2)),$$

where IC(s) is the information content of a sense s based on a corpus. Distance between two terms is calculated according to the following equation: $\mathrm{d}(t_1, t_2) = \max_{s_1 \in \mathrm{sen}(t_1), s_2 \in \mathrm{sen}(t_2)} d(s_1, s_2)$, where t_1 and t_2 are two terms, and $\mathrm{sen}(t_i)$ is the set of senses of t_i. The distance between two terms is usually defined as the minimum of the sense distances. We chose maximum because it ensures that only closely related terms will be placed to adjacent positions by the algorithm below.

Finding a semantic ordering of terms can be translated to a graph problem: a minimum-weight Hamiltonian path G' of G gives the ordering by reading the nodes from one of the paths 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 $G' = (V', E')$, where $V' = V$ and $E' \subset E$. 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.

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

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

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

Using WordNet, this seed term is *entity*, if the vocabulary of the text collection contains it.

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:

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

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

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 above algorithm can be thought of as a modified Prim's algorithm, but it does not find the optimal minimum-weight spanning tree.

2.4.2 Semantic Kernels in the L_2 Space

The L_2 space shares resemblance with a real vector space. Real-valued vectors are replaced by square-integrable functions, and the dot product is replaced by the following inner product: $(f_i, f_j) = \int f_i f_j dx$, for some f_i, f_j in the given L_2 space.

Lately, Hoenkamp has also pointed out that the L_2 space can be used for information retrieval when he introduced a Haar basis for the document space (Hoenkamp, 2003). He utilized a signal processing framework within the context of latent semantic indexing. In order to apply an L_2 representation for text classification, the problem is approached from a different angle than by Hoenkamp, taking discounting expansion terms as our point of departure.

Assigning a function $w(x-k)$ to the term in the kth position in a semantic order, a document j can be expressed as follows:

$$f_j(x) = \sum_{k=1}^{M} a_{kj} w(x-k), \qquad (3)$$

where x is in $[1, M]$, and it is the variable of integration in calculating the inner product of the L_2; x can be regarded as a "dummy" variable carrying no meaning in itself. The above formula will be referred to as a document function. In the experiments, the function $\exp(-bx^2)$ was used as $w(x)$, with b as a free parameter reflecting the width of the function expressing how many neighboring expansion terms are considered.

The inner product of the $L_2[1, M]$ space is applied to express similarity between two documents in similar vein as the dot product does in a real-valued vector space:

$$(f_i, f_j) = \int_{[1,M]} f_i(x) f_j(x) dx, \qquad (4)$$

where f_i and f_j are the representations of the documents in the L_2 space ($f_i, f_j \in L_2([1, M])$).

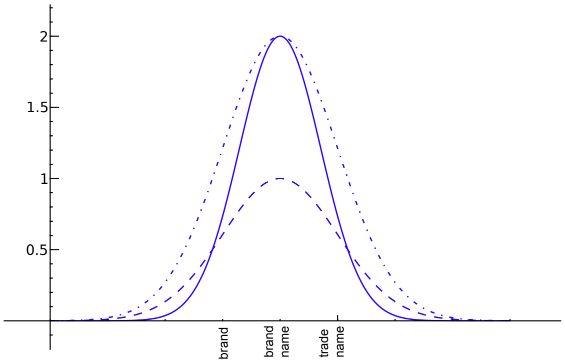

Figure 1: Two documents with matching term *brand name*. Dotted line: Document-1. Dashed line: Document-2. Solid line: Their product.

With the above formula, a matching term in two documents will be counted to its full term frequency or tfidf score, while semantically related terms will be counted less and less according their semantic similarity to the matching term. Assuming that the terms *brand*, *brand name*, and *trade name* follow each other in the semantic order, consider the following example. The first document has the term *brand name*, and so does the second document. In Figure 1, it can be seen brand name is counted the same way as it would be in a BOW model with its full term frequency score, brand and trade name are counted to a lesser extent, while other related terms are considered even less.

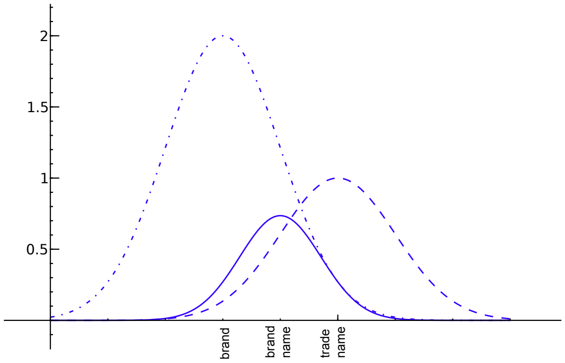

Figure 2: Two documents with no matching term but with related terms *brand* and *trade name*. Dotted line: Document-1. Dashed line: Document-2. Solid line: Their product.

Now if the two documents do not share the exact term, only related terms occur, for instance, *trade name* and *brand*, respectively, then the term *brand name*, placed between *trade name* and *brand* in the s semantic order, will be considered only to some extent for the calculation of similarity (see Figure 2).

3 Experimental Results

The most widely used benchmark corpus is the Reuters-21578 collection. For benchmarking purposes, the ModApte split was adopted. 9603 documents were used as the training set and 3299 as the test set in the experiments. Only those ninety text categories which had at least one positive example in the training set were included in the benchmark. Another benchmark data corpus we used was the 20

Newsgroups corpus, which is a collection of approximate 20,000 newsgroup documents nearly evenly divided among 20 discussion groups and each document is labeled as one of the 20 categories corresponding to the name of the newsgroup that the document was posted to.

In preparing the index terms, we restricted the vocabulary to the terms of WordNet 3.0 in order to be able to calculate the similarity score between any two terms. Stop words were removed in advance. Multiple word expressions were used to fully utilize WordNet. We used the built-in stemmer of WordNet, which is able to distinguish between different parts-of-speeches if the form of the word is unambiguous. For example, {accommodates, accommodated, accommodation} was stemmed to {accommodate, accommodate, accommodation}. We used term frequency as term weighting.

Prior to the semantic ordering, terms were assumed to be in alphabetic order. Measuring the Jiang-Conrath distance between adjacent terms, the average distance was 1.68 on the Reuters corpus. Note that the Jiang-Conrath distance was normalized to the interval $[0, 2]$. 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*.

The same average distance after reordering the terms with the proposed algorithm and the Jiang-Conrath distance was 0.56 on the same corpus. About one third of the terms had very little distance between each other. 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, that is, the leftmost terms and the rightmost terms. While it is easy to find close terms 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. The noise can be reduced by the appropriate choice of the parameter b in $\exp(-bx^2)$, so that the impact of adjacent but distantly related terms can be minimized.

Kernel	Reuters Micro	Reuters Macro	20News Micro	20News Macro
Linear	0.900	0.826	0.801	0.791
Poly	0.903	0.824	0.796	0.788
L_2	0.911	0.835	0.813	0.799

Table 1: Micro- and macro-average F_1 results

Table 1 shows the results on the two benchmark corpora with the baseline linear kernel. Precision and recall with regard to a class c_k, the F_1 score shown is their average. For all the kernels, the results with the best parameter settings are shown. Polynomial kernels were benchmarked between degrees 2 and 5. L_2 kernels were benchmarked with width b between 1 and 8, the performance peaking at 4 in all cases. The model is able to outperform the baseline kernels, and the differences in micro-averaged results are statistically significant. In all cases of the L_2 kernel, the increase of F_1 was due the increase in both precision and recall.

4 Conclusions

Information systems are in great need of automated intelligent tools, but existing algorithms and methods cannot be pushed much further. Most techniques in current use are impaired by the semantically poor but widespread representation of information and knowledge. For this reason, we propose a new formalism that combines Cruse's idea about a sense spectrum, approximated by semantic ordering, and its calculation by functions.

The suggested model combines term expansion with the semantic relations and semantic relatedness used in semantic smoothing kernels. This slightly unusual approach needs to transform the real vector representation to the L_2 space, and the experimental results show that this new representation can improve text classification effectiveness.

Our new model also blends insights from different approaches to lexical semantics theory at its different levels. First, during the semantic ordering of terms the distributional hypothesis meets handcrafted lexical resources of word meaning that relate to term occurrences as if they were their referents, a component external to term context. While high-quality lexical resources enable such an ordering in themselves, the procedure can benefit from data derived from the specific corpora in study – semantic relatedness measures such as the Jiang-Conrath similarity operate this way. Secondly, once the ordering is done and a sense spectrum is constructed, weights expressing statistical relationships between terms and documents are borrowed from the vector space model to form the basis for constructing hypothetical signals of content, documents as continuous functions.

5 Future research

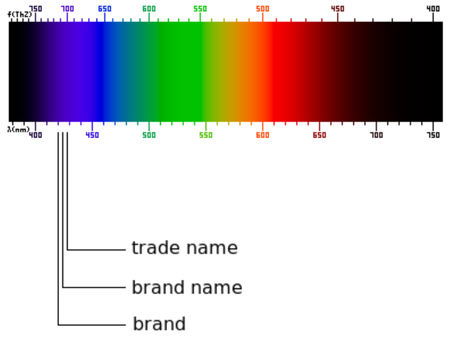

Figure 3: A hypothetical spectrum of terms.

As we have shown, a spectral interpretation of sense granularity can lead to improved text categorization results by utilizing L_2 space for information representation. Whether non-periodic functions other than the variant tested in this paper can be applied to the same end needs to be explored.

Turning back to the use of the spectrum of visible light for representing meaning, this raises new research questions. On the one hand, translating one-dimensional semantic ordering into colors is straightforward. Consider the following mapping. Assume that a language has a finite N number of terms, so the 1-dimensional result is an ordered list $o_1, o_2, \ldots, o_N$. Calculate the following: $\Delta = \sum_{i=1}^{N-1} d(o_i, o_{i+1})$, where Δ is the sum of distances between consecutive words. Further let $F : [0, \Delta] \rightarrow [400, 700]$ be the following mapping: $F(x) = 400 + x\frac{300}{\Delta}$. The visible spectrum is between 400 and 700 nm, F maps the cumulative distances of terms from $[0, \Delta]$ to the visible spectrum congruently, i.e. without distorting the distances. With this bijective (one-to-one) mapping, each term is assigned a physical wavelength and frequency. Figure 3 shows an example of such a term spectrum.

On the other hand, we have only begun to test the applicability of periodic functions in L_2 space (Wittek and Darányi, 2007), hence a well-established link to semantic computing by waves is missing for the time being. A possible compromise between the non-periodic vs. periodic approaches can be to apply wavelets instead of waves, a direction where our ongoing research shows promising results. These will be reported elsewhere. In a broader frame of thought, we are also working on the optical equivalents of the vector space model and the generalized vector space model as a first step toward coding more semantics in mathematical objects, and putting them to work in novel computing environments.

6 Acknowledgments

The authors are grateful to Martha Palmer (University of Colorado Boulder) for her inspiring suggestions and advice on sense granularity.

References

E. Agirre and O.L. De Lacalle. 2003. Clustering WordNet word senses. In *Proceedings of RANLP-03, 4th International Conference on Recent Advances in Natural Language Processing*, pages 121–130.

E. Agirre, E. Alfonseca, and O.L. de Lacalle. 2004. Approximating hierarchy-based similarity for WordNet nominal synsets using topic signatures. In *Proceedings of GWC-04, 2nd Global WordNet Conference*, pages 15–22.

R. Basili, M. Cammisa, and A. Moschitti. 2005. Effective use of WordNet semantics via kernel-based learning. In *Proceedings of CoNLL-05, 9th Conference on Computational Natural Language Learning*, pages 1–8.

S. Bloehdorn, R. Basili, M. Cammisa, and A. Moschitti. 2006. Semantic kernels for text classification based on topological measures of feature similarity. *Proceedings of ICDM-06, 6th IEEE International Conference on Data Mining*.

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

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

D.A. Cruse. 1986. *Lexical semantics.*

S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

C. Dorrer, P. Londero, M. Anderson, S. Wallentowitz, and IA Walmsley. 2001. Computing with interference: all-optical single-query 50-element database search. In *Proceedings of QELS-01, Quantum Electronics and Laser Science Conference*, pages 149–150.

E. Gabrilovich and S. Markovitch. 2005. Feature generation for text categorization using world knowledge. In *Proceedings of IJCAI-05, 19th International Joint Conference on Artificial Intelligence*, volume 19.

E. Hoenkamp. 2003. Unitary operators on the document space. *Journal of the American Society for Information Science and Technology*, 54(4):314–320.

A. Hotho, S. Staab, and G. Stumme. 2003. WordNet improves text document clustering. In *Proceedings of SIGIR-03, 26th ACM International Conference on Research and Development in Information Retrieval.*

J. Hu, L. Fang, Y. Cao, H.J. Zeng, H. Li, Q. Yang, and Z. Chen. 2008. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of SIGIR-08, 31st ACM International Conference on Research and Development in Information Retrieval*, pages 179–186.

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

T. Joachims. 1998. 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.

D. Lin. 1998. 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.

D. Mavroeidis, G. Tsatsaronis, M. Vazirgiannis, M. Theobald, and G. Weikum. 2005. Word sense disambiguation for exploiting hierarchical thesauri in text classification. *Proceedings of PKDD-05, 9th European Conference on the Principles of Data Mining and Knowledge Discovery*, pages 181–192.

S. Mohammad and G. Hirst. 2005. Distributional measures as proxies for semantic relatedness.

H. Paijmans. 1997. Gravity wells of meaning: detecting information-rich passages in scientific texts. *Journal of Documentation*, 53(5):520–536.

M. Palmer, H.T. Dang, and C. Fellbaum. 2006. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(02):137–163.

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

P. Resnik. 1995. 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.

J. Rodd, G. Gaskell, and W. Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245–266.

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

J. Shawe-Taylor and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis.*

S. Shi, J.R. Wen, Q. Yu, R. Song, and W.Y. Ma. 2005. Gravitation-based model for information retrieval. In *Proceedings of SIGIR-05, 28th ACM International Conference on Research and Development in Information Retrieval*, pages 488–495. ACM New York, NY, USA.

G. Siolas and F. d'Alché Buc. 2000. Support vector machines based on a semantic kernel for text categorization. In *Proceedings of IJCNN-00, IEEE International Joint Conference on Neural Networks.*

C. J. van Rijsbergen. 2004. *The Geometry of Information Retrieval.*

P. Wittek and S. Darányi. 2007. Representing word semantics for IR by continuous functions. In S. Dominich and F. Kiss, editors, *Proceedings of ICTIR-07, 1st International Conference of the Theory of Information Retrieval*, pages 149–155.

P. Wittek, C.L. Tan, and S. Darányi. 2009. An ordering of terms based on semantic relatedness. In H. Bunt, editor, *Proceedings of IWCS-09, 8th International Conference on Computational Semantics.*

S.K.M. Wong, W. Ziarko, and P.C.N. Wong. 1985. Generalized vector space model in information retrieval. In *Proceedings of SIGIR-85, 8th ACM International Conference on Research and Development in Information Retrieval*, pages 18–25.