Term Relevance Dependency Model for Text Classification

Meng-Sung Wu* and Hsin-Min Wang†
*Information and Communications Research Laboratories,
Industrial Technology Research Institute, Hsinchu, Taiwan
†Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail:wums@itri.org.tw, whm@iis.sinica.edu.tw

Abstract

Text classification (TC) has long been an important research topic in information retrieval (IR) related areas. Conventional language model (LM)-based TC is solely based on matching the words in the documents and classes by using a naïve Bayes classifier (NBC). In the literature, both the term association model (TA), which further considers word-to-word information, and the relevance model (RM), which further considers word-to-document information, have been shown to outperform a simple LM for IR. In this paper, we study a novel integration of TA with RM for LM-NBC-based TC. The new model is called the term relevance dependency model. In the model, the probability of a word given a class is represented by a term association LM probability learned by a RM framework. The results of TC experiments on the 20newsgroups and Reuters-21578 corpora demonstrate that the new model outperforms the standard NBC and several other LM-NBC-based methods.

1. Introduction

Text classification (TC) is the task of classifying documents into a set of predefined categories. It has long been an important research topic in information retrieval (IR). Many statistical classification methods and machine learning techniques have been developed for and applied to TC, such as the naive Bayes classifier (NBC), the support vector machines (SVM), the k-nearest neighbor method (*k*-NN), neural networks and the boosting method. Although several experiments have shown that SVM can produce better classification results than NBC, the latter is still widely used in text classification for its simplicity and efficiency. Most of the existing methods represent a document using a vector space model (VSM) or a language model (LM). For example, the bag-of-words (BoW) method is a widely used data representation in IR and TC. Under this scheme, each document is modeled as a vector with a dimension equal to the size of the dictionary, and each element of the vector denotes the frequency that a word appears in the document. Basically, all the words are treated independently based on the assumption that the association between words in sentences can be ignored. However, the assumption inevitably imposes a limitation on the classification performance.

The idea of using a term relationship [7] has been found to be useful in a wide range of applications. In the field of IR, integrating term relationships into LM has been studied and has attracted great interest in the past. Many researchers have revealed that modeling term associations could provide richer semantics of documents for language modeling and IR [2, 4, 12]. However, all the above methods only consider the relationships of words. We believe that the performance of these statistical methods can be further improved by considering more relevance information.

Incorporating relevance in IR systems has been broadly studied in the past. Relevance feedback [10] is a well-known method to the IR community. The relevance model (RM) has been shown to perform very well for estimating an accurate query LM in IR [11]. Different to the term relationship-based methods, in the RM method, the probability of a word is conditioned on a set of words rather than a sequence of historical words. For applications other than IR, the relevance model has been proposed for speech recognition [5] and text summarization[8].

This paper is focused on LM-based TC. We propose a novel term relevance dependency model (TRDM) for TC. TRDM incorporates the strengths of term associations into the model translation framework. Unlike previous studies, which only consider co-occurrences of words, we take a different approach. To discover the word-to-document relationship in each class, we learn

the model translation based on the relevance-based language models. We select a related document in the class, and then a document term is generated based on the observed document.

2. Background and Related Work

2.1. Applying LM as Text Classifier

The naive Bayes classifier (NBC) is a popular machine learning technique for TC. The method assumes a probabilistic generative model for text. LM was introduced to TC by Bai and Nie [1]. The score of a class c for a given document d can be decided as follows

$$c* = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} P(d|c)P(c) \quad (1)$$

By assuming that all words in d are independent of each other, $P(d|c)$ can be further decomposed into the product of individual feature (word) probabilities, and the decision can be rewritten as follows

$$c* = \arg\max_{c \in C} \prod_{t=1}^{|n_d|} P(w_t|c)P(c), \quad (2)$$

where $|n_d|$ is the number of words/terms in the d. We can linearly interpolate the class unigram LM with the collection unigram LM, $\mathscr{M}_B$, by using the *Jelinek-Mercer* smoothing method as follows

$$P(w|\mathscr{M}_C) = \lambda P(w|c) + (1-\lambda)P(w|\mathscr{M}_B), \quad (3)$$

where λ can be tuned empirically and $\mathscr{M}_C$ is the new LM estimated for class c. Then, $P(w_t|c)$ in (2) is replaced by $P(w_t|\mathscr{M}_C)$. The class prior probability $P(c)$ are estimated from the training documents with Laplace smoothing.

2.2. Modeling Relevance in IR

The goal of an IR system is to retrieve the documents to fulfill the user's information need formulated as a query $\mathbf{q} = \{w_1, \cdots, w_{n_q}\}$. Thus, it is important to be able to measure the *relevance* between a query and a document, and make use of it. One classic way to use relevance is the relevance feedback, where the relevant documents marked by the user are employed to refine the query representation. The relevance-based LM for IR have been proposed in [10]. Therein, a relevance model (RM) refers to $P(w|\mathscr{R}_\mathbf{q})$, the probability of seeing word w in a document in the relevant class denoted by $\mathscr{R}_\mathbf{q}$. It is assumed that $P(w|\mathscr{R}_\mathbf{q})$ can be approximated in the following way

$$P(w|\mathscr{R}_\mathbf{q}) \approx P(w|\mathbf{q}) = \frac{P(w,\mathbf{q})}{P(\mathbf{q})}. \quad (4)$$

It is important to point out that the relevance-based LM approaches discussed above are kind of relevance feedback schemes, in the sense that the probability of seeing a word in a document of the relevant class is modified by the retrieved documents.

3. Term Relevance Dependency Model

The unigram LM for TC is simply based on matching the literal words in the documents and classes. As a principled approach to capture the semantic relationships of words in IR, the term association model (TA)-based approaches have been shown to outperform the simple LM-based approaches [12]. The task of applying a TA in TC can be interpreted as calculating $P(w|c) = \sum_{t \in c} P(w|t)P(t|c)$, where $P(w|t)$ is the probability that word w is semantically translated to word t. However, TA only considers the word-to-word relationships, and some training data is required for training the model [12]. In order to apply the TA in TC, we incorporate the relevance information into the model translation framework. Therefore, the probability that a word w is sampled from a class c is not estimated directly based on the frequency of the word occurring in the class, but based on the frequency of the word in the relevant documents as well as the likelihood that the class generates the respective documents:

$$P(w|c) = \sum_{d' \in c} P(w|d')P(d'|c). \quad (5)$$

$P(w|d')$ is the probability of translating word w into document d'. The use of $P(w|d')$ allows us to score a class by counting the matches between a document word and semantically related documents in the class. $P(d'|c)$, which can be computed via the maximum likelihood estimate, reflects the distribution of training documents in the class c. The formulation of (5) is equivalent to the TM approach for IR proposed by Berger and Lafferty [2]. By replacing $P(w|c)$ in (3) with the one computed by (5), we have a new class unigram LM as follows,

$$P(w|\mathscr{M}_C) = \lambda \sum_{d' \in c} P(w|d')P(d'|c) + (1-\lambda)P(w|\mathscr{M}_B), \quad (6)$$

The model in (6) is obviously more computationally intensive than the model in (3). Therefore, we need to build a global term relevance dependency model for all classes and the word probability distribution for each class beforehand.

In order to explore the term relationship within this framework, we make use of RM [10] which has been

shown to be effective for IR in the past. The score of seeing a word in a given relevance document d' in the class (denoted by $\mathscr{R}_c$) can be estimated by (4). Recall that $\mathbf{d} = \{w_1, \cdots, w_{n_d}\}$. A relevance-document d' is generated in the source class. Therefore, we have

$$P(w|\mathbf{d}') = \frac{P(w,\mathbf{d}')}{P(\mathbf{d}')} = \frac{P(w,w_1,\cdots,w_{n_{d'}})}{P(w_1,\cdots,w_{n_{d'}})}. \quad (7)$$

The joint probability in the numerator is assumed to be

$$\begin{aligned} &P(w,w_1,\cdots,w_{n_{d'}}) \\ &= \sum_{\tilde{\mathbf{d}}\in\mathscr{R}_c} P(\mathscr{M}_{\tilde{\mathbf{d}}})P(w,w_1,\cdots,w_{n_{d'}}|\mathscr{M}_{\tilde{\mathbf{d}}}) \\ &\approx \sum_{\tilde{\mathbf{d}}\in\mathscr{R}_c} P(\mathscr{M}_{\tilde{\mathbf{d}}})P(w|\mathscr{M}_{\tilde{\mathbf{d}}})\prod_{i=1}^{n_{d'}} P(w_i|\mathscr{M}_{\tilde{\mathbf{d}}}) \end{aligned}, \quad (8)$$

where $\mathscr{M}_{\tilde{\mathbf{d}}}$ is the LM estimated from document $\tilde{d}$. More data should improve performance even further. We build the document model to combine evidence from multiple classes following [9]. The probability of a word w in the relevance-document d' is estimated by using a mixture of collection-specific RM as follows

$$P(w|\hat{d}') = \sum_{c\in C} P(w|d',c)P(c), \quad (9)$$

where C is a set of categories and $P(w|d',c)$ is the probability of a word given the relevance-document and the class collection , which is the RM computed using class c and can be estimated by (7). Following Larenko and Croft's work [10], we set $P(\mathscr{M}_d) = 1/|\mathscr{R}_c|$. Finally, we get the new document model as follows

$$P(w|\hat{d}') = \sum_{c\in C}\frac{P(c)}{|\mathscr{R}_c|}\sum_{\tilde{\mathbf{d}}\in\mathscr{R}_c} P(w|\mathscr{M}_{\tilde{\mathbf{d}}})\prod_{i=1}^{n_{d'}} P(w_i|\mathscr{M}_{\tilde{\mathbf{d}}}). \quad (10)$$

Finally, $P(w|\hat{d}')$ is used to replace $P(w|d')$ in (6). In this paper, the method that combines (10) and (6) is denoted as TRDM.

4. Experiments

4.1. Data sets

We evaluate the proposed TC methods on two standard document collections: 20 Newsgroups[1] (20NG) and Reuters-21578[2] (Reuters). For the 20NG data set, we use the sorted by date version. For each category, we randomly select 60% of the documents for training and the remaining 40% for testing. Following [3], 30 largest categories of the Reuters corpus are chosen for our experiments. The performance of text classification is evaluated in terms of the precision, recall and F-measure. To evaluate the average performance across classes, we examined the micro-averaged score and macro-averaged score [13].

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

[2] http://kdd.ics.uci.edu/databases/reuters21578/

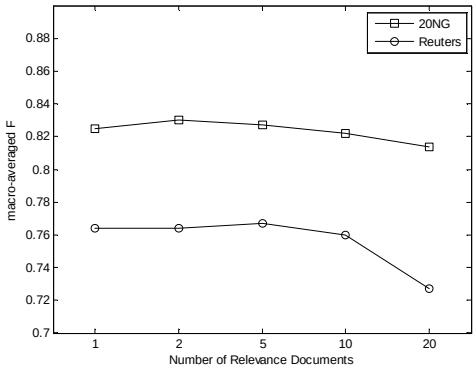

Figure 1. Results of TRDM models using different numbers of relevant documents on the 20NG and Reuters data sets

4.2. Effect of Relevant Document Size

First of all, we present results of our term relevance dependency model (TRDM) in Figure 1. The figure shows the macro-F performance with respect to different relevant document sizes evaluated on the 20NG and Reuters collections. For the 20NG dataset, the TRDM approach obtains the best macro-F of 0.83 when 2 relevant documents are used. For the Reuters dataset, the TRDM reaches the best macro-F of 0.767 at 5 relevant documents. From the figure, we observe that the performance first increases and then drops with the number of relevant documents. The best setup will be used in the following experiments.

4.3. Classification Performance for Different Methods

We compare the proposed method TRDM (the relevance dependency translation model with the naive Bayes classifier) with three existing methods: NBC (the naive Bayes classifier with Laplace smoothing), UN (the naive Bayes classifier with the unigram language model) and TA (the naive Bayes classifier with the term association model). The results of the micro-averaged score of the precision, recall and F-measure for different methods evaluated on the 20NG dataset are shown in Figure 2. In Figure 2, the micro-F is

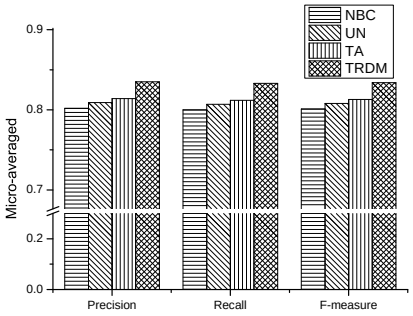

Figure 2. A comparison of micro-averaged scores of precision, recall and F-measure for different methods evaluated on the 20NG data set

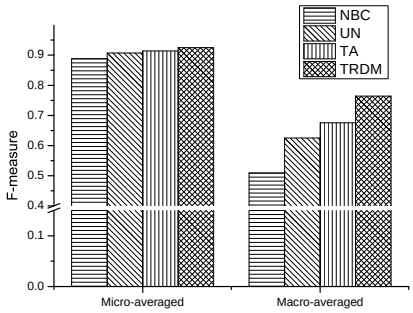

Figure 3. A comparison of micro-averaged and macro-averaged F-measure for different methods evaluated on the Reuters data set.

0.834 for TRDM, which is better than that obtained by NBC (0.801), UN (0.808), and TM (0.813). The relative improvements are 4.12%, 3.22%, and 2.58%, respectively. The improvements of TRDM over NBC and UN are statistically significant according to the t-test. Figure 3 shows the micro-averaged and macro-averaged F-measure of different classification methods evaluated on the Reuters dataset. In this experiment, TRDM obtains a micro-F of 0.925, which is better than that of NBC (0.888), UN (0.907), and TM (0.914). For macro-F, the results are 0.764, 0.509, 0.625, and 0.676. It is obvious that TRDM also achieves a higher macro-F than the other three methods. The improvement in macro-F is more significant than that in micro-F. Because the class labels in the Reuters dataset are highly skewed, micro-F is dominated by the performance of some common categories.

5. Conclusion and Future Work

In this paper, we have proposed a novel term relevance dependency modeling approach for TC. The advantage of incorporating more word relationship information in language modeling has been confirmed by several previous studies. The results of TC experiments evaluated on two datasets demonstrate that the new model outperforms the standard NBC and several other LM-NBC-based text classification methods. In our future work, we will train our model in a dynamic adaptation manner [6] so that it can effectively include the domain knowledge from the newly arrived data.

ACKNOWLEDGMENT

This work was sponsored by Ministry of Economic Affairs, Taiwan, R.O.C. through project No. B3522P1200 conducted by ITRI.

References

[1] J. Bai and J.-Y. Nie. Using language models for text classification. In *Proc. of the AIRS*, 2004.

[2] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proc. of the ACM SIGIR*, pages 222–229, 1999.

[3] D. Cai, X. He, and J. Han. Document Clustering Using Locality Preserving Indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1624–1637, 2003.

[4] G. Cao, J.-Y. Nie, and J. Bai. Integrating word relationships into language models. In *Proc. of the ACM SIGIR*, pages 298–305, 2005.

[5] K.-Y. Chen and B. Chen. Relevance language modeling for speech recognition. In *Proc. of the ICASSP*, pages 5568–5571, 2011.

[6] J.-T. Chien and M.-S. Wu. Adaptive Bayesian latent semantic analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):198–207, 2008.

[7] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[8] H. Daumé and D. Marcu. Bayesian query-focused summarization. In *Proc. of the ACL*, pages 305–312, 2006.

[9] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proc. of ACM SIGIR*, pages 154–161, 2006.

[10] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of ACM SIGIR*, pages 120–127, 2001.

[11] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *Proc. of the CIKM*, pages 1895–1898, 2009.

[12] X. Wei and W. B. Croft. Modeling term associations for ad-hoc retrieval performance within language modeling framework. In *Proc. of the ECIR*, pages 52–63, 2007.

[13] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization . *Journal of Information Retrieval*, 1(1-2):67–88, 1999.