Unnecessary Utterance Detection for Avoiding Digressions in Discussion

Riki Yoshida and Takuya Hiraoka and Graham Neubig and Sakriani Sakti
and Tomoki Toda and Satoshi Nakamura
Nara Institute of Science and Technology (NAIST), Nara, Japan
E-mail: {riki-y,takuya-h,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

***Abstract*—In this paper, we propose a method for avoiding digressions in discussion by detecting unnecessary utterances and having a dialogue system intervene. The detector is based on the features using word frequency and topic shifts. The performance (i.e. accuracy, recall, precision, and F-measure) of the unnecessary utterance detector is evaluated through leave-one-dialogue-out cross-validation. In the evaluation, we find that the performance of the proposed detector is higher than that of a typical automatic summarization method.**

I. Introduction

Discussions play a very important role in human activities. For example, we often make decisions regarding a particular problem through discussion. In contrast, we often waste time due to *digressions*, in which the conversation proceeds in a different direction than that specified by the original agenda. If, for example, a dialogue system could help us avoid digressions and keep the conversation on track, the discussion could proceed more efficiently. If we are to develop such a system it is important for the system to have the ability to determine which utterances are unnecessary.

In this paper, we propose a statistical unnecessary utterance detector for avoiding digression. To our knowledge, there is no previous research explicitly regarding avoiding digression, but research for automatic summarization of meetings [1], [2], [3], [4], [5], [6], [7] is somewhat related. Considering the definition of the summarizer as "a system whose goal is to produce a condensed representation of its input for human consumption" [8], traditional automatic summarization is used as the method for accessing the condensed content of a past dialogue efficiently. In contrast, our detector is used as a method for avoiding digression in an ongoing dialogue.

The criterion to determine whether an utterance is necessary is different for each of these applications. In our research, an "unnecessary" utterance is one that does not help the participants in the proceeding dialogue to achieve the goal of their discussion. In contrast, in traditional summarization, an "unnecessary" utterance is defined as an utterance that is not useful for a person reviewing past dialogue. Therefore, in our research, utterances necessary for ensuring the dialogue proceeds smoothly (e.g. back-channels, confirmations, or clarifications) are also considered as necessary, while traditional summarization tends to exclude these kinds of utterances.

In order to build and evaluate the proposed digression detector, we first collect a dialogue corpus assuming a situation where participants have a discussion with the goal of reaching a consensus, and annotate whether the utterances are necessary or unnecessary from the point of view of avoiding digressions. Next, we construct the detector for unnecessary utterances based on a statistical framework. Finally, we perform experimental evaluation of the detector. In the evaluation, we find that the performance of the proposed detector is higher than that of a traditional automatic summarization method.

II. Collection and Annotation of a dialogue corpus

As a target for our research, we collect a Japanese dialogue corpus where participants are playing the consensus game [9]. In the consensus game, participants discuss the importance of the some items (e.g. water or pistol) for survival in extraordinary situations: We assumed 3 extraordinary situations, disaster in the desert, moon, or jungle. In this research, two participants are considered as a one pair, and 8 pairs were recruited for the corpus collection. The agendas are different for each pair.

For constructing the digression detector, we annotate each of the utterances in the collected corpus as necessary or unnecessary. We obey the definition of digression as "a passage or section (in this research, utterance) that deviates from the central theme in speech or writing (in this research, agenda of discussion)" [10]. The corresponding utterances are considered as **unnecessary utterances**. Three annotators independently annotate tags for the transcribed utterances in the corpus.

The inter-annotator agreement of annotation is 71% (Fleiss's kappa [11] is 0.61). Considering the general criteria for agreement [12], we can see that the annotation substantially agrees. Most of the mismatched utterances contain sentences referring the usage of items. An example of such an utterance is "How do we eat the food while wearing a space suit?" The annotation also tended to vary when one of the participants was talking him/herself.

Note that the annotation of the unnecessary tag for avoiding digressions is quite different from that for excluding utterances in summarization. We compare the annotation of the unnecessary utterances for avoiding digressions with those for summarization by also having the annotators annotate **non-summarization** tags for the utterances which are unnecessary to create a summary. As there is no standardized method for creating summaries [2], we just give annotators the abstract

Utterance	Summarization	Digression Detection
If there is the space food, we don't need the milk.	+	+
Hmm, is milk useful?	-	+
(The milk) is useful if we have an accident in a snow-covered mountain.	-	+
I see.	-	+
Milk becomes energy.	-	+
I don't like milk.	-	-
I see.	-	-

+:Necessary utterance
-:Uncessary utterance

Fig. 1. Necessary and unnecessary utterances for summarization and digression detection.

instructions that they should extract only "utterances which represent an essential point." According to a t-test comparing the annotation results, we find that avoiding digressions and summarization are quite different ($p < 0.001$). An example of the each annotation result is shown in Fig 1. From this result we can see that, in summarization, utterances refering to decisions regarding the importance of an item are considered as necessary utterances, and the remaining utterances are considered as unnecessary utterances. In contrast, in avoiding digression, utterances expressing consent or providing information for decisions are also considered as necessary utterances.

III. Detection of unnecessary utterances

In this section, we propose an unnecessary utterance detector based on a statistical framework. In unnecessary utterance detection, utterances are the input, and the classifier determines whether the utterances are unnecessary or not. Therefore, we can deal with this detection problem as sequence labeling. In our research, we use conditional random fields (CRF) to learn the classifier [13]. CRFs are widely used for sequence labeling, where output symbols $\boldsymbol{y}$ are predicted from input symbols $\boldsymbol{x}$. Specifically, we consider a string of utterances as $\boldsymbol{x}$, and necessity labels (unnecessary or not) as $\boldsymbol{y}$. In this research, as the features of utterances, we use 1) surface information (Section III-A) and 2) information based on topic shift (Section III-B).

A. Surface features

Humans often use particular linguistic expressions as a cue for detecting unnecessary utterances. For example, a sentence containing "This is totally off the topic, but" is probably considered an unnecessary utterance. In contrast, the topic might return to the correct agenda after an utterance like "Let's get back to the original topic" appears.

Therefore, we use the frequency of words as surface linguistic features. The transcribed utterances are first tokenized, and word level 1-grams are calculated. In addition, we also use the number of the words as feature, as we can assume the length of the utterance has some relation to whether it is an unnecessary utterance.

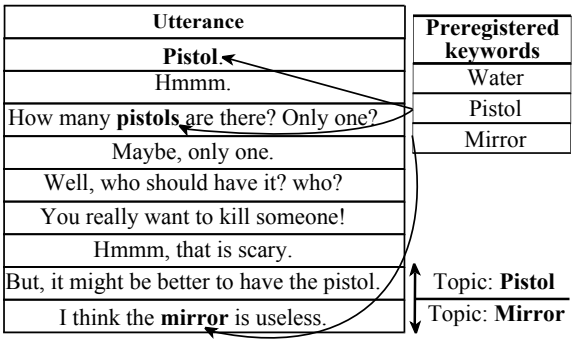

Fig. 2. An example of the keyword based topic tracking (translated from Japanese)

B. Features based on topic shifts

In addition to the cues described in the previous section, humans probably use the topic (or agenda) to which the current utterance belongs for detecting digression. To capture this intuition, we track topics to which each utterance belongs, and determine whether each topic is necessary or unnecessary.

We propose features based on topic shift for detecting digression. Topic shift is tracked with two tracking methodologies described in the following sections.

1) Keyword based topic tracking: This method tracks the topic shift based on a preregistered keyword's appearance in the utterance [14]. If the keywords appear in the utterance, later utterances are determined to belong to the topic corresponding to the keyword until another new keyword appears. In addition, if the new keyword appears in the utterance, this method determines that the topic has shifted to a new topic corresponding to the new keyword. As the preregistered keywords, we use noun phrases automatically extracted from the documents explaining the topic of the discussion. In the research of [14], preregistered keywords are manually determined. For reducing the bias of keyword registration, we use an automated method. We use noun extracted from the documents for instruction of corpus collection (Section II) as keyword. An example of the keyword-based topic tracking is shown in Figure 2.

As the features for keyword based topic tracking, the infomation about "whether a preregistered keyword appears or not" is encoded as a binary variable.

2) Topic tracking based on lexical chains: This method tracks the topic shift based on lexical chains between utterances. Lexical chains are semantic connections of words in a series of utterances. In this research, similar to previous research [15], [16], nouns are used for the lexical chain. If a lexical chain exists between the utterances, these utterances are tied to one topic. Otherwise, the method determines that the topic has shifted.

Specifically, the topic tracking algorithm based on lexical chains is described below. We call an utterance group which has the same topic as "chained utterances." In addition, we manually preregister connectives (e.g. back-channel) that tend

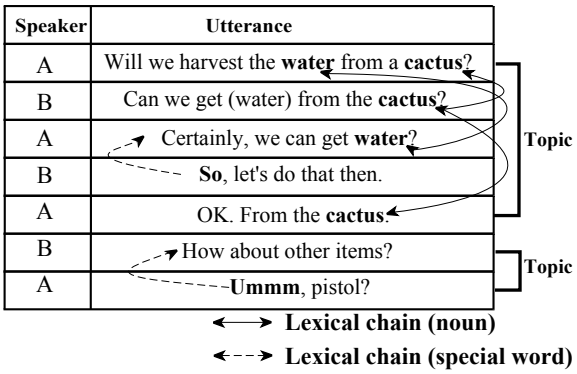

Fig. 3. An example of the topic tracking based on lexical chains

to semantically connect to the previous utterance as "special words."

- Morphological analysis is performed for the current utterance, and nouns and special words are extracted.
- If the current utterance contains a word extracted in the previous utterance, or a special word, the current utterance is added to the chained utterances.
- If the current utterance is not added to the chained utterances, this method determines that topic has shifted. In addition, the list of chained words is emptied.

An example of the topic tracking based on the lexical chain is shown in Figure 3.

As the features representing topic tracking based on lexical chains, four states representing "a new topic is started," "contains a word in the chain," "contains a word included in the partners previous utterances," and "contains a special word," are encoded as binary variables.

IV. Experimental evaluation of the unnecessary utterance detector

A. Experimental conditions

The performance of the unnecessary utterance detector is evaluated through leave-one-dialogue-out cross-validation. We use the 8 dialogue corpus annotated in Section II. The dialogue corpus consists of 1743 necessary utterances and 1429 unnecessary utterances. In each fold of cross-validation, we use one dialogue for test, and the remaining dialogues for training. As evalution criteria, we use accuracy (A), recall (P), precison (P), and F-measure (F). We use MeCab [17] for tokenizing utterances, and CRFSuite [18] as the CRF classifier.

For evaluating the effectiveness of the proposed method described in Section III, we compare with the performance of the following two baseline methods.

Maximal marginal relevance (MMR):
: MMR is widely known as an automated summarization method [19]. MMR needs to specify the number of sentences to be extracted beforehand. We set this to the average number of the sentences annotated as "necessary" in the traning dialogue corpus.

AllUnnecessary:
: AllUnnecessary classifies all utterances as unnecessary. This method represents a system that can not detect necessary utterances at all.

TABLE I
Experimental result of each method. The ground-truth label is decided by the majority of annotators.

Method	*A*	*R*	*P*	*F*
MMR	0.62	0.54	0.53	0.53
AllUnnecessary	0.40	1.00	0.40	0.57
1-gram	**0.71**	0.58	0.65	0.61
Chain	0.64	0.21	0.71	0.32
Key	0.65	0.40	0.58	0.46
Chain+Key	0.66	0.42	0.60	0.49
1-gram+Chain	**0.71**	0.57	0.65	0.60
1-gram+Key	**0.73**	0.59	0.68	0.63
1-gram+Chain+Key	**0.73**	0.60	0.66	0.63

TABLE II
Experimental result of each method for utterances in which all annotators agreed.

Method	*A*	*R*	*P*	*F*
MMR	0.65	0.56	0.55	0.54
AllUnnecessary	0.39	1.00	0.39	0.55
1-gram	**0.76**	0.66	0.70	0.67
Chain	0.68	0.28	0.78	0.40
Key	0.69	0.45	0.62	0.51
Chain+Key	0.70	0.47	0.64	0.54
1-gram+Chain	**0.76**	0.65	0.71	0.66
1-gram+Key	**0.78**	0.65	0.73	0.68
1-gram+Chain+Key	**0.78**	0.66	0.72	0.68

Further, the effectiveness of the features described in Sections III-A and III-B are examined. To do so, combinations of surface features (**1-gram**), features based on the registered keyword topic tracking (**Key**), and lexical chains (**Chain**) are evaluated.

B. Experimental results and discussion

Experimental results are shown in Tables I (with all utterances) and II (with utterances where all annotators agreed). These results indicate that our proposed method more accurately identifies unnecessary utterances than the baselines. Fisher's exact test indicates that the accuracy of all proposed methods significantly improved compared to MMR and AllUnnecessary ($p < 0.05$).

Focusing on comparison of the feature combinations in the proposed method, we can see that most effective feature set is the 1-gram features. In Tables I and II, the proposed methods that are not significantly different from 1-gram+Chain+Key according to Fisher's exact test are emphasized with bold. From this result, we can see that if the proposed method does not use 1-gram features, accuracy is significantly decreased.

Focusing on the difference between MMR and the proposed methods, we can see that proposed method correctly classifies utterances that are necessary to maintain the flow of the dialogue more than MMR. An example of a dialogue where the proposed methods works effectively is shown in Figure 4. Humans consider back-channels (e.g. "yes," "right") necessary for proceeding with the dialogue. However, MMR classifies these back-channels as unnecessary utterances. MMR considers utterances containing new information as necessary, but back-channeling appears many times in the dialogue. In

Speaker	Utterance	Human	1-gram	Chain+Key	1-gram+ Chain+Key	MMR
A	Light ranks ninth, and heater ranks eighth.	+	+	+	+	+
B	Yes.	+	+	+	+	-
A	And rope ranks seventh, uhh.	+	+	+	+	+
B	Yes.	+	+	+	+	+
A	FM ranks seventh, and the compass ranks fifth. Right?	+	+	+	+	+
A	Is it better to inverse the order?	+	+	+	+	+
B	Right.	+	+	+	+	-
A	Hmm, so, sixth is the compass.	+	+	+	+	+
B	OK.	+	+	+	+	-

+:Necessary utterance
-:Uncessary utterance

Fig. 4. An example of the dialogue where the proposed method works effectively. *Human* represents the evaluation result of the human annotator.

Speaker	Utterance	Human	1-gram	Chain+Key	1-gram+ Chain+Key	MMR
A	If the amature shot the pistol, he never hit.	+	+	+	+	+
B	He is the captain and guide for the jungle.	+	+	+	+	+
A	He is useless.	-	-	+	-	+
B	Certainly, he mishandled the ship.	-	+	+	+	-
A	Yes, he made the ship run into a stump.	-	+	+	+	+
B	Useless, he is.	-	-	+	+	+
A	He should do it because he is useless.	-	-	+	+	+
B	He be useless until the end.	-	-	+	-	-
A	Useless until the end.	-	-	+	-	-

+:Necessary utterance
-:Uncessary utterance

Fig. 5. An example of a dialogue where Chain+Key does not work well.

contrast, the proposed method correctly classifies these backchannels. From an error analysis, we found that the methods based on the topic tracking do not work well for topics consisting of both necessary and unnecessary utterances. An example of a dialogue where Chain+Key does not work well is shown in the Figure 5. The human annotators consider the second utterance as necessary, and most of the later utterances as unnecessary. However, Chain+Key misclassifies these later utterances. At first, Chain+Key classifies this utterance as necessary, and determines the second and third utterances to belong to the same topic because of the lexical chain of "captain". Chain+Key determines the later utterances and third utterance belong to the same topic because of the lexical chain of "captain" and "useless." Thus, these later utterances are also classified as necessary, resulting in misclassification.

V. Conclusion

In this paper we proposed a method for avoiding digressions by detecting unnecessary utterances based on linguistic features and features based on topic shifts. Experimental results indicated that the proposed detector is more accurate in identifying human annotations of necessary/unnecessary utterances than MMR. In addition, in our proposed detector, 1-gram features were most effective in improving the performance.

As future work, we plan to apply the proposed detector to a dialogue corpus in a different domain and investigate the effect on accuracy. We also plan to investigate with corpora automatically transcribed with speech recognition. Further, we plan to compare other supervised summarization methods. In addition, based on the proposed detector, we plan to develop and deploy a full end-to-end system that avoids digressions in conversation.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number 24240032 and by the Commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

[1] Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur, "Long story short-global unsupervised models for keyphrase based meeting summarization," *Speech Communication*, 2010.
[2] Gabriel Murray, Steve Renals, and Jean Carletta, "Extractive summarization of meeting recordings," *Proceeding of INTERSPEECH'2005 - Eurospeech*, 2005.
[3] Yang Liu and Shasha Xie, "Impact of automatic sentence segmentation on meeting summarization," *Proceedings of ICASSP*, 2008.
[4] Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore, "Evaluating automatic summaries of meeting recordings," *Proceedings of ACL*, 2005.
[5] Lu Wang and Claire Cardie, "Focused meeting summarization via unsupervised relation extraction," *Proceedings of the SIGDIAL*, 2012.
[6] Klaus Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, 2014.
[7] Gabriel Murray and Steve Renals, "Towards online speech summarization," *Proceedings of INTERSPEECH*, 2007.
[8] Inderjeet Mani, "Automatic summarization (vol. 3)," *John Benjamins Publishing*, 2001.
[9] J.C Lafferty, P.M. Eady, and J. Elmers, *The desert survival problem*, Experimental Learning Methods, 1974.
[10] K Dictionaries Ltd., "RANDOM HOUSE KERNERMAN WEBSTER'S college dictionary," *Random House and Inc*, 1993.
[11] Joseph L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological Bulletin*, 1971.
[12] J. Richard Landis and G. Koch Gary, "The measurement of observer agreement for categorical data.," *Biometrics*, 1977.
[13] John Lafferty, Andrew McCallum, and Fernando CN Pereira., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.," *Proceedings of the ICML*, 2001.
[14] Hiroshi Ichikawa and Takenobu Tokunaga, "An empirical study on detection and prediction of topic shifts in information seeking chats," *Proceedings of SemDial*, 2007.
[15] Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Honyan Jing, "Discourse segmentation of multi-party conversation.," *Proceedings of ACL*, 2003.
[16] Pei-Yun Hsueh and Johanna D Moore, "Combining multiple knowledge sources for discourse segmentation.," *Proceedings of ACL*, 1995.
[17] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto, "Applying conditional random fields to Japanese morphological analysis.," *Proceedings of EMNLP*, 2004.
[18] Naoaki Okazaki., "CRFsuite: A fast implementation of Conditional Random Fields (CRFs)," 2007.
[19] Jaime Carbonell and Jade Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries.," *Proceedings of ACM SIGIR*, 1998.