Leveraging Social Annotation for Topic Language Model Adaptation

Youzheng Wu, Kazuhiko Abe, Paul Dixon, Chiori Hori, Hideki Kashioka

Spoken Language Communication Laboratory,
National Institute of Information and Communications Technology,
Kyoto, Japan

{youzheng.wu,kazuhiko.abe,paul.dixon,chiori.hori,hideki.kashioka}@nict.go.jp

Abstract

Social annotations such as Yahoo! Answers[1] already define broad coverage of hierarchical topic categories and include millions of documents annotated by web users. For topic language model (LM) adaptation, we present a novel approach of performing topic segmentation of LM training corpus via leveraging such social annotations, which may be more effective than unsupervised methods. Experimental results on the IWSLT-2011 TED ASR data sets demonstrate that we can achieve modest improvements when compared with the unsupervised methods such as clustering-based and LDA-based algorithms, e.g., the absolute WER improvement over the LDA-based method on the test set reaches 0.5 points.

Index Terms: Topic language model adaptation, social annotations, speech recognition.

1. Introduction

Topic language model adaptation is concerned with identifying a topic of document and adapting a language model toward that topic. Many supervised and unsupervised machine learning approaches have been previously proposed. However, one of the difficulties inherent in supervised topic LM adaptation is the lack of training data. Training data can be collected by manually annotating a suitable corpus for a set of predefined topics, but this is an extremely expensive and labor-intensive task. Moreover, predefined topics need to be modified when the target task shifts to a different domain, for example, from TED lectures [2] to academic lectures [3, 12] or broadcast news. To avoid this problem, unsupervised approaches [4, 8, 9] such as clustering and LDA (Latent Dirichlet Allocation) have been proposed. However, unsupervised approaches do not always result in optimal clustering due to a lack of prior knowledge about the topics used, and it is difficult for users to interpret the clustering results. This paper investigates a novel approach that can achieve better performance while not requiring extensive labor to annotate the training corpus.

Recently, social annotations have gained increasing popularity in many research areas such as text analysis [6] and information retrieval [15]. Social annotations, also called collaborative tagging or folksonomy, are created by users freely annotating objects such as web pages. The Yahoo! Answers service, for example, provides an interactive platform for users to post questions and answers. The users' questions and answers are also organized into hierarchical topic categories (as shown in Figure 1) and can be accessed through navigation and filtering. The Open Directory Project (ODP)[2], a widely used social annotation, has $4.8M$ web pages and $712K$ categories with each web page classified by human experts into categories. In this paper, we present an approach of leveraging millions of social-annotated data in social annotation services as training data of topic classifier, and then employing the learned classifier to partition the entire LM training data into topics for topic LM adaptation. Our approach has the following advantages: (1) topics and a plethora of training corpus can be automatically obtained for supervised topic modeling; (2) our method can be expected to achieve better performance than unsupervised methods because it uses social-annotated data about topics; (3) it is easy to adapt our method to a new domain because the social annotations' broad coverage of topics. To the best of our knowledge, there is little literature from this perspective.

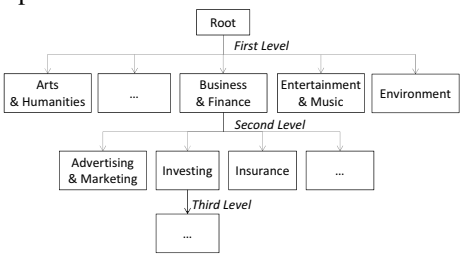

Figure 1: *An example of Yahoo! Answers taxonomy.*

2. Related studies

Topic language models have been explored by many researchers and can be classified into two categories: static

[1] http://answers.yahoo.com/dir/index

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

and dynamic topic LMs. To build static topic LMs, for example, Shi and Chu [13] adopted supervised approach that defined 129 topics and manually annotated more than 20,000 news articles in order to build an SVM classifier. Because it is expensive to annotate training data, the most common techniques in static topic modeling are unsupervised clustering-based and LDA-based approaches [4, 8, 9] that will be used in this paper for comparison. Dynamic topic LM adaptation [12, 13] usually employs a search engine to retrieve similar documents in the collection to build topic LMs.

2.1. Clustering

The clustering algorithm finds a predefined number of clusters based on a specific criterion. In this paper, we chose the following function that maximizes intra-cluster similarity:

$$maximize \sum_{i=1}^{K} \sqrt{\sum_{v,u \in s_i} sim(v,u)} \qquad (1)$$

where K is the desired number of clusters, S_i is the set of documents assigned to the $i\text{-}th$ cluster, v and u represent two documents, and $sim(v,u)$ is the cosine similarity between the two documents. This paper employs the CLUTO toolkit [7] to perform hard clustering and assign a document to one cluster. Based on the sentences belonging to each topic cluster, K topic n-gram LMs are constructed.

2.2. LDA

LDA [1] models documents as mixtures over K topics, each topic being a multinomial distribution over words. Specifically, LDA training finds document-topic distributions, i.e., θ_{d_i,t_k}, and word-topic distributions, i.e., ϕ_{w_j,t_k}. LDA provides soft clustering of each document into multiple topics. We, however, assign one optimal topic t_i^* to a document d_i using Equation (2). Then, each document contributes to a topic LM. After LDA training, an n-gram topic LM is constructed for each topic using the documents within it.

$$t_i^* = \arg\max_{1 \le k \le K} \theta_{d_i,t_k} \qquad (2)$$

Note: K in Equation (1) and (2) is empirically set to 50. During testing, cross entropy measure is used to determine topics for a test document.

3. Topic LM via social annotations

Unlike prior work in topic LM adaptation, this paper explores and analyzes social annotations with regard to classification of corpus collection to build topic language models. This paper argues that topic classification based on such social annotations, while possibly noisy, may be more accurate than the unsupervised clustering-based and LDA-based algorithms introduced above. Our algorithm is shown in Figure 2. which includes, SVM learning via social annotation, and performing topic segmentation of the LM corpus using the learned SVM.

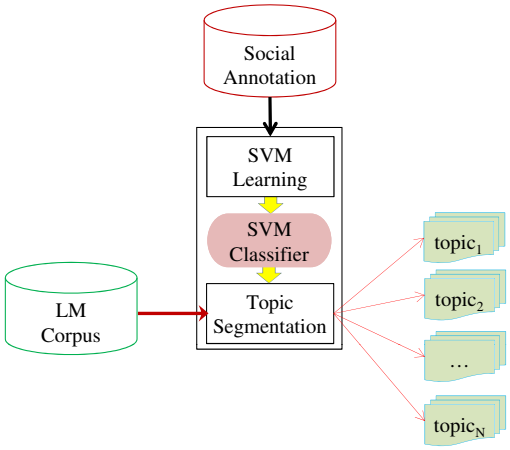

Figure 2: *Social annotation for topic language model.*

3.1. SVM Learning

Many social annotations can be used for our study. This paper employs the Yahoo! Answers service, which has 26 categories at the first level and 1,262 categories at the leaf level. It contains millions of questions and answers in 26 languages. In the experiments, we remove two categories that are very special for Yahoo! Answers. Accordingly, 24 categories in the first level (shown in Table 1) are used[3].

Table 1: *Categories used for topic language models.*

Arts & Humanities	Beauty & Style	Business & Finance
Cars & Transportation	Computers & Internet	Consumer Electronics
Dining Out	Education & Reference	Entertainment & Music
Environment	Family & Relationships	Food & Drink
Games & Recreation	Health	Home & Garden
News & Events	Pets	Politics & Government
Pregnancy & Parenting	Science & Mathematics	Social Science
Society & Culture	Sports	Travel

Since the entire dataset is too large, we randomly extract 6,000 (question,answer) pairs for training each category and held out 1,000 for testing. We then extract up to trigrams with their frequencies. To reduce the number of nuisance features, we first sort n-grams in the decreasing order of Chi-square statistics and then take the highest 20,000 n-grams as classification features. A multi-class SVM is used to perform the classification. We use the freely available libsvm[4] and set it up to the linear kernel

[3]The Yahoo! Webscope Dataset Yahoo! Answers Comprehensive Questions and Answers version 1.0.2, available at http://reseach.yahoo.com/Academic_Relations.

[4]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

and default parameters. The classification performance, measured by $accuracy = \frac{\#Correctly\ classified\ test\ data}{\#Total\ number\ of\ test\ data}$, on the held-out test data is 65.0%. The performance is not high because of noise in the social annotated data.

This multi-class SVM is then used to partition the entire LM training data into 24 topic clusters and 24 topic n-gram LMs are then constructed. The SVM also outputs a classification confidence score that represents how confident the classifier is on each classified data. During the test, the confidence score is used to determine topics for a test document. Note that they may not necessarily represent the successful classification of LM corpora and positively impact the recognition results. We will validate these results in the experiments.

4. Experiments

This paper uses the data sets for the IWSLT-2011 ASR shared task in order to recognize the TED talks published on the TED website. The talks include environment, photography and psychology, while not adhering to a single genre. This task reflects the recent increase of interest in automatically transcribing lectures, in order to make them either searchable or accessible.

Table 2: *Summary of data sets and LM training data.*

test sets			
data	#talks	#utterances	#words
development set	11	1664	6,911
test set	8	818	3,200
LM training data			
	#sentences		#words
in-domain	$124K$		2,063K
general-domain	115,101K		2,458,626K

For LM training, the IWSLT-2011 evaluation campaign defines a closed set of publicly available English texts, including a small collection of TED transcriptions (called *in-domain* corpus) and a large collections of news sentences (called *general-domain*). Statistics of the test sets and LM training data are shown in Table 2. All of the training data are preprocessed by a non-standard-word-expansion tool that is used to convert non-standard words (such as CO2 or 95%) to their pronunciations (CO two, ninety five percent). The most frequent $100K$ words were extracted from the preprocessed corpora, which, together with the CMU.v7 pronunciation dictionary, are used as the LMs' vocabulary. Finally, our vocabulary contains $157K$ entries and has an OOV rate of 0.78% on the development data set. To build an effective language model, a data filtering technique proposed in [11] is used to filter out three quarters of the general-domain data. For the in-domain and filtered general-domain corpora, modified Kneser-Ney smoothed trigram LMs are constructed using the MITLM [5], and then interpolated to form a generic LM by optimizing the perplexity of the development set.

The acoustic model (AM) is a tied-state Hidden Markov Model trained on 198h of HUB4 speech, featuring $26K$ triphone units and $465K$ Gaussians. It has $77K$ states and 6 mixtures per state. The feature vectors have 39 dimensions comprising 13 static MFCCs and their first- and second-order derivatives. In the decoding phase, the generic LM and the AM are first used to generate 1-best hypotheses. For each talk, the 1-best hypothesis is used to choose n (empirically set to 3) topic LMs. We finally interpolate the generic LM with the chosen n topic LMs by minimizing the perplexity of the 1-best hypothesis to obtain the final topic-adaptive LM. Table 3 compares the experimental results by word-error-rate (WER) and perplexity. Figure 3 shows the improvements for each talk in the test set. In the final system, RNNLM [10] trained on the in-domain corpus is used to rescore n-bests generated from our model. Note that numbers of hidden layers and classes in the RNNLM are set to 480 and 300, respectively.

Table 3: *Comparison of topic LM approaches.*

LM	**WER (%)**		**Perplexity**	
	test	dev	test	dev
Generic	25.7	26.0	132.9	133.5
Clustering	25.6	26.0	132.6	133.1
LDA	25.7	26.0	132.9	133.5
Social annotation	25.2	25.8	130.7	131.9
RNNLM	24.2	24.9	117.2	116.8

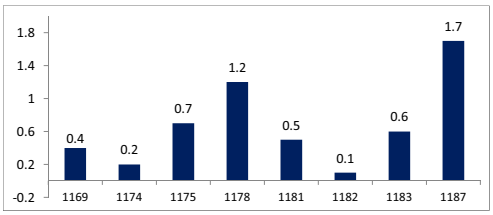

Figure 3: *Absolute improvement over the clustering-based method.* 1169,...,1187 *denote talks' ids.*

The experiments show that our approach outperforms the generic LM, the clustering-based and LDA-based approaches. For example, the absolute WER improvement over the generic LM on the test set reaches 0.5%, the improvement over the clustering-based method on talk1187 is 1.7%. However, the improvement on the development set is not significant compared with the RNNLM. The main reasons of negatively impacting our topic LMs may lie in: (1) The LM training corpora provided by the IWSLT-2011 evaluation campaign are sentences without document boundaries. Because of the short lengths, the sentences do not provide enough keywords for topic

classification. (2) The majority of the LM corpora is in the general-domain, i.e., newspapers. The coverage of the spontaneous TED transcriptions is quite limited with topic LMs. (3) The performance of the multi-class classifier using social annotations is not so satisfactory in this paper. Better classifiers, such as a hierarchical SVM, could be expected to further improve ASR results. However, one important advantage of our approach is that it assigns meaningful topics to TED talks that are almost consistent with human labeling. For example, the top two assigned topics for talk1169 are "Arts & Humanities" and "Social Science".

In order to better evaluate the proposed approach, the following experiments are conducted based on the HUB4 language model data (LDC98T31) and the English Gigaword Fourth Edition (LDC2009T13), respectively. The LDC98T31 contains data from transcribed news broadcasts. We use the documents in the "st_train" to train generic and topic language models. The LDC2009T13 is a comprehensive archive of newswire data that includes the six distinct international sources of English newswire texts. The "xin_eng" part of the LDC2009T13 is used in this experiment. Table 4 shows the perplexities on the IWSLT-2011 data sets. The numbers in subscripts are the relative improvements over the generic language model.

Table 4: *Perplexities of Language models learned from the HUB4 and Gigaword corpora.*

		Generic	**Clustering**	**Social annotation**
HUB4	dev	172.7	$168.7_{2.3\%}$	$161.6_{6.4\%}$
	test	171.2	$166.3_{2.9\%}$	$156.7_{8.5\%}$
Giga	dev	415.6	$373.9_{10.0\%}$	$369.7_{11.1\%}$
	test	446.1	$385.3_{13.6\%}$	$379.9_{14.8\%}$

This experiment indicates that our proposed approach significantly outperforms the generic LM and modestly outperforms the clustering-based topic LMs. For example, the improvements of our approach over the clustering-based approach on the LDC98T31 are 4.2% and 5.8%, respectively. The improvements on the LDC2009T13, however, are not significant. This is because the LDC98T31, transcribed broadcast news, is more similar to the spontaneous TED transcriptions than the LDC2009T13, the newswire texts. From this experiment, we can conclude that our approach may be quite promising given that the topic LMs are trained from a large in-domain corpus.

5. Conclusions

Social annotations such as the ODP define hierarchical topic categories and millions of user-annotated data. This paper investigates the possibility of leveraging such social annotations for topic language modeling and compares it with the common clustering-based and LDA-based algorithms. The experiments have shown that the proposed topic LMs can improve the WER compared with the other approaches using a recognition system. Though the difference between the topic modeling methods is not significant, one important advantage of our approach is that it can assign meaningful topics to TED talks compared with unsupervised techniques.

As Figure 1 shows, social annotations define a hierarchical categories. This paper just exploited 24 categories in the first level. More large-scale, finely tuned topic language models will be investigated by leveraging deep-level categories of social annotations in our future work. In addition, we will work on improving classification performance based on social annotations. Our proposed approach may become more promising if better classifiers can be obtained.

6. References

[1] Blei, D., Ng, A. and Jordan M., "Latent dirichlet allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[2] Federico, M., Bentivogli, L., Paul, M. and Stuker, S., "Overview of the IWSLT 2011 evaluation campaign", in Proc. of IWSLT, pp. 11-27, 2011.

[3] Glass, J., Hazen, T.J., Cyphers, S., Malioutov, I., Huynh, D., and Barzilay, R., "Recent Progress in the MIT Spoken Lecture Processing Project", in Proc. of Interspeech, pp. 2553-2556, 2007.

[4] Hsu, B.J. and Glass, J., "Style and topic language model adaptation using HMM-LDA", in Proc. of EMNLP, pp. 373-381, 2006.

[5] Hsu, B.J, and Glass J., "Iterative language model estimation: efficient data structure & algorithms", In Proc. of Interspeech, 2008.

[6] Iwata, T., Yamada, T. and Ueda N., "Modeling social annotation data with conent relevance using a topic model", In Proc. of NIPS, 2009.

[7] Karypis, G., "Software for clustering high-dimensional datasets", http://glaros.dtc.umn.edu/gkhome/views/cluto.

[8] Lei, X., Wu, W., Wang, W., Mandal1, A. and Stolcke A., "Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversation", in Proc. of Interspeech, 2009.

[9] Liu, Y. and Liu, F.F., "Unsupervised language model adaptation via topic modeling based on named entity hypotheses", in Proc. of ICASSP, 2008.

[10] Mikolov, T., Kombrink, S., Burget, L., Cernocky, J., Khudanpur, S., "Extensions of recurrent neural network language model", In Proc. of ICASSP 2011.

[11] Moore, R., and Levis, W., "Intelligent selection of language model training data", in Proc. of ACL, pp. 220-224, 2010.

[12] Nanjo, H. and Kawahara, T., "Language model and speaking rate adaptation for spontaneous presentation speech recognition", IEEE Transactions on Speech and Audio Processing, 12(4): 391-400, 2004.

[13] Shi, Q., Chu, S.M., Liu, W., Kuo, H-K., Liu, Y. and Qin, Y., "Search and classification based language model adaptation", in Proc. of Interspeech, 2008.

[14] Tur, G. and Stolcke, A., "Unsupervised language model adaptation for meeting recognition", In Proc. of ICASSP, pp. 173-176, 2007.

[15] Zhou, D., Bian, J., Zheng, S.Y., Zha, H.Y. and Giles, C. L., "Exploring social annotations for information retrieval", In Proc. of WWW, pp. 715-724, 2008.