Labelling Topics in Weibo Using Word Embedding and Graph-based Method

Zhipeng Jin[1] Qiudan Li[1] Can Wang[1] Daniel D. Zeng[1,2] Lei Wang[1]
[1]The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences Beijing 100190, China
[2]Department of Management Information Systems, University of Arizona ,Tucson, Arizona, USA
e-mail: {jinzhipeng2013, qiudan.li, wangcan2015, dajun.zeng, l.wang }@ia.ac.cn

***Abstract*—Nowadays, in China, Weibo is becoming an increasingly popular way for people to know what is happening in the world. Labelling topics is of much importance for better understanding the semantics of topics. Existing works mainly focus on deriving candidate labels by exploring the use of external knowledge, which may be more appropriate for well formatted and static documents. Recently, it has been a new trend to generate labels for sparse and dynamic microblogging environment using summarization method. The challenges of labelling topics are how to obtain coherent candidate labels and how to rank the labels. In this paper, based on the latest research work in deep learning, we propose a novel and unified model for labelling topics in Weibo, which firstly adopts word embedding and clustering method to learn dense semantic representation of topic words and mine the coherent candidate topic labels, then, generates interpretable labels using a graph-based model. Experimental results show that topics labels discovered by our model not only have high topic coherence, but also are meaningful and interpretable.**

Keywords- Weibo; microblogs; deep learning; labelling topics; graph

I. INTRODUCTION

Weibo, as a popular microblogging service in China, is playing an important role in our daily life. We use Weibo to socialize with friends, to keep up to date with events and to talk about the newly released electronic products[1]. By labelling topics using dense and coherent semantic topic labels, we have opportunities to understand the semantics of topics, further gain insights into the events. Take the iPhone 6 announcement as an example, by learning the dense semantic representation of topic words, we can know that some people talk about the cellphone's technical specifications, others may concern about the ways of purchase. Furthermore, from the interpretable labels such as processor, hardware, camera, chip, memory, we can clearly understand that the underlying semantic of the topic is about technical specifications. Thus, it's necessary to develop a unified topic labelling model to help people understand the semantics of topics in Weibo.

Existing work mainly focuses on deriving candidate labels by exploring the use of external knowledge, which may be more appropriate for well formatted and static documents [2]-[4]. Recently, it has been a new trend to generate labels for sparse and dynamic microblogging environment using summarization method [5]. It proposes to apply summarization algorithms which are independent of external sources to generate topic labels in Twitter and proves its better performance than LDA. The challenges of labelling topics are how to obtain coherent candidate labels and how to rank the labels. To improve the coherence, deep neural networks [6] provides a successful way to measure the syntactic and semantic word similarities by using distributed representations of words.

In this paper, we focus on mining interpretable and coherent topic labels from dynamic Weibo environment. Based on the latest research work in deep learning, we propose a novel and unified model for labelling topics in Weibo, which firstly adopts word embedding to learn dense semantic representation of topic words and then mines the coherent candidate topic labels with clustering method, finally, generates interpretable labels using a graph-based model.

Our contributions are as follows: 1) we unify word embedding and graph-based model to detect interpretable labels for topic labelling; 2) We evaluate the efficacy of the proposed model in Weibo environment.

The rest of this paper is organized as follows: In section II, we discuss relevant studies in the literature. The detailed procedure of our model is presented in Section III. We empirically evaluate our model in Section IV. Section V sums up our study and discusses future research directions.

II. LITERATURE REVIEW

Our work is related to topic labelling and deep learning. In this section, we review the related works.

The most generic topic labelling method used the top n words in a topic distribution learned by LDA as topic labels [7][8]. However, the top terms are not enough for interpreting the coherent meaning of a topic [9]. [9] proposed an unsupervised probabilistic methodology to automatically assign a label to a topic model, which aims to minimize the KL divergence between a given topic and the candidate labels while maximizing the mutual information between these two word distributions. [10] adopted different ranking mechanisms including pointwise mutual information and conditional probabilities to select top-n terms as topic labels.

Some topic labeling methods focus on deriving candidate labels by making use of external knowledge. [2] used the hierarchy obtained from the Google Directory service and the OpenOffice English Thesaurus to derive candidate topic labels for topics induced by LDA. [3] generated label candidates for a topic using top ranking topic terms and titles of Wikipedia articles, a Support Vector Regression (SVR)

2160-1291/16 $31.00 © 2016 IEEE
DOI 10.1109/ICISE.2016.15

model was further built to rank the label candidates. [4] employed a structured data source (DBpedia) and graph centrality measures to obtain semantic concept labels. These approaches may be more appropriate for well formatted and static documents.

As shown in [5], the sparse and dynamic characteristics of microblogging urgently need new topic labelling method. [5] proved that summarization algorithms can be successfully used to label topics. [11] employed PageRank to weigh the words in the graph and score the candidate labels.

The main challenges of labelling topics lie in two aspects: mining coherent candidate labels and ranking the labels.

Recently, deep-learning techniques, have proven successful in various tasks including topic classification, sentiment analysis, question answering and language translation [12]. The characteristics of automatically discovering multiple levels of representations from raw data make deep learning techniques more appropriate for mining coherent topic labels.

Inspired by the latest research work in deep learning, we aims to mine coherent labels by word embedding and rank the labels using graph-based method.

III. METHOD

Figure 1 depicts the proposed methodology of topic labelling. The model firstly learns topic word embeddings, and then groups them into several subtopics to obtain coherent topic labels candidate. Finally, a graph based method is employed to capture meaningful and interpretable words as the labels of the topic.

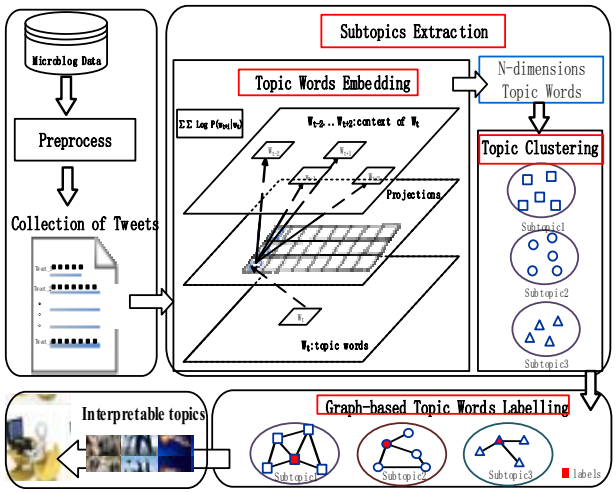

Figure 1. The framework of topic labelling model in Weibo

A. Mining topic labels candidate

Intuitively, we divide the task of mining topic labels candidate into two main parts, topic words embedding learning and subtopics clustering. Recently, inspired from neural-network language modeling, dense vectors are used to represent words to catch the semantic and syntactic similarities between words. Therefore, we employ the Skip-gram model proposed in [13] which is a state-of-the-art word-embedding method. Given a sequence of training words $w_1, w_2, ..., w_N$, the model aims to maximize the average log probability:

$$\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{j=-k}^{k} log\ p(w_{n+j} \mid w_n)\right] \qquad (1)$$

where k is the size of the training window and $p(w_{n+j} \mid w_n)$ denotes the probability of correctly predicting the word w_{n+j}, in which w_n represents the middle word in the training window. In this way, each topic term c_m in the candidate topic words collection $C=\{c_1, c_2, c_3, ..., c_M\}$ which is obtained through tokenization, stop words removal, term frequency and POS filtering is denoted as a low-dimensional vector $c_m = \{c_{m1}, c_{m2}, c_{m3}, ..., c_{md}\}$, where d is the number of dimensionality. Subsequently, we apply Euclidean distance to measure the similarities of candidate topic words. To get more stable subtopics, k-means++, which has been shown to perform better than k-means in speed and accuracy, is used to group the candidate topic words into several subtopics.

B. Graph-based Topic Words Labelling

To obtain better interpretation of subtopics, we develop a TextRank [5] based labelling model. The model first builds an undirected text graph for each subtopic cluster, where the vertices are the candidate topic words and the edges denote the pair words co-occurring in a text window. The importance score of each topic word is defined by:

$$Score(c_i) = (1-\beta) + \beta * \sum_{j \in Edge_\alpha(c_i)} \frac{1}{|Edge_\alpha(c_j)|} Score(c_j) \qquad (2)$$

where β is a damping factor that can be set between 0 and 1, $Edge_\alpha(c_j)$ is a set of vertices which connect to the vertex c_j, and α is the size of co-occurrence window size. After a number of iterations, we obtain the rank list of the candidate words. Finally, words with higher scores are selected as the labels of each subtopic.

IV. EXPERIMENTS

A. Dataset and Parameter Settings

To evaluate the performance of topic labelling model, we crawl three evaluation datasets from 1st April 2014 to 20th September 2014 from Weibo. The events include "Huawei Honor 6", "Apple" and "A Bite of China". For each event, we report the evaluation results using a 5-fold cross validation. The details of this collection are shown in Table I.

TABLE I. BASIC DATASETS INFORMATION

Event Name	# Microblog	Description
Huawei Honor 6	5106	A popular cellphone in China
Apple	19143	Apple Inc. or a kind of fruit
A Bite of China	23340	A popular TV show about food in China

The parameter k and d in word embedding learning are set to be 5, 200 respectively. In addition, we select nouns, verbs or adjectives as candidate topic words and the numbers of clusters of the three events are empirically set to be 4, 4 and 3 respectively. We set the damping factor β to be 0.85 empirically. As to the co-occurrence window size parameter α, we set it as the best performance value of 5.

B. Results and Discussions

Coherence is a very common metric to evaluate the generated topic labels. To demonstrate the validity of the proposed model, we employed the measure proposed in [14] which has been proved to be the state-of-the-art approach to evaluate the subtopic coherence. Let $T=\{\vec{w}_1,\vec{w}_2,...,\vec{w}_n\}$ be the labels of one subtopic, each label is represented as a vector of n dimensions and each vector is weighted using NPMI. The coherence is defined by:

$$Coherence_{Sim}(T)=\frac{\sum_{1\le i\le n-1,\ i+1\le j\le n} Sim_{cos}(\vec{w}_i,\vec{w}_j)}{0.5n(n-1)} \qquad (3)$$

where $Sim_{cos}(\vec{w}_i,\vec{w}_j)$ is a cosine similarity measure. We employ the mean value of all subtopics' coherence in one event as the event's coherence. The higher event's coherence means more interpretable labels of the subtopics.

Latent Dirichlet allocation (LDA) was conducted as the baseline, which is a popular topic model. Figure 2 and Figure3 show the performance of our method compared to the LDA model. We choose the top-n labels of each subtopic to calculate the event coherence.

From Figure 2 and Figure3, we can see that our unified model outperform the baseline LDA model in most cases.

The improvement is due to two reasons: (i) we use deep neural networks to uncover the association of words. (ii) A graph-based model is used to ensure important words to be ranked at the top of the list. The result reveals the validity of integrating the two models for topic labelling in Weibo.

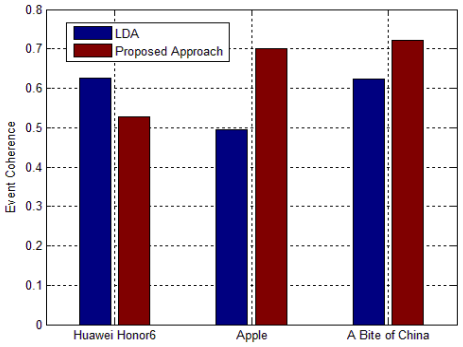

Figure 2. Event coherence on top-5 labels

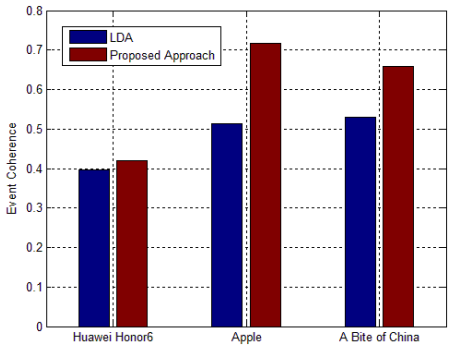

Figure 3. Event coherence on top-20 label

Besides, to provide deeper insight into the advantages of the proposed topic labelling approach, we further illustrate the mined topic labels. The top-5 labels for 4 subtopics of "Apple" event and " Huawei Honor 6" event identified by our method and the LDA method are shown in Table II and Table III, respectively. Table IV shows the top-5 lables for 3 subtopics of " A Bite of China " mined by our model and LDA.

TABLE II. TOP 5 TOPIC LABELS FOR 4 SUBTOPICS OF "APPLE"

Our method	LDA method
● Fruit, banana, food, honey, nutrition	● cellphone, download, fruit, address, news
● America, message, Apple Inc. , report, company	● promotion, cellphone, sales promotion, cheap, look
● product launch, announce, product, China, time	● win, accounts, free, journalist, girl
● cellphone, friends, Samsung, promotion, microblog	● product launch, cellphone, announce, China, America

TABLE III. TOP 5 TOPIC LABELS FOR 4 SUBTOPICS OF " HUAWEI HONOR 6"

Our method	LDA method
● price, reservation, J ingdong, selling price, cost performance ● Xiao Mi, performance test, configuration, expect, product ● Hisilicon, Kirin, announce, chip, product launch ● cellphone, support, experience, performance, flagship	● chip, announce, adopt, Hisilicon, Kirin ● cellphone, Xiao Mi, performance test, announce , product launch ● cellphone, Sina, flagship, experience, announce ● cellphone, reservation, configuration, share, the globe

TABLE IV. TOP 5 TOPIC LABELS FOR 3 SUBTOPICS OF " A BITE OF CHINA "

Our method	LDA method
● CCTV, documentary, broadcast ,start to broadcast, director ● greedy, delicious, drool, tasty, recommend ● life, culture, flavour, story, love	● start to broadcast, CCTV, channel, instant noodle, broadcast ● greedy, drool, love, video, hot pot ● director, Shanghai, documentary, culture, food

"Apple" event and " Huawei Honor 6" event are most about cellphone product, users often discuss the subtopics such as price, performance, product launch, promotion, experience, etc. Since apple is also a kind of fruit, therefore, some posts about fruit may contain in the data set. It can be seen from the above results that the proposed model can well identify different subtopics by using deep neural networks to learn the dense vector representation of topic words. For example, as shown in Table II, the model distinguishes fruit Apple subtopic from cellphone Apple subtopic. The mined topic labels including fruit, banana, food, honey and nutrition intuitively describe the semantic of the fruit subtopic.

We can also observe that the labels in each subtopic discovered by our method are obviously correlative. However, the labels discovered by the LDA approach are relatively hard to explain. For example, in Table II and Table III, the word "cellphone" occurs in three subtopics mined by LDA simultaneously which is somewhat confusing. In Table III, different views of price are intuitively explained by topic labels identified by the proposed model, these labels include price, reservation, Jingdong, selling price, cost performance, which can give users a comprehensive understanding of the price of the product.

"A Bite of China " event is about a food documentary which shows the food ecology around China, and helps better understand the long history of Chinese diet culture. Users often share their favorite food with each other and express their love to delicious food in Weibo, the topic labels including greedy, delicious, drool, tasty and recommend obtained by our model vividly convey users' feelings about delicious food.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we present a topic labelling method by training neural networks to discover the deep correlations of words and extracting representative words based on a graph model. Preliminary experiments have demonstrated the effectiveness of our model. Labels discovered by our approach are much meaningful and interpretable. In the future, we would like to discover the phrase labels of topics based on our existing work and evaluate the effectiveness of the method in other popular microblogging system.

ACKNOWLEDGMENT

This research is supported in part by National Natural Science Foundation of China under Grant No. 91224008, 61172106, 71402177. The Important National Science & Technology Specific Projects under Grant No. 2012ZX10004801, 2013ZX10004218.

REFERENCES

[1] Gao, H., Li, Q., Bao, H.,Song, S. 2012. How shall we catch people's concerns in micro-blogging? In WWW'12, pp. 505-506.

[2] Davide Magatti, Silvia Calegari, Davide Ciucci, and Fabio Stella. Automatic labeling of topics. In Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications, ISDA '09, pp. 1227–1232, Washington, DC, USA.

[3] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT'11, pp.1536-1545, Stroudsburg, PA, USA.

[4] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised graph-based topic labelling using dbpedia. In Proceedings of the sixth ACM international conference on Web search and data mining, WSDM '13, pp. 465–474, New York, NY, USA.

[5] Cano Basave, A. E., He, Y. and Xu, R. 2014. Automatic labelling of topic models learned from Twitter by summarisation. In ACL'14, pp. 618-624.

[6] Baroni, M., Dinu, G. and Kruszewski, G. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In ACL'14, pp. 238-247.

[7] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. PNAS, 101(suppl. 1), 2004, pp. 5228–5235.

[8] David Meir Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. In J. Mach. Learn. Res.2003, 3, pp. 993–1022.

[9] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '07, pp. 490–499, New York, NY, USA.

[10] Jey Han Lau, David Newman, Karimi Sarvnaz, and Timothy Baldwin. Best Topic Word Selection for Topic Labelling. CoLing 2010.

[11] Nikolaos Aletras and Mark Stevenson. Labelling Topics using Unsupervised Graph-based Methods. In ACL'14, pp. 631–636, Baltimore, Maryland, USA.

[12] LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. *Nature, 2015,521*(7553), pp. 436-444. doi:10.1038/nature14539

[13] Mikolov, T., Chen, K., Corrado, G.,Dean, J. Efficient estimation of word representations in vector space. In ICLR'13 Workshop.

[14] Aletras, N. and Stevenson, M. Evaluating topic coherence using distributional semantics. In IWCS'13, pp. 13-22.