1998 International Computer Symposium
Workshop on Artificial Intelligence
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

ABSTRACT GENERATION FOR NEWSPAPER ARTICLES

Kenichi Naitou, and Susantha Herath

Graduate School of Computer Science and Engineering
The University of Aizu, Aizuwakamatsu 965-8580, Japan
Email: m5011119,herath@u-aizu.ac.jp

ABSTRACT

Extracting necessary information is hard and time consuming in the information-oriented society. An abstract generation system for newspaper that represents a large volume of information is vital. This paper presents an experimental system developed for abstracting newspaper articles on traffic accidents without applying complicated natural language processing techniques. The level of abstraction can be selected by the user. The user saves time significantly by excluding unwanted information in the article.

1 INTRODUCTION

There are too much information. It is hard and time consuming to extract necessary information. Sentence-oriented written information is in various forms such as books, magazines and newspapers. The interest or usefulness of such information differs from user to users. Some may needs a full details of the information while others not. User can save time and effort in extracting needy information if (s)he can control selecting the information according to his/her interest.

2 NEWSPAPER ARTICLES

TV, radio and newspaper are the main media of general information dissemination. News disappears as soon as it broadcasts by radio or TV, however, news in newspaper does not disappear and the reader has more control on accessing them.

Newspaper is the most disseminate and common. News in a newspaper is categorized into various sections such as social, political, economic, sports, and TV/radio guide, etc, making the reader easier to look news on his preference. There are two ways of reading the newspaper; one is just screening all the news from the start and read any interesting news in detail, the other is first select the interesting section or interesting article skipping other sections or articles and read it from the start. Each news item carries a title and a subtitle giving the main idea of the context. In most cases, the information provided by title and subtitle are insufficient in understanding the context. If some one is interested on the news item, (s)he has to read it to the extend of his/her needs. People skip some paragraph or parts of the article while reading to get only the important news that the user wants. Therefore, a system that assists the reader to get only the information that (s)he looks for is very helpful.

3 ABSTRACT GENERATION

Summarizing is difficult as it varies from parson to person. Basically, morphology (structure of the word), syntax (grammar), semantics (meaning), and pragmatic knowledge (real situation information) are necessary to pick the essence of a text to generate a summary[1, 2]. Applying such techniques in summerization is not in success.

There are many ways to abstract a text. The following two are more general.

1. Analyze the sentence into part of speech, such as subject, object, verb etc., and remove modifiers[3].
2. Separate the sentence into words and give a weight to each word. Then, rearrange the sentence according to the significant of the weight.

In the second method, people give different weights according to their interest. So, a unique result is not possible. The abstracts can be different from one to another. Abstract generation by computer has been researched for a long time, however, due to lack of sufficient techniques in natural language processing, successful systems are not available yet.

4 IMPLEMENTATION

An experimental abstract generation system is developed for newspaper articles on traffic accidents without using general language processing techniques. Asahi, a major Japanese newspaper is chosen for the experiment.

4.1 News Items on Traffic Accidents

There are about 100 news items on accidents in a year in both morning and evening editions of Asahi newspaper. About 80 of them are traffic related accidents. Each such news items carries a title, a subtitle, and a few paragraphs. The longest article had four paragraphs and about 1500 characters. In general, three paragraphs is most common, while many are with two paragraphs.

Basically, the first paragraph gives date and place in two patterns. One starts with date, time and place in that order. The other pattern gives place, date, and time. If it is a two paragraph article, the victims names are also included in the first paragraph. In longer articles, the names of the victims appear in second paragraph. The cause of accident is given in second paragraph of two paragraph articles, while it is in third paragraph in four paragraph articles. More details of the accidents are given in last paragraph of four paragraph articles.

4.2 Identification of Significant Parts

Necessary information in reading a traffic accident can be different from person to person. However, in general, people prefer to know:

1. seriousness of the accident (deaths, injuries etc)
2. place and details of the victims (depend on readers relation)
3. type of accident (what involved)
4. cause of the accident
5. the detail of the accident

On average, 1-3 above are the most relevant information for any reader of an accident. Mental power of people to extract only necessary information is higher than a machine. This system tries to imitate the human ability and let the reader only reads what (s)he needs in abstract form.

5 SYSTEM

A technique different form natural language processing is considered here. The system consists of four choices of information.

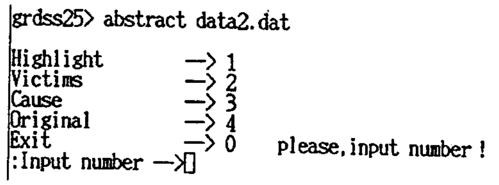
The first choice gives the highlight with five elements:

1. date
2. place
3. kind of accident
4. number of victims
5. kind of injury

The above elements will be selected from the article and place them in the following skeleton.

(date) (place) で[de] 、(kind of injury) 事故が起こりました [jikouga okorimashita]。その結果 [sono kekka]、 (number of victims) が [ga] (kind of injury) でした [deshita]。

Second choice gives the names of victims. Third Choice shows the cause of accident. Fourth choice produces the full text (original article). Figure 1 shows the first screen with select option of the system.



```
grdss25> abstract data2.dat

Highlight        —> 1
Victims          —> 2
Cause            —> 3
Original         —> 4
Exit             —> 0      please, input number !
:Input number —>□
```

Figure 1: The first screen

5.1 Algorithm

First Choice: Highlights

In choice 1, five elements; date, place, kind of accident, number of victims, and kind of injury are extracted.

The table 1 shows the patterns(P) of the sentences beginning with *date* and *place*.

P	Form	No	%
p1	(date) *keyword*, (place)	419	88.4
p2	··· (date) *keyword*,(place)	2	00.4
p3	(place) de,(date)	20	04.2
p4	(place) de (date)	30	06.3
p5	(place) dewa (date)	1	00.2
p6	no rule	2	00.4
	total	474	100.0

Table 1: Date and Place Patterns in Sentences

In table 1, *keyword* represents words such as ごろ [goro] (about), 朝 [asa] (morning), 前 [mae] (before) etc. Pattern 1-2 (**p1-2**) with *date* and *place* represents 88.8% of the total article, while **p3-5** 10.7% and **p6** 0.4%. The following algorithm is used to get the necessary information to fill the skeleton of the abstract.

First, read the news item from left to right and top to bottom character by character.

- When the article is in **p1-2**:
  - *Date*: Pick up the first three characters that represent the *date*.

  - *Place:* Continue reading and pick up characters between *keyword* and 都 [to]・道 [dou]・府 [fu]・県 [ken] or 市 [shi] representing the *place* including 都・道・府・県・市.

- When the article is in **p3-5**:
  - *Place*: Pick up characters up to 都 [to]・道 [dou]・府 [fu]・県 [ken] or 市 [shi] from the beginning.
  - *Date:* Continue reading and pick up the characters between で [de]、 and 日 [nichi](date).

- *Kind of accident*: Continue reading and pick up any of 衝突 [syoutotsu] (collision), 激突 [gekitotsu] (rear-end collision), 転落 [tenraku] (crash), or 追突 [tsuitotsu] (fell off) in the text.

- *Number of victims*: Count for numbers in *kanji* that comes with 人 [nin](human); eg. 三人 [sannin](three people). It needs to convert Japanese numbers into Arabic numbers, then addition and return them to Japanese numbers to include in the skeleton. If number of persons are not given, the system counts the number of names. If no name is given, * is inserted.

- *Kind of injury involved*: The system finds necessary information as in *table 2*.

Second Choice: Victims

The system prints all victims names with age. Table 3 shows how the victims name and age appear in the text.

P	Form	No	%
p1	*Title1* name さん [san] ()	968	73.1
p2	*Title1* name *Title2* ()	100	07.6
p3	name さん [san] ()	184	13.9
p4	name *Title2* ()	15	01.1
p5	*Title2* ()	57	04.3
	total	1324	100.0

Table 3: Victims

Title represents 運転手 [untensyu] (driver), 会社員 [kaisyain] (office worker), 無職 [musyoku] (jobless) etc, and sometimes with some symbols such as, 、, △, =, etc. The system looks for parenthesis, ( ), with maximum two numerics representation the age, pick up 15 characters appear before the parenthesis, and remove *Title1* with the characters appear before it. In the case of p3, p4 and p5 nothing is removed as all 15 characters will appear in the skeleton, which is not desirable.

Third Choice: Cause

Cause of accident is not given in 34%of the article. Table 4 shows the pattern representing cause of accident in the news item.

P	Form	No	%
p1	*Keyword*、 cause	312	65.8
p2	no *Keyword*	1	00.2
p3	cause not given	161	34.0
	total	474	100.0

Table 4: Cause

Keyword represents such as 調べでは [shirabedeha], 調べだと [shirabeniyoruto], によると [niyoruto] etc. The system finds the *Keyword* and pick up characters comes after it to the end of sentence.

Fourth Choice: Full Text

The original article is printed.

6 EXPERIMENT

The system, written in C is implemented on Sun WS, on a unix system. The articles with no titles (both head and sub) is assumed to be on line, in left to right, and top to bottom writing. The system reads from the first paragraph and pick up the important information, as explained in *section 5.1*.

Input

Figure 2 shows an input.

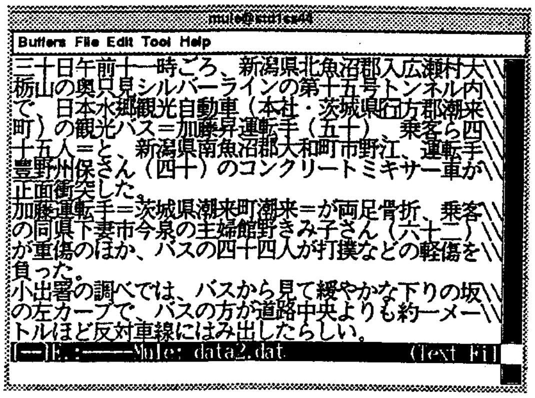
[sanjyunichi gozen jyuuichiji goro, niigataken kitagyonumagun irihirosemura ootochiyama no okutadami shiruba-rain no daijyugo gou tonnerunai de, nihonsuigoukankou jidousya (honsya・Ibarakiken yukuegun shioraityou) no kankoubasu = katou noboru untensyu(50), jyoukyaku ra 45nin = to, niigataken minamigyonumagun yamatomachi ichinoe, untensyu toyono syuuho san(40)no konkuri-to mikisa-sya ga syoumensyoutotsu shita.
katou untensyu = ibarakiken shioraityou shiorai = ga ryouashi kossetsu, jyoukyaku no douken shimotsumashi imaizumi no syufu tateno kimiko san(62) ga jyuusyou nohoka, basu no 44nin ga daboku nadono keisyou wo otta.
koidesyo no shirabedeha, basu karamite odayakana kudari no saka no ka-bu de, basu nohouga dourotyuuou yorimo yaku 1me-toru hodo hantaisyasen ni hamidashitarashii.]

(Sightseeing bus with forty-five passengers driven by Noboru Katou (50 age) and concrete mixer driven by Syuuho Toyono(40 age) collied head-on in Niigata on 30th at 11am.

Input	Output for (5)		
死亡 [shibou] (death)	死亡		死傷 [shisyou] (casualties)
即死 [sokushi] (instant death)			
重傷 [jyuusyou] (serious injury)	重傷	重軽傷 (two more elements)	
重体 [jyuutai] (serious condition)	重体		
重軽傷 (serious and slight injuries) (jyuukeisyou)	重軽傷		
軽傷 [keisyou] (slight injuries)	軽傷		
No element	怪我 [kega](injury)		

Table 2: Kind of injury

Katou's both legs were broken. Passenger Kimiko Tateno, a housewife was seriously injured. Forty-four persons got miner injuries.
Koide police who investigate the accident suspects that the bus ran about a meter left into opposite traffic lane at a curve causing the accident.)



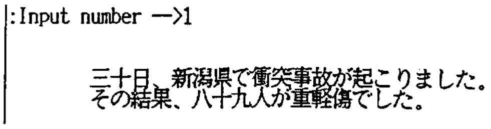
```
三十日午前十一時ごろ、新潟県北魚沼郡入広瀬村大\
栃山の奥只見シルバーラインの第十五号トンネル内\
で、日本水郷観光自動車（本社・茨城県行方郡潮来\
町）の観光バス＝加藤昇運転手（五十）、乗客ら四\
十五人＝と、新潟県南魚沼郡大和町市野江、運転手\
豊野州保さん（四十）のコンクリートミキサー車が\
正面衝突した。
加藤運転手＝茨城県潮来町潮来＝が両足骨折、乗客\
の同県下妻市今泉の主婦館野きみ子さん（六十二）\
が重傷のほか、バスの四十四人が打撲などの軽傷を\
負った。
小出署の調べでは、バスから見て緩やかな下りの坂\
の左カーブで、バスの方が道路中央よりも約一メー\
トルほど反対車線にはみ出したらしい。
[--]E.:----Mule: data2.dat        (Text Fil
```

Figure 2: Input data

Output

Level One:

Figure 3 shows the output at level 1.



```
:Input number -->1

        三十日、新潟県で衝突事故が起こりました。
        その結果、八十九人が重軽傷でした。
```

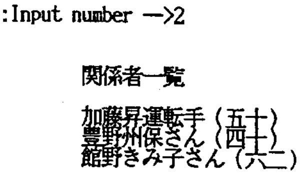
Figure 3: Choice 1 Output

[Sanjyuunichi, Niigataken de syoutotsujiko ga okorimashita. sonokekka, 89nin ga jyuukeisyou deshita.]

(There was a head-on collision at Niigata on 30th. Eighty-nine people were injured)

Level Two:

Figure 4 shows the output at level 2.



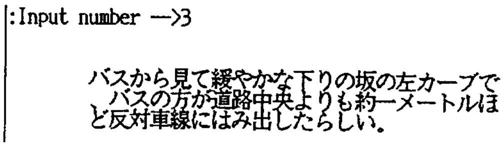
```
:Input number -->2

        関係者一覧
        加藤昇運転手（五十）
        豊野州保さん（四十）
        館野きみ子さん（六二）
```

Figure 4: Choice 2 Output

List of victims: Noboru Katou driver (50 age), Syuuho Toyono (40 age), Kimiyo Tateno (62 age)

Level Three:

Figure 5 shows the output at level 3.



```
:Input number -->3

        バスから見て緩やかな下りの坂の左カーブで
          バスの方が道路中央よりも約一メートルほ
        ど反対車線にはみ出したらしい。
```

Figure 5: Choice 3 Output

[basukaramite odayakana kudari no saka no hidari ka-bu de, basu nohouga dourotyuuou yorimo yaku 1metoru hodo hantaisyasen ni hamidashitarashii.]

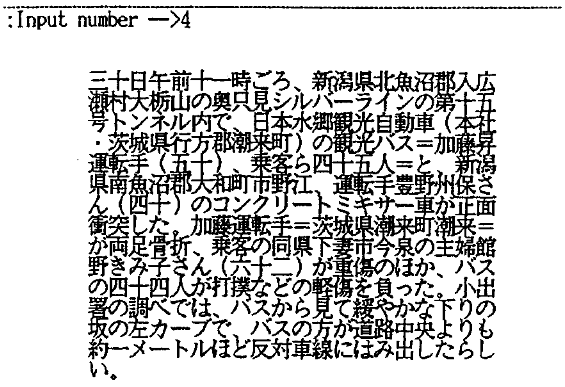
(Cause of accident is that the bus run over contrary traffic lane on one meter at left curve.)

Level Four:

Figure 5 shows output at level 4.

Total number of traffic accident article appeared in Asahi newspaper in the last 10 years was 474. All of them were experimented in this system. The table 5 shows results of successful rates.

Out of 474 total articles, 470 articles were abstracted successfully. Four articles were deviated from the uni-



```
:Input number —>4

三十日午前十一時ごろ、新潟県北魚沼郡入広
瀬村大栃山の奥只見シルバーラインの第十五
号トンネル内で、日本水郷観光自動車（本社
・茨城県行方郡潮来町）の観光バス＝加藤昇
運転手（五十）、乗客ら四十五人＝と、新潟
県南魚沼郡大和町市野江、運転手豊野州保さ
ん（四十）のコンクリートミキサー車が正面
衝突した。加藤運転手＝茨城県潮来町潮来＝
が両足骨折、乗客の同県下妻市今泉の主婦館
野きみ子さん（六十二）が重傷のほか、バス
の四十四人が打撲などの軽傷を負った。小出
署の調べでは、バスから見て緩やかな下りの
坂の左カーブで、バスの方が道路中央よりも
約一メートルほど反対車線にはみ出したらし
い。
```

Figure 6: Choice 4 Output

Table 5: In highlight successful

Item	Success %	Unsuccess %
Date	99.2	0.8
Place	99.2	0.8
Kind of accident	100	0
Number of victim	100	0
Kind of injury	100	0

form structure, so that they were not abstracted satisfactorily. The original articles with about 300 - 1500 characters are abstracted to 30 characters, 3-20%, by this system.

7 CONCLUSION

An experimental abstract generation system is developed and successfully implemented for newspaper articles on traffic accidents without using natural language processing techniques.

The system needs further tuning and expansion to include other news.

References

[1] Based theory of Natural Language knowledge information treatment series 4, Kazuhiro Fuchi, Kouichi Furukawa, Fumio Mizoguchi, Kyouritsu publishing company, Vol.5, pp.145–170, 1986.

[2] Natural Language Processing, Harry Tennant, Kenichi Mori, Tsutomu Kawata, Shinya Amano, Sangyou liblary, Vol.3, No.4, pp.51–135, 1984.

[3] An Experimental System Generating Summary of Japanese Editorials by Combining Multiple Discourse Characteristics, Kazuhide Yamamoto, Shigeru Masuyama, Shouzo Naito.

[4] ASHAHI newspaper pocked edition ASHAHI newspaper company, 1988 – 1992

[5] ASHAHI newspaper(morning edition), 1993 – 1996