NLP-NITMZ @ CLScisumm-18

Dipanwita Debnath[1], Amika Achom[1] and Partha Pakray

Department of Computer Science & Engineering
National Institute of Technology Mizoram[1]
Aizawl, India
{parthapakray,ddebnath.nita,achomamika01}@gmail.com

Abstract. This paper report NLP-NITMZ @ CL-Scisumm 2018 system participation for the shared task task 1A, task 1B and task 2 at the BIRNDL 2018 Workshop. We developed our system based on the previous years data provided by the organizer. For task 1A and 1B, we apply various rule based approaches and trained the K- Nearest Neighbors Classifier (KNN) using different features identified from the input citation text and the reference Text. We achieved an overall accuracy score of 42.75 (%) and 78.75 (%) score for task 1A and task 1B. For task 2, We built our summary generation system using OpenNMT tool. We developed the model using the training and development datasets released from previous year tracks and validated the results of our system using previous years test data set. We have evaluated our system using Recall-Oriented Understudy for Gisting Evaluation, (ROUGE) score on some test data of CL-Scisumm 2017. For task 2, we achieved an overall accuracy score of 37.75 (%).

Keywords openNMT, citation based summarization, rule based method, K- Nearest Neighbors Classifier , KNN, ROUGE metrics

1 Introduction

Generating automatic summaries of scientific papers is one of the most important challenging tasks in the field of summarization. Research articles need to be summarized to provide the reader a brief glance of the paper. citation sentences provide all the useful information about the reference paper. There are a lot of shared tasks like TAC 2014 Biomedical Summarization Track, CL-Scisumm 2016 , CL-SciCumm 2017 and CL-Scisumm 2018. These all shared tasks provide new challenges every year. It motivated Researchers to share their own theories and methodology in the field of scientific paper summarization.

This short paper will highlight the participation of our NLP-NITMZ @ CL-Scisumm 2018 system for the shared task at the BIRNDL 2018 Workshop. We would elaborately explain the methodologies, techniques and the achievement of our system in this Shared Task of scientific summarization. We first identified various cited text span or citation. For example the cited text like (Ceylan et al 2010) or (Clarke et al 2010) or (Navigli, 2009, WSD) are identified. These all citations contains year as common . So we extract the phrase or the sentence containing the citation using the NLTK regular expression pattern matching. The second task of our strategy involves the identification of the reference Paper. From the extracted citances, we identify the topics name of the reference paper. This is done through the LDA model of Word2vec using Gensim package. For example, consider the cited text spans, ”In addition, dis-discriminative weighting methods were proposed to assign appropriate weights to the sentences from training corpus (Matsoukas et al, 2009) or the phrase pairs of phrase table (Foster et al, 2010)”, reference paper was identified to be (Matsoukas et al, 2009), so we extract the phrase ”In addition, discriminative weighting methods were proposed to assign appropriate weights to the sentences” from training corpus . We then extract feature from the input citation texts and the reference paper which are then trained on the K Nearest Neighbor Classifier (KNN) to group the similar sentences in one cluster that cited the reference paper. It is reported in [5], the authors applied the automatic identification of cited text spans using a multi-classifier approach. So we identified the facet of our citation text accordingly into different classes or categories as aim section, method section, implementation section, aim and method section, result section and implication section. Finally we trained the Open NMT system to summaries the reference paper and to generate a brief summary of the reference paper using the citances. Each year the open task on scientific paper summarization enhances important feature. The following section discussed the Task presented at CL-Scisumm 2018. Following are the task assigned at BIRNDL 2018 Workshop.

Task 1A: For each citances in the citing papers (i.e. text spans containing a citation), identify the cited spans of text in the reference paper that most accurately reflect the citance.

Task 1B:

For each cited text span, identify which discourse facet it belongs to, among the following facet, namely Aim Citation, Result Citation, Method Citation, Implementation Citation.

Task 2: Finally, an optional task consists on generating a structured summary of the reference paper with up to 250 words from the cited text spans.

In this work we report and present the systems developed at NLP-NITMZ in participate at the CL-SciSumm 2018. We further explain the methodology and architecture developed for this shared task. We evaluate our system and compared with the previous different runs of participation for the shared Task on the previous year dataset.

The paper is organized as follows: Section 1 presents the literature survey on this field of work. Section 2 reports the system description and the architecture for developing the system. Section 3 shows an elaborate explanation on the approaches used for developing the system on scientific paper summarization. Section 4 discusses the evaluation results and section 5 and section 6 concludes our work with an aim for the future work.

Related Works

Determining the similarity between sentences is one of the crucial task which have a wide impact in many natural language processing applications. In scientific document summary generation task, summary generation largely depends on the accuracy of sentence similarity. More accurately we identified the correct sentences of Reference papers (RPs) by giving a query or keyword from the cited text span (CPs), more accurately so that we can generate the summary. The system NJUST@CLscisumm17 used different similarity measure as features for example LDA, JACCARD, TF-IDF, DOC2VEC similarity and used SVMBBF , SVM linear Decision tree and logistic regression to extract the best accurate sentences from RPs for Task 1A. For task 1B they have used dictionary based facet selection and for task 2 they used bisecting M-Mean and maximal marginal relevance (MMR) for summarization. PKC@CLscisumm17 used search based methods with features like threshold based TF-IDF, word2vec from genism and word mover distance to find the sentence similarity and used conditional probability to identify the RPs text span. Metzler et al. evaluate the performance of statistical translation models in identifying topically related sentences compared to several simplistic approaches such as word overlap, document fingerprinting, and TF-IDF measures. Lei Liyuan Mao et in [2], implemented task 1A and 1B using rule-based methods with various features of lexicons and similarities and trained the system using SVM classifier. For Task 2, hLDA topic model is adopted for content modeling, which provides us the knowledge about sentence clustering (subtopic) and word distributions (attractivenesses) for summarization. We further study the implementation of an unsupervised summarization technique, TextSentenceRank in [1] that helps in finding the similarity of sentences to the citation on a textual level. This paper employed the classification method method to select the original text sentences from the candidates texts using the TextSentenceRank algorithm. Also in [1] works has proven the implementation of unsupervised summarization of the relevant sub-part of the document that was previously selected in a supervised manner. In [4], the authors applied the Learning to Rank algorithm with multiple features, including lexical features, topic features, knowledge-based features and sentence importance, to task 1A by regarding the reference span. They viewed the approach as Information Retrieval method. They viewed the task 1B as the discourse facet identification which falls under the text classification problem by considering features of both citation contexts and cited spans. In [3], CIST@ CLSciSumm-17 used multiple features based on citation linkage for the classification and summarization approaches.

System Architecture

This section describes the description of our system used for our work. The Extractor module extract the cited text span. In this module the cited text span like (Ceylan et al 2010) or Clarke et al (2010) or (Navigli, 2009, WSD) are identified. We identified the common word. In these examples year is common. so this serves as a matching pattern to identify and extract the corresponding texts from the reference Text. So we extract all the sentences containing the year as a key terms from the Reference Text. This is done through the RegexpParser(pattern) module imported from NLTK. The next step involves the identification of feature from the input cited text span and the Reference text. We used Lexical and syntactic features for this. We used the n-gram matching paradigm like the uni-gram matching, bi-gram matching and n gram matching. We used the Word2Vec model of the Gensim. Cosine similarity, LDA are also being employed. The identified feature are then feed into the K-Nearest Neighbor Classifier (KNN) classifier and then trained the dataset. After training the classifier and using some rules based method, we classified the citation text into the possible facet like Aim citation, Aim and method citation, implication citation and method citation, method citation, Implication citation, and Result citation. We employ a rule based approach here. If the text is generated from the beginning section of the RPs, it is classified and assigned the facet Aim Citation. Similarly, if the sentence from RPs is extracted both from the Aim and Method

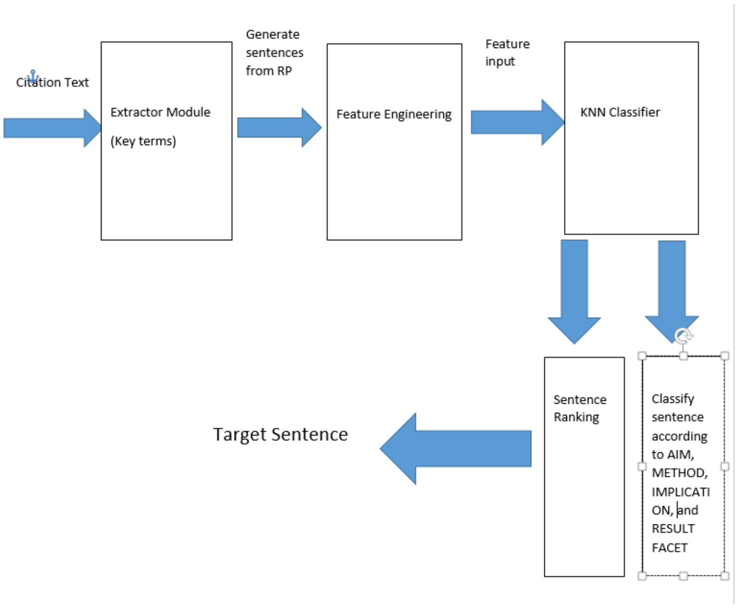

Fig. 1 – System Architecture for Task 1A and Task 1B

Citation, We classified the text sentence into the Aim and Method Citation. Also correspondingly, if the identified sentence from the RPs is generated from the implication section, it is assigned the implication facet. If the generated sentence from the reference text describes about the aim or the method, we assign the facet of the sentence as the method facet. If the generated sentence is generated from any section other than abstract and result section, we classify the sentence into the method and implication facet. Lastly the generated text sentence as classified into the result and method citation if the sentences describes the method and result description.

Methodology

The main step of any work is cleaning or pre processing the datasets. we pre processed the datasets released from the previous year from CL-Scisumm 17, CL-Scisumm 18. For this we used the NLTK preprocessing module. We clean, tokenized, extract pattern from the reference paper to extract those sentences that contain the cited text. The various task undertaken to complete Task 1A, Task 1B and Task 2 are enumerated:

For Task 1A: In task 1A of CL-Scisumm 18, we find the clean cited text span to identify cited spans of text in the reference paper that most accurately reflects the citation. Here we choose only those sentences or portion of sentence that contain citation to the actual reference paper. We applied different preprocessing steps to identify the key contributing term so as to extract the cited text span. The steps are undermentioned:

Extraction of cited text span

Identification of Cited text span: Here the cited text spans containing the citation like (Ceylan et al 2010) or (Clarke et al 2010) or (Navigli, 2009, WSD) are identified. These all citation text contains year 2010 as common . So we first extracted only those lines containing the citation using Regular expression regexp(pattern) from NLTK.

Identification of the actual reference

After the extraction and analysis of citation text, we identified the reference papers topic name and then extract only those sentence or portions of sentences that belong to the actual reference paper. For this we used the LDA model in Gensim. For example, considering the cited text spans In addition, discriminative weighting methods were proposed to assign appropriate weights to the sentences from training corpus (Matsoukas et al, 2009) or the phrase pairs of phrase table (Foster et al, 2010), we find that the reference paper was (Matsoukas et al, 2009), so we extract the phrase In addition, discriminative weighting methods were proposed to assign appropriate weights to the sentences from training corpus (Matsoukas et al, 2009). We only choose one referenced sentence at a time and for each folder we made a text file containing all the sentence of the cited papers that contains citation to the Reference paper. We compute the TF-IDF score of the query citation text and the reference paper. The top scoring key terms are used as an index query keywords to identify and extract the sentences from the reference paper. From the extracted sentences and the input citation sentences, we engineered out different feature like the n gram matching and apply a rule based method on it. The distance similarity function like Cosine similarity function, LDA score, Word Mover Distance are being used as different feature. We then trained the K Nearest Neighbor (KNN) classifier using the extracted features. The classifier group the similar sentences on cluster by cluster basis. We then further classified the sentences and identified their facet using the rule based approaches.

Top scoring sentences are selected from the reference paper based on their Jaccard similarity score and TF-IDF score. For task 1B, we classify the classification and facet identification into the following subclasses.

1. Aim Citation : If the text is present on the location of the beginning of the paper.
2. Aim Citation, Method Citation: Similar sentence in the aim and method section we used both the term related to method and in present future tense.
3. Implication Citation : In the Introduction or Method section we used some terms related to implementation of a method or technique or dataset.
4. Method Citation : In the aim or method section we used some terms related to method or technique.
5. Method Citation, Implication Citation : Any section other than abstract and result section. Details about the method and its implementation
6. Result Citation, Method Citation : If the text is present in either the method or result section and contains some result and method describing term.
7. Result Citation : In the result or any other section containing some numbers with percentage or the term accuracy, performance, score etc.

Task 2 : Summary generation

From the sentences generated from the task 1A we combined all the sentence of reference sentence that are cited in the cited papers to create an extractive summary and abstractive summary.

Extractive text Summarization:

In Extractive text summarization approach, we applied three simple rules to generate summary as the text itself is short. We have ranked the generated sentences from reference paper a score based on Jaccard similarity score between all the cited text and reference text. We also considered sentence length and location, where in summary there should be at least a sentence from introduction, Implementation, methods and results. And after Task 1A we ranked the sentences as they are refereed in cited papers. Based on these three criteria we finally selected each sentence from all section i.e. Introduction, Implementation, Methods and Results. And if the length is not exceed to 250 we added more sentence based on similarity score.

Abstractive text Summarization:

We build our system model based on CL-Scisumm 16 and CL-Scisumm 16 and CL-Scisumm 17 data and DUC1 dataset and DUC1 dataset. We made four file i.e. training text data, training summary data, validation text data, and validation summary data. We preprocessed all the data using OpenNMT. For summary generation, we used preprocessed output of Task 1A's as test data and summary is generated from this by translating using OpenNMT . After we interpreted the summary generated by the OpenNMT based on CL-Scisumm 16 and CL-Scisumm 17 data. The DUC1 dataset is not an accurate one because of less data. For NMT to work properly and give result accurately, we need to train the OpenNMT system with more data.

System Result and Observation

We build our system using the dataset released from CL-SciSumm-18, CL-Scisumm 17 and CL-Scisumm-16 datasets and tested the result of our system on the Scisumm-17 datasets. We evaluated the result of our system runs using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score or metrics.

Table 1 – Result of the run on test dataset 2017

Sl no	Sentence Id	Score	Average similarity Score	Accuracy
1	Sentence1	0.0526	0.130256	42.105%
2	Sentence2	0.0625		
1	Sentence3	0.42105		
2	Sentence4	0.0625		
1	Sentence5	0.05263		

From the table given above, we can conclude that for Task 1A, we got accuracy 42.105 score in terms of % on some the data from CL-Scisumm 17 dataset. For Task 1B we got accuracy 78.75 % score. For Task 2, using OpenNMT we got an accuracy score of 37.75 %.

Table 2 – Result of the run on test dataset 2016

Sl no	Sentence Id	Score	Average Similarity Score	Accuracy
1	Sentence1	0.02225	0.124352	36.721%
2	Sentence2	0.1021		
1	Sentence3	0.36721		
2	Sentence4	0.10240		
1	Sentence5	0.0278		

Conclusion and Future Works

In the succeeding future track, We would like to train and build our system incorporating more lexical, syntactic and semantic feature. We would extend our work by training multi classifiers and thereby improving the accuracy score of our developed model. Moreover, We would like to study the dataset imbalanced problem that we encountered while experimenting the system and would like to handle and address the solution to this problem.

Acknowledgments We would like to express our deepest appreciation to the Department of Computer Science and Engineering, National Institute of Technology, Mizoram for providing and assisting us all the required and necessary financial assistance, laboratory, and the technical facilities for conducting out our full experimental research works on this research paper work for the participation at CL-SciSumm Shared Task 2018 @SIGIR2018: The Scientific Summarization Shared Task.

References

1. Klampfl, S., Rexha, A., Kern, R.: Identifying referenced text in scientific publications by summarisation and classification techniques. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). pp. 122–131 (2016)
2. Li, L., Mao, L., Zhang, Y., Chi, J., Huang, T., Cong, X., Peng, H.: Cist system for cl-scisumm 2016 shared task. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). pp. 156–167 (2016)
3. Li, L., Zhang, Y., Mao, L., Chi, J., Chen, M., Huang, Z.: CIST@ CLSciSumm-17: Multiple features based citation linkage, classification and summarization. In: Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL2017). Tokyo, Japan (August 2017) (2017)

4. Lu, K., Mao, J., Li, G., Xu, J.: Recognizing reference spans and classifying their discourse facets. In: Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL). pp. 139–145 (2016)
5. Ma, S., Xu, J., Zhang, C.: Automatic identification of cited text spans: a multi-classifier approach over imbalanced dataset. Scientometrics pp. 1–28 (2018)