Semantic classification with WordNet kernels

Diarmuid Ó Séaghdha
Computer Laboratory
University of Cambridge
United Kingdom
do242@cl.cam.ac.uk

Abstract

This paper presents methods for performing graph-based semantic classification using kernel functions defined on the WordNet lexical hierarchy. These functions are evaluated on the SemEval Task 4 relation classification dataset and their performance is shown to be competitive with that of more complex systems. A number of possible future developments are suggested to illustrate the flexibility of the approach.

1 Introduction

The estimation of semantic similarity between words is one of the longest-established tasks in Natural Language Processing and many approaches to the problem have been proposed. The two dominant lexical similarity paradigms are distributional similarity, which compares words on the basis of their observed co-occurrence behaviour in corpora, and semantic network similarity, which compares words based on their position in a graph such as the WordNet hierarchy. In this paper we consider measures of network similarity for the purpose of supervised classification with kernel methods. The utility of kernel functions related to popular distributional similarity measures has recently been demonstrated by Ó Séaghdha and Copestake (2008); we show here that kernel analogues of WordNet similarity can likewise give good performance on a semantic classification task.

2 Kernels derived from graphs

Kernel-based classifiers such as support vector machines (SVMs) make use of functions called *kernel functions* (or simply *kernels*) to compute the similarity between data points (Shawe-Taylor and Cristianini, 2004). Valid kernels are restricted to the set of *positive semi-definite (psd) functions*, i.e., those that correspond to an inner product in some vector space. Kernel methods have been widely adopted in NLP over the past decade, in part due to the good performance of SVMs on many tasks and in part due to the ability to exploit prior knowledge about a given task through the choice of an appropriate kernel function. In this section we consider kernel functions that use spectral properties of a graph to compute the similarity between its nodes. The theoretical foundations and some machine learning applications of the adopted approach have been developed by Kondor and Lafferty (2002), Smola and Kondor (2003) and Herbster et al. (2008).

Let G be a graph with vertex set $V = v_1, \ldots, v_n$ and edge set $E \subseteq V \times V$. We assume that G is connected and undirected and that all edges have a positive weight $w_{ij} > 0$. Let $\mathbf{A}$ be the symmetric $n \times n$ matrix with entries $A_{ij} = w_{ij}$ if an edge exists between vertices v_i and v_j, and $A_{ij} = 0$ otherwise. Let $\mathbf{D}$ be the diagonal matrix with entries $D_{ii} = \sum_{j \in V} A_{ij}$. The *graph Laplacian* $\mathbf{L}$ is then defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{1}$$

The normalised Laplacian is defined as $\hat{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$. Both $\hat{\mathbf{L}}$ and $\mathbf{L}$ are positive semi-definite, but they are typically used as starting points

for the derivation of kernels rather than as kernels themselves.

Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of $\mathbf{L}$ and $u_1, \ldots, u_n$ the corresponding eigenvectors. Note that $u_n = 0$ for all graphs. $\mathbf{L}$ is singular and hence has no well-defined inverse, but its pseudoinverse $\mathbf{L}^+$ is defined as

$$\mathbf{L}^+ = \sum_{i=1}^{n-1} \lambda_i^{-1} u_i u_i^{\mathrm{T}} \quad (2)$$

$\mathbf{L}^+$ is positive definite, and its entries are related to the *resistance distance* between points in an electrical circuit (Herbster et al., 2008) and to the *average commute-time distance*, i.e., the average distance of a random walk from one node to another and back again (Fouss et al., 2007). The similarity measure defined by $\mathbf{L}^+$ hence takes information about the connectivity of the graph into account as well as information about adjacency. An analogous pseudoinverse $\hat{\mathbf{L}}^+$ can be defined for the normalised Laplacian.

A second class of graph-based kernel functions are the *diffusion kernels* introduced by Kondor and Lafferty (2002). The kernel $\mathbf{H}_t$ is defined as $\mathbf{H}_t = e^{-t\hat{\mathbf{L}}}$, or equivalently:

$$\mathbf{H}_t = \sum_{i=1}^{n-1} \exp(-t\hat{\lambda}_i)\hat{u}_i\hat{u}_i^{\mathrm{T}} \quad (3)$$

where $t > 0$, and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ and $\hat{u}_1, \ldots, \hat{u}_n$ are the eigenvalues and eigenvectors of $\hat{\mathbf{L}}^+$ respectively. $\mathbf{H}_t$ can be interpreted in terms of heat diffusion or the distribution of a lazy random walk emanating from a given point at a time point t.

3 Methodology

3.1 Graph construction

WordNet (Fellbaum, 1998) is a semantic network in which nodes correspond to word senses (or *synsets*) and edges correspond to relations between senses. In this work we restrict ourselves to the noun component of WordNet and use only hyponymy and instance hyponymy relations for graph construction. The version of WordNet used is WordNet 3.0.

To evaluate the utility of the graph-based kernels described in Section 2 for computing lexical similarity, we use the dataset developed for the task on Classifying Semantic Relations Between Nominals at the 2007 SemEval competition (Girju et al., 2007). The dataset comprises candidate example sentences for seven two-argument semantic relations, with 140 training sentences and approximately 80 test sentences for each relation. It is a particularly suitable task for evaluating WordNet kernels, as the candidate relation arguments for each sentence are tagged with their WordNet sense and it has been previously shown that a kernel model based on distributional lexical similarity can attain very good performance (Ó Séaghdha and Copestake, 2008).

3.2 Calculating the WordNet kernels

The noun hierarchy in WordNet 3.0 contains 82,115 senses; computing kernel similarities on a graph of this size raises significant computational issues. The calculation of the Laplacian pseudoinverse is complicated by the fact that while $\mathbf{L}$ and $\hat{\mathbf{L}}$ are very sparse, their pseudoinverses are invariably dense and require very large amounts of memory. To circumvent this problem, we follow Fouss et al. (2007) in computing $\mathbf{L}^+$ and $\hat{\mathbf{L}}^+$ one column at a time through a Cholesky factorisation procedure. Only those columns required for the classification task need be calculated, and the kernel computation for each relation subtask can be performed in a matter of minutes. Calculating the diffusion kernel involves an eigendecomposition of $\hat{\mathbf{L}}$, meaning that computing the kernel exactly is infeasible. The solution used here is to approximate $\mathbf{H}_t$ by using the m smallest components of the spectrum of $\hat{\mathbf{L}}$ when computing (3); from (2) it can be seen that a similar approximation can be made to speed up computation of $\mathbf{L}^+$ and $\hat{\mathbf{L}}^+$.

3.3 Experimental setup

For all kernels and relation datasets, the kernel matrix for each argument position was precomputed and normalised so that every diagonal entry equalled 1. A small number of candidate arguments are not annotated with a WordNet sense or are assigned a non-noun sense; these arguments were assumed to have self-similarity equal to 1 and zero similarity to all other arguments. This does not affect the positive semi-definiteness of the kernel matrices. The per-argument kernel matrices were summed to give the kernel matrix for each relation subtask. The ker-

	Full graph		$m = 500$		$m = 1000$	
Kernel	Acc	F	Acc	F	Acc	F
B	72.1	68.4	-	-	-	-
$\mathbf{L}^+$	73.3	69.4	73.2	70.5	73.6	70.6
$\hat{\mathbf{L}}^+$	72.5	70.0	72.7	70.0	**74.1**	**71.0**
$\mathbf{H}^t$	-	-	68.6	64.7	69.8	65.1

Table 1: Results on SemEval Task 4

nels described in Section 2 were compared to a baseline kernel B. This baseline represents each word as a binary feature vector describing its synset and all its hypernym synsets in the WordNet hierarchy, and calculates the linear kernel between vectors.

All experiments were run using the LIBSVM support vector machine library (Chang and Lin, 2001). For each relation the SVM cost parameter was optimised in the range $(2^{-6}, 2^{-4}, \ldots, 2^{12})$ through cross-validation on the training set. The diffusion kernel parameter t was optimised in the same way, in the range $(10^{-3}, 10^{-2}, \ldots, 10^{3})$.

4 Results

Macro-averaged accuracy and F-score for each kernel are reported in Table 1. There is little difference between the Laplacian and normalised Laplacian pseudoinverses; both achieve better performance than the baseline B. The results also suggest that the reduced-eigenspectrum approximations to $\mathbf{L}^+$ and $\hat{\mathbf{L}}^+$ may bring benefits in terms of performance as well as efficiency via a smoothing effect. The best performance is attained by the approximation to $\hat{\mathbf{L}}^+$ with $m = 1,000$ eigencomponents. The heat kernel $\mathbf{H}^t$ fares less well; the problem here may be that the optimal range for the t parameter has not been identified.

Comparing these results to those of the participants in the 2007 SemEval task, the WordNet-based lexical similarity model fares very well. All versions of $\mathbf{L}^+$ and $\hat{\mathbf{L}}^+$ attain higher accuracy than all but one of 15 systems in the competition and higher F-score than all but three. Even the baseline B ranks above all but the top three systems, suggesting that this too can be a useful model. This is in spite of the fact that all systems which made use of the sense annotations also used a rich variety of other information sources such as features extracted from the sentence context, while the models presented here use only the graph structure of WordNet.[1]

5 Related work

There is a large body of work on using WordNet to compute measures of lexical similarity (Budanitsky and Hirst, 2006). However, many of these measures are not amenable for use as kernel functions as they rely on properties which cannot be expressed as a vector inner product, such as the lowest common subsumer of two vertices. Hughes and Ramage (2007) present a lexical similarity model based on random walks on graphs derived from WordNet; Rao et al. (2008) propose the Laplacian pseudoinverse on such graphs as a lexical similarity measure. Both of these works share aspects of the current paper; however, neither address supervised learning or present an application-oriented evaluation.

Extracting features from WordNet for use in supervised learning is a standard technique (Scott and Matwin, 1999). Siolas and d'Alche-Buc (2000) and Basili et al. (2006) use a measure of lexical similarity from WordNet as an intermediary to smooth bag-of-words kernels on documents. Siolas and d'Alche-Buc use an inverse path-based similarity measure, while Basili et al. use a measure of "conceptual density" that is not proven to be positive semi-definite.

6 Conclusion and future work

The main purpose of this paper has been to demonstrate how kernels that capture spectral aspects of graph structure can be used to compare nodes in a lexical hierarchy and thus provide a kernelised measure of WordNet similarity. As far as we are aware, these measures have not previously been investigated in the context of semantic classification. The resulting WordNet kernels have been evaluated on the SemEval Task 4 dataset and shown to attain a higher level of performance than many more complicated systems that participated in that task.

Two obvious shortcomings of the kernels discussed here are that they are defined on senses rather than words and that they are computed on a

[1] Of course, information about lexical similarity is not sufficient to classify all examples. In particular, the models presented here perform relatively badly on the ORIGIN-ENTITY and THEME-TOOL relations, while scoring better than all SemEval entrants on INSTRUMENT-AGENCY and PRODUCT-PRODUCER.

rather impoverished graph structure (the WordNet hyponym hierarchy is quite tree-like). One of the significant benefits of spectral graph kernels is that they can be computed on arbitrary graphs and are most powerful when graphs have a rich connectivity structure. Some potential future directions that would make greater use of this flexibility include the following:

- A simple extension from sense-kernels to word-kernels involves adding word nodes to the WordNet graph, with an edge linking each word to each of its possible senses. This is similar to the graph construction method of Hughes and Ramage (2007) and Rao et al. (2008). However, preliminary experiments on the SemEval Task 4 dataset indicate that further refinement of this approach may be necessary in order to match the performance of kernels based on distributional lexical similarity (Ó Séaghdha and Copestake, 2008).
- Incorporating other WordNet relations such as meronymy and topicality gives a way of kernelising semantic association or relatedness; one application of this might be in developing supervised methods for spelling correction (Budanitsky and Hirst, 2006).
- A WordNet graph can be augmented with information from other sources, such as links based on corpus-derived similarity. Alternatively, the graph-based kernel functions could be applied to graphs constructed from parsed corpora (Minkov and Cohen, 2008).

References

Roberto Basili, Marco Cammisa, and Alessandro Moschitti. 2006. A semantic kernel to classify texts with very few training examples. *Informatica*, 30(2):163–172.

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

Chih-Chung Chang and Chih-Jen Lin. 2001. LIBSVM: A library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. 2007. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 Task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-07)*.

Mark Herbster, Massimiliano Pontil, and Sergio Rojas Galeano. 2008. Fast prediction on a tree. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS-08)*.

Thad Hughes and Daniel Ramage. 2007. Lexical semantic relatedness with random graph walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-07)*.

Risi Imre Kondor and John Lafferty. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning (ICML-02)*.

Einat Minkov and William W. Cohen. 2008. Learning graph walk based similarity measures for parsed text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*.

Diarmuid Ó Séaghdha and Ann Copestake. 2008. Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*.

Delip Rao, David Yarowsky, and Chris Callison-Burch. 2008. Affinity measures based on the graph Laplacian. In *Proceedings of the 3rd TextGraphs Workshop on Graph-based Algorithms for NLP*.

Sam Scott and Stan Matwin. 1999. Feature engineering for text classification. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*.

John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Georges Siolas and Florence d'Alche-Buc. 2000. Support vector machines based on a semantic kernel for text categorization. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*.

Alexander J. Smola and Risi Kondor. 2003. Kernels and regularization on graphs. In *Proceedings of the the 16th Annual Conference on Learning Theory and 7th Workshop on Kernel Machines (COLT-03)*.