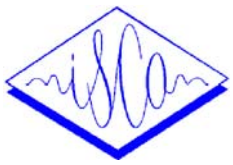

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TOPIC SPOTTING AND ITS DESCRIPTION OF SUMMARY FROM SPONTANEOUS SPEECH

Masayuki Nakazawa, Jianxin Zhang, and Ryuichi Oka

Real World Computing Partnership,
Tsukuba Mitsui Building 13F, 1–6–1 Takezono, Tsukuba, Ibaraki, Japan
{nakazawa, oka}@rwcp.or.jp, chou@mediadrive.co.jp

10.21437/Eurospeech.1999-538

ABSTRACT

This paper proposes two new methods of how to carry out automatic topic spotting in continuous speech and how to describe its summary in the form of slot expression. The topic spotting works independently of both speakers and topics. The method of topic spotting is based on a dialogue model to segment a topic using only a surface feature of continuous speech. We have evaluated our methods using the speech of the broadcasting news and the spoken dialogue speech. The result was that detection quality was over 80% on the broadcasting news. This paper shows some experimental results for Japanese spontaneous speech.

1. INTRODUCTION

Advances in speech technology and computing power have created a surge of interest in the practical application of speech recognition. However, the most accurate speech recognition systems in the research world are still far too slow and expensive to be used in practical, large vocabulary continuous speech applications. Their main goal has been recognition accuracy, with emphasis on acoustic and language modeling.

The purpose of this paper is to propose a new interface using a speech recognition technology and a natural language processing technology, and to examine a feasibility of its interface using a surface feature of continuous speech. To realize a new speech interface, this paper proposes two new methods of how to carry out automatic topic spotting in continuous speech and how to describe its summary in the form of slot expression.

The topic spotting in these methods works independently of both speakers and topics. The method of topic spotting is based on a dialogue model to segment a topic using only a surface feature of continuous speech. The surface feature is a dip of an envelope in running histogram obtained from speech waveforms that have a similar duration to each other. The description of summary for a spotted topic is obtained by filling out a slot expression with words that have a similar interval speech and are also included in the thesaurus and the co–occurrence dictionary.

2. RELATED WORKS

In recent years there has been renewal of interest in topic spotting [1],[2],[3],[4]. All of above researches need training speech data and language resources, because they use HMM for word recognition. Moreover, these approaches only classify topic without generating a topic summary. This means their methods are training–dependent and need large quantity of language resources. The methods we propose are topic–independent, because a surface feature of spontaneous speech is used for keyword spotting and the thesaurus and co–occurrence dictionary are used for classification of topics.

3. KEYWORD SPOTTING

Important keywords that characterize a topic (e.g. a proper noun, a normal noun and a verb) repeatedly appear in a topic, and are expected to have a long duration. We suppose the word appears frequently, and the important element that characterizes a topic itself. Actually, we have got a good result when we compare the important words and phrases which we extracted from the textbook by hand with the word that a plural time appears[5],[6]. In other words, important words and phrases which reflect a topic is supposed "keyword to be that a plural time appears in a topic".

In this research, we make use of this feature to spot a keyword, and we define important keywords as the words that have similar intervals and appear multiple times in a topic. Kiyama et al.[7] propose Incremental Reference Interval-free Continuous Dynamic Programming (IRIFCDP) method as a technique for extracting a resemblance section in the long time speech. In this research, we will make use of this technique for spotting resemblance speech. *Fig 1* shows basic ideal of IRIFCDP method.

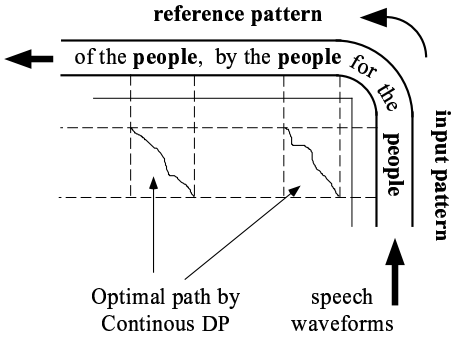

Fig 1: IRIFCDP basic idea: Speech input as input pattern changes to reference pattern. This method compares input and reference patterns using Continuous Dynamic Programming. Then optimal path of these patterns is calculated.

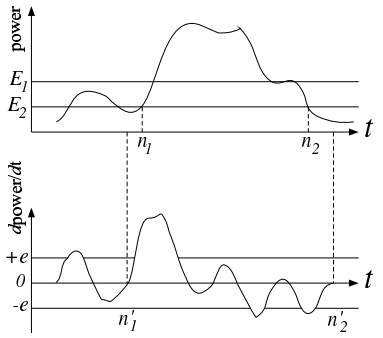

Fig 3: Extraction interval of topic: n_1 and n_2 show a voiced section by the ordinary method of voiced selection. n'_1 and n'_2 show a sub-topic section by the this method that use differential coefficients of its histogram.

4. TOPIC MODELING

4.1. Topic Segmentation

The segmentation of the topic begins with making a running histogram using a extraction time of the important keyword section. Then, this method estimate the boundary of a topic from the contour of the histogram. As for this histogram, we can estimate that the appearance frequency in the continuance of a topic is high, and we can also estimate that the appearance frequency in the transition of topics is low. Because another keywords and phrases appear frequently if the topic changes. In other words, we can estimate that the hollow of the histogram is a boundary of the topic by this method.

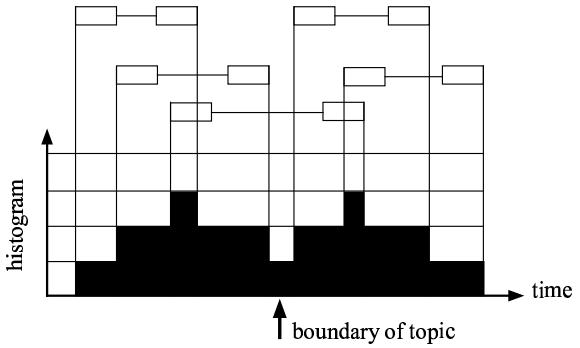

Fig 2: Topic boundary: A white square shows keyword pairs, a black square shows keyword histogram. The keywords connect a direct line each other. A hollow of histogram is the boundary of a topic.

The extraction of boundaries in the histogram is computed as a part of the histogram where changed from minus to plus in the front. This method can estimate the point where the histogram begins to increase rapidly, and it can estimate the point which the histogram finishes decreasing with a time to the reverse. *Fig 2* shows basic idea of topic segmentation.

The extraction method using the speech power estimates a voiced section (n_1,n_2) using two thresholds: E_1 and E_2 ($E_1 > E_2$). $E(n)$ is a acoustic power (n is a frame number). The start frame of a voiced section is searched as the point of over E_2 that isn't lower than E_1 and E_2. The end frame of a voiced section is searched a time to the reverse in the same method. In this E_1 and E_2 are derived from the experimental results. The method of this paper don't use its histogram just the state, but we use differential coefficients of its histogram. This is an advantageous to divide a topic into fine sub-topics, because it can estimate the point where the histogram begins to increase rapidly. $h(t)$ is the frequency in the cumulative similar pairs by the phonetic on time t. m is a parameter for averaging the histogram, $\hat{h}(t)$ is results of the moving average. $v(t)$ indicates the histogram change on time t, α is a parameter for regularizing a histogram change. The threshold value for segmentation in a topic is e, the time of each sub-topics that divided a topic into is indicated by H. E_1 and E_2 are correspond to 0 and e by the ordinary method of voiced selection. T is a total time of the speech.

$$
\begin{aligned}
\hat{h}(t) &= \frac{1}{m}\sum_{k=1}^{m} h(t+k) \qquad (1 \le t \le T), \\
\alpha &= \max_{k=1,\cdots,T} h(\hat{k}), \\
v(1) &= 0, \\
v(t) &= \alpha \cdot \left(\hat{h}(t) - \hat{h}(t-1)\right) \qquad (2 \le t \le T), \\
H &= \left\{ t \mid v(t) \ge e, -v(T-t) \ge e \right\}, \qquad (1)
\end{aligned}
$$

5. EXTRACTION OF THE TOPIC

Usually, it is an ideal of a summary that extracts or generates 5W1H information such as "when, where, who, what, whom, how". When we try to select a meaning from the structure of the sentence,

or to build a grammar (especially, correspondence to a spoken language), sometime occurs analytic impossibility from a correcting speech and a non-sentence (ill-formed sentence). In this research, we classify meanings by the thesaurus, and we dissolve the ambiguity of the word sense. And, we use the topic slot which we extracted from the thesaurus for this method.

In method of describing a summary, similar interval pairs are segmented by dips of the histogram. These interval pairs are recognized as the important keywords in a topic. Topic information is generated from these keywords using the thesaurus and co–occurrence dictionary. This thesaurus has two information: 1) headword notation, 2) concept information. The concept information has many headwords notation that is same meaning, and the headword information connect to several concept information. The concept information classifies 395,013 headword notations in 6,000 intermediate concepts. The co–occurrence dictionary describes 113,966 words only for Japanese. The slot is defined by a intermediate concept in this thesaurus. At present, following slots are used:

Time, Location, Change, Acts, Movement, Phenomenon, Condition, Subject, Thing

5.1. Algorithm

This section shows how to extract a topic information using slots. W means a word set, and w_k $(1 \leq k \leq L)$ means a element of W. C is a concept set in the thesaurus, and c_k $(1 \leq k \leq M)$ is a element of C. $\rho(w_k)$ is subsets of concept set, and it correspond to concepts by word w_k. U_k is subsets of concept C, and it represented by $\rho(w_k)$. And R_h represented subsets of a upper concept by $\xi(c_h)$. In particular, $\hat{c}$ indicated a root concept in the thesaurus. S is a slot set, and it is correspondent to the concept, because S is a subset of C. W_k is subsets of a word in W, and it is classified by $\tau(s_k)$. $\mu(w_i,w_j)$ gives subsets of co–occurrence relations.

If there isn't a co–occurrence relation given by words w_i and w_j, it is shown by $\mu(w_i,w_j)= \phi$. Generally, $\mu(w_i,w_j)$ isn't equal to $\mu(w_j,w_i)$. For example: If w_i={eat} and w_j={bread} are exist, a co–occurrence relation is represented as $\mu(w_i,w_j)$={eat–bread}. And, there isn't such a relation as $\mu(w_i,w_j)$={bread–eat}, because there isn't such a sentence, and to avoid miss-analyzing by a grammar. But we deal with a this relation as $\mu(w_i,w_j)=\mu(w_j,w_i)$. This relation has an effect on the coordinate relations such as "Mountain and River".

Retrieval algorithm for slot information

```
function trace( w:W ) : integer;
  var i, j : integer;  var o : W;  var c : C;
begin
  o := φ;
  for i = 1 to |w| then begin
    for j = i to |w| then begin
      if μ(w_i,w_j) <> φ then o := o ∪ w_i ∪ w_j;
      if μ(w_j,w_i) <> φ then o := o ∪ w_j ∪ w_i;
    end;
  end;
  for i = 1 to |o| then begin
    c := ρ(o_i);
    repeat
      c := ξ(c);
      if c ∩ S <> φ then τ(c ∩ S) := τ(c ∩ S) ∪ o_i;
    until c <> ĉ;
  end;
end;
```

In above the algorithm, it searches not only the co–occurrence relations but also the concept information by words w_i and w_j. And ambiguity of meaning is canceled by the concept information. The summary of each sub-topics is represented by the slot contexts.

6. EXPERIMENTAL RESULTS

We have evaluated these methods using the speech of the actual broadcasting news (Evening news: 297.47 *sec*), and a spoken dialogue speech (255.13 *sec*). *Fig 4* and *Fig 5* show distribution of keywords. Evening news has 5 sub-topics: 1) a affair of arsenic poisoning, and 2) a nuclear power station, and 3) a sale of new car, and 4) a weather forecast, and 5) an exchange rage. A spoken dialogue speech is a consultation of travel, and it has 3 sub-topics: 1) a destination of travel, and 2) a circumstance of destination, and 3) a procedure of travel.

6.1. Evening News

This speech data was divided into 7 sub-topics, and generated 4 slots information. The first slot was about "a affair of arsenic poisoning", the second and third was "a nuclear power station", the forth was about "a sale of new car". Table 1 shows a result of these slots information. The detection rate was 90.3% and detection quality was 89.3% about the spotting of important keyword. We couldn't get about a weather forecast sub-topic, because of a lack of co–occurrence entry such as "shine", "cloudy" and "rain".

6.2. Spoken Dialogue Speech

This speech data couldn't be divided into sub-topic, because a stammer, a stutter, a slip, a rephrase, and a demonstrative pronoun occurred. The reason why this speech data is a spoken dialogue of travel, it was too difficult to spot a keyword. Moreover, we couldn't get slot information, because of lack of travel co–occurrence relations. If this dialogue wasn't made face to face, the important keywords would have appeared repeatedly.

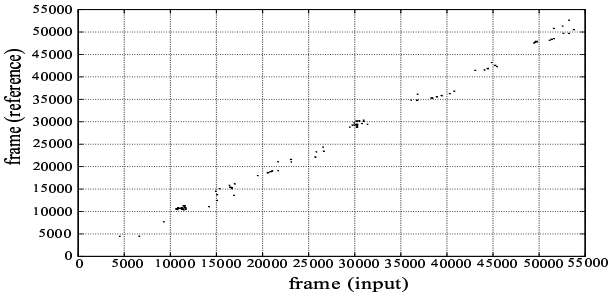

Fig 4: Keyword distribution on Evening news: A x and y axis are time domain. A dot in the graph is important keyword pairs.

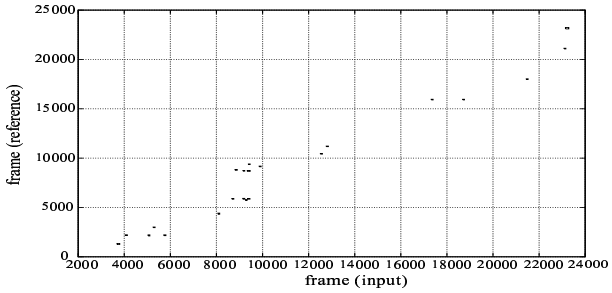

Fig 5: Keyword distribution on Spoken dialogue speech: A x and y axis are time domain. A dot in the graph is important keyword pairs.

7. CONCLUSION

We proposed the technique about the topic summary based on the surface feature of spontaneous speech. And we attempted to make clear the feasibility of these methods. In the broadcasting news, these methods carried out a high detection accuracy and quality. And we have attempted another broadcasting news (541.224 *sec*), where detection accuracy was 88.6%, detection quality was 50.6% of a word spotting. The slot information of its speech data was satisfying for us. But, in the spoken dialogue, we couldn't get an adequate result.

From now on, we intend to improve the segmentation precision in a topic, the recognition accuracy, a lack of the co–occurrence relations.

ACKNOWLEDGEMENTS

The authors appreciate to the members of our laboratory who had made eager discussions. This research used the speech database "ETL-WD I,II" of Electronic Laboratory and EDR dictionary.

Table 1: Slot information of Evening News

Topic No 1		
SLOT	HEADWORD	MEANING
Acts	detection investigation	detect a component investigate a suspected person about affair
Movement	detection	detect a component
Subject	police head office	public office for public safety and control center of organization
Thing	compound kind	material by chemically combine conclusion in nature of the matter
Topic No 2, 3		
Acts	arrange gather	put into order bring together, cause to assemble
Condition	safety	freedom from danger or risks
Subject	MITI	the Ministry of International Trade and Industry
Thing	subject problem	what is discussed or described or represented doubtful or difficult question or task
Topic No 5		
Time	last year this year	last year year in this time
Location	domestic	of one's own country
Movement	sales	exchange for commodity for money or other consideration
Topic No 7		
Time	last week end	before this week end of term

REFERENCES

[1] T. Imai, A. Kobayashi, and A. Ando, "Topic Spoting from Japanese Broadcast News by Using a Topic Mixture Model," in *Techinical Report of JSAI*, pp. 99–104, 1997.

[2] M. J. Carey and E. S. Parris, "Topic Spotting with Task Independent Models," in *Proc. EUROSPEECH*, vol. 1, pp. 2133–2136, 1995.

[3] J. H. Wright, M. J. Carey, and E. S. Parris, "Improved Topic Spotting through Statistical Modelling of Keyword Dependencies," in *Proc. ICASSP*, vol. 1, pp. 313–316, 1995.

[4] R. Kuhn, P. Nowell, and C. Drouin, "Approaches to Phoneme-Based Topic Spotting: An Experimental Comparison," in *Proc. ICASSP*, vol. 3, pp. 1819–1822, 1997.

[5] M. Nakazawa, K. Furukawa, J. Toyoura, and R. Oka, "A study on speech summary using demi-phoneme symbols generated from speech waves," in *Techinical Report of IEICE, SP96–28*, pp. 61–68, 1996.

[6] M. Nakazawa, Z. Jianxin, and R. Oka, "A study on topic segmentation and topic group generation for speech and topic summary," in *Techinical Report of JSAI*, vol. 1, pp. 97–98, 1997.

[7] J. Kiyama, Y. Ito, and R. Oka, "Topic-Independent Speech Summary and Atuomatic Topic Boundary Detection Using Incremental Reference Interval-free Continuous Dynamic Programming," in *Techinical Report of IEICE*, vol. J79–D–II No. 9, pp. 1464–1473, 1996.