MOOCon: A Framework for Semi-supervised Concept Extraction from MOOC Content

Zhuoxuan Jiang(✉), Yan Zhang, and Xiaoming Li

School of Electronics Engineering and Computer Science, Peking University, Beijing, China
{jzhx,lxm}@pku.edu.cn, zhy@cis.pku.edu.cn

Abstract. Recent years have witnessed the rapid development of Massive Open Online Courses (MOOCs). MOOC platforms not only offer a one-stop learning setting, but also aggregate a large number of courses with various kinds of textual content, e.g. video subtitles, quizzes and forum content. MOOCs are also regarded as a large-scale 'knowledge base' which covers various domains. However, all the contents generated by instructors and learners are unstructured. In order to process the data to be structured for further knowledge management and mining, the first step could be concept extraction. In this paper, we expect to utilize human knowledge through labeling data, and propose a framework for concept extraction based on machine learning methods. The framework is flexible to support semi-supervised learning, in order to alleviate human effort of labeling training data. Also course-agnostic features are designed for modeling cross-domain data. Experimental results demonstrate that only 10% labeled data can lead to acceptable performance, and the semi-supervised learning method is comparable to the supervised version under the consistent framework. We find the textual contents of various forms, i.e. subtitles, PPTs and questions, should be separately processed due to their formal difference. At last we evaluate a new task: identifying needs of concept comprehension. Our framework can work well in doing identification on forum content while learning a model from subtitles.

Keywords: Concept extraction · MOOC · Semi-supervised · CRF

1 Introduction

With the prosperity of Massive Open Online Courses (MOOCs), tens of millions of students all over the world have been beneficial in recent years. An important characteristic of MOOC is providing a one-stop online learning setting which is composed by: (1) video clips, (2) homework, (3) email notification, (4) discussion forum, and (5) quiz/assignment. As more and more disciplines are offered on MOOC platforms, a massive number of cross-domain knowledge have been aggregated in a form of textual content, e.g. subtitles of videos, questions in quizzes, and posts in forums. However, most of the textual content are

Z. Bao et al. (Eds.): DASFAA 2017 Workshops, LNCS 10179, pp. 303–315, 2017.
DOI: 10.1007/978-3-319-55705-2_24

unstructured and diverse. For example, subtitles are well-organized and formal, since they are generated by instructors. While the contents of posts are written by various learners, thus they are colloquial and informal. So there is an issue that how to make the textual content structured, in order to facilitate subsequential knowledge management and mining.

As MOOCs are in fact an educational and unstructured knowledge base, an intuitive and first-step idea is to extract knowledge points, i.e. concepts, from textual content. Concept extraction from MOOC data may be faced with several difficulties: (1) MOOC data is multi-discipline so the method should be instructor- and course-agnostic, (2) obtaining labeled training dataset is extremely expensive since usually domain expertise is required. However, once concepts are well extracted, many subsequential applications are feasible, e.g. building course- or domain-specific concept map, structured cross-domain concepts management, and even personalized learning.

In this paper, we explore the feasibility of concept extraction from MOOC textual content. We regard this task as a problem of sequence labeling by natural language processing method. Since conditional random fields (CRFs) has been proved successful in many sequence labeling tasks like part of speech (POS), named entity recognition, and word segment [19], we propose a framework called MOOCon by adapting CRFs to a semi-supervised version in order to alleviate the human effort of labeling training dataset. The framework is shown as Fig. 1.

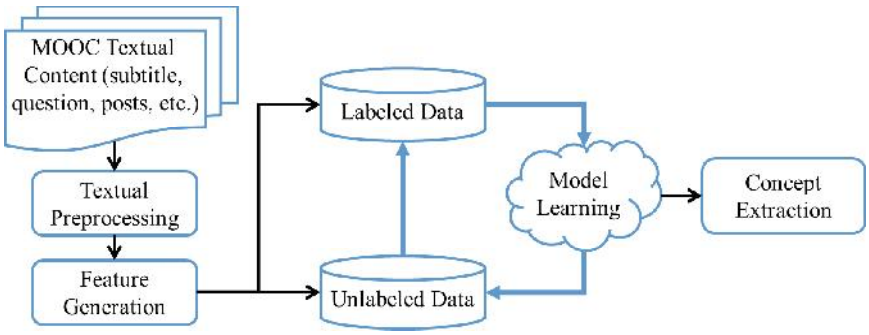

Fig. 1. Framework of semi-supervised concept extraction from MOOC textual content.

As shown in Fig. 1, after collecting raw MOOC textual content, our basic idea is to first preprocess the data with natural language processing tools. We define instructor- and course-agnostic features during feature generation. The thoroughly-considered features should be applied to various courses. Then by obtaining a handful of labeled data, an initial model can be learned. Some unlabeled data are selected based on a measure and can be tagged by the model. The newly-labeled data join the 'Labeled Data', and all labeled data are used again to learn a new model. Cyclically until there is no unlabeled data, the final

model can be used to do concept extraction. Note that the data used during the 'Concept Extraction' are usually unobserved samples during 'Model Learning'.

Experimental results demonstrate only 10% labeled data can lead to satisfactory performance. And semi-supervised model is comparable to the supervised version. On the other hand, data in various forms, e.g. subtitle, question and post, should be separately learned, because they have different model ability when being regarded as training data. Especially, an experiment of identifying threads about need of concept comprehension is conducted. The result shows the proposed framework can indirectly run binary classification on MOOC forum content.

2 Related Work

Concept extraction in MOOC settings is a little different from traditional information extraction, e.g. key phrase extraction [7], terminology extraction [4], and named entity recognition (NER) [14,17]. Key phrases (or key words) are usually involved in search engine. Terminology is domain-specific. And named entity is usually person name, location, time and address. Concepts in MOOC settings are not only domain-specific but also cross-domain. For example, the word of *Network* may be a concept in both courses of *Criminal Laws* and *Introduction to Computer Network*. But the two '*Networks*' are actually different concepts.

In the past decades, methods for sequence labeling have been largely studied in the field of information extraction [5]. For example, [10] proposes a rule-based method to extract terms, and [2,6] propose statistical methods. Recently some machine learning methods are proposed for terminology extraction [15,16]. But [15] is designed for software document domain where terminologies are domain-specific proper nouns, e.g. *User ID field*. While [16] only considers nouns when doing extraction. By the way, [15] also proves that directly applying existing methods of NER to terminology extraction will not perform well.

In terms of MOOC data mining, large of studies have been proposed in recent years. For example, [1] studies a badge system to produce incentives for activity and contribution in the forum based on behavior patterns. [8] analyzes the behavior of superposter in 44 MOOCs forums and finds MOOCs forums are mostly healthy. [22] studies the sentiment analysis in MOOCs discussion forums and finds no consistent influence of expressed sentiment to dropout exists. [21] studies the learning gain reflected through forum discussions. [9] conducts an analysis from the perspective of influence in MOOCs forum. In summary the previous studies with MOOCs data focus more on how to improve learning efficacy.

3 Semi-supervised Concept Extraction

In this section, we introduce our framework for concept extraction from MOOC content. Firstly, we introduce the CRFs model which can be applied to our task. Then we state the feature engineering. Then the method of learning a CRFs model is described. At last we introduce the semi-supervised framework which

can alleviate human effort for labeling. Note that the framework can be adapted to other probabilistic graphical models, rather than only CRFs.

3.1 Conditional Random Fields Model

As a sequence labeling problem, concept extraction is similar to other sequence labeling tasks, e.g. named entity recognition and part-of-speech annotation. Probabilistic graphical models are the corresponding solution to this kind of tasks, and conditional random fields (CRFs) can obtain the state-of-the art performance [19]. We leverage the CRFs framework to our task.

The problem of sequence labeling can be formally described as solving the conditional probability $P(Y|X)$. The random variable X is features of each sentence which consists of a word sequence $x = \{x_1, x_2, ..., x_T\}$, and the random variable Y is a label sequence of the sentence $y = \{y_1, y_2, ..., y_T\}$.

As to our task, we concern more on the conditional probability of labeling sequence Y, i.e. $p(Y|X)$, rather than their joint probability $p(Y, X)$, so linear chain CRFs framework [11] is the natural choice. The conditional distribution over label sequence y given an observation word sequence x can be defined as:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x_t)\right) \tag{1}$$

where $Z(x) = \sum_y \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x_t)\right)$ and $\mathscr{F} = \{f_k(y_{t-1}, y_t, x_t)\}_{k=1}^{K}$ are the set of feature functions defined on given x; $\Theta = \{\lambda_k\} \in \Re^K$ are parameter vector. N is the length of sentence and K is the number of features. Thus in order to fulfil the task of concept extraction, the next steps are defining feature functions $\mathscr{F}$ and learning the model $\Theta = \{\lambda_k\}$.

3.2 Feature Engineering

A crucial part of CRFs framework is the definition of feature functions. Based on our observation, we define five kinds of features which are adapted to our educational data. All the features are course-agnostic and make our framework flexible for scalability.

Text Style Features

- Whether the target word are English.
- Whether the two neighbor words are English.
- Whether the word is the first word in a sentence.
- Whether the word is the last word in a sentence.
- Whether the target word is in a quotation.

Text style features capture the stylistic characteristics. Some concepts usually appear at the beginning or the last of a sentence in instructor's language, e.g. "*Netwrok* means..." or "...This is the definition of *Network*." Since our data are from a Chinese MOOC, we regard whether the word is English as a feature. Obviously, when it comes to English MOOCs, capitalization is the key feature of English concepts. So this kind of features are flexible to different situations.

Structure Features

- Part-of-speech tag of the target word.
- Part-of-speech tag of the previous word.
- Part-of-speech tag of the next word.

We treat the part-of-speech as a feature because fixed combination of part-of-speech, e.g. adjective + noun or noun + noun, may indicate concept phrases. We utilize the Stanford Log-linear Part-Of-Speech Tagger (POS Tagger)[1] [20] to assigns parts-of-speech to each word. Note that as to the corresponding feature functions, we adopt binary value, 0 or 1, to every part-or-speech. For example, there is a function to capture whether target word is a noun or not, and so on.

Context Features

- TF-IDF value of the target word and two neighbor words.
- Normalized uni-gram BM25 score of the target word.
- Normalized bi-gram BM25 score of the target word.
- Normalized bi-gram BM25 score of the two neighbor words

Context features capture the importance of words and word-level information within the whole documents. The training set is partitioned to documents based on video clips. Statistical metric of normalized bi-grams BM25 scores [18] is used to quantify word relevance by default parameters.

Semantic Features

- Semantic similarity of the target word with the previous two words respectively.
- Semantic similarity of the target word with the next two words respectively.

Some frequent-co-occurrence words may be concept phrases. Also close words in the semantic space may be concepts. So by learning the word semantics, features of adjacent words can be captured. The similarity of two adjacent words in semantic space is calculated with the corresponding word vectors trained by Word2Vec[2] [13]. All textual content are used to learn the word embeddings. The corpus size is 145,232 words and the vector dimension is set as 100 by default.

Dictionary Features

- Whether the target word and two neighbor words are in dictionary.
- Whether the two neighbor words are in dictionary.

As in most tasks about natural language processing, a dictionary is useful, we design a run-time dictionary which is just a set of concepts in training dataset.

[1] Stanford Log-linear Part-Of-Speech Tagger: http://nlp.stanford.edu/software/tagger.shtml.

[2] Word2Vec: https://code.google.com/p/word2vec/.

3.3 Learning and Inference

Given a training dataset, the model $\Theta = \{\lambda_k\}_{k=1}^{K}$ could be learned by Maximum Likelihood Estimation (MLE). To avoid overfitting, we add a regularized term to the function. Then the log-likelihood function of $p(y|x, \lambda)$ based on the Euclidean norm of $\lambda \sim (0, \sigma^2)$ is represented as:

$$L(\Theta) = \sum_{x,y} \log p(y|x, \Theta) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2} \tag{2}$$

So the gradient function is:

$$\frac{\partial L}{\partial \lambda_k} = \sum_{x,y} \sum_{t=1}^{T} f_k(y_{t-1}, y_t, x_t) - \sum_{x,y} \sum_{t=1}^{T} \sum_{y,y'} f_k(y, y', x_t) p(y, y'|x) - \frac{\lambda_k}{\sigma^2} \tag{3}$$

The detail of learning the CRFs model can be referred to [19]. Then given a new word sequence x^* and a learned model $\Theta = \{\lambda_k\}_{k=1}^{K}$, the optimal label sequence y^* could be calculated by:

$$y^* = arg \max_{y \in \mathscr{Y}} p(y|x^*, \Theta) \tag{4}$$

where $\mathscr{Y}$ is the set of all possible label sequences for the given sentence x^*. We employ L-BFGS algorithm to learn the model and Viterbi algorithm to infer the optimal label sequence y^*.

3.4 Semi-supervised Learning Framework

Because the effort for labeling training data is extreme expensive, we propose a semi-supervised framework. We leverage the ideas of self training [12] and k nearest neighbors (k-NN). The intuition is that if an unlabeled sample is similar to a labeled sample in semantic space, the unlabeled sample is very probable to be successfully inferred by the model which is learned from all the current labeled data. Then the unlabeled sample is turned to a labeled one and can be added into the labeled dataset with model-inferred labels. A new model can be learned. We calculate the similarity by Cosine distance between two sentences. Sentence vector is denoted as:

$$VecSentence_i = \frac{1}{T} \sum_{t=1}^{T} VecWord_t \tag{5}$$

where $VecWord$ is learned by Word2Vec with default parameters. Algorithm 1 is the details of the semi-supervised version of training process.

The time complexity of Algorithm 1 is $O(NM^2) + \frac{M}{c} O(\text{TrainCRF})$ where N and M are the sizes of labeled set and unlabeled set respectively, and c is the number of unlabeled data which are selected to be inferred in each loop. The additional computing cost is deserved since human effort can be largely reduced, especially when N and M is not large.

Algorithm 1. k-NN Self Training for SSC-CRF

INPUT: labeled dataset $X_L = \{(x, y)\}$, unlabeled dataset $X_U = \{x\}$, number of candidates c
OUTPUT: model Θ

repeat
  $\Theta = \text{TrainCRF}(X_L)$
  $X_{c-nearest} = \varnothing$
  for i=1:c
    $x = \arg\min_{x \in X_U} \text{Cosine\_distance}(x, X_L)$
    $X_U = X_U - \{x\}$
    $X_{c-nearest} = X_{c-nearest} \cup \{x\}$
  $Y_{c-nearest} = \text{InferCRF}(X_{c-nearest}, \Theta)$
  $X_L = X_L \cup \{(X_{c-nearest}, Y_{c-nearest})\}$
until $X_U = \varnothing$
$\Theta = \text{TrainCRF}(X_L)$
return Θ

4 Experiment

In this section, firstly we evaluate the performance of supervised and semi-supervised framework for concept extraction. Then we conduct a task of identifying need of concept comprehension with forum content. Last we discuss the contribution of different features to the model.

4.1 Dataset Collection

We collect the corpus of an interdisciplinary course conducted in the fall of 2013 on Coursera. The course contains computer science, social science and economics. Textual content include video subtitles, PPTs, questions and forum contents (i.e. threads, posts and comments). Table 1 lists the statistics of the content. We invited the instructor and two TAs to help label the data. As seen in Table 1, the number of concepts in questions and PPTs are much smaller than that in subtitles. This implies subtitles may cover other kind of textual content. Based on our observation, during labeling the data, the instructor and TAs would spend much time on understanding each sentence. By statistic, labeling 3,000 sentences would spend about 8 hours per person (in average 10 seconds per sentence). So due to the complexity of labeling, we have not labeled concepts in forum content yet. However we conduct another task which also demonstrates the effectiveness of our model.

A preprocessing step of word segment for Chinese may be necessary. We adopt the Stanford Word Segmenter[3] [3]. All data are randomly shuffled before they are learned and validated. 5-cross-fold validation is adopted.

Label of Word. The label of a word is defined as three classes: *NO*, *ST* and *IN*. They respectively mean *not a concept*, *the beginning word of a concept* and *the middle word of a concept*. So the label variable is $Y \in \{NO, ST, IN\}$. The example sentence with labels is shown as Fig. 2.

[3] Stanford Chinese word segment:http://nlp.stanford.edu/software/segmenter.shtml.

Table 1. Corpus statistic of the course.

Source	# sentence	# word	# concept
Subtitles	3,036	69,437	402
PPTs	2,823	22,334	249
Questions	268	7,138	95
Threads (title and initial post)	213	12,759	-
Posts	704	28,095	-
Comments	691	27,803	-

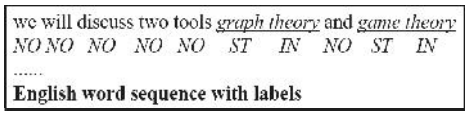

Fig. 2. Examples of word sequences with labels. Underlined words are concepts.

4.2 Baselines

We propose several baselines to extract concepts for comparison with our approach. The preprocessing is identical for baselines as for our method.

- **Term Frequency (TF)**: Words are ranked by their term frequency. If a word is a concept, the instructor may say it repeatedly in lecture.
- **Bootstraping (BT)**: Instructors may have personal language styles to give talks. So we design the rule-based algorithm by giving several patterns containing true concepts. This method is actually course- and instructor-dependent.
- **Stanford Chinese NER (S-NER)**: This is an exiting tool developed for named entity recognition, whose model is already trained and we just use it to infer concepts in our educational datasets[4] [14].
- **Terminology Extraction (TermExtractor)**: This is an exiting tool for terminology extraction[5]. The well-trained model is also only used to infer concepts in our datasets.
- **Supervised Concept-CRF (SC-CRF)**: This is a method of supervised learning based conditional random fields with all features as defined before.
- **Semi-supervised Concept-CRF (SSC-CRF)**: This is the semi-supervised version for concept extraction. The parameter of c, number of candidates, is empirically set as 20.

We adopt three metrics, precision, recall and F1-value, to measure the results.

[4] Stanford Chinese Named Entity Recognizer (NER): http://nlp.stanford.edu/software/CRF-NER.shtml.

[5] Terminology Extraction by Translated Labs: http://labs.translated.net/terminology-extraction/.

4.3 Task of Concept Extraction

In this subsection, we use subtitles, PPTs and questions to evaluate the proposed models.

Table 2 shows the comparison of performance between baselines. We use 30% data of subtitles as training data for SC-CRF and SSC-CRF, and the rest are for evaluation. Especially for SSC-CRF, half of the training data are unlabeled. The statistic-based methods (TF@500 and TF@1000) are unreliable due to many stopwords may degrade the performance. The rule-base method (BT) is highly dependent on human experience, and the low precision means plenty of subsequent work for filtering the outputs is required. On the other hand, Stanford Chinese NER and TermExtractor do not perform well maybe because of two reasons: (1) named entity and terminology are different from concepts, (2) the models are not learned from our dataset. The semi-supervised CRF is comparable to the supervised version.

Table 2. Performance of baselines. SC-CRF and SSC-CRF use 30% data of subtitles for training. Half of the training data as unlabeled for SSC-CRF.

	Precision	Recall	F1
TF@500	0.402	0.500	0.446
TF@1000	0.600	0.746	0.665
BT	0.099	0.627	0.171
S-NER	0.131	0.080	0.099
TermExtractor	0.202	0.107	0.140
SC-CRF	**0.914**	**0.897**	**0.905**
SSC-CRF	0.889	0.825	0.856

Figure 3 manifests that the semi-supervised learning would be comparable to the supervised version, especially when less than 20% data are used for training. Half of training data is identically regarded as unlabeled by SSC-CRF. Note that the amount of labeled data when using 10% training data by SC-CRF are equivalent to that of using 20% training data by SSC-CRF, but SSC-CRF performs better than SC-CRF. This result means the semi-supervised framework can obtain satisfactory performance by only labeling a handful of data.

Now we evaluate the different model ability among various MOOC textual content. As shown in Table 3, the items in row are training dataset while those in column are testing dataset. This table can explain some common situations of educational settings. Subtitles can cover almost all the concepts. They are ideal to be regarded as the training data. PPTs is also decent to be as training data seeing from the precisions, but the recalls are low. Maybe due to usually in PDF format, PPTs may cause incomplete sentences when being converted to text. Questions could lead to lower recalls than PPTs, because not all concepts

are present in questions as shown in Table 1. In summary, different kinds of MOOC textual content have different model ability, so they should be separately considered.

Feature Contribution. We analyze how the different kinds of features contribute to the model. The result is shown as Table 4. Dictionary Feature has a predominant influence on the final results, and Structure Feature is second important. Other features are also contributive but the difference is small. Even so, every kind of features contribute to the model positively.

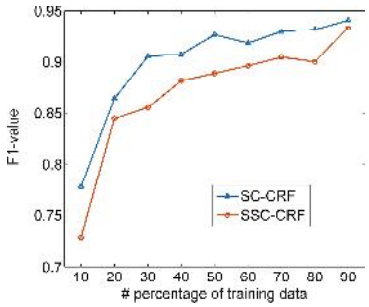

Fig. 3. Performance between supervised and semi-supervised models.

Table 3. Mutual learning between various content. The rows are training data and the columns are testing data.

	Subtitles			PPTs			Questions		
	P	R	F1	P	R	F1	P	R	F1
Subtitles	-	-	-	0.816	0.838	0.827	0.860	0.800	0.829
PPTs	0.868	0.764	0.813	-	-	-	0.857	0.685	0.761
Questions	0.846	0.349	0.494	0.722	0.360	0.480	-	-	-

4.4 Task of Concept Identification

Since we do not obtain the labels of forum content, we conduct a task of concept identification. The model is learned from subtitles, and then the model is used to infer concepts on forum content. We use 30% of subtitles to learn the semi-supervised model. Unlike directly identifying concepts, we use our model to identify the *need of concept comprehension*. The new task is actually a binary classification of forum threads, that is to identify whether a thread is about concept comprehension. Only threads title and the initial post are inferred by our model, instead of all the posts. The classification result is post-evaluated which means: as to each thread, that the assessed score is '1' means two situations:

– If no concept is identified and this thread is not about need of concept comprehension.
– If at least one concept is identified and the definition of identified concepts can at least partially answer the thread.

Other situations are assessed as '0'. The result is shown as Table 5. The accuracy is not bad. And the relatively high recall is useful. It can suggest which threads instructors should intervene. Our model applied in this task can not only identify whether a thread is about need of concept comprehension, but also identify which concept should be explained.

Table 4. Efficacy of features. 10% of data are used for training by SSC-CRF.

	Precision	Recall	F1
All	**0.780**	**0.775**	**0.777**
Without text style feature	0.768	0.776	0.772
Without structure feature	0.722	0.683	0.702
Without context feature	0.757	0.753	0.755
Without semantic feature	0.772	0.757	0.764
Without dictionary feature	0.689	0.235	0.350

Table 5. Result of identifying threads about need of concept comprehension by SSC-CRF.

Accuracy	Precision	Recall	F1
0.822	0.523	0.784	0.627

5 Discussion and Conclusion

Along with the coming of MOOCs, large-scale online educational resources are unprecedentedly produced. Instructors can provide videos, subtitles, lecture notes, questions and etc. While learners can generate forum content, Wiki, log of homework and etc. How to process these data from unstructured to structured is a challenging problem. In this paper, we explore the task of concept extraction on MOOC resources.

Concept extraction can benefit a lot of subsequential applications. Firstly, it is a kind of annotation for MOOC resources. The annotation can be used for studying machine learning methods for MOOC-related natural language processing tasks, such as information extraction, information retrieval and question answering. Secondly, concept extraction can pick up domain-specific or cross-domain knowledge points from complex text. This result can be further processed to build knowledge graph or concept map. With the graph (or map), instructors can better organize the course, and learners can plan their own learning paths

more easily. Then by collecting the feedback from learners, the whole teaching and learning process can be a virtuous cycle.

Getting back to the topic of this paper, we are faced with two challenges: (1) MOOCs are cross-domain, (2) labeling training data is extremely expensive. So we propose a flexible framework, called MOOCon, based on semi-supervised machine learning with domain-agnostic features. Experiments demonstrate the efficacy of our framework. Using very a little labeled data can achieve decent performance. We find that various kinds of MOOC content, e.g. subtitles and PPTs, have different modeling ability for concept extraction. So they should be separately treated in future work. Our framework also can be applied to the task of concept identification on MOOC forum content.

In the future, methods of transfer learning and deep learning may be better for extracting cross-domain concepts. External resources of knowledge, e.g. Wikipedia, may be helpful. The relationship between concepts is deserved to be paid more attention for building a domain-specific or even cross-domain concept map.

Acknowledgments. This research is supported by NSFC with Grant No. 61532001 and No. 61472013, and MOE-RCOE with Grant No. 2016ZD201.

References

1. Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J.: Engaging with massive online courses. In: WWW 2014, pp. 687–698 (2014)
2. Bin, Y., Shichao, C.: Term extraction method based on mutual information with threshold interval. In: Zhang, J. (ed.) ICAIC 2011. CCIS, vol. 227, pp. 186–194. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23226-8_25
3. Chang, P.C., Galley, M., Manning, C.: Optimizing Chinese word segmentation for machine translation performance. In: WMT 2008, pp. 224–232 (2008)
4. Collier, N., Nobata, C., Tsujii, J.: Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. Terminology **7**(2), 239–257 (2002)
5. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Zhang, S.S.W.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: KDD 2014, pp. 601–610 (2014)
6. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the c-value/nc-value method. Int. J. Digit. Libr. **3**(2), 115–130 (2000)
7. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: a survey of the state of the art. In: ACL 2014, pp. 1262–1273 (2014)
8. Huang, J., Dasgupta, A., Ghosh, A., Manning, J., Sanders, M.: Superposter behavior in MOOC forums. In: L@S 2014, Atlanta, GA, pp. 117–126, March 2014
9. Jiang, Z., Zhang, Y., Liu, C., Li, X.: Influence analysis by heterogeneous network in MOOC forums: what can we discover? In: EDM 2015, Madrid, Spain, pp. 242–249, June 2015
10. Justesona, J.S., Katza, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Nat. Lang. Eng. **1**(1), 9–27 (1995)

11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML 2001, pp. 282–289 (2001)
12. Liu, A., Jun, G., Ghosh, J.: A self-training approach to cost sensitive uncertainty sampling. Mach. Learn. **76**(2–3), 257–270 (2009)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop at ICLR 2013, pp. 1–12 (2013)
14. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investig. **30**(1), 3–26 (2007)
15. Nojiri, S., Manning, C.D.: Software document terminology recognition. In: AAAI Spring Symposium, pp. 49–54 (2015)
16. Qin, Y., Zheng, D., Zhao, T., Zhang, M.: Chinese terminology extraction using EM-based transfer learning method. In: Gelbukh, A. (ed.) CICLing 2013. LNCS, vol. 7816, pp. 139–152. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37247-6_12
17. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL 2009, pp. 147–155 (2009)
18. Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: CIKM 2004, pp. 42–49 (2004)
19. Sutton, C., McCallum, A.: An introduction to conditional random fields. Mach. Learn. **4**(4), 267–373 (2011)
20. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: HLT-NAACL 2003, pp. 252–259 (2003)
21. Wang, X., Yang, D., Wen, M., Koedinger, K., Rosé, C.P.: Investigating how students cognitive behavior in MOOC discussion forums affect learning gains. In: EDM 2015, Madrid, Spain, pp. 226–233, June 2015
22. Wen, M., Yang, D., Rose, C.: Sentiment analysis in MOOC discussion forums: what does it tell us? In: EDM 2014, pp. 130–137 (2014)