The 3rd International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2019)
December 17-19, 2019. Phu Quoc, Vietnam

A Study on Building Reliable Corpus for Punctuation and Quotation Mark Filling Task

Seunggyu Han[1], Kisu Yang[1], Heuiseok Lim[1]

[1] Korea University, Dept of Computer Science and Engineering
{seunggyu-han, willow4, limhseok}@korea.ac.kr

Abstract. Punctuation mark and quotation marks are used to clarify the meaning of sentence. For the tasks such as dialogue system or Speech-To-Text(STT), using appropriate quotation mark can improve the quality of result in perspective of human understandability. In this paper, we propose a corpus based on textualized speech of Blue House of Republic of Korea, which is a suitable and reliable corpus for such task.

Keywords: Corpus, Spoken language corpus, Punctuation and quotation mark.

1 Introduction

Punctuation mark and quotation marks are used to clarify the meaning of sentence. Number of marks used varies from language to language. For Korean language, 6 marks are the mostly used(‘ ’ “ ” , .). In most cases, punctuation mark and quotation marks are removed from sentence when processing for natural language processing. This may help the performance of encoding a sentence, but not in the generating stage. Dialogue system and speech-to text are the example of tasks focusing on generating sentence. In these two tasks, usually only period(.) is used at the end of the sentence. These sentence are understandable, but when correct quotation and punctuation marks are filled, it becomes much more understandable. Most models used in natural language generation are complete in its own. Therefore modifying the architecture of model to generate correct marks needs high cost. In order to lower the cost, adding another model which takes mark-less sentence and generates punctuation and quotation mark filled sentence is a good option.

In order to make a model that does the work described above, appropriate corpus is needed. These two tasks are deeply related to spoken language. There are some corpora that has been preprocessed for quotation and punctuation mark filling task, but these corpora are not based on spoken language. In this paper, we propose a corpus suitable for filling punctuation and quotation mark for spoken language. The sentences are crawled from speeches of the president and government officials of the Blue House of Republic of Korea.

2 Related work

[1] proposes a model based on Transformer architecture, which corrects uncapitalized characters and fills punctuation mark of English language. This model also proposes a method that slices the sentence in sliding-window manner, with overlap. Accuracy was measured in F1-Score. [2] uses LSTM, and [3] uses Bi-LSTM with Attention method for filling punctuation mark. [4] uses Convolution Neural Network(CNN) for completing punctuation mark of IWSLT data.

3 Constructing Corpus

As mentioned above, this paper focuses on gathering spoken language for corpus. This is a lot harder than searching for written language. Most of them exists as a form of audio, and the quality is not constant to use as a corpus. Fortunately, government all around the glove are trying to make a digital document of speech that they have spoken. In Republic of Korea, the Blue House keeps the record of speech of the President, the prime minister, and head of ministries. These are well-written, grammatically correct, has consistent quality, and has a lot of data. By crawling all the sentences from website of the Blue House and filter or fix the sentence to be suitable for Natural Language Processing, it can be a fine quality corpus. Table 1 shows the statistics of corpus gathered from the site mentioned above. All sentences are written in Korean.

Table 1. Statistics of gathered sentences

Description	Count or Number
Total sentences	126,795
Total words	1,633,817
.	121,256
,	67,097
!	4,422
?	1,177
‘ ’ (each)	7,321
“ ” (each)	1,230
Maximum words in sentence	107
Average words in sentence	12.89
Maximum marks in sentence	19
Average marks in sentence	1.66
Sentence containing only 1 mark	75.802

Sentences are fixed if (1) one sentence is separated into multiple lines, (2) multiple sentence is written in one same line, (3) sentence contains parentheses, (4) sentence has a number with comma, (5) sentence has spacing error.

Sentences are deleted if (1) quotation marks are not paired, (2) contains other than Korean letter, number and pre-assigned marks, (3) contains comma in the middle, (4) not a spoken word.

4 Conclusion

In this paper, we have constructed a fine-quality spoken language corpus with appropriate punctuation and quotation marks. This may be used in future work that builds a model for mark filling task. Also, other countries also has a government owned website that has all the speeches, so this sentence gathering can be expanded in to other languages as well.

Acknowledgments. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & communications Technology Promotion) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189)

5 References

1. B. Nguyen, V. B. H. Nguyen, H. Nguyen, P. N. Phuong, T.-L. Nguyen, Q. T. Do, and L. C. Mai, "Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging," arXiv preprint arXiv:1908.02404, 2019.
2. O. Tilk and T. Alum¨ae, "Lstm for punctuation restoration in speech transcripts," Sixteenth annual conference of the international speech communication association, 2015. Appendix: Springer-Author Discount
3. O. Tilk and T. Alum¨ae, "Bidirectional recurrent neural network with attention mechanism for punctuation restoration." Interspeech, pp. 3047–3051, 2016.
4. X. Che, C. Wang, H. Yang, and C. Meinel, "Punctuation prediction for unsegmented transcript based on word vector," Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 654–658, 2016.