INFORMATIZED CAPTION ENHANCEMENT BASED ON IBM WATSON API AND SPEAKER PRONUNCIATION TIME-DB

Yong-Sik Choi, YunSik Son and Jin-Woo Jung

Department of Computer Science and Engineering,
Dongguk University, Seoul, Korea

ABSTRACT

This paper aims to improve the inaccuracy problem of the existing informatized caption in the noisy environment by using the additional caption information. The IBM Watson API can automatically generate the informatized caption including the timing information and the speaker ID information from the voice information input. In this IBM Watson API, when there is noise in the voice signal, the recognition results are not good, causing the informatized caption error. Especially, it is more easily found in movies such as background music and special sound. Specifically, to reduce caption error, additional captions and voice information are entered at the same time, and the result of the informatized caption of voice information from IBM Watson API is compared with the original text to automatically detect and modify the error part. Based on the database containing the average pronunciation time, each word for each speaker is changed into the informatized caption in this process. In this way, more precise informatized captions could be generated based on the IBM Watson API.

KEYWORDS

Informatized caption, Speaker Pronunciation Time, IBM Watson API, Speech to Text Translation

1. INTRODUCTION

Recently, artificial intelligence technology is being researched and developed in various fields. Artificial intelligence refers to the intelligence created by a machine, and is the intelligence that a computer program behaves and calculates, such as human thinking. However, since artificial intelligence that does not understand human language is useless, the most important thing in artificial intelligence technology is natural language processing technology and speech recognition technology. Typical speech recognition technologies include speech to text conversion. Among captions in which speech is converted into characters, captions including timing information and speaker ID information [1] are referred to as informatized captions [2]. Such an informatized caption can be generated using the IBM Watson API or the like [3]. However, the IBM Watson API is more susceptible to clipping errors due to poor recognition results if there is noise in the audio signal, especially in movies such as movies where background music and special sounds are used. In order to solve this problem, there has been proposed a method of predicting the timing information of the informatized caption based on a linear estimation [2] formula proportional to the number of alphabets. In this paper, we use the IBM

Natarajan Meghanathan et al. (Eds) : NLP, JSE, CST, SIP, ARIA - 2018
pp. 105– 110, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80211

Watson API, which provides basic functions of informatized caption including timing information, speaker ID information, and so on, to generate a word information list based on the proposed method.

2. SPEAKER PRONUNCIATION TIME-DB (SPT-DB)

2.1. Structure

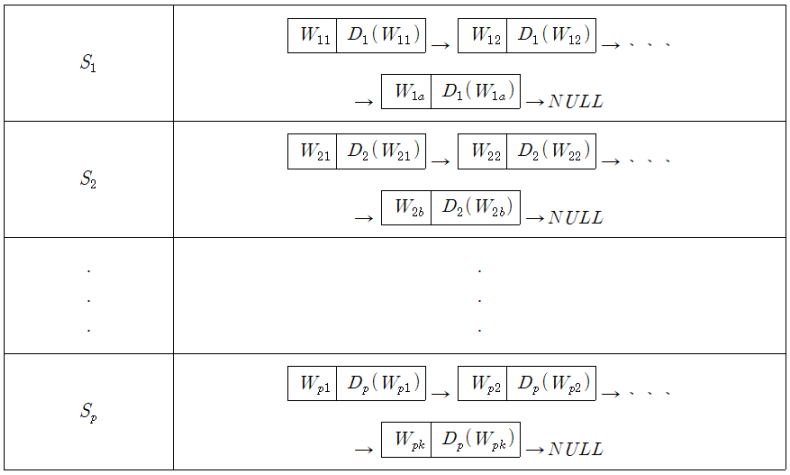

Figure 1. Structure of SPT-DB

SPT-DB consists of each node for each speaker(S_p) as shown in Fig. 1.The nodes consist of the average pronunciation times(D_p) of each word(W_{pk}).The nodes of the speaker are arranged in ascending order based on the average pronunciation time, and are connected to each other, and a null value is present at the end. When SPT-DB searches for a word spoken by the speaker, it searches based on the pronunciation time.

2.2. Assumption

Before proceeding with the study, the following assumptions are based on SPT-DB. [Assumption] SPT-DB is already configured for each speaker.

3. PROPOSED ALGORITHM

3.1. Algorithm modifying incorrectly recognized word based on SPT-DB

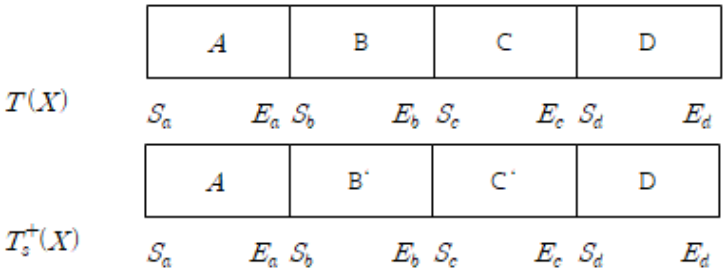

Figure 2. Original caption T(X) and informatized caption $T_s^+(X)$

Basically, original caption,T(X), and informatized caption from speech recognition result, $T_s^+(X)$, are input together.

Here, S_x and E_x mean the start time and end time of pronunciation for the word X, respectively.

[Step 1] Judge whether there is an incorrectly recognized word by comparing T (X) with $T_s^+(X)$.If there is no incorrectly recognized word, it terminates. If there is an incorrectly recognized word, go to the next step.

[Step2] Judge whether there are several consecutive words in the sequence, and pass the parameter to the case.

[Step3] Modify the words in the SPT-DB based on the start and end points of the cases.

[Step4] If there is an incorrectly recognized word in the following word, repeat steps 1 to 3 and terminate if there is no incorrectly recognized word.

3.2. Case 1: There is only one incorrectly recognized word.

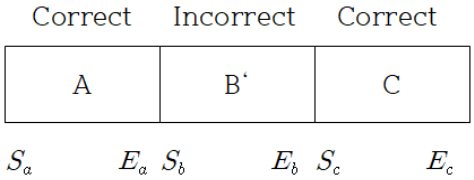

Figure3. There is one incorrectly recognized word

[Step1] Find the point at which the signal of a specific volume(dB) T or more starts for E_a to S_c and determine S_b.

[Step2] If there is a minimum time t' in S_b to S_c at which the signal intensity falls below a certain volume T and then remains below T until S_c, E_b = t' is determined. If there is no t' satisfying the above condition, $E_b = S_c$.

[Step3] Returns the start time and end time.

3.2. Case 2: There are more than two incorrectly recognized word.

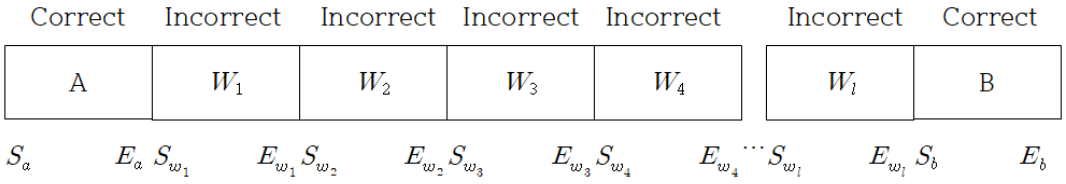

Figure 4. More than three incorrectly recognized word

[Step1] Find the point at which the signal of a specific volume(dB) T or more starts for E_a to S_{w2} and determine S_{w1}.

[Step2] If there is a minimum time t' in $S_{w1}$to S_{w2} at which the signal intensity falls below a certain volume T and then remains below T until S_{w2}, E_{w1} = t' is determined. If there is no t' satisfying the above condition,$E_{w1} = S_{w2}$.

[Step3] The ending point of the current word is obtained by multiplying the start time of the current word by the ratio of the pronunciation time of the incorrectly recognized words to the average pronunciation time of the current word. The following are summarized as follows.

$$E_{w_i} = \left(E_{w_l} - S_{w_1}\right) \times \frac{D(W_i)}{\sum_{i=1}^{l} D(W_i)}$$

[Step4] Returns the start time and end time.

4. CASE STUDY

The case was tested based on English listening assessment data. Fig.5 shows a problem of the English listening evaluation for university entrance examination. In a noisy environment like Fig.6, the accuracy dropped significantly. For reference, the original voice source was synthesized with raining sound using Adobe Audition CC 2017 to create a noisy environment.If we improve the proposed algorithm with noise, we can obtain the same result as Table1. The accuracy of speech recognition is 100% by the help of original caption and each word includes its own start time and end time.

W: Dad, I want to send this book to Grandma. Do you have a box?
M: Yeah. I've got this one to put photo albums in, but it's a bit small.
W: The box looks big enough for the book. Can I use it?

Figure 5. Original caption

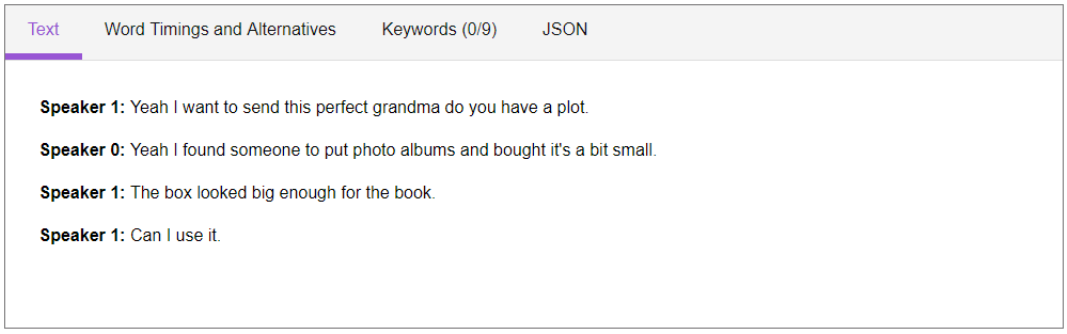

Figure 6. Recognition of mixed voice with rain noise by IBM Watson system

Table1.Informatized caption modified by the proposed algorithm

Sentence \ Word		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Speaker 0	Dad	I	want	to	send	this	book	to	Grandma	Do	you	have	a	box	
		0.03-0.58	0.74-0.87	0.87-1.19	1.19-1.35	1.35-1.66	1.66-1.83	1.83-2.1	2.1-2.24	2.24-2.89	3.21-3.45	3.45-3.67	3.67-3.95	3.95-4.03	4.03-4.75	
B	Speaker 1	Yeah	I've	got	this	one	to	put	photo	albums	in	but	it's	a	bit	small
		5.22-5.7	6.01-6.27	6.27-6.62	6.62-6.86	6.86-7.15	7.15-7.26	7.26-7.48	7.48-7.88	7.88-8.29	8.29-8.59	8.69-9.1	9.28-9.48	9.48-9.55	9.55-9.81	9.81-10.51
C	Speaker 0	The	box	looks	big	enough	for	the	book	Can	I	use	it			
		10.86-10.99	10.99-11.41	11.41-11.67	11.67-11.96	11.96-12.26	12.26-12.47	12.47-12.6	12.6-13.16	13.46-13.71	13.71-13.79	13.79-14.12	14.12-14.42			

5. CONCLUSIONS

In this paper, we propose an algorithm to find and modify incorrectly recognized words based on the SPT-DB, which stores the average pronunciation times and appearance frequencies of the corresponding words in the speaker to correct the errors in the informatized caption obtained through the IBM Watson API. However, the proposed algorithm has a limitation that SPT-DB should be created first because it is assumed that the information of the corresponding words already exists in SPT-DB. Future research will be conducted to modify incorrectly recognized words while performing speech recognition and to update the SPT-DB in real time.

ACKNOWLEDGEMENTS

Following are results of a study on the "Leaders in INdustry-university Cooperation+" Project, supported by the Ministry of Education and National Research Foundation of Korea, and partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2015R1D1A1A09061368),and also partially supported by this work was supported by the Korea Institute for Advencement of Technology(KIAT) grant funded by the Korean government (Motie: Ministry of Trade, Industry &Energy, HRD Program for Embedded Software R&D) (No. N001884).

REFERENCES

[1] CheonSun Kim, "Introduction to IBM Watson with case studies." Broadcasting and Media Magazine, Vol.22, No. 1,pp24-32.

[2] Yong-Sik Choi, Hyun-Min Park, Yun-Sik Son and Jin-Woo Jung, “Informatized Caption Enhancement based on IBM Watson API,” Proceedings of KIIS Autumn Conference 2017, Vol. 27, No. 2, pp105-106.

[3] IBM Watson Developer’s Page, https://www.ibm.com/watson/developer

AUTHORS

Yong-Sik Choi has been under M.S. candidate course at Dongguk university, Korea, since 2017. His current research interests include machine learning and intelligent human-robot interaction.

YunSik Son received the B.S. degree from the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea, in 2004, and M.S. and Ph.D. degrees from the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea in 2006 and 2009, respectively. He was a research professor of Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Korea from 2015-2016. Currently, he is an assistant professor of the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea. Also, His research areas include secure software, programming languages, compiler construction, and mobile/embedded systems.

Jin-Woo Jung received the B.S. and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1997 and 1999, respectively and received the Ph.D. degree in electrical engineering and computer science from KAIST, Korea in 2004. Since 2006, he has been with the Department of Computer Science and Engineering at Dongguk University, Korea, where he is currently a Professor. During 2001~2002, he worked as visiting researcher at the Department of Mechano-Informatics, University of Tokyo, Japan. During 2004~2006, he worked as researcher in Human-friendly Welfare Robot System Research Center at KAIST, Korea. During 2014, he worked as visiting scholar at the Department of Computer and Information Technology, Purdue University, USA. His current research interests include human behaviour recognition, multiple robot cooperation and intelligent human-robot interaction.