Topic Detection Based on Dialogue History

Takayuki NAKATA, Shinichi ANDO, Akitoshi OKUMURA

Multimedia Research Laboratories, NEC Corporation

4-1-1, Miyazaki, Miyamae-ku, Kawasaki, KANAGAWA, 216-8555, JAPAN

t-nakata@bk.jp.nec.com, s-ando@cw.jp.nec.com, a-okumura@bx.jp.nec.com

Abstract

In this paper, we propose a topic detection method using a dialogue history for a speech translator. The method uses a k-nearest neighbor method for the algorithm, automatically clusters target topics into smaller topics grouped by similarity, and incorporates dialogue history weighted in terms of time to detect and track topics on spoken phrases. From the evaluation of detection performance using test data comprised of realistic spoken dialogue, the method has shown to perform better with clustering incorporated, and when combined with dialogue history of three sentences, gives detection accuracy of 72.1%.

1 Introduction

In recent years, demand for international information exchange has rapidly increased with the advance of globalization and cross-border utilization of information. Consequently, machine translation technologies have achieved significant progress, and speech translator has enhanced its practical value for certain settings such as a travel domain (Watanabe et al., 2000). Further improvement in quality of translation technology will expand its applicable domains and the user market.

As the translation system is improved to increase its applicable situations and topics, higher translation quality using dialogue context and scene knowledge becomes a crucial factor. For example, we take an example phrase in a travel conversation, "It's rare". In a restaurant, the word "rare" probably means that the meat is cooked for a short period of time. In other situations, the word "rare" most likely means uncommon and not widely known. The word should be appropriately translated if the target language has different expressions for these two different situations. User's previous input sentences can help identify the sentence topic, thus disambiguate polysemy and improve the quality of speech translation.

We have conducted studies on domain adaptation techniques to enable speech translation system embodied with understanding of what a topic is. Domain adaptation techniques are comprised of 1) dynamic replacing of the topic knowledge according to the theme transition to process direct translation; 2) semi-automatic extracting and accumulating of knowledge on respective topics; and 3) topic detecting and tracking of input data.

This paper proposes and examines the topic detection on speech translation system, which can link previous dialogue sentences to identify topics that are relevant to disambiguate polysemy and improve the speech recognition accuracy. It also investigates the method's limitation and its possible solutions.

The remainder of this paper is organized as follows. Section 2 describes the constraints in detecting a topic from dialogue utterances. Section 3 describes our topic detection algorithm to overcome these constraints. Section 4 explains the evaluation of our method using a travel conversation corpus and Section 5 presents the evaluation result. Section 6 discusses the effect of our method from a comparison of the results on typical dialogue data and on real situation dialogue data. We conclude in Section 7 with some final remarks and directions for future work.

2 Topic detection

The topic detection module uses one spoken phrase in a dialogue as an input. It dynamically tracks the topic transitions and outputs most appropriate topic as a detected topic. We use the term "topic" to define an abstract scene determined by a purpose and an occurrence at a given place covered by the conversation, and any surrounding knowledge relevant to the scene. For example, in a speech translation system for travel conversation, topics may include "Hotel", "Restaurant", "Sightseeing", and "Emergency".

Among conventional topic detection methods, one uses compound words that features certain topic as trigger information for detecting a topic (Hatori et al., 2000), and another uses domain-dependant dictionaries and thesauruses to construct knowledge applicable to a certain topic (Tsunoda et al., 1996). In the former method, a scene-dependant dictionary provides the knowledge relevant to the scene and compound words in the dictionary are used for detecting a topic. In the latter method, words appearing in a scene are defined as the knowledge relevant to the scene and superordinate/subordinate relation and synonyms provided by thesauruses are used to enhance the robustness.

These conventional methods are suitable for written texts but not for dialogue utterances in a speech translator. A speech translator requires:

- Topic detection for each utterance in a dialogue;
- Prompt topic detection in real time processing;
- Dynamic tracking of topic transition.

The following two major constraints make the topic detection for dialogue utterances more difficult.

(1) Constraint due to single sentence process
   - Sentences in a dialogue are usually short with few keywords.
   - In a dialogue, the frequency values of the word in a sentence are mostly one, making it difficult to apply a statistical method.

(2) Constraint due to the nature of spoken dialogue
   - In a dialogue, one topic is sometimes expressed using two or more sentences.
   - The words appearing in a sentence are sometimes replaced by anaphora or omitted by ellipsis in the next sentence.
   - Topics frequently change in a dialogue.

To make topic detection adaptive to the speech translator, we propose a topic detection method which accepts single utterance as an input, detects the topic transitions dynamically and outputs most appropriate topic for the latest utterance. The k-nearest neighbor method (Yang, 1994) is used with the clustering method linked with the dialogue history as a topic detection algorithm for dialogue utterance. The k-nearest neighbor method is known to have high precision performance with less restriction in the field of document categorization. This method is frequently used as a baseline in the field and also applied to topic detection for stories but not for a single sentence (Yang et al., 1999). We incorporated two new methods to the k-nearest neighbor method to overcome the constraints mentioned above.

To overcome the first constraint we clustered a set of sentences in training data into subsets (called subtopics) based on similarity between the sentences. A topic is detected by calculating the relevance between the input sentence and these subtopics. Clustering sentences on the same subtopic increases the number of characteristic words to be compared with input sentence in calculation.

To overcome the second constraint, we grouped an input sentence with other sentences in the dialogue history. A topic is detected by calculating the relevance between this group and each possible topic. Grouping the input sentence with the preceding sentences increases the number of characteristic words to be compared with topics in calculation. We consider the time sequence of sentences in a dialogue in calculating the relevance to avoid the influence of topic change in the dialogue.

3 Topic Detection Algorithm

This section explains three methods used in the proposed topic detection algorithm: 1) k-nearest neighbor method, 2) the clustering method using TF-IDF, and 3) the application of the dialogue history.

3.1 k-nearest neighbor method

We denote the character vector for a given sentence in the training data as D_j, and that for a given input sentence as X. Each vector has a TF-IDF value of the word in the sentence as its element value (Salton, 1989).

The similarity between the input sentence X and the training data D_j is calculated by taking the inner product of the character vectors.

$$Sim\ (X, D_j) = \frac{\sum_i x_i \times d_{ij}}{\| X \|_2 \times \| D_j \|_2}$$

The conditional probability of topic C_l being related to the training data D_j is calculated as:

$$Pr(C_l \mid D_j) = 1 \Big/ \begin{pmatrix}\text{The number of topics } C_l \\ \text{being related to the } D_j\end{pmatrix}$$

The relevance score between the input sentence X and each topic C_l is calculated as the sum of similarity for k sentences taken from the training data in descending order of similarity.

$$Rel(C_l \mid X) = \sum_{D_j \in \{k \text{ top ranking sentence}\}} Sim(X, D_j) \times Pr(C_l \mid D_j)$$

3.2 Topics clustering method

This method clusters topics into smaller subtopics. The word "topic" used in this method consists of several subtopics representing detailed situations. The topic "Hotel" consists of subtopics such as "Checking In" and "Room Service". Sentences in training data categorized under the same topic are further grouped into subtopics based on their similarity. Calculating the relevance between the test data input and these subsets of training data provides more keywords in detecting topics. Our method to create the subtopics identifies a keyword in a sentence set, and then recursively divides the set into two smaller subsets, one that includes the keyword and one that does not.

TF-IDF Clustering Method

(1) Find the word that has the highest TF-IDF value among the words in the sentence set;
(2) Divide the sentence set into two subsets; one that contains the word obtained in step (1) and one that does not;
(3) Repeat steps (1) and (2) recursively until TF-IDF value reaches the threshold.

Subtopics created using this method consist of keywords featuring each subtopic and their related words.

3.3 Application of the dialogue history

We incorporated the dialog history to improve the topic detection. The method interprets a current input sentence and the sentences prior to the current input as a dialogue history subset, and detects topics by calculating the relevance score between the dialogue history subset and the each topic. The method increases the number of keywords in the input for calculation. Each sentence in the dialogue history subset is weighted to control the effect of time sequence.

The relevance score combined with the dialog history is calculated as:

$$Rel(C_l \mid X, Xr_1, ..., Xr_n) = \lambda\, Rel(C_l \mid X) + \lambda r_1\, Rel(C_l \mid Xr_1) + ... + \lambda r_n\, Rel(C_l \mid Xr_n)$$

Here the similarity is calculated with the input sentence X and the sentence in the dialog history subset Xr_i, taking λ and λr_i as the weights for the input sentences and the sentences in the dialogue history, respectively.

4 Evaluation

To evaluate the proposed method, we prepared training data and test data from a travel conversation corpus. We also prepared three different thresholds for clustering and two sets of weight values for dialogue history.

4.1 Training Data

In the evaluation, we used approximately

25,000 sentences from our original travel conversation corpus as our training data. The sentences are manually classified into four topics: 1) Hotel, 2) Restaurant, 3) Shopping, and 4) Others. The topic "Others" consists of sentences not categorized into the remaining three. Topics such as "Transportation" or "Illnesses and injuries" are placed into this "Others" in this evaluation. Sentences which fall under multiple topics can be categorized into any of the candidate topics and sentences which may fall under all four topics are specially categorized as "General conversations". Variations of the four topics produce 15 probable combinations: "Hotel", "Hotel and Restaurant", "Restaurant and Shopping" for example.

4.2 Test Data

We prepared two sets of test data. One set consists of 62 typical travel dialogues comprising 896 sentences (hereafter called "typical dialogue data"). The other set consists of 45 dialogues comprising 498 sentences, which may include a variety of expressions but closely representing daily spoken language (hereafter called "real situation dialogue data"). Each dialogue consists of about ten conversation sentences carried on in travel circumstances. All sentences are manually classified into the topics once with their preceding context presented and once without it.

Sentences in "typical dialogue data" are often heard in travel planning and travelling situations, and form a variety of initiating dialogues as the travel conversation unfolds. The data includes words and phrases often used in the topics listed above, and each sentence is short with little redundancy. On the other hand, "real situation dialogue data" consists of spoken dialogue phrases which are likely to appear in real situations in the travel domain. Some phrases may be typically used, while others may consist of more colloquial expressions and words and phrases that are redundant. Some of the words may not appear in the training data.

4.3 Clustering the topics

We applied the clustering with the aforementioned method to 8,457 sentences from training data which are categorized into one or more of the three topics: 1) Hotel, 2) Restaurant, and 3) Shopping. Clusters are created on three arbitrarily defined different thresholds: 8,409 clusters (small-sized cluster), 3,845 clusters (medium-sized cluster) and 2,203 clusters (large-sized cluster). To implement the clustering, we created a cluster using the TF-IDF value of each word in sentences. We set one sentence as one cluster if the sentence does not contain a word whose TF-IDF value is not equal to or greater than the threshold. We excluded data that falls only under the topic "Others" and data that falls under all four topics, which are considered as "General conversations" in clustering. Excluding these two topics produces 13 combinations to be clustered. The number of clusters for the above 8,457 training data is smallest (13) when we set one topic as one cluster and largest (8,457) when we set one sentence as one cluster.

4.4 Use of the dialogue history

We are interested in the effect of the dialogue history from two different perspectives, the contribution of each sentence in the dialogue history to the detection accuracy and the number of sentences in the dialogue history required to reach the sufficient detection accuracy.

To evaluate the contribution of each sentence in the dialogue history, we use an input sentence, the most preceding and the next preceding sentence (hereafter "sentence 0", "sentence -1", and "sentence -2") as a dialogue history. Two types of sentence weights are applied to these three sentences, one of equal weights, and one of weights based on a time series. These sets are:

$$(\text{sentence } 0, \text{sentence } -1, \text{sentence } -2) = (0.33, 0.33, 0.33)$$

$$(\text{sentence } 0, \text{sentence } -1, \text{sentence } -2) = (0.5, 0.3, 0.2)$$

The first weighting is considered a base line which weights each sentence with equal value. The second weighting is a modification that weights sentences according to their importance, which is assumed greater for sentences that appear closer to "sentence 0". Sentences that appear closer are more pertaining to the current topic than earlier sentences.

To evaluate the number of sentences in the dialogue history required to reach sufficient

detection accuracy, we use a certain number of preceding sentences in the dialogue history with the equal weight for each sentence.

5 Results

We performed the detection test on 13 kinds of topic combinations described in 4.3 using typical dialogue data and real situation dialogue data to examine the effect of clustering. Moreover, we performed the detection test for two weight sets described in 4.4 using typical dialogue data and real situation dialogue data.

5.1 Test result on typical dialogue data and real situation dialogue data

We evaluated the effect of the clustering for typical dialogue and real situation dialogue. All sentences for test data are classified with context and without it. The result without context is used as the answer since the dialogue history isn't implemented at this time. The detected topic is evaluated as correct only when the detected topic exactly matches the answer. That is, for the answer "Hotel and Restaurant", the detected topic "Hotel and Restaurant" is correct, but not the "Hotel".

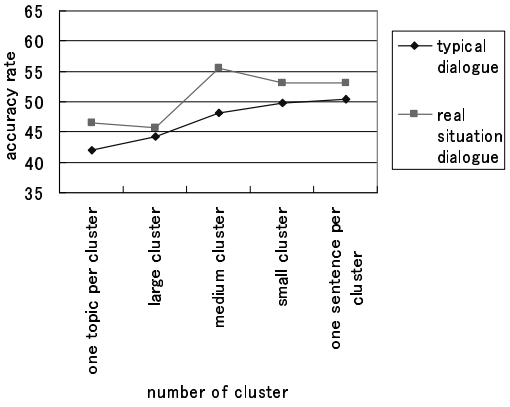

Figure 1 : The result for the clustering

Figure 1 shows the results of topic detection on typical dialogue data and real situation dialogue data for a varying number of clusters. The figure shows that the accuracy is highest when one sentence is set as one cluster (one sentence per cluster) in each topic, and lowest when one whole topic is set as one cluster for typical dialogue. The figure also shows that the accuracy of the medium cluster is slightly better than that for one sentence per cluster for real situation dialogue data. This indicates that sentences grouped in terms of some criterion heighten the validity of similarity calculation between input sentences and the training data, and consequently the detection accuracy rate is improved.

5.2 Results of dialogue history application

We evaluated the effect of the dialogue history for typical dialogue test data, and compared the case of one sentence per cluster with the case of medium cluster. This time all sentences are categorized taking the context into account in order to evaluate the impact of dialogue history. Incorporating sentences in the dialogue history improved the accuracy rate as we expected, and for one-sentence-per-cluster case, the equally weighted aggregate achieved the best accuracy rate of 72.1% with three sentences as a dialogue history (Figure 2).

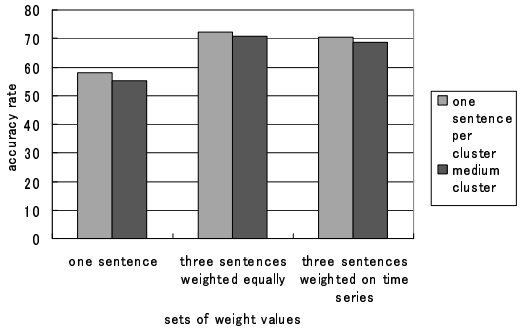

Figure 2 : The result for the weighted dialogue history

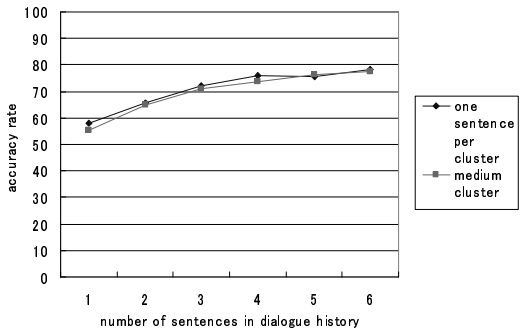

Figure 3 : The result for the number of dialogue history

Figure 3 shows the result of topic detection with change in number of sentences in a dialogue history. The accuracy rate improved as the number of incorporated sentences increased and the best accuracy rate achieved was about 80%.

6 Discussion

The approach using clustering implementation was observed with threshold variance. Our results using the typical dialogue data demonstrated that the one-sentence-per-cluster case achieved the best accuracy rate of all clustering cases. This is due to nature of typical dialogue and its short sentences, frequently allowing the input data's feature words to exactly match those of the training data.

On the same detection results, the accuracy rate of one-topic-per-cluster case was substantially poorer than all other clustering cases, primarily due to large number of sentences merged into one cluster, making each feature word less effective. For example, the training data sentence "Is it all right to pick it up with my hand?" may be categorized under the topic "Shopping" as this conversation phrase can be used to a shop clerk. The one-sentence-per-cluster case correctly categorizes the input sentence: "Is it all right to pick up this with my hand?" to the same topic as it nearly matches the above sentence. The case using one-topic-per-cluster appears to degrade by miscategorizing the same input sentence into the topic "Others". This topic contains phrases like "My hand hurts" or "I feel all right lately" according to their topic similarity in describing physical conditions. One-topic-per-cluster puts the input sentence under the same topic based on matching words "my hand" and "all right". Topic detection performs better with a large topic divided into smaller groups or even into single sentences when the approach is applied to a typical dialogue.

The results using the real situation dialogue data showed that the one-sentence-per-cluster case achieved higher accuracy rate than one-topic-per-cluster case. Moreover, the medium-cluster case performed better than one-sentence-per-cluster case in accuracy. This attributes to the nature of sentences in real situation dialogue. The sentences are often redundant, incomplete or short of feature words that the feature words of an input sentence and those of a training data sentence seldom match exactly when using one-sentence-per-cluster case. The medium-cluster case seems to cover this shortcoming by incorporating subtopics. The case groups sentences using the clustering method described in 3.2, so that each group (called subtopic) is composed of words used to represent the subtopic and its related words. These subtopics seem to form a kind of context and help to improve the topic detection. Clustering single sentences into small groups improves the topic detection performance than leaving them as singleton for the real situation dialogue.

We found that the hit/miss performance of topic detection on a same sentence varied significantly between the one-sentence-per-cluster case and the medium-cluster case. We had 12 input sentences that are incorrectly detected in one-sentence-per-cluster case but correctly detected in the medium-cluster case. All these input sentences do not entirely match the training data sentences. The clustering is advantageous when there is no strong feature word and the topics are determined by sets of words in the sentence. We had 9 input sentences that are correctly detected in one-sentence-per-cluster case but not in the medium-cluster case. One-third of these input sentences exactly matched the training data sentences. In the case where the strong feature words exist in input sentences, clustering does not improve and sometimes hurt the performance of topic detection.

It seems clear from the results that the one-sentence-per-cluster is advantageous for the input sentences in the typical dialogue and the medium-cluster is favorable for the input sentences in the real situation dialogue. When compared to one-sentence-per-cluster, the medium-cluster topic detection performance is almost equal for the typical dialogue data and somewhat superior for the real situation dialogue. A practical improvement in the performance would result from a better estimation of optimal topic clusters applied to both typical and real situation dialogues derived from large amount of travel corpus.

We examined the effect of implementing a dialogue history in topic detection. Equally weighted dialogue history demonstrated moderately better than the time series weighted history in terms of accuracy rate.

We examined the sentences in the dialogue history and found that the sentences categorized into single topic is 45.1% of all sentences. 39.6% of all sentences are categorized as "General conversation", which means it cannot specify a relevant topic. When we apply the weight set based on time series and the later sentence is categorized into "General conversations" or into two or more topics, this sentence will likely lead to a false topic detection. On the other hand, when the sentences in a dialogue history is equally weighted, the topic detection performance for the sentence representing only one topic is relatively better than other sentences, leading to the correct result.

The effect of varying the number of sentences in the dialogue history was also examined. The performance improved as the number of dialogue history sentences increased. The accuracy rate productively increase up to 3 or 4 preceding sentences, then reach about 76% with 4 preceding sentences, then there is limited benefit in adding sentences after this point. This result appears adequate for the future practical use.

7 Conclusions

In this paper, we proposed a topic detection method using a dialogue history for a speech translator. We investigated its limitation in dialogue utterances and provided solutions by clustering training data and utilizing dialogue history. Our method showed topic detection accuracy of at least 50% for both typical and real situation dialogues in 13 topic combinations and 72.1% with three sentences in dialogue history. For typical dialogues, we found that the best results were obtained when one sentence is used for one cluster, and for real situation dialogues, slightly better results were obtained when clustering was introduced. It seems clear to us that the topic detection accuracy is improved for both typical and real situation sentences if an appropriate size cluster is introduced.

We intend to use our topic detection technique for specifying a scene condition in our speech translator (Ikeda et al., 2002). Topic detection also helps improve accuracy of the speech translator by disambiguating polysemy and selecting a correct word dictionary and resources, which are organized according to the topic.

Our future work will focus on linking the dialogue history and successful clustering to improve the topic detection accuracy.

References

H. Hatori, Y. Kamiyama (2000) *Web translation by feeding back information for judging category*, Information Processing Society of Japan 63rd. Annual Meeting, Vol. 2, pp. 253-254.

T. Ikeda, S. Ando, K. Satoh, A. Okumura, T. Watanabe (2002) *Automatic Interpretation System Integrating Free-style Sentence Translation and Parallel Text Based Translation*, ACL-02 Workshop on Speech-to-speech Translation (to appear).

G. Salton (1989) *The vector space model, automatic text processing — the transformation, analysis, and retrieval of information by computer*, Addison-Wesley Publishing Company Inc., pp.312-325.

T. Tsunoda and H. Tanaka (1996) *Evaluation of Scene Information as Context for English Noun Disambiguation*, Natural Language Processing, Vol.3 No.1, pp. 3-27.

T. Watanabe, A. Okumura, S. Sakai, K. Yamabana, S. Doi, K. Hanazawa (2000) *An Automatic Interpretation System for Travel Conversation*, The Proceeding of the 6th International Conference on Spoken Language Processing Vol. 4, pp. 444-447.

Y. Yang (1994) *Expert Network, Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval*, Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94) 1994:11-21.

Y. Yang, J.G. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu (1999) *Learning approaches for detecting and tracking news events*, IEEE Intelligent Systems, 14(4), pp. 32-43.