A Knowledge-based Semantic Kernel for Text Classification

Jamal Abdul Nasir[1], Asim Karim[1], George Tsatsaronis[2], and Iraklis Varlamis[3]

[1] School of Science and Engineering, LUMS Lahore, Pakistan
[2] Biotechnology Center (BIOTEC,) Technische Universität Dresden, Germany
[3] Department of Informatics and Telematics Harokopio University of Athens, Greece
{jamaln,akarim}@lums.edu.pk,
george.tsatsaronis@biotec.tu-dresden.de,
varlamis@hua.gr

Abstract. Typically, in textual document classification the documents are represented in the vector space using the "Bag of Words" (BOW) approach. Despite its ease of use, BOW representation cannot handle word synonymy and polysemy problems and does not consider semantic relatedness between words. In this paper, we overcome the shortages of the BOW approach by embedding a known *WordNet*-based semantic relatedness measure for pairs of words, namely *Omiotis*, into a semantic kernel. The suggested measure incorporates the TF-IDF weighting scheme, thus creating a semantic kernel which combines both semantic and statistical information from text. Empirical evaluation with real data sets demonstrates that our approach successfully achieves improved classification accuracy with respect to the standard BOW representation, when *Omiotis* is embedded in four different classifiers.

Keywords: Text Classification, Thesaurus, Semantic Kernels

1 Introduction

The key steps in text classification are document representation and classifier training using a corpus of labeled documents. In the commonly used 'Bag of Words' (BOW) representation, documents are represented by vectors whose components are weights given to different words or terms occurring in the document. Weights indicate the importance of each word, typically quantified by measures like TF-IDF. However, the BOW representation has some significant limitations: (1) It disregards the sequential order of words in documents. (2) It considers synonyms as distinct components of the vector (synonymy problem). (3) It disregards polysemy of words (i.e. words having multiple senses or meanings – polysemy problem). The lack of semantics in the BOW representation limits the effectiveness of automatic text classification methods.

In the absence of external semantic knowledge, corpus-based statistical methods, such as Latent Semantic Analysis (LSA) [1] can be applied to alleviate the synonymy problem, but the problem of polysemy still remains. The application

of Word Sense Disambiguation (WSD) techniques [2] during document preprocessing can be helpful; however, this is usually computationally expensive, and the performance of the unsupervised techniques is poor while use of supervised techniques requires large amounts of hand-annotated text documents. The use of external semantic knowledge provided by word thesauri or ontologies to adjust or "smooth" the *BOW* representation has shown much promise [3, 4]. However, the embedding of semantic information is usually computationally expensive.

In this paper, we present and evaluate a semantically-enriched *BOW* representation for text classification. We adopt a recently proposed semantic relatedness measure called *Omiotis* [5] for building a smoothing matrix and a kernel for semantically adjusting the *BOW* representation. *Omiotis* is constructed from the word thesaurus and lexical ontology *WordNet*, and is capable of handling the synonymy and polysemy problems. We evaluate four popular text classification methods on four different data sets with and without *Omiotis*-based semantic smoothing of *BOW* representation. The results demonstrate that our semantic kernel produces significant improvement in text classification performance.

The paper is organized as follows. Section 2 discusses the related work. Section 3 presents the *Omiotis* measure for the semantic relatedness between pairs of terms. In Section 4, we develop our semantic kernel and semantic smoothing matrix, and discuss its computational complexity. Section 5 presents our experimental results. Finally, Section 6 discusses our next steps.

2   Semantics in Text Mining and Information Retrieval

The importance of embedding semantic relatedness between two text segments for text classification was initially highlighted in [6] where semantic similarity between words has been used for the *semantic smoothing* of the *TF-IDF* vectors.

Semantic-aware kernels have been proposed by Mavroeidis et al. [4] who propose a generalized vector space model with *WordNet* senses and their hypernyms to improve text classification performance. Bloehdorn at al. [7] propose smoothing kernels for text classification by implicitly encoding a super concept expansion and achieve satisfactory results under poor training data and data sparseness. In [8] authors use the Latent Semantic Indexing (LSI) approach for capturing semantic relations between terms and embed them into their semantic kernel. Basili et al.[9] propose kernel functions to use prior knowledge in learning algorithms for document classification by means of the term similarity based on the *WordNet* hierarchy (conceptual density). Results show the benefit of the approach for Support Vector Machines when few training examples are available.

In this work, we present a new semantic smoothing matrix and kernel for text classification, based on a semantic relatedness measure that takes into account all of the available semantic relations in *WordNet*, by embedding the *Omiotis* measure introduced by Tsatsatonis et al. [5]. Our experimental evaluation offers an additional empirical evidence towards the claim that embedding semantic information from a knowledge base, such as *WordNet*, through a semantic kernel, improves the text classification performance.

3 Semantic Relatedness and the *Omiotis* Measure

Lexical relatedness measures can be roughly classified in three categories: (1) knowledge-based measures; (2) corpus-based measures; and (3) hybrid measures. In this work, we are using the *Omiotis*[5] knowledge-based measure for computing the relatedness between terms or words. *Omiotis* is based on a sense relatedness measure, called SR. Due to space limitations, we suggest readers to consult [5] for the details of SR, which given a pair of senses s_1, s_2, finds all the paths that connect s_1 to s_2 in the WordNet's graph and defines the pair's relatedness as:

$$SR(s_1, s_2) = \max_{P = \langle s_1, \ldots, s_2 \rangle} \{SCM(P) \cdot SPE(P)\}$$

where P ranges over all the paths that connect s_1 to s_2, SCM and SPE capture respectively the notions of the *value* of the path connecting two senses in WordNet, as well as of the *depth* of path's edges in the path with respect to the *height* of the used thesaurus/ontology. If no path exists, then $SR(s_1, s_2) = 0$.

The measure can be expanded to measure the semantic relatedness between terms, by selecting the maximum for each of the pairwise sense combinations for a pair of terms. More precisely, given a pair of terms $T : (t_1, t_2)$ for which there are entries in O, let X_1 be the set of senses of t_1 and X_2 be the set of senses of t_2 in O. Let $S : \{S_1, S_2, \ldots, S_{|X_1| \cdot |X_2|}\}$ be the set of pairs of senses, $S_k = (s_i, s_j)$, with $s_i \in X_1$ and $s_j \in X_2$. Then $SR(T, S, O)$ is defined as:

$$\max_{S_k}\{\max_P\{SCM(S_k, O, P) \cdot SPE(S_k, O, P)\}\} = \max_{S_k}\{SR(S_k, O)\} \forall k = 1..|X_1| \cdot |X_2|. \tag{1}$$

Semantic relatedness between two terms t_1, t_2 where $t_1 \equiv t_2 \equiv t$ and $t \notin O$ is defined as 1. Semantic relatedness between t_1, t_2 when $t_1 \in O$ and $t_2 \notin O$, or vice versa, is considered 0. This latter definition of SR for a pair of terms is the definition of the *Omiotis* measure that we are using in our case. [4]

4 *Omiotis*-based Semantic Kernel

4.1 Semantic Smoothing Matrix and Semantic Kernel Design

A document d is represented in the BOW representation as follows:

$$\phi : d \mapsto \phi(d) = [tf\text{-}idf(t_1, d), tf\text{-}idf(t_2, d), \ldots, tf\text{-}idf(t_D, d)]^T \in \Re^D$$

where $tf\text{-}idf(t_i, d)$ is the TF-IDF weight of term t_i in document d, and D is the total number of terms (e.g. words) in the dictionary (the superscript T denotes the transpose operator). In the above expression, the function $\phi(d)$ represents the document d as a TF-IDF vector. This function, however, can be any other mapping from a document to its vector space representation.

---

[4] A Web service implementation of *Omiotis* with pre-computed SR scores for all *WordNet* sense pairs is made available by the authors in [5], at `http://omiotis.hua.gr/`.

To enrich the BOW representation with semantic information, we construct the semantic relatedness matrix R using the $Omiotis$ semantic relatedness measure. Specifically, the i, j element of matrix R is given by $SR(T, S, O)$ (refer to Eq. 1), which quantifies the semantic relatedness between terms $T : (t_i, t_j)$. Thus, R is a $D \times D$ symmetric matrix with 1's in the principal diagonal. This smoothing matrix can be used to transform the documents' vectors in such a way that semantically related documents are brought closer together in the transformed (or feature) space (and vice versa). Mathematically, the semantically enriched BOW representation of a document d is given as:

$$\bar{\phi}(d) = (\phi(d)^T R)^T$$

Although the feature space defined above can be used directly in many classification methods, it is sometimes helpful to define the feature space implicitly via the kernel function. This is particularly important in kernel-based methods or kernel machines when the feature space is very large or even infinite in size. By definition, the kernel function computes the inner product between documents d_i and d_j in the feature space. For our case, this can be written as:

$$\kappa(d_i, d_j) = \bar{\phi}(d_i)^T \bar{\phi}(d_j) = \phi(d_i)^T R R^T \phi(d_j) \qquad (2)$$

For this to be a valid kernel function, the Gram matrix G (where $G_{ij} = \kappa(d_i, d_j)$) formed from the kernel function must satisfy the Mercer's conditions [8]. These conditions are satisfied when the Gram matrix is positive semi-definite. It has been shown in [8] that the matrix G formed by the kernel function (Eq. 2) with the outer matrix product $R R^T$ is indeed a positive semi-definite matrix.

4.2   Computational Aspects

The computational complexity of the suggested semantic kernel depends on two main factors: (1) the similarity measure between two documents d_1 and d_2, which requires the evaluation of all the unique term pairs' relatedness values and has a theoretical complexity of $O(|d_1| \cdot |d_2|)$, where $|d|$ denotes the total number of distinct terms in document d; (2) the computational complexity of $Omiotis$ for all $|d_1| \cdot |d_2|$ term pair combinations, which comprises the construction time of the semantic network to compute the paths connecting the senses of two words, and the time needed to execute the $Dijkstra$'s algorithm in order to find the optimal path connecting two senses. The complexity of the former is $O(2 \cdot k^{l+1})$ [10], where k is the maximum branching factor of the used thesaurus nodes and l is the maximum semantic path length in the thesaurus, and of the latter is $O(nL + mD + nE)$, where n is the number of nodes in the network, m the number of edges, L is the time for insert, D the time for decrease-key and E the time for extract-min. Using The use of Fibonacci heaps reduces the cost of extract-min to $O(\log n)$ and $L = D = O(1)$, thus significantly reducing the cost of execution. The pre-computation of all the pairwise sense and term relatedness values, which are publicly available through the $Omiotis$ service[5] makes the semantic kernel computation applicable even for large data sets.

[5] http://omiotis.hua.gr/WebSite/wsinfo.html

5   Empirical Evaluation

We evaluate the performance of our semantic smoothing approach by using four classification methods on four popular text classification data sets (Ohsumed[6], 20 Newsgroups[7], WebKB[8] and Movie Reviews[9]). All data sets are preprocessed via tokenization, stop word removal, and TF-IDF vector construction (the standard *BOW* representation).

Supervised text classification methods can be based on a generative or a discriminative model of the problem. We employ two discriminative methods, Support Vector Machines (SVM) and Balanced Winnow (BW), and two common generative methods, Naive Bayes (NB) and Maximum Entropy (ME). We perform our experiments using the software RapidMiner[10] (for SVM and NB) and the Mallet toolkit[11] (for ME and BW). For each method, we evaluate its performance under two settings: (1) standard *BOW* representation and (2) semantically smoothed *BOW* represented using the *Omiotis* measure. We report the performance with average classification accuracy obtained from an 10-fold cross-validation process.

Table 1 shows the results of our empirical evaluation. It gives the percent accuracy obtained from 10-fold cross-validation by each method on the four data sets. The methods identified with the *Omiotis* subscript are the ones using our *Omiotis*-based semantic kernel (or semantic smoothing approach). These results demonstrate that enriching the *BOW* representation with our semantic smoothing approach improves text classification performance. This improvement is seen across different classification methods and different data sets. From among the 16 pairs of results, the performance of the *Omiotis*-based methods is better than the standard methods in 14 pairs.

Table 1. Text classification performance in percent accuracy

	$MovieReview$	$Ohsumed$	$20Newsgroups$	$WebKB$
SVM	83.30	55.15	90.08	86.37
$SVM_{Omiotis}$	**91.97**	**57.17**	**92.93**	84.58
NB	77.41	50.32	87.27	84.17
$NB_{Omiotis}$	84.13	51.29	90.44	88.52
ME	79.11	51.47	85.31	91.02
$ME_{Omiotis}$	81.86	50.17	87.35	**91.52**
BW	76.23	50.93	81.66	81.42
$BW_{Omiotis}$	79.25	51.83	84.58	85.34

---

[6] http://ir.ohsu.edu/ohsumed/ohsumed.html

[7] http://people.csail.mit.edu/jrennie/20Newsgroups/

[8] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

[9] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[10] http://www.rapid-i.com/

[11] http://mallet.cs.umass.edu/

To verify the consistency of the observed results, we applied the Wilcoxon signed-ranks test, which is recommended for our case [11], on the observed differences in performances of all methods on all the data sets. In our test, we found that the observed differences are statistically significant and the null hypothesis is rejected (having achieved a very low p-value of only 0.0023). This test confirms that our semantic kernel produces consistent and statistically significant improvement in text classification performance.

6   Conclusions and Future Work

In this paper, we present a semantic kernel for smoothing the *BOW* representation. We evaluate the impact of our semantic kernel on text classification problems using four popular classifiers on four commonly-used text corpora. We find that the *Omiotis* enhanced representation produces significant improvement in classification accuracy for all classifiers. As a next step, we will extend the *BOW* representation by incorporating discrimination information for text classification and evaluate and compare our representation approaches for text clustering tasks.

References

1. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by Latent Semantic Analysis. JASIS **41**(6) (1990) 391–407
2. Navigli, R.: Word sense disambiguation: A survey. ACM Computing Surveys **41**(2) (2009) 10:1–10:69
3. Basili, R., Cammisa, M., Moschitti, A.: A semantic kernel to exploit linguistic knowledge. In: Proc. of the AI*IA 2005. (2005) 290–302
4. Mavroeidis, D., Tsatsaronis, G., Vazirgiannis, M., Theobald, M., Weikum, G.: Word sense disambiguation for exploiting hierarchical thesauri in text classification. In: Proc. of the 9th PKDD. (2005) 181–192
5. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research **37** (2010) 1–39
6. Siolas, G., d'Alché Buc, F.: Support vector machines based on a semantic kernel for text categorization. In: Proc. of IEEE IJCNN'00, Washington, DC, USA (2000)
7. Bloehdorn, S., Basili, R., Cammisa, M., Moschitti, A.: Semantic kernels for text classification based on topological measures of feature similarity. In: Proc. of ICDM'06. (2006) 808–812
8. Cristianini, N., Taylor, J.S., Lodhi, H.: Latent Semantic Kernels. In: Proc. of the Eighteenth International Conference on Machine Learning. (2001) 66–73
9. Basili, R., Cammisa, M., Moschitti, A.: A Semantic Kernel to classify texts with very few training examples. In: Informatica, 30 (2). (2006) 163–172
10. Tsatsaronis, G., Vazirgiannis, M., Androutsopoulos, I.: Word sense disambiguation with spreading activation networks generated from thesauri. In: Proc. of IJCAI. (2007) 1725–1730
11. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30