ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

A semi-supervised approach to question classification*

David Tomás[1] and Claudio Giuliano[2]

1- Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, Spain

2- Human Language Technology group
FBK-Irst, Italy

Abstract. This paper presents a machine learning approach to question classification. We have defined a kernel function based on latent semantic information acquired from unlabeled data. This kernel allows including external semantic knowledge into the supervised learning process. We have combined this knowledge with a bag-of-words approach by means of composite kernels to obtain state-of-the-art results. As the semantic information is acquired from unlabeled text, our system can be easily adapted to different languages and domains.

1 Introduction

Question classification is one of the main tasks carried out in a question answering (QA) system. The goal is to assign labels to questions based on the expected answer type. For example, a question like "Who was the first American in space?" could be classified as *person*. In a QA system, this information allows to narrow down the set of expected answers to those that match the class identified (a name of a *person* in the previous example).

In this paper, we present a semi-supervised machine learning approach to question classification based on kernel methods [1]. Classical n-gram models are unable to deal with the problem of ambiguity and variability of questions. We extend the traditional bag-of-words representation, offering an effective way to integrate external semantic information in the question classification process by means of semantic kernels. As a result, we obtain a generalized similarity function between questions. This function can incorporate semantic relations between words acquired from unlabeled data, such as Wikipedia. The result is a flexible system easily adaptable to different languages and domains.

We tested our approach on a corpus of 6,000 questions. We obtained state-of-the-art results combining bag-of-words with unlabeled data, showing a further improvement when this information is combined with other lexical resources.

The rest of the paper is organized as follows. Section 2 describes the kernels defined for this task and how the semantic information is included in the system. Section 3 shows the experiments carried out and the results obtained. Section 4 presents related work. Conclusions and future work are discussed in Section 5.

*This work has been supported by the QALL-ME project (6th EU Framework Research Programme, contract number FP6-IST-033860), the X-Media project (sponsored by the European Commission as part of the Information Society Technologies program under EC grant number IST-FP6-026978) and the FIRB research project (N. RBIN045PXH).

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

2 Semantic kernels for question classification

Kernel methods are a popular machine learning approach within the natural language processing community. The strategy adopted by kernel methods consists of splitting the learning problem in two parts. They first embed the input data in a suitable feature space, and then use a linear algorithm to discover nonlinear pattern in the input space. Typically, the mapping is performed implicitly by a so-called *kernel function*. The kernel function is a similarity measure between the input data that depends exclusively on the specific data type and domain.

Formally, the kernel is a function $k : X \times X \rightarrow \mathbb{R}$ that takes as input two data objects and outputs a real number characterizing their similarity. That is, for all x_i, $x_j \in X$, it satisfies

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

where ϕ is an explicit mapping from X to an (inner product) feature space $\mathcal{F}$.

The simplest method to estimate the similarity between two questions is to compute the inner product of their vector representations in the vector space model (VSM). Formally, we define a space of dimensionality N in which each dimension is associated with one word from the dictionary, and the question q is represented by a row vector

$$\phi(q) = (f(t_1, q), f(t_2, q), \ldots, f(t_N, q)),$$

where the function $f(t_i, q)$ records whether a particular token t_i is used in q. Thus, we can define the *bag-of-words kernel* $K_{BOW}(q_1, q_2)$ between questions as

$$\langle \phi(q_i), \phi(q_j) \rangle = \sum_{l=1}^{N} f(t_l, q_1) f(t_l, q_2).$$

However, such an approach does not deal well with lexical variability and ambiguity. To address these shortcomings, we introduce the class of semantic kernels in order to define an effective semantic VSM using external knowledge.

In the field of question classification, semantic information has demonstrated to be fundamental for improving accuracy [2]. In the context of kernel methods, semantic information can be integrated considering linear transformations of the type $\tilde{\phi}(q_j) = \phi(q_j)\mathbf{S}$, where $\mathbf{S}$ is a $N \times k$ matrix [1]. The matrix $\mathbf{S}$ can be rewritten as $\mathbf{S} = \mathbf{WP}$, where $\mathbf{W}$ is a diagonal matrix determining the word weights, while $\mathbf{P}$ is the *word proximity matrix* capturing the semantic relations between words. This matrix $\mathbf{P}$ can be defined by setting non-zero entries between those words whose semantic relation is inferred from an external source of domain knowledge. The *semantic kernel* takes the general form

$$\tilde{k}(q_i, q_j) = \phi(q_i)\mathbf{SS}'\phi(q_j)' = \tilde{\phi}(q_i)\tilde{\phi}(q_j)'.$$

We have defined two alternatives approaches to define the proximity matrix. The first makes use of manually built lists of semantically related words and it is defined for comparative purposes only, while the second exploits unlabeled data and represents the main contribution of this work.

Explicit semantic kernel Manually constructed lists of semantically related words typically provide a simple and effective way to introduce semantic information into the kernel. To define a semantic kernel from such resources, we can explicitly construct the proximity matrix $\mathbf{P}$ by setting its entries to reflect the semantic proximity between the words i and j in the specific lexical resource.

We used the class-specific word lists manually constructed by [2] to define $\mathbf{P}$.[1] The corresponding explicit kernel, called *semantic related kernel* $K_{SemRel}(q_i, q_j)$, is defined as

$$\phi(q_i)\mathbf{P}\mathbf{P}'\phi(q_j)' = \tilde{\phi}(q_i)\tilde{\phi}(q_j)'.$$

Latent semantic kernel An alternative approach to define a proximity matrix is by looking at co-occurrence information in a (large) corpus. Two words are considered semantically related if they frequently co-occur in the same texts. Latent semantic indexing (LSI) [3] is an effective vector space representation of corpora being able to acquire semantic information using co-occurrence information. This second approach is more attractive because it allows us to automatically define semantic models for different languages and domains.

We use singular valued decomposition (SVD) to automatically define the proximity matrix $\mathbf{\Pi}$ from Wikipedia texts, represented by its term-by-document matrix $\mathbf{D}$, where the $\mathbf{D}_{i,j}$ entry gives the frequency of term t_i in document d_j. SVD decomposes the term-by-document matrix $\mathbf{D}$ into three matrices $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices whose columns are the eigenvectors of $\mathbf{D}\mathbf{D}'$ and $\mathbf{D}'\mathbf{D}$ respectively, and $\mathbf{\Sigma}$ is the diagonal matrix containing the singular values of $\mathbf{D}$.

The selection of a representative corpus is an important part of the process of defining a semantic space. The use of Wikipedia allows us to define a open-domain statistical model. Under this setting, we define the proximity matrix $\mathbf{\Pi}$ as

$$\mathbf{\Pi} = \mathbf{U}_k\mathbf{\Sigma}_k^{-1},$$

where $\mathbf{U}_k$ is the matrix containing the first k columns of $\mathbf{U}$ and k is the dimensionality of the latent semantic space and can be fixed in advance.

The matrix $\mathbf{\Pi}$ is used to define a linear transformation $\pi : \mathbb{R}^N \rightarrow \mathbb{R}^k$, that maps the vector $\phi(q_j)$, represented in the standard VSM, into the vector $\tilde{\phi}(q_j)$ in the latent semantic space. Formally, π is defined as

$$\pi(\phi(q_j)) = \phi(q_j)(\mathbf{W}\mathbf{\Pi}) = \tilde{\phi}(q_j),$$

where $\vec{q_j}$ is represented as a row vector, $\mathbf{W}$ is a $N \times N$ diagonal matrix determining the word weights such that $\mathbf{W}_{i,i} = idf(w_i)$, where $idf(w_i)$ is the *inverse document frequency* of w_i.

Finally, the *latent semantic kernel* is explicitly defined as

$$K_{LS}(q_i, q_j) = \langle \pi(\phi(q_i)), \pi(\phi(q_j)) \rangle.$$

Note that we have used a series of successive mappings each of which adds some further improvement to the question representation.

[1] The word lists are freely available at http://l2r.cs.uiuc.edu/ cogcomp/Data/QA/QC/.

ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

3 Experiments

The number of studies carried out the last years in the field of question classification presents this task as a non-trivial and well-defined subtask of the QA process. In this section, we describe the evaluation framework and the results obtained with our approach to question classification. The purpose of these experiments is to test the effect of the different semantic kernels in the classification, and the robustness of this approach in dealing with different languages. We compare our results with other state-of-the-art systems.

3.1 Description of the data set

The UIUC corpus has become a *de facto* standard for the evaluation of question classification systems. It was first described in [2] and contains a training set of 5,452 questions and a test set of 500 questions. These questions are labeled with a two level hierarchical taxonomy of classes. The first level consists of 6 coarse-grained classes (like *human*, *location* or *numeric*) that are subclassified on a second level of 50 fine-grained classes (refinements like *city*, *country* or *mountain* for the coarse class *location*). This hierarchy allows classifying questions at different degrees of granularity.

We evaluated our system on both coarse- and fine-grained classification. The latter is a touchstone for machine learning approaches, as the number of samples per question class is drastically reduced with respect to the coarse classification.

3.2 Experimental setup

We have defined three composite kernels in our experiments to combine and extend the individual ones. We did it by means of the closure properties of the kernel functions: $K_{BOW} + K_{LS}$ combines the bag-of-words with semantic information automatically acquired from Wikipedia; $K_{BOW} + K_{SemRel}$ combines the bag-of-words with semantic information acquired from manually constructed lists of words semantically related to specific answer types; $K_{BOW} + K_{LS} + K_{SemRel}$ combines elements from both previous kernels.

To define the proximity matrix, we performed the SVD using 400 dimensions ($k = 400$) on the term-by-document matrix obtained from 50,000 pages randomly selected from the English version of Wikipedia. The statistical significance of all the results was checked by means of the *approximate randomization* procedure [4], with significance levels of 0.05 and 0.01.

3.3 Experimental results

Table 1 shows the accuracy on the benchmark. The results obtained employing only the latent semantic kernel K_{LS} (70.4% for coarse and 71.2% for fine), show that the semantic information induced by this kernel is not enough for the task of question classification. The importance that *wh-words* and stopwords have in question classification cannot be captured with this model.

Kernel	Coarse	Fine
K_{BOW}	86.4	80.8
K_{LS}	70.4	71.2
$K_{BOW} + K_{LS}$	90.0	83.2
$K_{BOW} + K_{SemRel}$	89.4	84.0
$K_{BOW} + K_{LS} + K_{SemRel}$	90.8	85.6

Table 1: Results for coarse and fine classes.

Using the composite kernel $K_{BOW} + K_{LS}$, we improve the results when compared with the baseline K_{BOW}, achieving 90.0% for coarse classes and 83.2% for fine classes. This difference is statistically significant ($p < 0.01$) in both coarse and fine classification. In this case, the composite kernel allows to successfully complementing the information from the bag-of-words with the semantic knowledge obtained by means of the K_{LS}.

On the other hand, the difference between $K_{BOW} + K_{LS}$ and $K_{BOW} + K_{SemRel}$ is not statistically significant in both coarse and fine classification. This means that the improvement achieved with both resources is equivalent. The advantage of the approach with the composite kernel $K_{BOW} + K_{LS}$ is that we do not need any handcrafted resources.

Finally, we combined both semantic resources in the kernel $K_{BOW} + K_{LS} + K_{SemRel}$. This composite kernel further improves the results obtained with the previous kernels at a significance level of $p < 0.05$ for the fine-grained classification, obtaining 85.6% precision. This result reveal that K_{LS} and K_{SemRel} capture different semantic relations and can complement each other.

The learning curves for both coarse- and fine-grained experiments (not included here due to space limitations), revealed that the use of the composite kernel $K_{LS} + K_{BOW}$ increased the generalization skills of the system. On average, this composite kernel achieved the same accuracy that the baseline K_{BOW} with just half of the training samples.

4 Related work

The work by [2] was one of the first serious attempts to evaluate the performance of question classification systems in isolation. They developed a hierarchical classifier based on SNoW. They employed several resources to obtain a linguistically rich feature space, including head chunks, named entities and a handcrafted list of semantically related words. Their system obtained 91% precision for coarse classes and 84.2% for fine classes.

The work developed in [5] was a first attempt to apply SVD for dimensionality reduction in question classification, but they did not obtain any improvement with this technique. They achieved 79.8% precision for fine-grained classification with SVD and 2000 dimensions, obtaining worse results than the original n-gram representation. The main difference with our approach is that they built the

statistical model using a very small corpus of questions (i.e., the training and test sets), instead of exploiting a large unlabeled corpus of documents.

Another proposal based on SVM is the one developed in [6]. They obtained 91.8% precision for coarse classes employing bag-of-words and a tree kernel with parsing information. They did not perform any experiment for fine-grained classification.

In [7], they obtained 92.6% performance with log-linear models for coarse classes. This system learned from lexical and syntactic information obtained from a parser specially trained for tagging questions. For the fine-grained classification, they employed features extracted from WordNet, named entities and gazetteers, obtaining 86.6% precision.

5 Conclusions and future work

We have presented an approach to question classification based on kernel methods. We employed composite kernels to incorporate semantic information and extend the bag-of-words representation. We employed a latent semantic kernel to obtain a generalized similarity function between questions. The model was acquired from unlabeled text from Wikipedia, resulting in a flexible system easily adaptable to different languages and domains. We further improved the system including an explicit semantic kernel based on lists of semantically related words.

We tested the system on the UIUC data set, a corpus of questions widely employed in question classification research, in order to compare the performance of our proposal with other systems. We obtained results comparable to the state-of-the-art in both coarse and fine classification. We surpassed many other systems that make an intensive use of linguistic resources and tools.

For future work, we want to investigate the effect of varying the corpus, the number of documents, and dimensions used to define the semantic space and test our approach in different languages and restricted domains.

References

[1] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[2] Xin Li and Dan Roth. Learning question classifiers. In *Proceedings of COLING'02*, pages 1–7, 2002.

[3] Scott C. Deerwester, Susan T. Dumais, Thoms K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] Eric W. Noreen. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, New York, NY, 1989.

[5] Kadri Hacioglu and Wayne Ward. Question classification with support vector machines and error correcting codes. In *Proceedings of NAACL '03*, pages 28–30, 2003.

[6] Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of ACL'07*, pages 776–783, 2007.

[7] Phil Blunsom, Krystle Kocik, and James R. Curran. Question classification with log-linear models. In *Proceedings of SIGIR '06*, pages 615–616, 2006.