So, H. J. et al. (Eds.) (2020). Proceedings of the 28th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Automatic Entity Recognition based on BERT in Computer Supported Collaborative Learning

Yuanyi ZHEN & Lanqin ZHENG*
School of Educational Technology, Beijing Normal University, China
*bnuzhenglq@bnu.edu.com

Abstract: Knowledge building plays a critical role in promoting knowledge acquisition and facilitating the retention of target knowledge in computer supported collaborative learning (CSCL). The interactive texts in CSCL environment provide a valuable opportunity for instructors to understand and evaluate the knowledge building process and results. Entity recognition for interactive texts is the first vital step in evaluating the level of knowledge building. However, the methods of manual recognition and key-term matching are widely applied, which not only time consuming and lack semantic understanding for interactive texts, but also the accuracy of recognition is hardly guaranteed. We proposed an automatic, accurate combination method to recognize knowledge entity based on a state-of-the-art natural language processing model-BERT (Bidirectional encoder representation from transformers) to understand semantic meaning of interactive texts in CSCL. Text classification and entity recognition are employed in this study. Adopting BERT automatically classify the whole interactive texts into knowledge and non-knowledge types. Levenshtein Distance (LD) and semantic matching based BERT are used to recognize entity from literal and semantic similarity between student interactive texts and entity corpus provided by teachers. Using 16047 interactive texts produced by 51 groups of college students around the strategies of problem-solving in educational psychology are analyzed. The classification accuracy is 90.07%. 7025 knowledge interactive texts were used to automatic entity recognition and F1 value of concept entity and principle entity recognition are 72.02% and 61.18% respectively, while processes entity is 48.75% and examples entity is 44.32%. The automatic combination method shows potential value in assisting teachers in understanding the level of knowledge building and provide feedback timely in CSCL context.

Keywords: Computer supported collaborative learning, entity recognition, text classification, bidirectional encoder representation from transformers, natural language processing

1. Introduction

It has been widely acknowledged that collaborative learning can facilitate knowledge construction, high-rank thinking ability and communication skills (Harassim & Xiao, 2015). Based on network and technological platforms, CSCL supports students to share and construct knowledge through social interaction (Tchounikine, 2019), conducted by interactive texts that are messages students sent to chat groups or discussion board. However, CSCL could not occur spontaneously, which requires teachers supervise student learning process and results to ensure their learning gain. It is really a challenging problem for teachers to track interactive texts in time when they organize large scale online collaborative learning, that is because large number of interactive texts contains many complex topics are produced by students. So that learning results of each group are hard to handle timely and even which knowledge entities are discussed and which are not in each group is hard to know for teachers. Moreover, evaluating students' learning gain in a specific task activity is usually judged by the similarity of the knowledge entity to the knowledge entity provided by teachers. Since teachers usually have the comprehension of domain knowledge closely approximate the true representation of that domain, the similarity to an established knowledge entity can be considered as an indicator for measuring the level of knowledge building (Clariana et al., 2009). Therefore, how to automatically and accurately recognize the knowledge entity referred by students in interactive texts has become a vital research problem in CSCL.

There are three problems need to be solved for automatically and accurately recognizing the knowledge entity referred by students in CSCL. The first one is how to extract the interactive texts which contains knowledge entity. Since CSCL is a complicated process of knowledge construction and social interaction, students would generate task orders for organizing leaning activities and express their emotions when they face difficulties and even conduct social regulation. Besides off-task interactive texts will also be produced during learning process (Ding, 2009). Therefore, how to eliminate the interference of irrelevant knowledge topics is an important task for accurately recognize knowledge entity.

Another problem is that since students would express colloquial terms of one concept or explaining one specific knowledge entity in CSCL, which need us to mine literal and sematic meaning of interactive texts at the same time so that we can recognize entities more comprehensively. Otherwise we would miss the knowledge entities are contained in interactive texts. It is important for students to explain their own understanding about one knowledge entity that is because knowledge building positively affects knowledge acquisition (Draskovic et al., 2004). However, it is hard for machines to understand a sentence whose sematic meaning is equivalent to a knowledge entity.

The last problem is that entity recognition is mainly performed on unstructured texts in knowledge graphs. There are so many data can be used for training model and getting the better result. As for the interactive texts in CSCL, the scale of interactive texts corpus is small, which will cause a negative outcome. Besides the manual labeling of training set is a time-consuming task. Therefore, how to understand the target knowledge entity from literal and semantic level in CSCL based on a training-free method is also a difficulty.

Based on the description above, this work would adopt the combination method of text classification and entity recognition to recognize the knowledge entity in interactive texts in CSCL. This work can contribute to the development of automatic learning monitoring and assessment in CSCL.

2. Related work

2.1 Entity recognition algorithm

In computer science domain, entity recognition methods can be divided into two different cases based on the existence of a knowledge base. When knowledge base exists, entity connection can be used for entity recognition. Otherwise, named entity recognition (NER) is involved. The process of entity recognition has changed from extracting noun like time and names of people, locations, institutes in a single field to an open field (Chinchor & Marsh, 1998). Early entity recognition mainly based on the combination of heuristic algorithms and artificial rules (Rau, 1991), or basing on statistical machine learning (Liu et al., 2011). Now, due to the small scale of data in domain entity extraction, iteration is used for extending entity corpus.

In education domain, term extraction plays an important role in entity recognition, which is determined by the educational entity characteristics. Because of the long-tail of term in education domain, general extracting method is not accurate and comprehensive. Term extraction focuses on the simple terms formed by a single word and compound terms formed by several words. The process of extraction mainly contains two steps. Firstly, obtain the candidate terms based on the unity of strings. Secondly, select the real terms by the entity terminology of candidate entities. Unity is to measure the stability of string collections, while terminology is to measure the speciality of the string combinations in specific field (Kageura & Umino, 1996). The terminology of terms can be judged based on some features, including TF (Term Frequency) method, TF-IDF (Term Frequency-Inverse Document Frequency) method, Information Gain and mutual information. Li et al. (2018) came up with the DRTE method to extract terms from unstructured texts automatically, which is based on the sentence pattern mining using term definition and term relation, associating with morphological rules and boundary detection to extract terms. First thing is text pre-processing, then term definition and term relation are used for selecting terms. Patterns are used for performing definition extraction from texts, which can generate initial candidate terms. Then, use morphological rules and boundary detection to select terms,

limiting the length of terms within 2 to 6 words. According to the part-of-speech table, reduce the strict restriction on part-of-speech matching to obtain a boundary word table and use part-of-speech matching analysis to get the final terms, which will make the number of elements of sentence less then 4. Finally, update the term collections and word segmentation results.

2.2 Entity recognition method in CSCL

In order to measure the level of knowledge building, some researchers have conducted studies on how to recognize knowledge entity based on interactive texts in CSCL. The methods of manual recognition and key-term matching are widely used. Zheng et al. (2015) based on knowledge entity provided by teachers and used the manual recognition method to segment information flow generated by students in CSCL processes. Then the values of proposed knowledge building indicators were automatically calculated. Besides, many attempts have been overcome some limitations of manual recognition, which mainly use automatic method to recognize entity by key-term matching. Hong and Scardamalia (2014) used key terms matching extract knowledge entity from interactive texts in CSCL, which are used to indicate and assess level of knowledge building. Zheng et al. (2018) used the key-term-based method to recognize entity and the specific process is as follows: split student's interactive texts in Chinese by open-source splitting software ICTCLAS (Zhang et al., 2013); replace the key terms by terms in synonyms dictionary and extract the key terms based on knowledge entity provided by the teachers.

Previous studies have indicated that the manual method and the key-term matching method are widely used to recognize knowledge entity in CSCL. However, manual method not only strongly depend on artificial coders, but it is also a post event approach and its results can only be seen when students finish their collaborative learning, which cannot provide real time feedback for teachers and students. Key-term-based method cannot mine semantic meaning of interactive texts in CSCL, which cannot detect interactive texts without key-terms although they have same meaning, and that will lead to the recall rate of the entity recognition is low. In conclusion, it can be noticed that an efficient, real-time and accurate method for automatically recognizing the knowledge entity produced by students in CSCL is lacking.

3. Methodology

The combination method of entity recognition is as *Figure 1*. In order to conduct entity recognition for knowledge interactive texts. Text classification is conduct firstly, which divide whole interactive texts into knowledge texts and non-knowledge texts. Text classification and entity recognition are included in research method. The specific process in each part can be seen in 3.1 and 3.2.

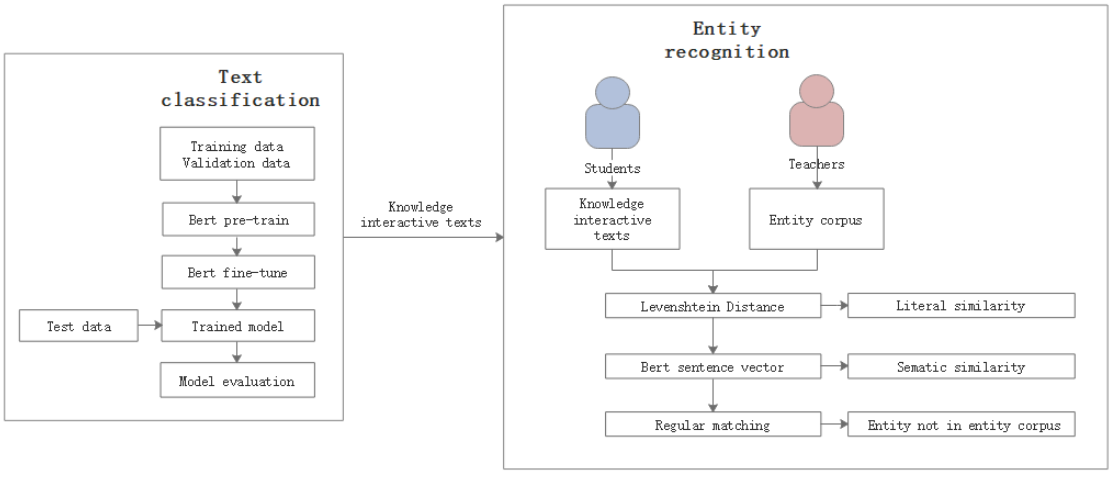

Figure 1. Entity Recognition Method

3.1 Text classification

3.1.1 Classification rules

In order to eliminate the inference of non-knowledge interactive texts, we need to conduct the text classification firstly. All the interactive texts in CSCL can be divided into two kinds, including knowledge kind and non-knowledge kind.

From existed researches, knowledge interactive texts generated in the process of knowledge building which has been defined as organizing, restructuring, interconnecting and integrating new information with prior knowledge (Kalyuga, 2009). Therefore, knowledge texts include explanation, examples and application about knowledge entity. While greeting, manage instruction, confused expression and off-task texts are belong to non-knowledge. There are some examples of knowledge texts and non-knowledge texts in Table 1.

Table 7. *Knowledge and Non-knowledge Interactive Texts*

Type	Interactive texts
Knowledge	"Problem solving strategies include algorithmic and heuristic." "In the whole problem-solving process, we should seek more possible methods after analyzing the problem." "For example, take a beautiful photo, it is considered a poor-structure problem."
Non-knowledge	"Hello every, my name is Yixin." "Let's turn to the next question." "We haven't solved the second problem." "I'm so hungry."

3.1.2 BERT model

BERT model is pre-trained by deep bidirectional representations in unlabeled text and can be fine tuned with one output layer. And it creates state-of-the-art on many natural language processing tasks including text classification (Devlin, 2018). Multi-layer bidirectional transformer encoder can be seen in the model architecture since transformers have better ability to save training time and pay more attention to important section.

BERT model can be divided into Pre-training part and Fine-Tuning part. Masked Model and Next Sentence Prediction are used to train word vector in Pre-training part. The encoder part of the

transformer model is used for training unit, the start characters *[cls]* and stop characters *[sep]* are labeled, then the word vector is output. Words in a sentence is randomly covered or replaced in the Masked Model and then the model predicts the masked words by the context. In the replacement part, Loss calculation only calculates the loss of the covered part. In the NSP task, it mainly calculates whether two sentences are matching.

3.2 Entity recognition

Knowledge texts can be extracted by text classification and then entity recognition is performed on them. According to previous research (Yang, 2011), entity have four types which are concept, principle, process and example. The corpus of knowledge entities and entity types is provided by teachers in advance and students need to engage them. Therefore, the main task is how to recognize knowledge entity and their type from knowledge interactive texts based on entity corpus provide by teachers.

Since the generative characteristics of CSCL, students are required to produce new opinions contain new knowledge entities to show their knowledge building which maybe not include in entity corpus provided by teachers. It is reasonable that entity corpus may not completely cover all entities generated by students during the learning process. Our approach is that judging a knowledge interactive text whether contain one or more knowledge entities in the corpus firstly. Then if it contains, we replace the interactive texts with the matched knowledge entities. Otherwise, we need to mark the entity type based on the language template and text length appearing in the text, then it was saved as an entity.

Among them, our approach mainly includes text similarity calculation and regular matching two parts. Text similarity calculation is used to match the knowledge entity with the entity corpus provide by teachers in advance. While the regular matching is used to recognize the knowledge entity which is not exit in entity corpus. Besides, text similarity calculation includes distance method and semantic similarity calculation, which the first one is used to match entity on literal meaning and the second one is used to match them based on sematic meaning.

3.2.1 Text similarity calculation

Levenshtein Distance. Levenshtein Distance (LD) was proposed by Soviet mathematician Vladimir Levenshtein and it is also known as Edit Distance, which is mainly used to compare the similarity of two string. Levenshtein distance refers to the mini-mum number of operations required to convert a string of sequences through insertion, deletion, replacement into another string. The smaller the edit distance, the greater the similarity between the two strings (Li & Liu, 2007). It is widely used in comparing the literal similarity of two short texts.

Semantic matching. Only literal text similarity calculation cannot completely match the entity described by the students with the knowledge corpus, so the method of semantic similarity calculation between two texts needs to be introduced. In this work, BERT sentence vector is chosen as the text representation, which can map a variable length sentence to a fixed length vector. The first text is come from the entity corpus provided by teachers, while the second text is a knowledge interactive text in learning process. Then two text vectors are calculated respectively. Sentence vector similarity is measured using the cosine value. If the cosine value between two sentence vectors is greater than 0.85, it indicates that two sentence vectors are similar (Zhang et al., 2011), which means that the knowledge interactive texts contains the corresponding knowledge entity, otherwise no knowledge entities are contained in it.

Steps of using BERT sentence vector refer to the service provided by BERT-Service[1], and Python is used for performing sentence vector encoding.

[1] BERT-as-service homepage, https://github.com/hanxiao/BERT-as-service#book-tutorial, last accessed 2020/8/30.

3.2.2 Regular match

Since knowledge semantic interactive texts are not similar to the knowledge entity in the entity corpus provided by the teacher through the text similarity calculation, that means that they are new knowledge entities generated by students during learning process. The processing method is to obtain the entity type of the entity through regular matching and text length calculation. Among them, the two types of entity are identified by concept and example when performing regular matching. The template used is "is/are" represents the elaboration and explanation of concepts, and "such as", "for example" represents an example of the facts is presented. After extracting the corresponding entity, we calculated the entity length and presented entities with its length to three domain experts. Experts decided that if the text length is greater than 15 and less than 30, it is classified as a process. If the text length is less than 15, it is classified as a principle. If the text length is greater than 30, they are classified as examples.

4. Experiment and Results

4.1 Data

This research selected data from CSCL platform developed by our laboratory. There are 51 groups participating in learning activity and each group has 4 people. The learning task is to discuss problem-solving strategy in educational psychology, which is same for each group. The task includes 5 parts which include the strategy for problem-solving, the difference between experts and novice in problem-solving, how to develop students' capability in problem-solving, how to conduct knowledge construction based on problem-solving and the process of ill-structured problem-solving. Before the activity, group members are free to choose one of four roles (coordinator, interpreter, summarizer and information collector). The average time of this CSCL activity is 2 hours for each group.

There are 16047 interactive texts produced by 51 CSCL groups, and 315 texts for each group in average. In order to ensure our dataset to obey the real situation and improve the generalization ability of model, all the original data are preserved. Among all the 16047 texts the data are divided into training set, validation set and test set. 70% of the data in each category is selected as training set, 20% as validation set and 10% as test set. Then two parts of data are combined as the whole training set and test set. The experimental statistics of knowledge and non-knowledge are shown in Table 2.

Table 8. *Corpus Distribution of Experimental Dataset*

Type	Training set	Validation set	Test set	Number
Knowledge	4918	1405	702	7025
Non-knowledge	6317	1805	900	9022
Amount	11235	3210	1602	16047

Besides, 57 knowledge entities in entity corpus in Chinese are provided by teachers and it includes 13 concept entities (CN), 11 principle entities (PF), 31 process entities (PS) and 2 example entities (FC). And part of 57 knowledge entities are shown in Table 3.

Table 9. *Part of 57 Knowledge Entities Corpus*

Concept	Principle	Process		Example
Ill-structured problem	Unclear start	Knowledge construction based on problem-solving	Build a good knowledge architecture system	Bridge-crossing problem
Problem	Unclear end	Training of the ability of problem-solving	Differences in understanding and	ATM machine

			characterizing problems
Will-structured problem	Unclear method	Difference between expert and amateur	Differences in speed of problem-solving
Heuristic method		Strategy of problem-solving	Differences in focus in problem-solving
Algorithmic method		Process of ill-structured problem-solving	Differences in monitoring the problem-solving process
Mean-goal analysis		Consolidation of original knowledge and skills	Consider each situation and list each possibility

4.2 Experiments

The computer environment of research is 64 bites Ubuntu operation system, Intel(R) Xeon(R) CPU E5-2620 v4.

4.2.1 Text classification

Firstly, two annotators labeled one third data as 0 and 1 based on whether it belongs to knowledge text or not, while knowledge texts were labeled as 1 and non-knowledge texts as 0. Then we computed the consistency of them, Kappa-value is 0.90. At last, two thirds of the rest data were labeled by two annotators respectively.

Secondly, we modified BERT pre-trained model architecture based our experiment data and chose the pre-training model of Bert_chinese_L-12_H-768_A-12. In the fine-tune part, we set max sequence length is 256, train batch size is16 and learning rate is 1e-5 according to the previous studies. Thirdly, we predicted the test set by using trained BERT model. Finally, we got the classification result.

4.2.2 Entity recognition

Firstly, we calculated LD between the knowledge interactive texts and knowledge entity in entity corpus provide by teachers based on a Python package called FuzzyWuzzy, which is a fuzzy string matching package. If the value of LD greater than 85, that means knowledge text produced by students contains the correspond knowledge entity in entity corpus (Zhang & Cui, 2020). Secondly, we mined the sematic similarity between knowledge interactive texts by students and knowledge entity from teachers, which include two parts: got the vert sentence vector by BERT-service and calculated the cosine value between two vectors. Finally, we recognized the entity not exit in entity corpus but produced by students with rules and text length.

4.2.3 Evaluation

Evaluation of text classification and entity recognition typically employs the following three metrics: Precision, Recall and F1-measure (F1). The standard indicators of Song et al. (2014) are employed to measure the evaluation.

4.3 Results

4.3.1 Text classification

The text classification result can be seen in Table 10. The F1 value of knowledge text is 88.40% and non-knowledge texts is 91.33%. The total accuracy is 90.07% which is higher to make application in

real education situation. From the confusion matrix of BERT classification model (see *Figure 2*), we can see the value of diagonal is higher than others which means the classification result is great.

Table 10. *BERT Algorithm Classification Result*

Indicators	Types		Accuracy
	Knowledge	Non-knowledge	90.07
Precision	90.58	89.71	
Recall	86.32	93.00	
F1	88.40	91.33	

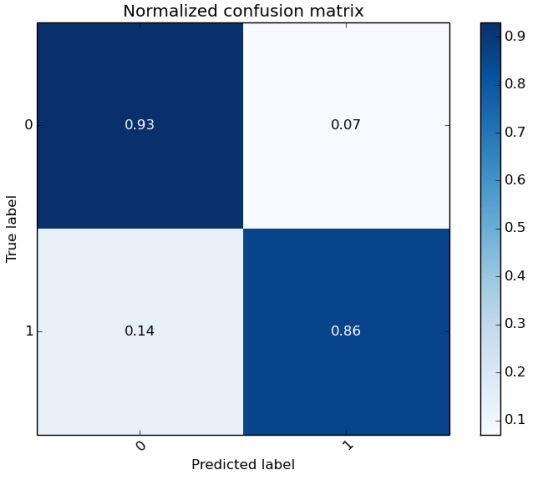

Figure 2. Confusion Matrix

4.3.2 Entity recognition

7025 knowledge interactive texts were used to further recognize entity and the result can be seen in Table 11. The F1 value of CN (concept) and PF (principle) are 72.02% and 61.18% respectively which is obviously higher than F1 value of PS (Process) and FC (Example) with 48.75% and 44.32%.

Table 11. *Entity Recognition Result.*

	Recall	Precision	F1
CN	81.77	64.35	72.02
PF	59.01	63.53	61.18
PS	60.56	40.79	48.75
FC	35.34	59.43	44.32

5. Discussion and Conclusions

Entity recognition is the most vital step for automatically assess students' knowledge building level, which can assist teacher monitor the discussion process and provide learning support timely. This study proposes an automatic, accurate combination method to recognize knowledge entity based on BERT from interactive texts in CSCL context. Text classification and entity recognition are employed in this study were found to be effective in recognizing entity. Text classification is conducted based on BERT, which create a state-of-art in lots of natural language processing tasks including text classification. Literal and semantic similarity between interactive texts and entity corpus provided by teachers are calculated to get matched entity. LD was used to calculate the literal similarity while BERT-service was

used to produce sentence vector and the cosine value between two sentence vectors can indicate the degree of semantic similarity. Besides we can also detect the knowledge entity generated by students themselves rather than required by teachers.

Based on 16047 interactive texts from an educational psychological content problem solving strategy from 51 groups. We adopt the combine method of text classification and entity recognition for recognize the entity in interactive texts. The result is shown that method is valid and we can automatically and accurately. Knowledge interactive texts can be classified accurately, and CN and PF entities can be detected precisely while PS and FC entities cannot be recognized properly. The reason of that is students take lots of examples in knowledge building and it is hard to recognize based on entity corpus while process entity is to explain the knowledge in a subjective procedure way and its similarity with entity corpus is lower than concept entity and principle entity. This method will be especially beneficial for teachers to handle the learning process or learning result of students when they are facing a large number of groups in CSCL context.

One limitation of this study is that the accuracy of process entities and example entities recognition is lower. Students will say more explanation or take more examples which are not in entity corpus or have far semantic distance with them. In all, our model's ability of detecting new generated entity produced by students is a little weaker and how to improve its' performance in processes and examples is our future works. For future studies, more attempts should be taken in the optimization of entity recognition model and the improvement of recognition accuracy. So that we can give more powerful supports for teachers to handle learning process of students accurately and timely.

Acknowledgements

This study is funded by the National Natural Science Foundation of China (61907003) and Faculty of Education, Beijing Normal University (1912102).

References

BERT-as-service homepage, https://github.com/hanxiao/BERT-as-service#book-tutorial, last accessed 2020/3/2.

Chinchor, N., Marsh, E. (1998). Muc-7 information extraction task definition. In *Proceeding of the seventh message understanding conference* (MUC-7), Appendices, 359-367.

Clariana, B., Wallace, E., Godshalk, M. (2009). Deriving and measuring group knowledge structure from essays: The Effects of anaphoric reference. *Educational Technology Research and Development*, *57*(6), 725-737.

Devlin, J., Chang, M., Lee, K., et al. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805.

Ding, N. (2009). Visualizing the sequential process of knowledge building in computer-supported collaborative problem solving. *Computers & Education*, *52*(2), 509-519.

Draskovic, I., Holdrinet, R., Bulte, J., Bolhuis, S., Van Leeuwe, J. (2004). Modeling small group learning. *Instructional Science*, *32*(6), 447-473.

Harassim, L., Xiao J. (2015). The fundamental guarantee of the quality of collaborative learning theory and practice online education. *Distance education in China*, *8*, 5-16.

Hong, H., Scardamalia, M. (2014). Community knowledge assessment in a knowledge building environment. *Computers & Education*, *71*, 279-288.

Kageura, K., Umino, B. (1996). Methods of automatic term recognition: A review Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication*, *3*(2), 259-289.

Kalyuga, S. (2009). Knowledge building: A cognitive load perspective. *Learning and Instruction*, *19*(5), 402–410.

Li, S., Xu, B., Yang, Y. (2018). DRTE: Terminology Extraction Method for Primary Education. *Journal of Chinese Information Processing*, *32*(3),101-109.

Li, Y., Liu B. (2007). Normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6): 1091-1095 (2007).

Liu, X., Zhang, S., Wei, F., Zhou, M. (2011). Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Vol. 1)*. Association for Computational Linguistics. 359-367.

Rau, L. (1991). Extracting company names from text. In *Proceedings of The Seventh IEEE Conference on Artificial Intelligence Application (Vol. 1)*, IEEE, 29-32.
Song, G., Ye, Y., Du, X., Huang, X., Bie, S. (2014). Short text classification: A survey. *Journal of Multimedia*, *9*(5), 635-643.
Tchounikine, P. (2019). Learners' agency and CSCL technologies: towards an emancipatory perspective. *International Journal of Computer-Supported Collaborative Learning*, *14*(2), 237-250.
Yang, K. (2011). On comprehensibility of curriculum and knowledge modeling technology. *Research on audio-visual education*, *6*, 12-16.
Zhang, H., Liu, Q., Cheng, X., Zhang, H., Yu, H. (2013). Chinese lexical analysis using hierarchical hidden Markov model. Paper presented at *The Second SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.
Zhang, L., Cui, R. (2020). A word order sensitive similarity measure based on edit distance. *Journal of Yanbian University (Natural Science)*, *46*(2): 140-144.
Zhang, Z., Xu, X., Long, J., Yuan, X. (2011). Parameters correlation and optimization in text similarity measurement. *Journal of Chinese Computer Systems*, *32*(5), 983-988.
Zheng, L., Huang, R., Hwang, G. J., Yang, K. (2015). Measuring knowledge building based on a computer-assisted knowledge map analytical approach to collaborative learning. *Educational Technology & Society*, *18*(1), 321-336.
Zheng, Y., Xu, C., Li, Y., Su, Y. (2018). Measuring and Visualizing Group Knowledge building in Online Collaborative Discussions. *Educational Technology & Society*, *21*(1), 91–103.