Aligning Articles in TV Newscasts and Newspapers

Yasuhiko Watanabe  
Ryukoku University  
Seta, Otsu  
Shiga, Japan  
Yoshihiro Okada  
Ryukoku Univ.  
Seta, Otsu  
Shiga, Japan  
Kengo Kaneji  
Ryukoku Univ.  
Seta, Otsu  
Shiga, Japan  
Makoto Nagao  
Kyoto University  
Yoshida, Sakyoku  
Kyoto, Japan  
watanabe@rins.ryukoku.ac.jp

Abstract

It is important to use pattern information (e.g., TV newscasts) and textual information (e.g., newspapers) together. For this purpose, we describe a method for aligning articles in TV newscasts and newspapers. In order to align articles, the alignment system uses words extracted from telops in TV newscasts. The recall and the precision of the alignment process are 97% and 89%, respectively. In addition, using the results of the alignment process, we develop a browsing and retrieval system for articles in TV newscasts and newspapers.

1 Introduction

Pattern information and natural language information used together can complement and reinforce each other to enable more effective communication than can either medium alone (Feiner 91) (Nakamura 93). One of the good examples is a TV newscast and a newspaper. In a TV newscast, events are generally reported in the following modalities:

- image information,
- speech information, and
- text information (telops).

In TV newscasts, the image and the speech information are main modalities. However, it is difficult to obtain the precise information from these kinds of modalities. The text information, on the other hand, is a secondary modality in TV newscasts, which gives us:

- explanations of image information,
- summaries of speech information, and
- information which is not concerned with the reports (e.g., a time signal).

In these three types of information, the first and second ones represent the contents of the reports. Moreover, it is not difficult to extract text information from TV newscasts. It is because a lot of works has been done on character recognition and layout analysis (Sakai 93) (Mino 96) (Sato 98). Consequently, we use this textual information for aligning the TV newscasts with the corresponding newspaper articles. The method for extracting the textual information is discussed in Section 3.1. But, we do not treat the method of character recognition in detail, because it is beyond the main subject of this study.

2 TV Newscasts and Newspapers

2.1 TV Newscasts

In a TV newscast, events are generally reported in the following modalities:

- image information,
- speech information, and
- text information (telops).

A text in a newspaper article may be divided into four parts:

- headline,
- explanation of pictures,
- first paragraph, and
- the rest.

In a text of a newspaper article, several kinds of information are generally given in important order. In other words, a headline and a first paragraph in a newspaper article give us the most important information. In contrast to this, the rest in a newspaper article gives us the additional information. Consequently, headlines and first paragraphs contain more significant words (keywords) for representing the contents of the article than the rest.
On the other hand, an explanation of a picture in an article shows us persons and things in the picture that are concerned with the report. For example, in Figure 2, texts in bold letters under the picture is an explanation of the picture. Consequently, explanations of pictures contain many keywords as well as headlines and first paragraphs.

In this way, keywords in a newspaper article are distributed unevenly. In other words, keywords are more frequently in the headline, the explanation of the pictures, and the first paragraph. In addition, these keywords are shared by the newspaper article with TV newscasts. For these reasons, we align articles in TV newscasts and newspapers using the following clues:

- location of keywords in each article,
- frequency of keywords in each article, and
- length of keywords.
Summary of this article: On Apr 4, the Croatian Government confirmed that Commerce Secretary Ronald H. Brown and 32 other people were all killed in the crash of a US Air Force plane near the Dubrovnik airport in the Balkans on Apr 3, 1996. It was raining hard near the airport at that time. A Pentagon spokesman said there are no signs of terrorist act in this crash. The passengers included members of Brown’s staff, private business leaders, and a correspondent for the New York Times. President Clinton, speaking at the Commerce Department, praised Brown as ‘one of the best advisers and ablest people I ever knew.’ On account of this accident, Vice Secretary Mary Good was appointed to the acting Secretary. In the Balkans, three U.S. officials on a peace mission and two U.S. soldiers were killed in Aug 1995 and Jan 1996, respectively.

( Photo) Commerce Secy Brown got off a military plane Boeing 737 and met soldiers at the Tuzla airport in Bosnia. The plane crashed and killed Commerce Secy Brown when it went down to Dubrovnik.

Figure 2: An example of newspaper articles (Asahi Newspaper; April, 4, 1996)
Figure 4: An example of title texts: “santei yosanan asu shu-in tsuka he (The House of Rep. will pass the provisional budget tomorrow)"

Figure 5: Information explained by telops in TV Newscasts

Figure 6: An example of a quotation of a speech: “kono kuni wo zenshin saseru chansu wo atae te hoshii (Give me a chance to develop our country)”

1. explanation of contents of a TV image
   (a) explanation of a scene
   (b) explanation of an element
      i. person
      ii. group and organization
      iii. thing
   (c) bibliographic information
      i. time of photographing
      ii. place of photographing
      iii. image data
2. quotation of a speech
3. explanation of a fact
   (a) titles of TV news
   (b) diagram and table
   (c) other
4. information which is not concerned with a report
   (a) current time
   (b) broadcasting style
   (c) names of an announcer and reporters

Step 2 Divide the texts extracted in Step 1 into lines. Then, segment these lines at the point where the size of character or the distance between characters changes. For example, the text in Figure 3 is divided into “Okinawa ken (Okinawa Prefecture)”, “Ohta (Ohta)”, and “chiji (Governor)”.

Step 3 Segment the texts by the morphological analyzer JUMAN (Kurohashi 97).

Step 4 Analyze telops in TV images. Figure 5 shows several kinds of information which are explained by telops in TV Newscasts (Watanabe 96). In (Watanabe 96), a method of semantic analysis of telops was proposed and the correct recognition of the method was 92%. We use this method and obtain the semantic interpretation of each telop.

Step 5 Extract nouns from the following kinds of telops.
   • telops which explain the contents of TV images (except “time of photographing” and “image data”)
   • telops which explain a fact

It is because these kinds of telops may contain adequate words for aligning articles. On the contrary, we do not extract nouns from the other kinds of telops for aligning articles. For example, we do not extract nouns from telops which are categorized into a quotation of a speech in Step 4. It is because a quotation of a speech is used as the additional information and may contain inadequate words for aligning articles. Figure 6 shows an example of a quotation of a speech.

3.2 Extraction of Layout Information in Newspaper Articles

For aligning with articles in TV newscasts, we use newspaper articles which are distributed in the Internet. The reasons are as follows:
Taking advantage of the HTML tags, we divide newspaper articles into four parts:

- headline,
- explanation of pictures,
- first paragraph, and
- the rest.

The procedure for dividing a newspaper article is as follows.

1. Extract a headline using tags for headlines.
2. Divide an article into the paragraphs using tags for paragraph breaks.
3. Extract paragraphs which start “《写真》(shashin, picture)” as the explanation of pictures.
4. Extract the top paragraph as the first paragraph. The others are classified into the rest.

### 3.3 Procedure for Aligning Articles

Before aligning articles in TV newscasts and newspapers, we chose corresponding TV newscasts and newspapers. For example, an evening TV newscast is aligned with the evening paper of the same day and with the morning paper of the next day. We aligned articles within these pairs of TV newscasts and newspapers.

The alignment process consists of two steps. First, we calculate reliability scores for an article in the TV newscasts with each article in the corresponding newspapers. Then, we select the newspaper article with the maximum reliability score as the corresponding one. If the maximum score is less than the given threshold, the articles are not aligned.

As mentioned earlier, we calculate the reliability scores using these kinds of clue information:

- location of words in each article,
- frequency of words in each article, and
- length of words.

If we are given a TV news article $x$ and a newspaper article $y$, we obtain the reliability score by using the words $k(k = 1 \cdots N)$ which are extracted from the TV news article $x$:

$$
SCORE(x, y) = \sum_{k=1}^{N} \sum_{i=1}^{4} \sum_{j=1}^{2} w(i,j) \cdot f_{paper}(i,k) \cdot f_{TV}(j,k) \cdot length(k)
$$

where $w(i,j)$ is the weight which is given to according to the location of word $k$ in each article. We fixed the values of $w(i,j)$ as shown in Table 1. As shown in Table 1, we divided a newspaper article into four parts: (1) title, (2) explanation of pictures, (3) first paragraph, and (4) the rest. Also, we divided texts in a TV newscast into two: (1) title, and (2) the rest. It is because keywords are distributed unevenly in articles of newspapers and TV newscasts. $f_{paper}(i,k)$ and $f_{TV}(j,k)$ are the frequencies of the word $k$ in the location $i$ of the newspaper and in the location $j$ of the TV news, respectively. $length(k)$ is the length of the word $k$.

### 4 Experimental Results

To evaluate our approach, we aligned articles in the following TV newscasts and newspapers:

- NHK evening TV newscast, and
- Asahi newspaper (distributed in the Internet).

We used 143 articles of the evening TV newscasts in this experiment. As mentioned previously, articles in the evening TV newscast were aligned with articles in the evening paper of the same day and in the morning paper of the next day. Figure 7 shows the results of the alignment. In this experiment, the threshold was set to 100. We used two measures for evaluating the results: recall and precision. The recall and the precision are 97% and 89%, respectively.

One cause of the failures is abbreviation of words. For example, “shinyo-kinko (credit association)” is abbreviated to “shinkin”. In our method, these words lower the reliability scores. To solve this problem, we would like to improve the alignment performance by using dynamic programming matching method for string matching. (Tsunoda 96) has reported that the results of the alignment were improved by using dynamic programming matching method.

In this experiment, we did not align the TV news articles of sports, weather, stock prices, and foreign

---

### Table 1: The weight $w(i,j)$

|        | title | pict. expl | first par | the rest |
|--------|-------|------------|-----------|----------|
| T      | 8     | 4          | 4         | 2        |
| V      | 4     | 2          | 2         | 1        |
exchange. It is because the styles of these kinds of TV news articles are fixed and quite different from those of the others. From this, we concluded that we had better align these kinds of TV news articles by the different method from ours. As a result of this, we omitted TV news articles the title text of which had the special underline for these kinds of TV news articles. For example, Figure 8 shows a special underline for a sports news.

5 Browsing and Retrieval System for Articles in TV Newscasts and Newspapers

The alignment process has a capability for information retrieval, that is, browsing and retrieving articles in TV newscasts and newspapers. As a result, using the results of the alignment process, we developed a browsing and retrieval system for TV newscasts and newspapers. Figure 9 shows the overview of the system. The important points for this system are as follows:

- Newspaper articles and TV news articles are cross-referenced.
- A user can consult articles in TV newscasts and newspapers by means of the dates of broadcasting or publishing.
- A user can consult newspaper articles by full text retrieval. In the same way, user can consult TV newscasts which are aligned with retrieved newspaper articles. In other words, content based retrieval for TV newscasts is available.
- Newspaper articles are written in HTML. In addition to this, the results of the alignment process are embedded in the HTML texts. As a result, we can use a WWW browser (e.g. Netscape, Internet Explorer, etc) for browsing and retrieving articles in TV newscasts and newspapers.

A user can consult articles in newspapers and TV newscasts by full text retrieval in this way: when the user gives a query word to the system, the system shows the titles and the dates of the newspaper articles which contain the given word. At the same time, the system shows the titles of TV news articles which are linked to the retrieved newspaper articles. For example, a user obtains 13 newspaper articles and 4 TV news articles when he gives "saishutsu (annual expenditure)" as a query word to the system. One of them, entitled "General annual expenditure dropped for three successive years" (June, 4, 1997), is shown in Figure 10. The newspaper article in Figure 10 has an icon in the above right, looks like an opening scene of a TV news article. The icons shows this article is linked to the TV news article. When the user select this icon, the system shows the TV news article "Public work costs were a seven percent decrease" (the top left window in Figure 10).

References

Feiner, McKeown: Automating the Generation of Coordinated Multimedia Explanations, IEEE Computer, Vol.24 No.10, (1991).

Nakamura, Furukawa, Nagao: Diagram Understanding Utilizing Natural Language Text, 2nd International Conference on Document Analysis and Recognition, (1993).

Kurohashi, Nagao: JUMAN Manual version 3.4 (in Japanese), Nagao Lab., Kyoto University, (1997) 1.

Mino: Intelligent Retrieval for Video Media (in Japanese), Journal of Japan Society for Artificial Intelligence Vol.11 No.1, (1996).

1The source file and the explanation (in Japanese) of Japanese morphological analyzer JUMAN can be obtained using anonymous FTP from ftp://pine.kuee.kyoto-u.ac.jp/pub/juman/juman3.4.tar.gz
Figure 10: An output of the reference system for articles in TV newscast and newspapers: "Public work costs were a seven percent decrease" and "General annual expenditure dropped for three successive years"

Sakai: A History and Evolution of Document Information Processing, 2nd International Conference on Document Analysis and Recognition, (1993).

Sato, Hughes, and Kanade: Video OCR for Digital News Archive, IEEE International Workshop on Content-based Access of Image and Video Databases, (1998).

Tsunoda, Ooishi, Watanabe, Nagao: Automatic Alignment between TV News and Newspaper Articles by Maximum Length String between Captions and Article Texts (in Japanese), IPSJ-WGNL 96-NL-115, (1996).

Watanabe, Okada, Nagao: Semantic Analysis of Telops in TV Newscasts (in Japanese). IPSJ-WGNL 96-NL-116, (1996).