Short Text Classification Using Very Few Words[*]

Aixin Sun
School of Computer Engineering, Nanyang Technological University, Singapore
axsun@ntu.edu.sg

ABSTRACT

We propose a simple, scalable, and non-parametric approach for short text classification. Leveraging the well studied and scalable Information Retrieval (IR) framework, our approach mimics human labeling process for a piece of short text. It first selects the most representative and topical-indicative words from a given short text as query words, and then searches for a small set of labeled short texts best matching the query words. The predicted category label is the majority vote of the search results. Evaluated on a collection of more than 12K Web snippets, the proposed approach achieves comparable classification accuracy with the baseline Maximum Entropy classifier using as few as 3 query words and top-5 best matching search hits. Among the four query word selection schemes proposed and evaluated in our experiments, term frequency together with clarity gives the best classification accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Keywords

Short Text Classification, Search and Vote

1. INTRODUCTION

We are now dealing with much more short texts. Examples are snippets in search results, tweets, status updates, comments, and reviews from various social platforms. Short texts are in general much shorter, nosier, and sparser, which calls for a revisit of many fundamental text mining techniques including text classfiication.

Short text classification is to assign a piece of short text one or more predefined categories. Most existing approaches try to *enrich* the representation of a short text using additional semantics. The semantics could be derived internally from the short text collection [3], externally from a collection of much longer documents in a similar domain as the short texts [5], or from much larger external sources such as Wikipedia and WordNet [1, 3, 4]. In [1, 4], the classification accuracy is significantly improved by enriching short text feature vector with relevant hidden topics derived from Wikipedia pages using topic model.

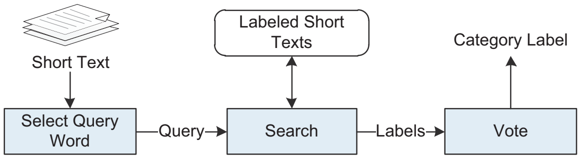


Figure 2: The search and vote framework

In the opposite direction of enriching short text representation, we propose to *trim* a short text representation to get a few most representative words for topical classification. This approach is to mimic the human labeling process. Due to the length of short text (*e.g.,* a search result snippet is usually fewer than 20 words, a tweet has at most 140 characters), one finishes reading a piece of short text at a glance; the category label is assigned mainly based on the few keywords observed from the short text. More specifically, given a short text, we identify one or more words that best represent the short text and formulate a weighted word query using each selected word and its associated weight. Illustrated in Figure 2, this query is submitted to a local search engine for the best matching labeled examples. Analogous to the widely used k-nearest-neighbor (kNN) classifier, the predicted category is the one receiving highest votes from the search results (*i.e.,* labeled short texts).

Leveraging the well studied IR techniques, the proposed approach is extremely scalable. More importantly, the lazy and non-parametric approach well accommodates fast updating of labeled examples, changes to the classification scheme (*e.g.,* adding a new category), as well as giving specific attention to certain groups of labeled examples by changing the scoring function in the search (*e.g.,* a search may favor more recently added labeled examples). Note that, different from kNN classifiers, our approach requires no computation of the nearest neighbors in the labeled data. Instead, we search for the top-H best hits using few query words.

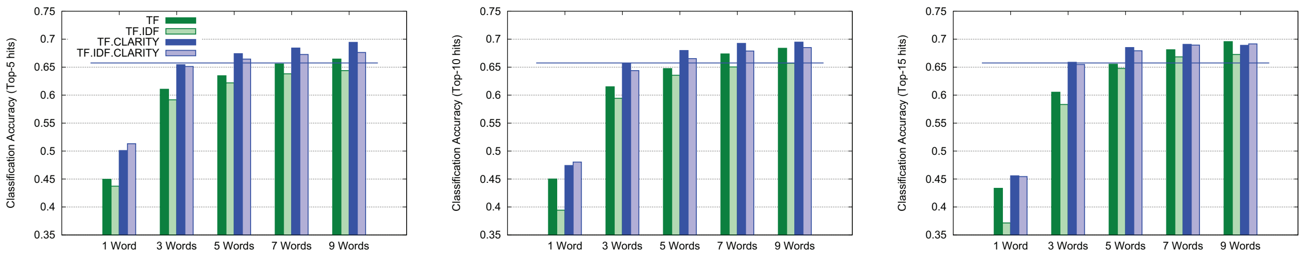
2. SEARCH AND VOTE

Query Word Selection. The most critical step in the proposed approach is the selection of representative words from a short text as query words. Ideally, the query words should be (i) well representing the main content of the short text, and (ii) topically indicative. A word is topically indicative if it is about a specific topic.

With bag-of-words representation, the most widely used TF or $TF.IDF$ weighting scheme is sufficient for the first requirement. For the second requirement, we propose to use $clarity$ to measure the topical-specificity of a word. Proposed for predicting query performance in [2], the clarity score of a query is the Kullback-Leibler (KL) divergence between the query language model (*i.e.,* unigram distribution of word occurrences) and collection language model. The former is inferred from a set of documents best matching a given query and the latter from the entire collection. Given a short text data collection $\mathcal{D}$. Let word w be a candidate word for selection, we use w as a single-term query to retrieve its top-N most

[*] This work was supported by MINDEF-NTU-DIRP/2010/03.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.



(a) Category voting by top-5 hits (b) Category voting by top-10 hits (c) Category voting by top-15 hits

Figure 1: The classification accuracy voted by top-{5, 10, 15} hits respectively using {1, 3, 5, 7, 9} words as query

Table 1: Number of training/test snippets in the 8 categories.

Category	#Train	#Test	Category	#Train	#Test
Business	1200	300	Culture-Arts-Ent	1880	330
Computer	1200	300	Education-Science	2360	300
Health	880	300	Engineering	220	150
Sports	1120	300	Politics-Society	1200	300

relevant documents ($N = 20$ in our experiments[1]), denoted by Q_w. The clarity score of word w is computed using the equation below, where $\mathcal{V}$ denotes the vocabulary.

$$Clarity(w) = \sum_{w' \in \mathcal{V}} P(w'|Q_w) \log \frac{P(w'|Q_w)}{P(w'|\mathcal{D})}$$

Intuitively, if w is specific to a topic, then the documents matching w share a common topic indicated by a few words with very high probabilities of occurrences against their probabilities in the entire collection. On the other hand, if the documents matching w do not share a common topic, then this set of documents is analogous to a random sample from the document collection, with similar distributions of word occurrences.

Scoring and Voting. In literature, many scoring functions for relevance ranking have been proposed for document search. To search for the most relevant labeled examples for a given query, we use the default scoring function implemented in Lucene[2]. Note that, Lucene allows query word boosting, *i.e.,* to assign weight to each query word in a multi-term query. The score computed in query word selection step is used to weigh each selected query word. After query execution, each category label of the returned documents serves as an evidence that supports the likelihood of the querying document belonging to that category. Majority voting shows better results than its weighted counterpart in our experiments.

3. EXPERIMENTS

We conducted experiments on the Web snippet dataset that has been used in [1, 4]. The dataset consists of 10,060 train and 2,280 test snippets from 8 categories, shown in Table 1. On average, each snippet has 18.07 words after indexing by Lucene.

We evaluate four schemes to select query words: TF, $TF.IDF$, $TF.CLARITY$, and $TF.IDF.CLARITY$. The latter two are the product of TF and $TF.IDF$ respectively with clarity score. For each scheme, the top {1, 3, 5, 7, 9} words with highest scores are selected as query words. The category label of a test snippet is voted by the top-{5, 10, 15} hits respectively. The classification accuracies are reported in Figures 1(a), 1(b), and 1(c). Note that, as each snippet has exactly one category label, the reported accuracy is the same as Micro-Precision/Recall/F_1. For comparison, the horizontal line plotted in the three figures indicates the accuracy of 0.6575 achieved by Maximum Entropy (MaxEnt) classifier as baseline [4].

From the results, we make the following observations. First, more words (from 1 word to 5 words) in query generally lead to better classification accuracy. The improvement becomes very minor when more than 5 words are used in query. Observe that using 3 words, the classification accuracy voted by top-5 hits matches MaxEnt, an advanced model that has shown good classification accuracies in many applications. Second, using 1 word in query, more hits lead to poorer results. When more than 3 words are selected in query, more hits yield slightly better accuracy. Third, among all word selection schemes, $TF.CLARITY$ outperforms the others in most runs when 3 or more words are used in query. For one word query, $TF.IDF.CLARITY$ is the best scheme.

We note that the best classification accuracy achieved is better than MaxEnt, but poorer than the results reported on the same dataset using enriched short text representation [1]. The enriched representation is a combination of *word feature* from the short text, and *topic feature* from a topic space derived from Wikipedia. We argue that our approach is directly applicable to any enriched representation because each enriched feature (*e.g.,* a topic feature) can be selected as a query word. That is, our approach complements the approach enriching text representation.

4. SUMMARY AND FUTURE WORK

We propose a simple, scalable, and non-parametric approach for short text classification. Using the simple $TF.CLARITY$ to select as few as 3 words, the default Lucene scoring function, and majority voting from 5 search hits, our approach achieves comparable classification accuracy with MaxEnt classifier. We also note that our approach complements the research on enriching short text representation. Moreover, the non-parametric approach offers much more flexibility in handling updates in labeled examples and classification schemes.

The proposed approach has great potential to achieve much better results with more research on (i) query word or phrase selection, (ii) relevance ranking techniques specifically for short texts, and (iii) short text representation enrichment.

5. REFERENCES

[1] M. Chen, X. Jin, and D. Shen. Short text classification improved by learning multi-granularity topics. In *IJCAI*, pages 1776–1781, 2011.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.

[3] X. Hu, N. Sun, C. Zhang, and T.-S. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *CIKM*, pages 919–928, 2009.

[4] X. H. Phan, M. L. Nguyen, and S. Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW*, pages 91–100, 2008.

[5] S. Zelikovitz, W. W. Cohen, and H. Hirsh. Extending WHIRL with background knowledge for improved text classification. *Information Retrieval*, 10(1):35–67, 2007.

[1] If a word appears in fewer than 20 documents, its clarity score is set to 1.0.

[2] `http://lucene.apache.org/core/`