Proceedings of the Workshop on Speech-to-Speech Translation:
Algorithms and Systems, Philadelphia, July 2002, pp. 9-14.
Association for Computational Linguistics.

Topic Detection Based on Dialogue History

Takayuki NAKATA, Takahiro IKEDA, 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, t-ikeda@di.jp.nec.co.jp, 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 selecting a scene in the automatic interpretation system (Ikeda et al., 2002). 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 corpus comprised of realistic spoken dialogue, the method has shown to perform better with clustering incorporated, and combined with time-weighted dialogue history of three sentences, gives detection accuracy of 77.0%.

1 Introduction

In recent years, speech-to-speech translation systems have been developed that integrate three components: speech recognition, machine translation, and speech synthesis (Watanabe et al., 2000). However, these systems cannot guarantee accurate translation because the individual components do not always provide correct results. To overcome this restriction, we proposed a method to use parallel text based translation for supporting free-style sentence translation. In addition, we built a prototype automatic interpretation system for Japanese overseas travelers (Ikeda et al., 2002). With this system, the user searches for an appropriate sentence in source language from the registered parallel text by using the criteria of an utterance, a scene, and a situation, and then uses the target language sentence for a translation.

Although parallel text based translation provides guaranteed translation results, it has two problems as the user searches for the sentence. One is difficulty in searching an appropriate sentence from user's short utterance, which is often heard in travel conversation. Short phrases provide only a few keywords and make the search result too broad. Specifying the exact scene and action helps narrow down the result, but the task may cause user frustration in having to select the right option from the vast categories of scenes and actions.

The other problem is existence of nonadaptive sentences that may be inappropriate in some of the scenes. Users usually select sentences according to the scenes so they can exclude those inapplicable sentences, but some new users may accidentally select those nonadaptive sentences by failing to specify a scene.

Here, we propose a method to detect a topic for each utterance. We define a topic as corresponding to a scene that is a place or a situation in which the user converses. The proposed method is based on the k-nearest neighbor method, which is improved for dialogue utterances by clustering training data and using dialogue history. We use the detected topic for specifying a scene condition in parallel text based translation, and thereby solve the two problems described above.

Detecting topics also helps improve accuracy of the automatic interpretation system by disambiguating polysemy. Some words should be translated into different words according to the scene and context selection. Topic detection can enhance speech recognition accuracy by selecting the correct word

dictionary and resources, which are organized according to the topic.

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 by 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 mention of future work.

2 Topic detection

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 translation system. 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 with 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.

On the other hand, a speech translation system requires the following:
- Topic detection for each utterance in a dialogue;
- Prompt topic detection in real time processing;
- Dynamic tracking of topic transition.

To make topic detection adaptive to the speech translation system, we propose a method applicable to one utterance in a dialogue as an input, which can be used for tracking the topic transitions dynamically and outputting 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 story but not for a single sentence (Yang et al., 1999). This paper incorporates two new methods to the k-nearest neighbor method to overcome two constraints mentioned above.

To overcome the first constraint, we cluster 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 number of characteristic words to be compared with input sentence in calculation.

To overcome the second constraint, we group 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 number of characteristic words to be compared with topics in calculation. We consider the

order of the sentences in the 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 by this method consist of keywords featuring each subtopic and their related words.

3.3 Application of the dialogue history

The proposed method applies the dialog history in 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 number of keywords in the input for calculation. We assign a weight to each sentence in the dialogue history subset to control the effect of time-sequence in sentences.

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

types of clusters with different thresholds and two types of dialogue history with different weight values.

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.

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 irregular expressions but closely representing daily spoken language (hereafter called "real situation dialogue data").

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 user-specific situations in the travel domain. Some phrases may be typically used, while others may consist of 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 different thresholds: 8,409 clusters (small-sized cluster), 3,845 clusters (medium-sized cluster) and 2,203 clusters (large-sized cluster). In carrying out clustering, 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 to be general conversation. Variations of these topics produce 13 probable combinations.

The number of clusters 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

To evaluate the effect of 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:

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

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

5 Results

We performed the detection test described in 4.3 on 13 types of topic combinations using typical dialogue data and real situation dialogue data.

5.1 Test results on typical dialogue data

Figure 1 shows the results of topic detection on typical 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.

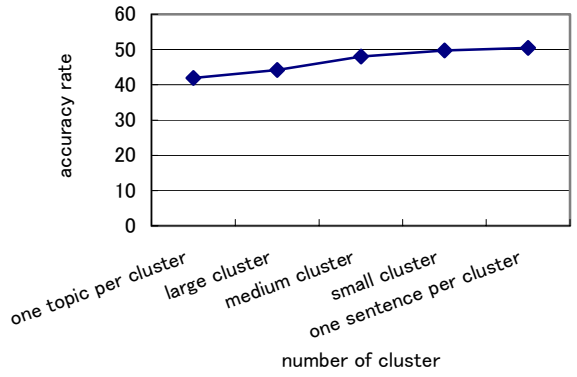
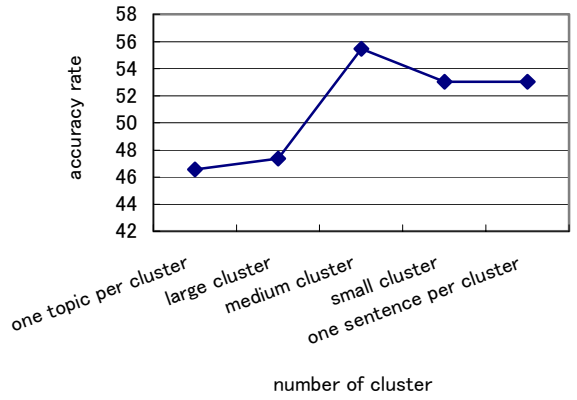


Figure 1: The result on typical test data

5.2 Test result on real situation dialogue data

Figure 2 shows the results of topic detection on real situation dialogue data for a varying number of clusters. The figure shows that the accuracy of the medium cluster is slightly better than that for one sentence per cluster. This indicates that sentences grouped in terms of similarity heighten the accuracy of similarity calculation between input sentences and the



training data.

Figure 2: The result on real situation test data

5.3 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. Using only the input sentence, the topic detection accuracy was 59.2% for the former and 56.0% for the latter. Using three sentences from the dialogue history, the respective figures were 72.0% and 70.0% with equal weights, 76.7% and 77.0% with time series weights.

6 Discussion

Looking at the results on the typical dialogue data, it can be argued that the one-sentence-per-cluster case shows the highest accuracy because the data is a typical dialogue and each sentence is short, so that feature words in the input sentences and those of the learning data are likely to match. On the other hand, it can be argued that the one-topic-per-cluster case shows the lowest accuracy because feature words become less effective when so many subtopics are in one cluster.

For example, let us look at the sentence in the learning data, “Is it all right to pick it up with my hand?” This sentence can be used when deciding what to buy, and so is categorized under the topic “Shopping”. When a cluster is one sentence, the result will likely be satisfactory if you input the sentence, “Is it all right to pick it up with my hand?” because the input sentence is similar to the cluster. However, when a cluster is one topic, this sentence might be categorized under the topic “Others”, along with sentences used to express physical conditions such as “My hand hurts” or “I am all right”. Therefore, it can be concluded that it is better to divide a large topic into smaller groups or even into single sentences.

Looking at the results on real situation dialogue data, we find the ratio of correct answers is almost the same for the one-sentence-per-cluster and the medium-cluster cases, but the actual sentences correctly detected topics differed significantly between them. In the former case, topics are identified correctly when there are strong feature words, while in the latter case, it works well when there is no strong feature word but the topics can be determined by sets of words. From this fact, we can conclude that typical input sentences can be compared easily with the one-sentence-per-cluster case, and real situation input sentences can be

compared with the medium-cluster case even though the sentences are different from those in typical dialogue in terms of content and expressions. We find that with typical dialogue data, the accuracy level is almost the same for the one-sentence-per-cluster and the medium-cluster cases, but with the real situation dialogue data, the accuracy level is slightly improved. Therefore, it might be possible to improve the practicality of topic detection by collecting a large amount of data, dividing the data into typical and real situation dialogues, and setting the appropriate clusters to each type.

7 Conclusions

In this paper, we proposed a topic detection method using a dialogue history to select a scene for the automatic interpretation system. 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. For typical dialogues, we found that the best results were obtained when one sentence is used for one cluster, and for real situation dialogues, we found slightly better results were obtained when clustering was introduced. Therefore, it can be argued that the topic detection accuracy is improved for both typical and real situation sentences if an appropriate size cluster is introduced.

We plan to use our topic detection technique for specifying a scene condition of parallel text based translation in our automatic interpretation system. Detecting topics also helps improve accuracy of the automatic interpretation system by disambiguating polysemy. Topic detection can enhance speech recognition accuracy by selecting the correct word dictionary and resources, which are organized according to the topic.

Our method is also applicable in determining time series behavior such as topic transition. Our future studies will focus on linking the dialogue history and clustering more closely 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.