CYUT Team2 SimpleText Shared Task Report in CLEF-2022

Shih-Hung Wu, Hong-Yi Huang

Chaoyang University of Technology, *Taichung, Taiwan (R.O.C)*

Abstract

This paper reports our approach to the SimpleText lab. For the task 1: what is in (or out)?, we designed a two-stage filtering scheme that utilizes the traditional keyword finding approach TF-IDF score to find the important documents in the first stage and the important sentences in the second stage. The result is comparable to manual run and ranked first in task 1. For the Task 3: Rewrite this!, our system adopts the T5 generation model to rewrite the original sentences. We fine-tuned the model to generate simplified sentence. The result ranked second in task 3. However, the simplified sentence cannot fully express the meaning of the original sentence, more fine-tuning is necessary.

Keywords 1

Simple Text Generation, TF-IDF, T5 model

1. Introduction

Interpreting scientific texts requires solid background knowledge and uses tricky terminology so that the scientific texts are hard to understand. How to simplify complex text in an automatic way is the key point of research. In CLEF-2022 SimpleText Lab [1] provides tasks to promote the research of text simplification. The goal of research is to make scientific texts more comprehensible to the general public in an automatic manner. SimpleText provides challenges of automatic text simplification in the following tasks:

- TASK 1: What is in (or out)? The goal of task 1 is given a query, a system has to find passages to include in a simplified summary.
- TASK 2: What is unclear? Given a passage and a query, a system has to rank terms that are required to be explained for understanding this passage.
- TASK 3: Rewrite this! Given a passage from scientific abstracts, a system has to rewrite it into a simplify passage.

SimpleText aims find the textual expression carrying information that should be simplified, the background information should be provided and the most relevant or helpful. Also system should try to improve the readability of a given short text.

In this year, we focus on Task1 and Task3 with the techniques from other related works.

2. Techniques in Our Approach

Our system uses the TF-IDF score to find the important sentences in a two-stage filtering scheme for task 1, and adopts the T5 generation model to rewrite the original sentences for task 3, the detail is given in section 4. Here, we will give a brief introduction to TF-IDF and T5 model.

[1]CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
EMAIL: shwu@cyut.edu.tw (A. 1); s11027604@gm.cyut.edu.tw (A. 2)
ORCID: 0000-0002-1769-0613 (A. 1)

2.1. Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency Inverse Document Frequency (TF-IDF) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. TF-IDF is calculated by multiplying two different metrics. The **term frequency (TF)** means the number of times the word appears in a document. The **inverse document frequency (IDF)** means, how common or rare a word is in the entire document set. The TF-IDF score for the word *t* in the document *d* from the document set *D* is calculated as follows:

$$tf\ idf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (1)$$

$$tf(t,d) = \log(+freq(t,d)) \quad (2)$$

$$idf(t,D) = \log\left(\frac{N}{count(d \in D: t \in d)}\right) \quad (3)$$

2.2. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer(T5)

Transformer-based models have achieved state-of-the-art performance for abstractive summarization [2][3][4]. T5, or Text-to-Text Transfer Transformer [2], is a Transformer based architecture that uses a text-to-text approach. T5 can convert all NLP tasks into Text-to-Text. The framework is shown in Figure 1. Our Task3 system is built on T5 model.

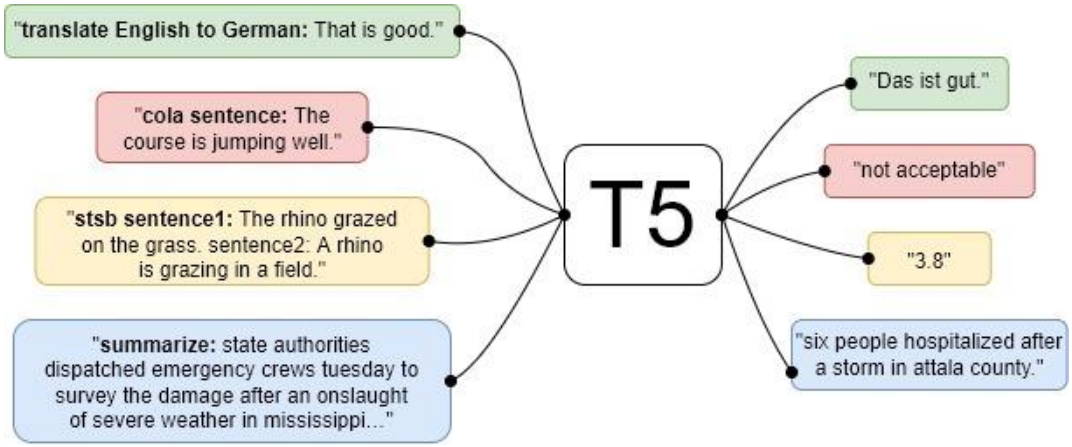

Figure 1: Text-to-text framework of T5 model [2]

3. Data set

SimpleText's data use the Citation Network Dataset: DBLP+Citation, ACM Citation network (12th version) as source of scientific documents to be simplified. The data is two-fold: Medicine and Computer Science. Scientific textual content and authorship on any topic related to computer science can be extracted from this corpus. Detail description please read the overview paper [5].

4. System

Since we focus only Task1 and Task3, here we give the detail of our system in task 1 in sub-section 4.1 and task 3 in sub-section 4.2.

4.1. Search passage using a two-stage TF-IDF filter

In Task1, the system uses TF-IDF score to filter the article and find the top 5 sentences matched by the query term. The flowchart is shown in Figure 2. The query term is normalized, and then the abstract matches it is extracted. If there are too many matched files, our system will ranking them by TF-IDF score to find the Top 5 files. Since only single sentence in the article is required, the TF-IDF score is calculated again after separating each sentence in the article, and the sentence with the highest TF-IDF score in each file is found, and the Top 5 sentence is obtained. Note that we limit article matching because we want to reduce the number of files, and when the conditions are true, the matching criteria are changed to abstracts and titles instead of just abstracts. In addition, when there is no matching document, we will split the query terms into single words to match the file, and when calculating the TFIDF scores, our system will calculate them separately and then take the sum.

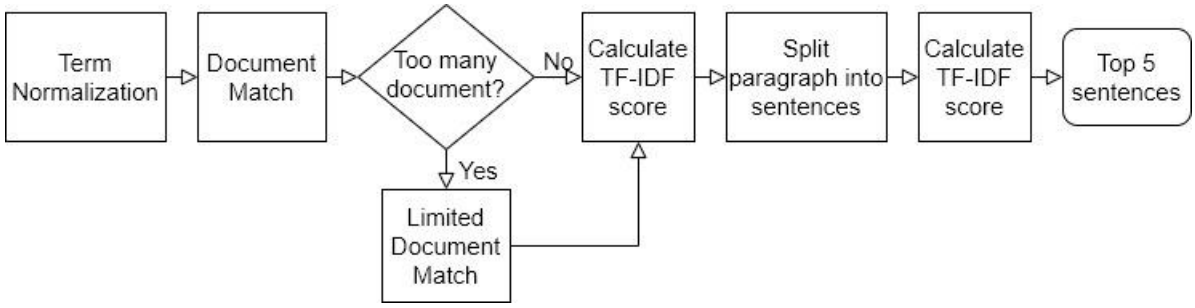

Figure 2: Flowchart for Task 1

4.2. T5 model for Summarization

In Task 3, our system adopt the T5 model to generate simplified sentences. We use the 648 data in the training set to fine-tune the T5 model with a ratio of 8:2 between the training set and the validation set, and the hyper-parameters are shown in Table 1. In addition, when generating sentences, we set the generated token to 0.78 times the source sentence token. This ratio is based on the average sample sentence token and source sentence token ratio of the data set, as shown in Table 2.

Table 1. Training parameters

Parameter	value
Model	t5-base
TRAIN_BATCH_SIZE	4
VALID_BATCH_SIZE	1
TRAIN_EPOCHS	3
LEARNING_RATE	1e-4

Table 2. The generated examples

source sentence	generated sentence
We describe a PDA (Personal Digital Assistant) based CSCW system called NewsMate, which provides mobile and distributed news journalists with timely information.	NewsMate provides mobile and distributed news journalists with timely information.
A CDA is a mobile user device, similar to a Personal Digital Assistant (PDA).	A CDA is a mobile user device, similar to a Personal Digital Assistant

Figure 3 shows the training flowchart of our Task3 system. The first step, we prepend the input sequence with '**summarize:**' (task_prefix) before encoding it. This will help in improving the performance, as this task prefix was used during T5's pre-training. Then uses the T5Tokenizer encoding sequence and train the model with parameters in Table 3. Finally generate the summary.

Figure 3. T5 model Training flowchart

5. Results and Discussion

We participated in the SimpleText challenge under the name "CYUT Team2". Our reported results in this section are obtained from the SimpleText official report [5].

For the Task1 evaluation, our team CYUT comes out on top of the ranking list by achieving a score of nDCG@5= 0.3322. Table 3 presents in more details the achievements of each run in more details. These values show that the automatic run made by CYUT and the manual run significantly outperform other automatic runs in terms of selecting the abstracts with a high relevance. Besides, it is important to note that the pooling method only kept articles chosen by at least two participants and gave a relevance score on a scale of 0 to 5. This method of evaluation will be detrimental to teams that find unique documents.

For the Task3 evaluation, we are ranked second with an score of 0.122 in table 4. Scores are evaluated by the average harmonic mean of normalized opposite values of Lexical Complexity, Syntactic Complexity and Distortion Level. In Table 5 shown information distortion in evaluated runs. It should be noted that most of the results generated by our method are truncated.

Table 3. SimpleText Task 1: Evaluation scores of official runs. Scores obtained by each run (Score), the number of returned documents with a score ≥ 1 (#Docs), the number of queries with at least one returned document (#Queries) and the average scores per document and query. [5]

Team	Score	#Docs	Doc Avg	#Queries	Query Avg	nDCG@5
CYUT	125	44	0.53	77	1.62	**0.3322**
UAMS-MF*	163	54	0.87	99	1.65	0.2761
UAMS	52	17	0.22	40	1.30	0.1048
NLP@IISERB	26	7	0.35	13	2.00	0.0290

⋆ *Manual run*

Table 4. SimpleText Task 3: Ranking of official submissions on combined score [5]

Run	Score
PortLinguE full	0.149
CYUT Team2	**0.122**
CLARA-HD	0.119

6. Conclusion and Future Works

This paper reports our approach to the SimpleText lab. In terms of information retrieval for Task 1, we achieve top of results using the TF-IDF filter. However, the polysemy problem of TF-IDF will cause difficult to find extended topic document. From our perspective, it would be more beneficial for Task

1 to have a better information retrieval model. In terms of generating sentences for Task 3, the result is not satisfactory. We expect the excess parts of the sentence should be removed, and finally the simplified sentence is obtained. However, our result shows that most of the sentences are truncated, but the simplified sentence cannot fully express the meaning of the original sentence. It is not suitable using the T5 model, or the number of training data is insufficient. In the future, we consider using other models and increase the size of the dataset to improve the performance.

Table 5. SimpleText Task 3: Information distortion in evaluated runs [5]

Run	Total	Unchanged	Truncated	Valid	Longer	Length Ratio	Evaluated	Uncorrect Syntax	Unresolved Anaphora	Minors	Syntax Complexity	Lexical Complexity	Information Loss
CLARA-HD	116,763	128	2,292	111,627	201	0.61	851	28	3	68	2.10	2.42	3.84
CYUT Team2	116,763	549	**101,104**	111,818	49	0.81	126	1		32	2.25	2.30	2.26
PortLinguE_full	116,763	42,189	852	111,589	3,217	0.92	564	7		5	2.94	3.06	1.50

7. Acknowledgements

This study was supported by the Ministry of Science and Technology under the grant number MOST 110-2221-E-324-011.

8. References

[1] Ermakova, L., Bellot, P., Kamps, J., Nurbakova, D., Ovchinnikova, I., SanJuan, E., Mathurin, E., Araújo, S., Hannachi, R., Huet, S., & Poinsu, N. (2022). Automatic Simplification of Scientific Texts: SimpleText Lab at CLEF-2022. In M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, & V. Setty (Eds.), Advances in Information Retrieval (Vol. 13186, pp. 364–373). Springer International Publishing. https://doi.org/10.1007/978-3-030-99739-7_46

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research. arXiv: 1910.10683. https://doi.org/10.48550/arXiv.1910.10683

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistic

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. (2020). BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension.

[5] Liana Ermakova, Eric SanJuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Silvia Araujo, Radia Hannachi, Elise Mathurin, and Patrice Bellot (2022). Overview of the CLEF 2022 SimpleText Lab: Automatic Simplification of Scientific Texts. In A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, & N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022)