Short Text Classification Improved by Learning Multi-Granularity Topics

Mengen Chen[‡] **Xiaoming Jin**[‡] **Dou Shen**[+]
[‡]Key Laboratory for Information System Security, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology
School of Software, Tsinghua University, Beijing, China
[+]Buzzlabs, Bellevue WA, 98007
cme0410@gmail.com xmjin@tsinghua.edu.cn dou@buzzlabs.com

Abstract

Understanding the rapidly growing short text is very important. Short text is different from traditional documents in its shortness and sparsity, which hinders the application of conventional machine learning and text mining algorithms. Two major approaches have been exploited to enrich the representation of short text. One is to fetch contextual information of a short text to directly add more text; the other is to derive latent topics from existing large corpus, which are used as features to enrich the representation of short text. The latter approach is elegant and efficient in most cases. The major trend along this direction is to derive latent topics of certain granularity through well-known topic models such as latent Dirichlet allocation (LDA). However, topics of certain granularity are usually not sufficient to set up effective feature spaces. In this paper, we move forward along this direction by proposing an method to leverage topics at multiple granularity, which can model the short text more precisely. Taking short text classification as an example, we compared our proposed method with the state-of-the-art baseline over one open data set. Our method reduced the classification error by 20.25 % and 16.68 % respectively on two classifiers.

1 Introduction

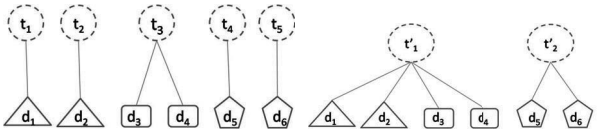
With the rapid development of the Internet, Web users and Web service are generating more and more short text, including tweets, search snippets, product reviews and so on. There is an urgent demand to understand the short text. For example a good understanding of tweets can help advertisers put relevant advertisements along the tweets, which makes revenue without hurting user experience. However, short text is very different from traditional documents, principally in its shortness. As a result, short text tends to be ambiguous without enough contextual information. Therefore, conventional machine learning and text mining algorithms cannot apply to short text directly. Each short text does not have enough content, or words specifically, while a set of text tends to span over a wide range of words. This makes it extremely hard to build a feature space directly for clustering and classification.

Existing work in the literature tries to address the aforementioned challenges from two directions. The first one is to fetch external text to expand the short text (e.g., [Sahami and Heilman, 2006]). Another direction is to discover a set of explicit or implicit topics and then connect the short text through these topics. For examples, in [Hu *et al.*, 2009a], the authors exploit the user-defined categories and concepts (e.g., Wikipedia[1]), while in [Phan *et al.*, 2008], the authors derive a set of hidden topics through topic model LDA [Blei *et al.*, 2003b] from one large existing Web corpus.

Clearly, fetching search snippets from search engines is not an ideal solution for some applications, since it is very time consuming and heavily depending on the quality of search engines. Using explicit pre-defined topics/taxonomy relaxes the dependence on search engines. But its adaptability can be an issue since the pre-defined topics and taxonomy may not be available for certain applications and in certain languages. For example, we can easily find well organized corpora like ODP[2]in English, however, it is hard to find comparable corpus in small languages like Akan. The solutions based on latent topics are preferable in that we can easily get a relatively large corpus relevant to the problem under consideration and then automatically generate the latent topics. Also, these solutions prove to achieve satisfying results in different problems [Phan *et al.*, 2008]. Therefore, we are putting forward new solutions along this direction in this paper.

So far, the latent-topic based solutions use topics at a single level. With a pre-defined number of topics, the well-known topic models like LDA [Blei *et al.*, 2003b] can extract the latent topics of that number from a given text corpus. Intuitively, if the number of topics is large, the discovered topics tend to have fine granularity. On the other side, if the number is small, the discovered topics tend to have coarse granularity. For a certain classification problem over a set of short text, it might be impossible to figure out the right number of latent topics and such number may not even exist. Figure 1 illustrates the situation. Let us assume there are 6 short texts, belonging to 3 categories, denoted by triangle, square and pentagon respectively. As a reasonable assumption, there are no overlapping terms among the short text, which makes it extremely hard to group them directly. According to the

[1]http://www.wikipedia.org
[2]http://www.dmoz.org/



(a) Fine granularity (b) Coarse granularity

Figure 1: Illustration of our motivation

latent-topic based solution, we need to derive hidden topics as the bridge. In Figure 1(a), 5 latent topics are generated. Here, we can see that the topics are of too fine granularity, such that d_1 and d_2 cannot connect. On the opposite side, in Figure 1(b), 2 latent topics are generated. This time, the topics are of too coarse granularity such that short texts of triangle and square mix together. Thus multi-granularity topic space is probably better than single-granularity one.

We validated our solution over an open data set [Phan *et al.*, 2008], on which two mainstream text categorization methods, Support Vector Machine (SVM) and Maximum Entropy (MaxEnt), were used as the classifiers. Our method reduced the classification error by 20.25 % and 16.68 % when using the two classifiers respectively. This validated the superiority of our method using multi-granularity topics over the state-of-the-art, single granularity, baseline.

The contribution of this paper is non-trivial in that we put forward a solution of exploiting topics of multi-granularity and present a systematic way to seamlessly integrate the topics and produce discriminative features for short text classification. This is the first time of leveraging multi-granularity topics for classification as far as we know.

The rest of this paper is organized as follows. We first review relevant works. After that, we present our approach based on multi-granularity topics. Finally we systematically validate our method over one open data set, which show that our method consistently outperforms the baseline.

2 Related Work

With the popularity of short text, some interesting work has appeared in the literature to study short text representation issues to facilitate short text clustering and classification. Overall, the work can be grouped into two directions: *Web search-based* methods and *Taxonomy/Topics based* methods.

2.1 Web Search-Based Methods

Getting the contexts for short text can provide more information to understand the short text. Intuitively, we can collect a large text corpus and then check under what context a certain short text usually shows up. Thereafter, we can leverage the collected context to enrich the short text. For example, we can make use of the search engines by treating short text as a query and submitting it to a search engine. The search results, presented in terms of Web page titles, URLs, summaries of Web pages (also called snippets), are widely used to enrich short text. In [Bollegala *et al.*, 2007], semantic similarity between words can be obtained by leveraging page counts and text snippets returned by search engine. Similarly titles and snippets are combined to enrich the original short text, which gains significant improvement on similarity measurement in [Yih and Meek, 2007; Sahami and Heilman, 2006]. In [Shen *et al.*, 2006], the authors use titles and snippets to expand the web queries and achieve much higher classification accuracy on query classification task compared to using queries alone.

2.2 Taxonomy/Topics-Based Methods

The Web search based methods have an efficiency problem when the set of short text under consideration is huge. Also, the performance of these methods heavily depends on the search engine's quality. To address these issues, researchers turn to use explicit taxonomy/concepts or implicit topics. These corpora (e.g., Wikipedia, ODP) have rich predefined taxonomy and human labelers assign thousands of Web pages to each node in the taxonomy. Such information can greatly enrich the short text. In [Hu *et al.*, 2009b], the authors use Wikipedia concept and category information to enrich document representation to address semantic information loss caused by bag-of-words representation. Similarly, Wikipedia is used in [Hu *et al.*, 2008] to build a concept thesaurus to enhance traditional content similarity measurement. A weighted vector of Wikipedia-based concepts is also used for relatedness estimation of short text in [Gabrilovich and Markovitch, 2007]. One possible shortcoming of using pre-defined taxonomy in the above ways is the lack of adaptability. Though it is easy to collect a huge text corpus, the taxonomy may not be proper for certain classification tasks. What's more, for some languages, we may not be able to obtain a corpus with pre-defined taxonomy. To overcome this shortcoming, in [Phan *et al.*, 2008] the authors derive latent topics from a set of texts from Wikipedia and then use the topics as appended features to expand the short text. Experiments show that their method using the discovered latent topics achieves the state-of-the-art performance. Note that the discovered topics in their method are of a single level, which may limit their approach.

3 Problem Specification

Short text is characterized by 1) shortness in the text length, and 2) sparsity in the terms presented, which results in the difficulty in managing and analyzing them based on the bag-of-words representation only. Short texts can be found everywhere, such as search snippets, product reviews etc. *Short text classification* is to classify the short texts and assign the short text a label from predefined taxonomy.

In this paper, we take web search snippets as a representative of short text as in [Hu *et al.*, 2009a; Phan *et al.*, 2008]. These search snippets are collected during web search transaction by using various phrases of different domains as issued queries (more details are provided in Section 5.1). Specifically, a classifier can be built for assigning web search snippets labels such as *Business*, *Computer* etc. Based on this classifier, search results can be organized effectively, and as a result web users can be navigated to the needed information.

One possible solution to handling sparsity of short text is to expand short texts by appending some words to the text based on *the semantic relatedness between words*[Hotho

et al., 2003; Strube and Ponzetto, 2006; Gabrilovich and Markovitch, 2007]. Such methods examine individual word only, without considering the context of the word, i.e., word co-occurrence within the short text, which usually be very meaningful for short text classification. To address this issue, a reasonable solution is to map the short text itself to topic space obtained from external corpus. But how to obtain the effective topic space? We put forward a multi-granularity framework in the following section.

4 Multi-Granularity Framework

Different from the single-granularity framework [Phan *et al.*, 2008], we enrich short text through topics of multi-granularity. The overall framework is shown in Figure. 2, with details presented in the following Sections.

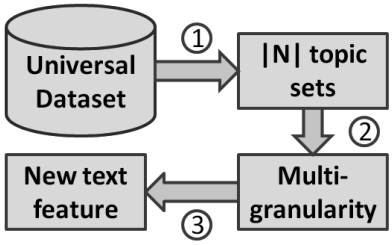


Figure 2: Multi-granularity Framework

1. Given a set of numbers empirically denoted by N, with the size $|N|$, we run LDA over the universal dataset to generate the topics with respect to each item in N. As a result, we obtain $|N|$ different sets of topics.
2. Choose a subset of all the generated topics T automatically to form the topic space of multiple granularity.
3. Combine topic feature, obtained from multi-granularity topic space, with word feature to form new text feature, based on which we build classifiers.

4.1 Generate $|N|$ Topic Sets

In our framework one of the hard problems is how to generate topic features of multi-granularity. There are different ways to obtain topics of multi-granularity. For example, we can use some hierarchical topic models [Blei *et al.*, 2003a] to derive topics with different granularities. In this paper, we use a straightforward way to obtain topics of multi-granularity. Intuitively the pre-defined number of topics will decide the granularity of the extracted topics. Therefore, we can use several different topic numbers to generate several sets of topics at different granularities. Clearly, this is not a principled approach to produce topics of multi-granularity. However, this will be good enough to validate whether topics of multi-granularity will help short text classification. Therefore we empirically choose a set of topic numbers N, and then run LDA over the universal dataset for each topic number in N.

4.2 Generate Multi-granularity Topic Space

In multi-granularity framework, one of challenges is to select the best subset from all the generated topic sets. Intuitively the quality of topics with a certain granularity depends on two aspects: one is their capability in helping discriminate short text with different class labels; the other is the distance between them and topics with other granularities. Considering the ambiguity of short texts, it is expected that the distance between selected topic sets is large enough.

Based on these two intuitions, we propose a selection algorithm for generating multi-granularity topic space. We denote $|N|$ different sets of topics as $T = \{T_1, T_2, \ldots, T_{|N|}\}$, where each entry T_i is a topic set in the form of $T_i = \{z_{i1}, z_{i2}, \ldots, z_{iK_i}\}$ where K_i is the number of topics and z_{ij} denotes a topic, which is a probability distribution over words. Thus the purpose of proposed algorithm is to select a subset of T, which is discriminative in terms of classification. $\mathcal{W} = \{w(T_1), w(T_2), \ldots, w(T_{|N|})\}$ is the weight vector, where $w(T_i)$ is the weight indicating the importance of topic set T_i. To get $\mathcal{W}$, a novel algorithm based on the key idea of Relief [Kononenko, 1994] is proposed. Specifically, for each short text x in data set D, the algorithm searches through D to find two nearest neighbors: one is from the same class ($nh(x)$ for short) and the other from different class

Table 1: **Algorithm 1**:Weighting Topic Sets

procedure WEISINGLETOPIC($D, Label$)
  Input: m training vectors of n attributes ($n = \sum_{i=1}^{|N|} K_i$) and the class label for each vector.
  Output: weight vector $\mathcal{W} = w(T_1), \ldots, w(T_{|N|})$.
  $w(T_i) \leftarrow 0$ for all topic set
  for each point $x \in D$ **do**
    Find $nm(x)$ and $nh(x)$
    for $i \leftarrow 1$ to $|N|$ **do**
      Update $w(T_i)$ using (1)
    end for
  end for
  Return $\mathcal{W}$
end procedure

Table 2: **Algorithm 2**:Generating Multiple Granularity topics

procedure GENMULTOPIC($Dis, \mathcal{W}, H, T$)
  $\mathcal{O} \leftarrow \emptyset$ ▷ a set of the selected topic set
  $T^{(1)} = \arg\max_{T_i \in T} w(T_i)$
  $\mathcal{O} = \mathcal{O} \cup \{T^{(1)}\}$
  $T = T - \{T^{(1)}\}$
  $Num \leftarrow 1$ ▷ the number of selected topic set
  while $Num < H$ **do**
    for $T_i \in T$ **do**
      Compute $Score(T_i)$ using (2)
    end for
    $Num = Num + 1$
    $T^{(Num)} = \arg\max_{T_i \in T} Score(T_i)$
    $\mathcal{O} = \mathcal{O} \cup \{T^{(Num)}\}$
    $T = T - \{T^{(Num)}\}$
  end while
  Return $\mathcal{O}$
end procedure

($nm(x)$ for short). Then the weight $w(T_i)$ is updated by

$$w(T_i) = w(T_i) + d(x_{T_i}, nm(x)_{T_i}) - d(x_{T_i}, nh(x)_{T_i}) \quad (1)$$

Where x_{T_i} is the topic distribution of x over T_i, inferred by Gibbs sampling in LDA. Specially $d(x_{T_i}, nm(x)_{T_i}) = \frac{1}{2}\sum_{z_k \in T_i}[p(z_k|x)\log p(z_k|x)/p(z_k|nm(x)) + p(z_k|nm(x))\log p(z_k|nm(x))/p(z_k|x)]$. So is the value $d(x_{T_i}, nh(x)_{T_i})$. Table 1 depicts the overall process.

To make multi-granularity topics selected more discriminative and less redundant, we tend to select the topic set which is most different from all the selected topic sets. Conventionally, KL-divergence is used to measure the distance between two probability distributions. In our work, distance between different topic sets $Dis(T_i, T_j)$ is set as the average KL-divergence value. In order to make sure $Dis(T_i, T_j) = Dis(T_j, T_i)$, we deploy $Dis(T_i, T_j) = \sum_{z_i \in T_i, z_j \in T_j} \mathcal{D}(p(w|z_i), p(w|z_j))/(K_i \times K_j)$ as the formula to compute the distance between topic sets T_i and T_j. Where $\mathcal{D}(p(w|z_i), p(w|z_j)) = \frac{1}{2} \times [D(p(w|z_i)\|p(w|z_j)) + D(p(w|z_j)\|p(w|z_i))]$ and $D(p(w|z_i)\|p(w|z_j))$ is the KL-divergence between distributions $p(w|z_i)$ and $p(w|z_j)$.

Let $\mathcal{O}$ represent the set of selected topic sets. We firstly obtain the score for each candidate topic set and then select topic set with highest score as a member of multi-granularity topics. As discussed before, whether a topic set of certain granularity T_i is chosen or not depends on its significance reflected by weight $w(T_i)$ and its distance from the selected topic set(s). Therefore we assign a score to the candidate topic set according to (2).

$$Score(T_i) = w(T_i) + \sum_{T^{(j)} \in \mathcal{O}} Dis(T_i, T^{(j)}) \quad (2)$$

The algorithm for generating multi-granularity topics is summarized in Table 2. This algorithm takes as input inter-topic distances as well as weight of topic sets, outputting the set of selected multi-granularity topics. Here we employ greedy strategy to select one topic set at each time (corresponding to the algorithm from 7 to 15). The topic set with the highest score is chosen at every step until the number of granularity exceeds the specified number H, whose impact will be investigated in experimental part.

4.3 Form New Feature for Short Text

Given a short text m and the H sets of topics $\mathcal{O}$, topic distribution of single granularity for m is inferred by Gibbs sampling [Phan *et al.*, 2008]. We denote the topic distribution against the i^{th} set of topics in $\mathcal{O}$ as $\vec{\Theta}_{m,i} = \{\vartheta_{m,1}, \cdots, \vartheta_{m,K_i}\}$. K_i is number of topics in the i^{th} set.

After we obtain topic distributions $\vec{\Theta}_{m,i} (i = 1..H)$, how to combine them? Here we adopt a simple but powerful way to combine them to form topic feature for short text. $\vec{F}_m = \left[\mu_1\vec{\Theta}_{m,1}, \ldots, \mu_i\vec{\Theta}_{m,i}, \ldots, \mu_H\vec{\Theta}_{m,H}\right]$. Note that μ_i denotes the weight for the i^{th} set of topics, which has similar meaning with $w(T_i)$ gained from *Algorithm 1*. Thus an automatic assignment of μ_i by means of $w(T_i)$ is deployed.

$$\mu_i = w(T_i)/w \quad (3)$$

where $w = \min_{T_i \in \mathcal{O}} w(T_i)$ represents weight of least important selected topic set. Finally new feature for short text m is obtained by appending the topic feature $\vec{F}_m$ to $\vec{w}_m$ as follows:

$$\vec{\Omega}_{\underline{m}} = \left[\vec{w}_{\underline{m}}, \beta \cdot \vec{F}_{\underline{m}}\right] \quad . \quad (4)$$

where β is a user specified parameter indicating the importance of topic features. With the new features, we can train classifiers in traditional ways. We choose SVM and MaxEnt as the classifiers for the comparison with the experiments in [Phan *et al.*, 2008], further demonstrating the advantages of multi-granularity topics over single granularity empirically.

5 Experiment

In this section, we compare our method with the state-of-the-art baseline over web-snippet data set. We validate our proposed algorithms effective. We also study the performance of our method under different parameter settings.

5.1 Data Sets

Search-snippet data set, collected by Xuan-Hieu Phan[Phan *et al.*, 2008], consists of two subsets, named *short text* and *Corpus*. We want to classify *short text*, drawing support from latent topics extracted from the *Corpus*.

Corpus The authors of [Phan *et al.*, 2008] prepared topic-oriented keywords. Take topic *Business* for example, some of the keywords are *advertising, e-commerce, finance*. For each keyword, the corresponding Wikipedia page and relevant pages by following outgoing hyperlinks were crawled by JWikiDocs.[3]

Search Snippets Search snippets consists of three parts: a URL, a short title and a short text description. The search snippets were selected from the results of web search transaction using predefined phrases of different domains. For each query phrase put into Google search engine, the top 20 or 30 ranked web search snippets were collected. Then the class label of the collected search snippets was assigned as the same as that of the issued phrase. Some basic statistics of both the corpus and the search snippets are summarized in Table 4.

5.2 Implementation

In the experiments, firstly, $|N|$ sets of topics were extracted from the corpus using LDA and then multi-granularity topics were selected according to our proposed algorithm. Secondly, topic features were constructed and combined with the original word features. Thirdly, the new features of training and test short text were used for training models and classification respectively. Finally, classification accuracy (the ratio of correctly classified data over all the test data) was used to measure the classification performance. Our method was run with the feature generating schema, i.e., linear combination of topics of different granularity (*Multi-L* for short).

5.3 Experimental Results

Firstly we drew 100 snippets for each class randomly from training data, as auxiliary for computing the weight of topic

[3] http://jwepro.sourceforge.net

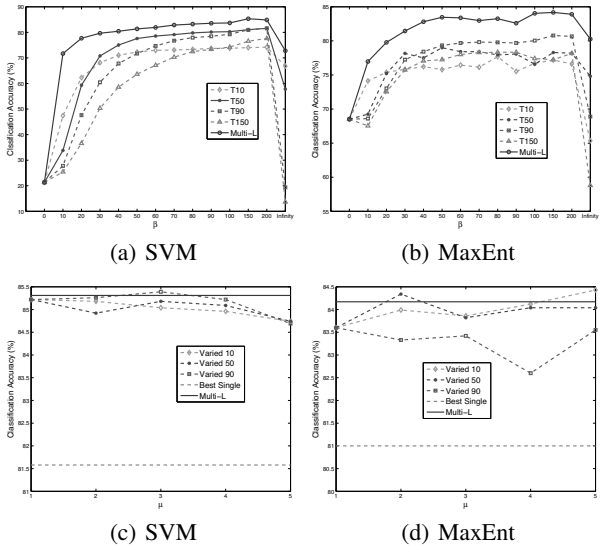
Table 3: Classification performance of SVM and MaxEnt based on both single and multi-granularity topics

Methods	T10	T30	T50	T70	T90	T120	T150	Multi-L
SVM	74.21	81.27	81.58	80.18	**81.58**	80.71	77.63	*85.31*
MaxEnt	77.68	**81.00**	79.00	80.39	80.79	78.95	78.33	*84.17*

Table 4: Statistics of crawled Wikipedia and search snippets

Raw Data: 3.5GB; $|Docs| = 461177$
Preprocess: removing duplicate documents, HTML tags,navigation links, stop and rare ($threshold = 30$) words
Final Data: 240MB; $|Docs| = 71986$; $|Vocabulary| = 60649$

Domain	# Train data	# Test data
Business	1200	300
Computer	1200	300
Culture-Arts-Ent	1880	330
Education-Science	2360	300
Engineering	220	150
Health	880	300
Politics-Society	1200	300
Sports	1120	300
Total	10060	2280



(a) SVM (b) MaxEnt

(c) SVM (d) MaxEnt

Figure 3: Accuracy of SVM and MaxEnt on various β and μ respectively

set $\mathcal{W} = \{w(T_1), w(T_2), \ldots, w(T_{|N|})\}$. The remaining training data was used to train the classifiers. Secondly we set parameters as following. β varied from 0 to 200. N was set as $N = \{10, 30, 50, 70, 90, 120, 150\}$ empirically. The number of Gibbs sampling in LDA equalled to 400. We set $H = 3$, the number of granularity of topics.

$\mathcal{O} = \{T_1(10), T_2(50), T_3(90)\}$ were selected to construct the multi-granularity topics after running *Algorithm 1* and *Algorithm 2*. $T_1(10)$ meant the selected topic set having 10 topics. For the parameters μ_i to join the topic features of different granularities, according to (3), we get $\mu_1^{(10)} = 2.41, \mu_2^{(50)} = 1.08, \mu_3^{(90)} = 1$. The results of classification of test data were shown in Table 3.

The results in Table 3 were the best for each method while β varied from 0 to 200. For example, $T50$ meant only 50 topics of single granularity were used. As we can see in Table 3, for both SVM and MaxEnt, our approach archived remarkable improvement in classification accuracy over any single granularity topics. Specifically, word-only approach (corresponding to $\beta = 0$ in Figure 3(a) and Figure 3(b)) performed very poor (21.32 % and 68.51 % for SVM and MaxEnt respectively). The performance was improved by introducing single-granularity topic features. When multi-granularity topics were used, more improvement in classification accuracy was achieved. Particularly, for SVM, *Multi-L* gained impressive improvement over single-granularity method, from 81.58% to 85.31% (reduced error by 20.25 %). For MaxEnt, *Multi-L*also has significant improvement, from 81.00 % to 84.17 % (reduced error by 16.68 %). Thus for classifying short text, the proposed multi-granularity framework could improve the classification performance remarkably and was superior to both word-only methods and single-granularity methods. Furthermore the improvement consistently showed the robustness of selected multi-granularity topics, which indicated the effectiveness and validation of our proposed algorithms for generating multi-granularity topics.

It was significant to notice that our method was always superior to $T150$ (single granularity framework with topic number 150). Therefore the superiority of our method cannot be interpreted by introducing more topics. Instead, the reason was the capability of our method in retrieving more useful information by exploring the multi-granularity topics. Another point needed to be aware was that when we only use topic feature (corresponding to $\beta = \infty$, i.e., neglecting the word feature), the classification accuracy was not the best. But we again found that our method had better results.

Note that the accuracy numbers for single-granularity methods were a little different from those reported in [Phan *et al.*, 2008]. This was due to the difference in combining topics features with word features. We did not exactly repeat the method in [Phan *et al.*, 2008] because we could not get their exact interval settings. However, the difference was tiny and we observed the exactly same trend, which supported the validity of our experiments.

In order to examine the effectiveness of automatic assignment of μ_i in (3), we enumerate some settings for μ. The results are illustrated in Figure 3. For each curve, the i^{th} granularity of topic varied and the others kept $\mu_j = 1 (j \neq i)$. Taking the red curve in Figure 3(d) for example, we kept $\mu = 1$ for $T = 10$ and $T = 50$ while varied the μ for $T = 90$

Table 5: Sensitivity of multi-granularity with respect to H

Classifier	1050	1090	5090	105090150	105090
SVM	84.12	83.68	84.34	84.43	85.22
MaxEnt	83.29	83.60	82.11	83.03	83.60

from 1 to 5. The curve, named after *Best Single*, was the best classification performance of single granularity topics. *Multi-L* was our approach with automatic assignment of μ_i. The classification performance of multi-granularity topics was always better than single granularity topics, whatever values μ took. What's more our approach (*Multi-L*) was comparable to the best one, proving the usefulness of our proposed method, which assigned weight automatically.

5.4 Sensitivity Analysis.

There is one important parameter, the number of granularity of topics H. We examined it experimentally to show the flexibility and robustness of our multi-granularity framework. The influence of different choices of multi-granularity topics was drawn in Table 5. 1050 meant the linear combination of 10 and 50 topics. So did the others. Single meant the best classification under single granularity framework. From Table 5, we found our multi-granularity framework is better than single whatever values H took, which demonstrated our method was flexible. On the other hand, considering too many granularities may introduce some noises as classification accuracy of 105090150 decreased compared with 105090 both on SVM and MaxEnt.

6 Conclusion

In this paper, we put forward a new method for short text classification. In order to handle its shortness, various ways have been tried to enrich short text to get more features, including using search snippets or implicit/explicit topics. These methods solve the problem to certain extent, but still leave much space for improvement. In this paper, we propose to use multi-granularity topics to generate features for short text. We compare our proposed method against the state-of-the-art baseline through two types of classifiers over one open data set. The experimental results show that our method can significantly reduce the classification errors by 20.25% and 16.68% over these two classifiers respectively.

Though we see the advantage of our proposed method clearly in this paper, we can further improve it by exploiting advanced approaches to generate topics of multiple granularities, such as using hierarchical topic models[Blei *et al.*, 2003a]. Also, we will try more methods to construct features based on the topics and merge them with the bag-of-word features.

Acknowledgments

The work was supported by National Natural Science Foundation of China (90924003, 60973103) and HGJ National Key Project (2010ZX01042-002-002).

References

[Blei *et al.*, 2003a] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of NIPS*, pages 17–24, 2003.

[Blei *et al.*, 2003b] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Mach. Learn. Res.*, 3:993–1022, 2003.

[Bollegala *et al.*, 2007] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Measuring semantic similarity between words using web search engines. In *Proceedings of WWW*, pages 757–766, 2007.

[Gabrilovich and Markovitch, 2007] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–1611, 2007.

[Hotho *et al.*, 2003] A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In *Proceedings of the Semantic Web Workshop of SIGIR*, 2003.

[Hu *et al.*, 2008] Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of SIGIR*, pages 179–186, 2008.

[Hu *et al.*, 2009a] Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceeding of CIKM*, pages 919–928, 2009.

[Hu *et al.*, 2009b] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of KDD*, pages 389–396, 2009.

[Kononenko, 1994] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proceedings of ECML*, pages 171–182, 1994.

[Phan *et al.*, 2008] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceeding of WWW*, pages 91–100, 2008.

[Sahami and Heilman, 2006] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of WWW*, pages 377–386, 2006.

[Shen *et al.*, 2006] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *Journal of ACM Trans. Inf. Syst.*, 24:320–352, 2006.

[Strube and Ponzetto, 2006] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *proceedings of the 21st national conference on Artificial intelligence*, pages 1419–1424, 2006.

[Yih and Meek, 2007] Wen-Tau Yih and Christopher Meek. Improving similarity measures for short segments of text. In *Proceedings of the 22nd national conference on Artificial intelligence*, pages 1489–1494, 2007.