Human Interaction and Emerging Technologies (IHIET 2022), Vol. 68, 2022, 392–400
https://doi.org/10.54941/ahfe1002757

The Effect of Topic-Shift Characteristics in Daily Conversation on Identification of Recognition Errors

Yotaro Iida, Hikaru Nishida, and Yumi Wakita

Osaka Institute of Technology, Kita-ku Chaya-machi 1-45, Osaka, Japan

ABSTRACT

To support senior and reclusive citizens' smooth conversations, we have developed a conversation support system named "associative board". it recognizes their conversation and provides several suitable topics for speakers when their conversation progress not so smooth. However. If there are too many recognition errors the system will not be able to present suitable words. The misrecognized words identification function is necessary for our associative board system. In this study, we clarify the problems with conventional misrecognized words identification methods for recognizing daily casual conversation. As results of evaluation, the conventional misrecognized words identification is effective for the conversations with predefined topics, however for casual conversations without predefined topic, the identification is difficult. The distribution of semantic similarity values among words for casual conversation are broader than that with predefined topics. When the semantic similarity values are under 0.3, despite the correct recognition utterances, that semantic similarity values of the recognition results are often lower than that of the misrecognition results. The 21.7% to all topics are that case. That means when the casual conversations in which the topic-shifting occurs frequently, the misrecognized words identification is difficult. The semantic similarity among recognized words should be calculated constantly, and when the semantic similarity values are high continuously or are low rarely, the identification method could be used. When the semantic similarity values are low continuously, the error words extraction and correction process should be stopped.

Keywords: Topic shift characteristics, Semantic similarity among words, Misrecognized words identification, Conversation support system

INTRODUCTION

With the recent increase in the number of senior citizens living either alone or in reclusive situations, many communities, companies, and schools have realized the importance of human-to-human communication. To support senior citizens' smooth conversations, we are developing a conversation support system that provides support for conversations at appropriate times, taking into account the smoothness of the conversation. We have already confirmed that it is possible to determine whether or not a conversation is smooth by using information on the distribution of the fundamental frequency (F0) and the speech power level (SPL) of each utterance (Wakita et al., 2016). Using

these parameters, we are developing a system that can estimate the degree of smoothness of a conversation and provide a new topic of conversation when the smoothness of the conversation begins to deteriorate, thereby ensuring smooth communication. The information provided as new topics is as follows:

- The content words imagined from recognition result words
- The web news topics related to the recognition results

To provide the information, the system comprises the following parts: (1) conversation atmosphere analysis (2) speech recognition (3) content words extraction from speech recognition results (4) similar word estimation from the content words (5) extracting news information provided from outside. We call this conversation support system "associative board."

In general, when speakers speak to dialogue system, their utterances are relatively clear, therefore, there are not so many misrecognitions in the speech recognition results. However, the "associative board" only listens to the conversations from the side, as a third party, therefore, the quality of input speech is poor and misrecognitions are frequent. If there are too many errors, the system will fail because it will not be able to present words that are in accordance with the topic. Several papers have been reported to identify speech recognition errors. The effectiveness of a method of identification of errors by training a spelling correction model (Guo et al., 2019) and neural error corrective language models (Tanaka et al., 2018) have been confirmed for spoken language. However, since these methods assume a relatively limited domain and conversational situations, the correct words are often regarded as error words in daily casual conversation where the topic is not clear or the topic changes rapidly.

In this study, we first explain how to estimate our conversation support system "associative board". Next, we clarify the problems with conventional methods of identifying misperceptions in daily casual conversation. We have already reported that the topic-shift characteristics are different in casual conversation between older adults and young people. Therefore, we evaluated the performance of the method in identifying misrecognition in casual conversation by comparing conversations between older adults and young people. In addition, we also compared the misrecognition identification performance when conversations with predefined topics were used. Finally, we report on the relationship between the semantic similarity among words and the conventional misrecognized words identification performance for casual conversation and discuss a possibility of improving the conventional identification performance.

CONVERSATION SUPPORT SYSTEM

Fig. 1 describes the structure of the conversation support system. The system uses a conversation atmosphere estimation function and additional functions to extract the following information from input utterances:

(1) Extraction of non-verbal information such as fundamental frequency (F0) and speech power level (SPL) to estimate the conversation atmosphere

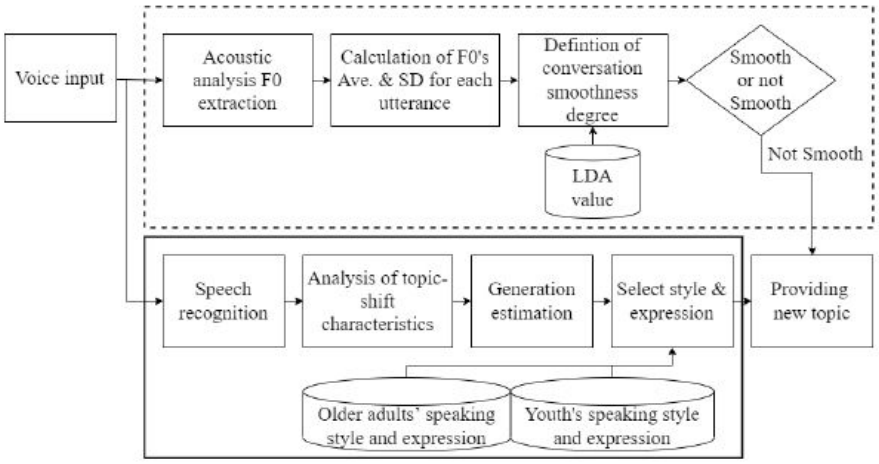

Figure 1: Process of the conversion support system.

(2) Speech recognition to understand the conversation
(3) Holding of a part of the news information provided from outside
(4) Extraction of content words from speech recognition words and information texts by the outsider.
(5) Estimation of similar words of extracted content words using our similar words dictionary.

The conversation smoothness degree is estimated using the average and variance values of F0 for each utterance. When our system estimates the conversation progress as not so smooth, the system recommends certain new topics on the display. Alternatively, from the content words included in the speech recognition result, the system estimates the conversation topic and topic shift characteristics. In consideration of these estimation results, the system extracts keywords related to the topic from recognition results and estimates certain related words or keywords and displays both keywords and related words on the monitor of the system. As described in the previous section, our "associative board" only listens to the conversations from the side as a third party, therefore, misrecognitions occur more frequently than when using speech dialogue system that is spoken into, by speakers directly. If there are too many errors, the system will fail because it will not be able to present suitable words that are imaged in line with the topic.

CASUAL CONVERSATION DATABASE

Casual Conversation Recording

We recorded several sets of 3-minute dyadic conversations. Fig. 2 depicts the schematic positional setup of these recordings. We used two microphones and a video camera for this purpose. The conditions of the recordings are listed in Table 1. Although the participating speakers were not meeting for the first time, they never had a mutual conversation before. We used fourteen conversation sets for the study, of which, seven sets were conversations

Table 1. Database conditions.

Number of Speakers	6 Older Speakers, 7 Young Speakers
Age	Older adults: 62–82 years old; Young people: 21–23 years old
Number of conversations	14 conversations (7 conversations each)
Conversation periods	3 min/conversation
Conversation condition	Free dyadic conversation

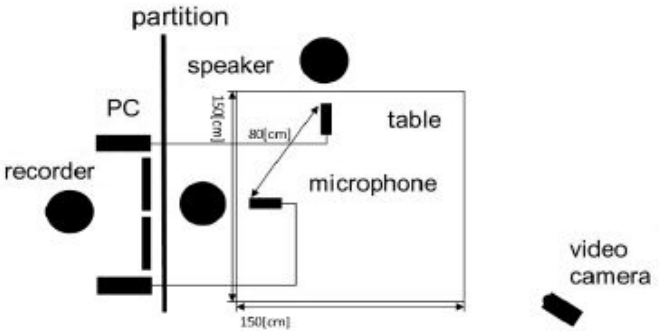

Figure 2: Positional setup of recording conversation.

between older adults aged between 62 and 82 years, and the other seven sets were conversations between youths aged between 21 and 23 years. We did not set any conditions regarding the topics and the participants conversed freely. After recording, we created a database of transcript text from all the recorded conversations.

Comparison of Casual Conversation Characteristics between Older Adults' Conversations and Youths' Conversations

After recording, we created a transcript database from the video recording data. Morphological analysis was performed on the transcript text database using "Mecab," which is a Japanese morphological analyzer engine. We asked three individuals to read the 14 transcripts to decide the various topic boundaries for each conversation. After deciding the boundaries, the individuals selected the most important word that expressed the topic for each part of the conversation separated by the boundary. The differences in boundaries given by the three individuals were within three utterances. This result explains that the changing point of the topics is clear. We decided the boundary according to the answers of the three individuals.

We already reported "the number of topics for each conversation" and "the number of content words for each utterance" comparing older adults' and youths' conversations (Iida and Wakita, 2021). The average of the number of topics in older adults' conversations is 2.29, and for youths' conversations, 4.0. The results reveal that the casual conversations of older people tended to include one long utterance with the speakers firmly expressing their knowledge or opinion. Conversely, the conversations of youths did not reveal

long utterances and they tended not to express their opinions at once, but only gradually, after watching the reaction of the other person. The average number of content words in an older adult's utterance is 2.34 and that in a youth's utterance is 1.45. This implies the topic shift characteristics are different between older people's and youths' conversation.

RECOGNITION ERROR IDENTIFICATION PERFORMANCE FOR CASUAL CONVERSATION

Our "associative board" uses speech recognition to identify the conversationalist's topic of conversation and presents associative words that are relevant to the topic. Speech recognition errors have a significant impact on the quality of the "associative board" and its ability to achieve its objectives. It is necessary to address speech recognition errors for our system performance. Several methods have already been proposed to identify speech recognition errors by calculating the semantic similarity among words in the recognition results. The method regards the words with low similarity to other words as misrecognition in some domain-limited conversations. However, we think the effectiveness of conventional speech recognition identification methods will decrease for casual conversation compared with limited domain conversation. In this chapter, we perform the conventional misrecognition identification in casual conversation and clarify the problems. As illustrated in the previous chapter, there are differences in the topic-shift characteristics among generations. We confirm that identification possibility for each generation.

Calculation of Semantic Similarity Degrees Among Words

In this analysis, the words are described by the variance representation obtained by word2vec, and the semantic similarity is calculated by the cosine similarity between the words represented by the word variance representation (Suzuki et al., 2018). Equation (1) demonstrates the similarity between words, where a, and b are the word vectors represented by the word variance representation. Using this formula, we calculated the cosine similarity of all word combinations for the content words appearing in a topic and obtained the mean and variance of the similarity for each topic.

$$\text{Semantic similarity degree} = \frac{\mathbf{a} \bullet \mathbf{b}}{\|\mathbf{a}\| \, \|\mathbf{b}\|} \tag{1}$$

Confirmation of Recognition Error Identification Performance Using Semantic Similarity Degree

In speech recognition, when speech segment detection errors or insertion of filler utterances occur, the first word of the utterance is often wrong. Therefore, assuming that the first word of each utterance was mistaken in our conversation database, we replaced the first phoneme of the first word with another word that has a similar pronunciation to it and considered this as the speech recognition error result. We calculated the inter-semantic similarity between the content words in both the original correct transcriptions

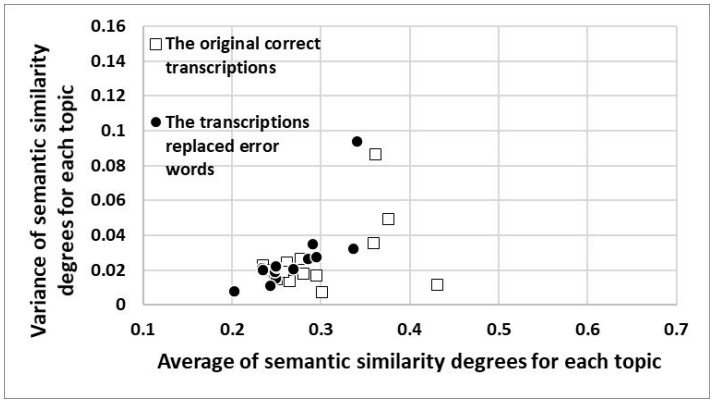

Figure 4-1: Relationship between the average and variance of the semantic similarity values for older adults' conversation.

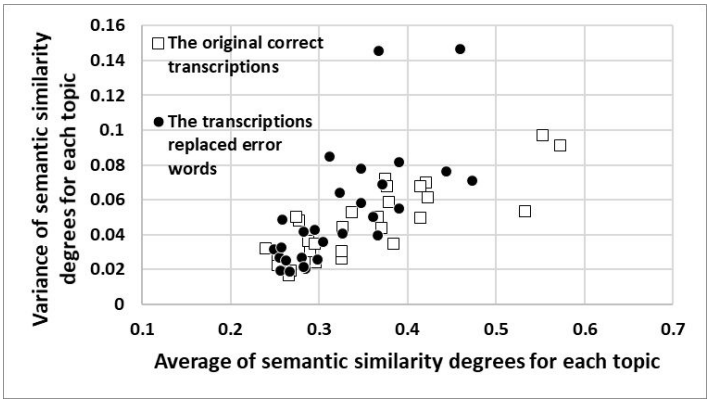

Figure 4-2: Relationship between the average and variance of the semantic similarity values for youths' conversation.

and the replaced error transcriptions. The inter-semantic similarity between the content words is calculated by each topic and compared their average values for each topic. The casual conversation data was free conversation without topic restriction. We added another data set in which the participants were conversing on a predetermined topic. In this experiment, we used 10 conversations from the "Spoken Language Database" provided by the National Institute for Japanese Language and Linguistics (NINJAL) (Maekawa, 2003) as limited topic conversation database. These conversations are the speakers' comments on their topics of lectures after their own lectures.

The Fig. 4-1 and Fig. 4-2 illustrate the relation average and variance values of semantic similarity values among words for each topic. Fig. 4-1 and Fig. 4-2 illustrate that for older adults' topics and youths' topics respectively. The figures illustrate the following:

- The distribution area of older adults' data is narrower than the area of youths' data. In the case of youth's conversation, both the average and variance of semantic similarity degrees of several topics are large, because

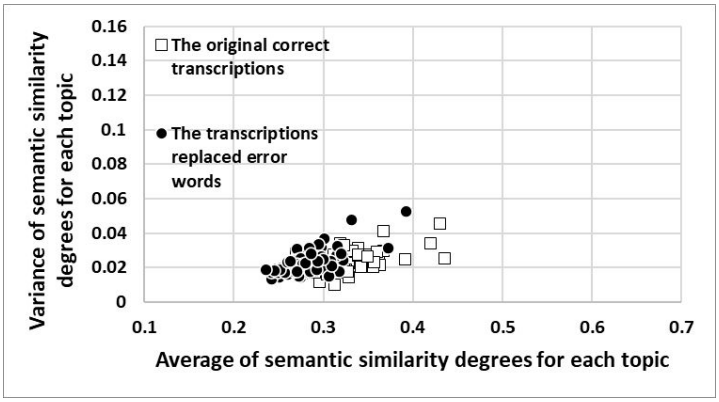

Figure 5: Relationship between the average and variance of the semantic similarity values for limited domain conversation.

the variety of expression is very narrow and they used the same words repeatedly.

Compering the semantic similarity degrees between the original correct transcriptions (Square dot) and the transcriptions replaced error words (Circle dot), when the semantic similarity values are large, the difference between original correct transcriptions and replaced error transcriptions are large. In such cases, it could be considered that the current word error identification method is effective. However, when semantic similarity values are small, the differences are very small. In such cases, the system cannot identify error words.

The Fig. 5 illustrates the relation average and variance values of semantic similarity values using ten limited topic conversations.

The difference between original correct transcriptions and replaced error transcriptions using limited topic conversations are sizeable than that using topic free conversation in Fig. 4. Especially the average values of the semantic similarity values illustrated in Fig. 5 are almost same in Fig. 4-1, however, the difference values are different. The current word error identification method would be effective for limited topic conversation, however, the effectiveness decreases for topic free conversation as casual conversation.

RECOGNITION ERROR IDENTIFICATION FOR CASUAL CONVERSATION

Our experiment results in the previous chapter suggest that the semantic similarity degree among content words in casual conversation is larger than that in domain-dependent conversation. The effectiveness of the current word error identification method would be enough in conversation when semantic similarity degree is high. However when semantic similarity degree is low, sometimes correct word are regarded as errors.

We calculated the difference values between the semantic similarity degree values of original transcription and that after replaced to error words by each utterance. The size of difference values mean the recognition error

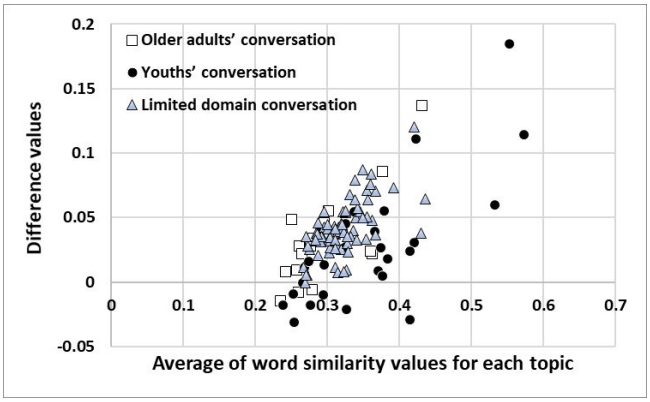

Figure 6: Relationship between recognition error identification and performance semantic similarity degrees.

identification performance. When the difference values are high, the identification performance would also be high. Fig. 6 illustrates the relationship between the semantic similarity degrees and the difference values.

Fig. 6 demonstrates the following:

- When the semantic similarity degree is large, the difference value is also large. The tendency is found for both limited-domain conversation and domain free casual conversation.
- The data area of the limited topic conversation is narrow and almost difference values (98.7%) are plotted over zero.
- In the case of casual conversation, the data area is broad. Especially the data of youths' conversation is most broad.
- In the case of casual conversation, the difference values are plotted under zero often. The rate is 21.7%.
- When the difference values are plotted under zero, the semantic similarity degree is also low. In all cases, the semantic similarity values are under 0.35.

Results suggest that when the topics where semantic similarity value are high, the current word error correction method would be effective. However, when the topics where semantic similarity values are low, even the topics are recognized correctly, sometimes the recognition words are regarded as "error," and the recognition performance decreases. To keep the error word extraction performance for casual conversation, the semantic similarity degrees should be calculated constantly. When the semantic similarity values are high continuously or are low rarely, the error words should be extracted. When the semantic similarity values are low continuously, the error words extraction and correction process should be stopped.

CONCLUSION

To confirm the effectiveness of using semantic similarity among words to identify misrecognized words for casual conversations, we compared the semantic similarity between correctly and incorrectly recognized fourteen

casual conversations (seven conversations by older adults and seven conversations by youths). In casual conversation, the distribution of semantic similarity values between words was wider, especially in conversations between young people, because topic-shifting was frequently performed. In these conversations, the semantic similarity was also low when the semantic similarity was low, for example, when the semantic similarity was less than 0.3, it was difficult to identify misrecognized single words. That suggest that in casual conversation, we believe it is necessary to constantly measure the semantic similarity among words in conversation and identify misrecognized words based on semantic similarity only in situations where the similarity value is higher than a threshold value. In future, we'd like to proposed the misrecognized words identification method for casual conversation even if the semantic similarity values are low.

ACKNOWLEDGMENT

This work is supported by JSPS KAKENHI Grant Number 22K04626.

REFERENCES

Guo, J., Sainath, T. N., & Weiss, R. J. (May, 2019). "A spelling correction model for end-to-end speech recognition", In ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5651–5655.

Iida, Y., & Wakita, Y. (July, 2021). "Topic-Shift Characteristics of Japanese Casual Conversations Between Elderlies and Between Youths", In International Conference on Human-Computer Interaction, pp. 418–427.

Maekawa, K. (2003) "Corpus of Spontaneous Japanese: Its design and evaluation", Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), Tokyo.

Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., & Inui, K. (2018). "A joint neural model for fine-grained named entity classification of wikipedia articles", IEICE Transactions on Information and Systems, pp. 73–81.

Tanaka, T., Masumura, R., Masataki, H., & Aono, Y. (2018). "Neural Error Corrective Language Models for Automatic Speech Recognition", In INTERSPEECH, pp. 401–405.

Wakita, Y., Yoshida, Y., & Nakamura, M. (July, 2016). "Influence of personal characteristics on nonverbal information for estimating communication smoothness", In International Conference on Human-Computer Interaction, pp. 148–157.