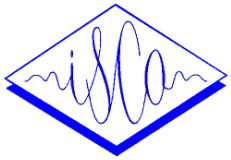

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

TOPIC DETECTION OF AN UTTERANCE FOR SPEECH DIALOGUE PROCESSING

Katsushi Asami, Toshiyuki Takezawa, and Genichiro Kikui

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{katsushi.asami, toshiyuki.takezawa, genichiro.kikui}@atr.co.jp

ABSTRACT

This paper proposes a method for topic detection of an utterance for speech dialogue processing including speech translation. There are three important points to consider in realizing topic detection of an utterance. The first is to obtain information from a short utterance. The second is to deal with a broad topic. The third is to have robustness against speech recognition errors. Our topic detection method is suitable as a solution to these problems. To deal with the first two problems, our method detects the topic of an utterance according to the relation factor between the topics and the words included in utterances. The third problem is solved by merging words with different surface forms. To verify the performance of our proposed method, we carried out experiments by using a broad-coverage travel conversation corpus. The experimental results show that the proposed method achieved 79.3% topic detection accuracy for correct transcriptions, and 75.6% accuracy for recognizer's outputs whose WER was 12.8%.

1. INTRODUCTION

A great deal of research on topic detection has been done to develop practical methods to perform text categorization for newspaper articles as written language [1], categorization of broadcast news as spoken language [2][3], and natural language call steering for conversational spoken language [4][5][6]. These methods are effective for information retrieval or automatic call steering tasks. Topic detection is also expected to improve speech recognition accuracy by automatically switching to the appropriate language models according to the topic. Furthermore, for speech translation, topic detection is expected to choose the appropriate translation according to the topic, and in a human-machine dialogue system, to anticipate a user's request and switch the dialogue scheme appropriately. We assume that topic detection must satisfy the requirements listed below to contribute to speech dialogue processing.

(1) Deal with a broad topic
(2) Detect a topic from a short utterance in spoken language.
(3) Detect a topic on utterance-by-utterance basis.
(4) Have robustness against speech recognition errors.

The word "*robustness*" in this paper means that topic detection is hardly influenced by speech recognition errors.

We studied topic detection by using a travel conversation corpus. The topics in travel conversations cover various categories, such as boarding procedures, lodging, dining, shopping, and meeting with troubles. This variety is common characteristic to newspaper articles or broadcast news. On the other hand, as conversational spoken language, it is also has common characteristic to the call steering task. Topic detection in travel conversations is interesting because of these characteristics. Moreover, an utterance in travel conversations is comprised of 10 or at most 20 words in contrast to at least tens of words in newspaper articles or broadcast news. From this standpoint, it is more difficult to categorize such an utterance than newspaper articles or broadcast news, and it is a challenging task.

In this paper, we propose a method for topic detection that satisfies the requirements mentioned above to get an idea about how to apply topic detection to speech dialogue processing. The method proposed in this paper is currently capable of detecting a topic from Japanese utterances only.

In section 2, we briefly introduce the broad-coverage bilingual basic expression corpus, which contains topic information for each utterance. This corpus consists of various spoken language expressions that are expected to occur in travel related scenarios. Section 3 describes our proposed topic detection method. This section explains the relation factor that is a ratio of mutual information of a word and a topic to entropy of the word, and our proposed method detects topic based on this factor. Section 4 reports the results of experiments that compare the performance of the proposed method and the support vector machine (SVM). Section 5 discusses the robustness and other characteristics of the proposed method and the SVM. Section 6 presents a summary of our conclusions.

2. THE BROAD-COVERAGE BILINGUAL BASIC EXPRESSION CORPUS

In this study, we used the broad-coverage bilingual basic expression corpus [7]. The ATR Spoken Language Translation Research Laboratories built this corpus. This corpus is a collection of Japanese sentences and their English translations usually found in phrasebooks for foreign tourists. The number of the sentences included in the corpus is over 200,000. In addition, it must be noted that we use only Japanese sentences in this study.

Each sentence contains topic information. This information is based on a description usually contained in a phrasebook. In general, a phrasebook has categories, such as "at airport", "on airplane", and "clearing immigration", for the convenience of tourists, but the categorizations differ for each phrasebook. The corpus integrates these different categorizations into 10, 20, and 252 categories for three layers. Each sentence has one topic and three labels that correspond to each layer. *Table 1* shows

Layer 0	Layer 1	Layer 2	Example
ACTIVITY	Shopping	Choose something Buy something	*May I see that? / May I pick it up?* *Can I buy it in Japanese yen? / How will you pay for this?*
	Sightseeing	Enjoy sport Take a picture	*Do you know where the baseball stadium is?* *Could you take a photo of us, please?*
TRANSPORT	Move	Use a rental car Buy a ticket	*Will you recommend one with good mileage?* *Is this the right platform for the train to Minneapolis?*
	Airport	Go through immigration Transfer	*Please show me your passport and immigration form.* *I have a connection to JAL flight seven three nine.*
	Airplane	Use the onboard services Have the onboard meal	*Which channel is the film on?* *What kind of drinks do you have?*
10 categories	20 categories	252 categories	: Total Number of Categories

Table 1: Examples of topic categorizations

examples of topic categorizations. We deal with topic labels for layer 0 and layer 1 (10 and 20 categories) in this paper.

This corpus includes non-topic specific sentences such as "*shouchishimashita*", "*kashikomarimashita*" ("sure" or "certainly" in English); nevertheless, they have the sole topic. The test set, however, listed all of the possible topics for each utterance. In the experiments discussed later, the detection result is counted as correct when the detected topic matches with those in a list.

3. TOPIC DETECTION

In the proposed method, an input is the text of an utterance that is segmented into a series of morphologically tagged words and an output consist of the topic labels for layer 0 and layer 1. Our topic detection method has two phases as follows:

(i) Training phase: Calculating a relation factor for the combinations of all topics and words in the training corpus.

(ii) Execution phase: Detecting topic of an input utterance by using the relation factors obtained in the training phase.

3.1. Relation factor

We define the relation factor, which represents the strength of relations between words and topics. Mutual information is often applied to determine such relations[5]. Usually, the mutual information of a word and a topic is considered to increase for a word that frequently occurs in utterances of a specific topic and to decrease for a word that occurs uniformly in utterances of various topics. As for the corpus that we used in this study, however, values of mutual information for words that occurred uniformly in utterances increased frequently in preliminary experiments when the entropy of such words was large. Moreover, these values often surpassed those of mutual information for words that occurred in utterances of a specific topic. That is to say, depending on the increase in the entropy of certain words, the value of their mutual information increased relatively to that of the other words.

In order to solve this problem, mutual information should be compensated to represent the relations between words and topics. Let T and W be information sources which have events in which a topic t_k occurs or not and a word w_i occurs or not. Then let $H(W)$ and $I(T;W)$ be the entropy of W and the mutual information between T and W respectively. Here, detecting the topic from the words included in an utterance requires having the values that represent the strength of relations as viewed from words to topics. Namely, it is the proportion of mutual information between a word and a topic to the entropy of the word. Thus the relation factor $r(t_k;w_i)$ is defined as follows.

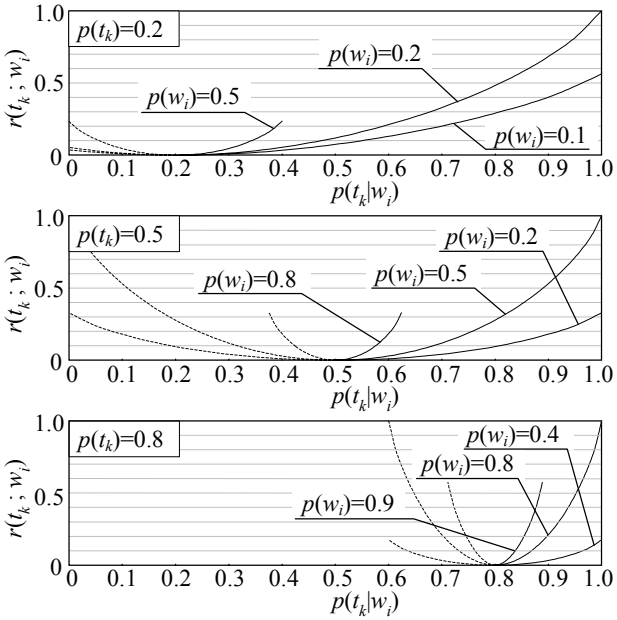

Figure 1: Probability and Relation factor.

$$r(t_k;w_i)=\begin{cases}\dfrac{I(T;W)}{H(W)} & \left(\dfrac{p_{TW}(t_k,w_i)}{p_W(w_i)}\geq p_T(t_k)\right)\\ 0 & \left(\dfrac{p_{TW}(t_k,w_i)}{p_W(w_i)}< p_T(t_k)\right)\end{cases} \quad (1)$$

Here, $p_T(t_k)$ and $p_W(w_i)$ are an occurrence probability of a topic t_k and a word w_i, and $p_{TW}(t_k,w_i)$ is a co-occurrence probability of t_k and w_i. Eq. (1) denotes the relation factor from a word w_i to a topic t_k. The range of $r(t_k;w_i)$ is [0,1]. When T and W are independent (i.e., $p_{T|W}(t_k|w_i)=p_T(t_k)$), there is no relation between T and W (i.e., $r(t_k;w_i)=0$). On the other hand, when $p_T(t_k)$ and $p_W(w_i)$ satisfy $p_{TW}(t_k,w_i)=p_T(t_k)=p_W(w_i)$, T and W relate to each other perfectly (i.e., $r(t_k;w_i)=1$). Normalizing $I(T;W)$ with $H(W)$ represents a correlation between the two information sources from the viewpoint of entropy.

Turning now to the conditions in Eq. (1), we added the condition to apply the normalization of $I(T;W)$ with $H(W)$ to our topic detection procedure described in section 3.2. If these conditions do not exist, $r(t_k;w_i)$ becomes symmetry by the center

axis of $p_{T|W}(t_k|w_i)=p_T(t_k)$ as shown in *Figure* 1. That is, $r(t_k;w_i)$ represents a relation between $T=t_k$ and $W=not\ w_i$ when $p_{T|W}(t_k|w_i)<p_T(t_k)$. A relation factor for a $W=not\ w_i$ is not needed simply because we consider the relation factors of words that occur in an input utterance at the execution phase. Therefore, the condition of Eq. (1) sets the value of the relation factor for word *not* w_i to 0. *Figure* 1 shows examples of the characteristics of $p_T(t_k)$, $p_W(w_i)$, $p_{T|W}(t_k|w_i)$, and $r(t_k;w_i)$.

Actually, since $p_W(w_i)$ is normally less than 0.5 in the corpus, the $H(W)$ rises proportionally as the occurrence of a word w_i increases. Thus, normalizing $I(T;W)$ with $H(W)$ has an effect similar to that of eliminating a frequently occurring word.

As mentioned in section 2, the object of detection is a topic label for layer 0 and layer 1. In calculating $r(t_k;w_i)$, a universal set is the upper layer topic. That is, regarding layer 0, a universal set is all of the training data, and regarding layer 1, a universal set is a set of utterances that has the same topic label in layer 0 as in the upper layer. For example, to calculate a relation factor of topic "Airport" of layer 1, the universal set is a set of sentences with a topic label "TRANSPORT" for layer 0.

3.2. Topic detection procedure

This section describes about the execution phase. The topic detection procedure of this paper is a simple matrix operation, as shown in Eqs. (2) ~ (4).

Vector $\boldsymbol{Z}$ represents the feature of an input utterance S. The elements of $\boldsymbol{Z}$ indicate that an input utterance includes word w_i or not as 1 or 0.

$$\begin{aligned}&\boldsymbol{Z}=[\mu_Z(w_1)\,\mu_Z(w_2)\cdots\mu_Z(w_m)]\\&\mu_Z(w_i)=\begin{cases}1: S \text{ includes a word } w_i\\0: S \text{ does not include a word } w_i\end{cases}\\&(i=1,2,\cdots,m)\end{aligned}\tag{2}$$

The elements of matrix $\boldsymbol{R}$ are the relation factors between words and topics.

$$\boldsymbol{R}=\begin{bmatrix}r(t_1;w_1)&\cdots&r(t_n;w_1)\\\vdots&\ddots&\vdots\\r(t_1;w_m)&\cdots&r(t_n;w_m)\end{bmatrix}\tag{3}$$

Vector $\boldsymbol{A}$, which indicates the relation factor of an input utterance – topics, is the product of vector $\boldsymbol{Z}$ and matrix $\boldsymbol{R}$.

$$\begin{aligned}\boldsymbol{A}&=\boldsymbol{Z}\cdot\boldsymbol{R}\\&=[\mu_Z(w_1)\cdots\mu_Z(w_m)]\begin{bmatrix}r(t_1;w_1)&\cdots&r(t_n;w_1)\\\vdots&\ddots&\vdots\\r(t_1;w_m)&\cdots&r(t_n;w_m)\end{bmatrix}\\&=[r_A(t_1)\cdots r_A(t_n)]\end{aligned}\tag{4}$$

In this paper, n=10 +20 (number of categories for layer 0 and layer 1).

Then, the next step is to multiply the relation factors of an input utterance to topics for layer 0 (i.e., upper layer) together those for the corresponding layer 1 (i.e., lower layer). For example, the relation factor of the topic "TRANSPORT" in layer 0 is multiplied together that of the topics "Move", "Airport", and "Airplane" in layer 1. This operation leads to the detection result vector $\boldsymbol{A}_{LOWEST}$. The subscript l indicates the number of topics in the lowest layer.

$$A_{LOWEST}=\left[r_{A_{LOWEST}}(t_1)\cdots r_{A_{LOWEST}}(t_l)\right]\tag{5}$$

Token	*Surface*	*Pronunciation*	*Base form*	*POS*
A	乗り換え	*Norikae*	乗り換える	Verb
B			乗り換え	Noun
C	乗換		乗換	

Regarding *Surface form* : A and B are the same word.
Regarding *Pronunciation* : A, B and C are the same word.
Regarding *Base form* : A, B and C are not the same word.
Regarding *POS* : B and C are the same word.
(The example word means "transfer" in English.)

Table 2: Example of word discrimination

Each topic in a lower layer has the sole topic in each upper layer. Thus, $\boldsymbol{A}_{LOWEST}$ leads to the detection result for all layers.

The 1-best detection result is acquired as follows.

$$\hat{t}=\arg\max_{t_k} r_{A_{LOWEST}}(t_k)\qquad(k=1,2,\cdots,l)\tag{6}$$

3.3. Properties of a word

A word has properties of "*Surface form*", "*Pronunciation*", "*Base form*", "*POS*", "*Inflection type*", and "*Euphonic*". Combinations of these properties, which are used for the discrimination of words, have an effect on the method's robustness against speech recognition errors.

Table 2 shows an example of word discrimination.

4. EXPERIMENTS

4.1. First experiment

The first experiment was a comparison of the results obtained by using the proposed method and those by using SVM. The input for this experiment was correct transcriptions and recognizer's outputs of utterances in the test set. The combinations of properties used for discriminating words were: (A) All properties, (B) *Pronunciation* + *POS*, and (C) *Surface form* + *POS*.

Table 3 shows the conditions of the training set and the test set and the word error rate of speech recognition.

Here, we briefly explain the training data for the SVM. Nouns and verbs were chosen as the attributes indicating the feature of an input utterance. The vector of training data $\boldsymbol{x}_i$ represents whether an input utterance includes these attributes (1) or not (0)[1]. For example,

$\boldsymbol{x}_i=(0,1,0,0,1,1,1,0,\cdots)$

Whether an input utterance is categorized to the target topic or not is represented as +1 or –1.

The kernel function of the SVM is the polynomial [1].

$$K_{poly}(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+1)^d\qquad(d=1)\tag{7}$$

	Training set	Test set
Number of Utterances	162320	1523
Word error rate of speech recognition		12.8(%)

Table 3: Experimental conditions

[1] We experimented with the input utterance vector $\boldsymbol{x}_i$ that consists of relation factors, but the result was worse than that described above.

Topic match rate (%)		(A) *All properties*		(B) *Pronunciation +POS*		(C) *Surface + POS*	
Input	Method	Layer 0	Layer 1	Layer 0	Layer 1	Layer 0	Layer 1
Transcriptions (WER 0%)	Proposed method	80.2	73.0	79.3	72.2	79.9	72.4
	SVM	60.3	51.1	58.2	50.7	---	---
Recognizer's outputs (WER 12.8%)	Proposed method (Decrement to the transcription input)	71.2 (–11.2)	63.2 (–13.4)	75.6 (–4.7)	67.4 (–6.6)	70.9 (–11.3)	63.0 (–13.0)
	SVM (Decrement to the transcription input)	56.0 (–7.1)	49.9 (–2.3)	54.0 (–7.2)	46.6 (–8.0)	---	---

Table 4: Topic match rate between detected topic and correct topic of test set (First experiment)

Table 4 shows the results of the first experiment. The topic match rates of the proposed method were higher than those of the SVM for every condition.

Let us compare the topic match rates from correct transcriptions with those from recognizer's outputs. For the SVM, the reduction in the topic match rate from transcriptions to recognizer's outputs was 7% or 8%, except for one case. This result is less than WER. On the other hand, for the proposed method, the reduction in the topic match rate was nearly equal to WER with property combinations (A) and (C), and it smaller in the case of property combination (B).

4.2. Second experiment

The second experiment was a comparison of the *n*-best (n=1, 2, 3) result by using the proposed method with the input of correct transcriptions and recognizer's outputs that include errors. The combination of properties for this experiment was (B), which had given the best results in the first experiment.

Figure 2 shows the results of the second experiment. An increment of *n*-best improves the topic match rate. It reaches the range of 94.6% (Layer 0 / Transcriptions) to 86.7% (Layer 1 / Recognizer's outputs) at 3-best.

5. DISCUSSION

The topic match rate of the proposed method is higher than that of the SVM; therefore, the proposed method surpasses SVM in topic detection ability. SVM is effective in topic detection (i.e., text categorizations) of newspaper articles or broadcast news that include many words. According to these results, it is likely that SVM is not effective for spoken language as a spoken utterance has fewer words; however, further investigation of these results is needed.

Turning now to the robustness against speech recognition errors; according to the comparison of WER (12.8%) and the reduction of the topic match rate from correct transcriptions to recognizer's outputs, the influence of speech recognition errors on SVM is small regardless of the combination of word properties. On the other hand, the robustness of the proposed method fluctuates according to the combination of word properties. The combination of word properties (B), which gives the best result, merges words that have the same pronunciation but different surface forms. The surface form difference often indicates a difference in meanings. The relation factors, however, reflect the difference in meanings appropriately by using a topic-by-topic calculation. By using other combinations of word properties, the relation factors reflect noise, such as slight differences of surface form. This is probably the reason for the effect of the word property combinations on the robustness of topic detection.

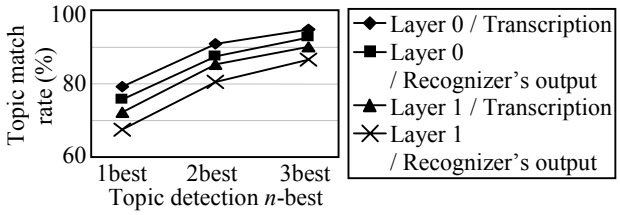

Figure 2: Topic match rate between detected topic and correct topic of test set (Second experiment)

6. CONCLUSIONS

In this paper, we first described the requirements for a topic detection method that can contribute to speech dialogue processing. We proposed a method that satisfies the prerequisites by using the relation factor of a word and a topic. The proposed method achieves about 65% to 80% at 1-best and about 85% to 95% at 3-best. This performance surpasses that of the SVM, which is an excellent classifier for text categorization. The proposed method also performs robustly against speech recognition errors by using appropriate combinations of word properties for discrimination of words.

A more detailed comparison of the proposed method and the SVM is needed, particularly in regard to an investigation of the effect of the combination of word properties on the method's robustness against speech recognition errors. A study of the effect of topic detection on an actual dialogue application is also needed.

7. ACKNOWLEDGEMENTS

This research was supported in part by the Telecommunications Advancement Organization of Japan.

8. REFERENCES

[1] Taira, H., Haruno, M., "Feature Selection in SVM Text Categorization", *AAAI 99*, pp. 480-486, 1999.

[2] Walls, F., Jin, H., Sista, S., Schwartz, R., "Topic Detection in Broadcast News", *Eurospeech 99,* pp. 2451-2454, 1999.

[3] Nakazawa, M., Zhang, J., Oka, R., "Topic spotting description of summary from spontaneous speech", *Eurospeech 99,* pp. 2447-2450, 1999.

[4] Wright, J.H., Gorin, A.L., Riccardi, G., "Automatic acquisition of salient grammar fragments for call-type classification", *Eurospeech 97*, pp. 1419-1422, 1997.

[5] Gorin, A.L., "Processing of semantic information in fluently spoken language", *Proc. ICSLP 96,* pp. 1001-1004.

[6] Chou, W., *et al.*, "Natural language call steering for service applications", *Proc. ICSLP 2000*.

[7] Takezawa, T., *et al.*, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world", *LREC 2002*, pp. 147-152, 2002.