LETTER

Personal Information Extraction from Korean Obituaries*

Kyoung-Soo HAN^{†a)}, Member

SUMMARY Pieces of personal information, such as personal names and relationships, are crucial in text mining applications. Obituaries are good sources for this kind of information. This study proposes an effective method for extracting various facts about people from obituary Web pages. Experiments show that the proposed method achieves high performance in terms of recall and precision.

key words: personal information extraction, named entity recognition, attribute extraction, automatic dictionary construction

1. Introduction

Named entities, such as persons and organizations, are essential elements that must be considered when performing natural language processing. It is necessary to construct a seed dictionary for the named entities in order to implement an effective named entity recognizer. One of the key tasks during text mining is to extract relationships between entities.

An obituary includes personal names, names of organizations, personal relationships, and other personal information, such as age, gender, and date of death. Figure 1 shows a set of sample obituaries. As shown in the figure, obituaries can be categorized into two types: summary and narrative.

- Summary obituary: This type of obituary is a compact death notice that contains information about the deceased and the bereaved. The summary obituary is presented in a structured format.
- Narrative obituary: This type of obituary is like a general news article that delivers a short biography about the deceased. A narrative obituary is mainly focused on the deceased, and the information about the bereaved is only mentioned optionally. The presentation format of the narrative obituary is not structured, but most obituaries of this type have similar styles.

A basic biographical dictionary can be constructed by extracting personal information from obituaries. This study proposes an effective method for extracting information about a person from obituaries.

Many studies have addressed personal information extraction. The WePS (Web People Search) Evaluation cam-

▲김인호 서울전자 사장 별세, 박나나 강남대 교수 남편상, 길동·창동·수희 부친상=22일22시

In-Ho Kim, president of Seoul Electronics, passed away; husband of Na-Na Park, professor at Gangnam Univ.; father of Gil-Dong, Chang-Dong, and Su-Hee=22:00 March 22.

(a) Summary obituary

김인호(65) 서울전자 사장이 22일 오후 10시 자택에서 심장마비로 별세했다. 지난달 뇌경색 증세로 서울병원에 ... (중략) ... 유족은 부인 박나나 강남대 교수와 아들 길동·창동, 딸 수희씨가 있다. 빈소는 일반인에게 공개되지 않았으나, ... (중략) ...

In-Ho Kim(65), president of Seoul Electronics, passed away by heart attack at home 10 p.m. March 22. He entered Seoul Hospital after suffering cerebral infarction last month ... (omitted) ... He is survived by his wife Na-Na Park, professor at Gangnam Univ., his son, Gil-Dong and Chang-Dong, his daughter, Su-Hee. His mortuary was closed to the public, ... (omitted) ...

(b) Narrative obituary

Fig. 1 Two types of obituary.

paign evaluated systems that extract personal attributes, including dates of birth, birth places, affiliations, occupations, and awards [1]. Ravichandran and Hovy proposed a pattern learning method that uses a training set that is constructed using question-answer pair in order to utilize the extraction result in a question answering system [2]. Garera and Yarowsky extracted biographical facts from free text using various models that rely on contextual patterns, relative attribute positions and sequences, and transitive attributes of co-occurring entities [3]. Chen et al. classified Web pages into a formal style or an informal style. They applied a rule-based extraction method and a machine learning method based on page type [4]. In order to extract profiles from researchers' home pages, Andruszkiewicz and Beata applied various machine learning methods, such as SVM (support vector machine), CRF (conditional random field), and MLN (Markov logic network) [5]. Culotta et al. extracted family relationships from Wikipedia pages by considering the relation extraction problem as a sequence labeling task. They also integrated the extraction model with contextual pattern learning.

All of these studies extracted personal information from general free text or Web pages, but only Embley et al. addressed extraction from obituaries [6]. They proposed a conceptual-model-based extraction method that can be easily applied to various domains. However, as the method did not use the characteristics of obituaries, such as the occupation patterns and the types of family relationships, the pre-

Manuscript received March 15, 2013.

Manuscript revised July 15, 2013.

[†]The author is with Div. of Computer Engineering, Sungkyul University, Anyang, Korea.

*This research was supported by the research fund of Sungkyul University under Grant No. 20112139.

a) E-mail: kshan@sungkyul.ac.kr

DOI: 10.1587/transinf.E96.D.2873

cision levels for some attributes were not high, about 74%. The types of family relationships such as relationships by blood or relationships by marriage, in particular, can be useful for restoring family names and extracting relationships.

Although the sentence structures of the obituaries are not very complex or diverse, it is not easy to extract information from the obituaries in a precise manner. This study proposes a precise method for extracting personal information from Korean obituary Web pages. Because most obituaries have similar description styles, this study uses a feature scoring method that applies prior knowledge about the style in a heuristic manner. The proposed method does not require training data annotation, which is a heavy burden for machine learning methods.

The remainder of this paper is organized as follows. Section 2 describes the details of our approach. Section 3 presents experimental results and analysis. The paper concludes with Sect. 4.

2. Personal Information Extraction

Because extraneous sections, such as menu and advertisements, are mixed into the original HTML pages of obituaries, it is necessary to separate the contents sections from the noise. The contents sections are extracted using regular expressions that are based on the HTML tags and the news structure, and the posterior extraction process is applied only to these contents sections.

2.1 Personal Name

It is very important to extract personal names correctly, because other personal information is associated with these names. If the personal name that is extracted is not correct, then the association of other information would be impossible or incorrect. Most of the people that are named in the obituaries are the deceased or the bereaved. In some cases, well-known bereaved persons are considered to be more important than the unknown deceased, as a result of which the name of the deceased person is not mentioned. The nameless person is called *null person* in this study. This person is extracted because there are relationships with other named people and the name could eventually be extracted from other sources.

A personal name is extracted using contextual patterns that occur in the text that surrounds the name. These contextual patterns are constructed from the training obituaries. Figure 2 shows the sample patterns that are used for extracting a personal name.

If it is impossible to determine the boundary of a personal name in a narrative obituary, then the name is selected

```
[▶▲]s*<B><DP_NAME></B>s*<DP_AFFIL>s+별세[s*]=]s*
[▶▲]?s*<B><NAME></B>s*<AFFIL>s+<RELATIONSHIP>상[s*]=]s*
[▶▲]?s*<B><NAME></B>s*<AFFIL>[.]
```

Fig. 2 Sample extraction patterns for personal name.

from a set of candidate words $W = \{w_1, w_2, \dots, w_n\}$ in the obituary text based on the following features. The feature scores for a candidate word w_i are calculated as follows:

- Title: The title of a narrative obituary usually includes the name of the deceased person. Therefore, the title feature plays a key role because the personal name can be restricted to the title words $T = \{t_1, t_2, \dots, t_m\}$.

$$Score_{title}(w_i) = \begin{cases} 1 & \text{if } w_i \in T \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

- Co-reference: Because a biography of the deceased person is included in narrative obituaries, several references to the deceased are also included. Analysis of the training obituaries shows that after the first mention of the deceased, it is common for the family name to be used in conjunction with the occupation, e.g., 김교수 *professor Kim*. Therefore, a candidate whose family name is the same as that of the co-reference r_i is likely to be a personal name.

$$Score_{coref}(w_i) = \begin{cases} 1 & \text{if } FN(w_i) = FN(r_i) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $FN(w_i)$ is a function that returns a family name of w_i .

- Affiliation pattern and occupation dictionary: This is a negative feature. As a result, the candidates that are matched to the pattern or the dictionary are not likely to be a personal name. Because the presentation format of summary obituaries is structured, affiliation names can be extracted from many of the training summary obituaries using the structural information. The affiliation naming patterns, such as (주) *Inc.* and 전자 *Electronics*, are compiled from the affiliation names. The occupation dictionary is explained in the following section.

$$Score_{affil}(w_i) = \begin{cases} -1 & \text{if } w_i \text{ starts with PREF} \\ -1 & \text{if } w_i \text{ ends with SUFFIX} \\ -1 & \text{if } w_i \text{ ends with OCCUP} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where PREF, SUFFIX, and OCCUP stand for prefix affiliation patterns, suffix affiliation patterns, and entries in an occupation dictionary respectively.

- Length: The length of Korean names is usually more than one and less than five Korean letters.

$$Score_{len}(w_i) = \begin{cases} 1 & \text{if } 2 \leq Length(w_i) \leq 4 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $Length(w_i)$ is the number of characters in w_i .

The score of a candidate word w_i is the sum of all the feature scores. The highest scored candidate word is selected as the personal name. If there are multiple candidates with the same high score, then the candidate word that occurs first in the obituary text is the one that is selected.

When the bereaved are listed, the family name is likely

to be omitted if the family name of the bereaved person is the same as that of the deceased. For example, in Fig. 1, the family name 김 *Kim* is omitted from the name of the bereaved persons: 길동 *Gil-Dong*, 창동 *Chang-Dong*, and 수희 *Su-Hee*. Therefore, the omitted family name can be restored using the characteristics of Korean names. According to the 2000 Population and Housing Census of Statistics Korea[†], Korean family names usually have one or two Korean letters and there are 13 two-letter family names. We try to restore the family name only when it is possible to restore it precisely. If a person has a family relationship with the deceased, and the type of the relationship indicates that the family name is the same among these related individuals, and the first letter of the names is different, and the length of the personal name is less than 3, then the family name of the deceased is used as that of the person.

2.2 Affiliation and Occupation

A job affiliation is expressed using the name of the organization that the person belongs to and the position or duty that the person holds. In some cases, it is presented with the occupation type only and it is not possible to distinguish the occupation type from the position or duty. For example, in the expression 박나나 교수 *Professor Na-Na Park*, the expression 교수 *professor* may refer to an occupation type or an academic rank, such as a full professor in a university. This study does not distinguish these details and refer to all of them as the occupation.

The last part of the occupation expression is likely to be an occupation type. For example, each of the expressions, 명예교수 *honorary professor*, 겸임교수 *adjunct professor*, and 초빙교수 *visiting professor*, refers to a sort of 교수 *professor*. Therefore, a list of occupation types is constructed using the Korean Standard Classification of Occupations^{††} and the ending letter n-gram patterns (e.g. 교수 *professor*) of the occupation expressions are extracted from the training summary obituaries. The occupation dictionary is constructed using the occupation types and the n-gram patterns.

When the personal name is extracted, the job expression that is associated with the person is also detected. The occupation expression is extracted based on the occupation dictionary, and the job expression except the occupation is considered an affiliation.

2.3 Relationship

This study extracts family relationships from the obituaries using a relationship thesaurus that is constructed manually. The relationship thesaurus includes the relationship name, the gender of the person who is in the relationship, and the direction of the relationship (directed or undirected). According to the thesaurus, the gender of each person, if un-

known, is estimated, and the appropriate relationship is selected based on the relationship name and the gender of the person. In the sample obituary that appears in Fig. 1, it is estimated that 김인호 *In-Ho Kim* is male and that 박나나 *Na-Na Park* is female, because the relationship name 남편 *husband* matches a thesaurus entry (남편, X:male, Y:female), meaning “X(male) is a husband of Y(female).” After this, the relationship (남편, 김인호, 박나나) is extracted.

If several people have the same relationship with the deceased person, then the relationship name is usually expressed only once. Therefore, people are grouped first, the relationship between the group and the deceased is determined, and then the relationship between each of the group members and the deceased is determined. Since this study is focused on accurate relationship extraction, the relationships are not inferred in cases where the relationships between the bereaved are not clearly specified in the obituary.

2.4 Other Attributes

In addition to the personal information mentioned in the previous sections, other attributes about a person are extracted as follows:

- Gender: The gender can often be estimated from the relation name (e.g., husband, mother) and the occupation (e.g., nun, housewife) even when it is not clearly specified in the obituaries.
- Occupation currency: It is estimated using occupation expression patterns whether one's occupation is currently a fact or not as of the date when the obituary page was written.
- Location and cause of death: The location and cause of death usually appear only in narrative obituaries. They are extracted based on expression patterns.
- Age and time of death: They are extracted using simple regular expressions.

3. Experimental Results

Korean obituary Web pages were crawled from The Chosun Ilbo Web site^{†††} for the evaluation. Obituary pages that were published between 2007 and 2012, a total of 2,268, were used for constructing the patterns and designing the extraction method during the development phase. Obituary pages that were published between January 1, 2013, and February 28, 2013, a total of 49, were used strictly for testing purposes. The former set is called the training set, and the latter set is the test set. One page typically contains many obituaries and there are about 9 personal names of deceased persons on average in a test obituary page. The relationship thesaurus contained about 150 family relationships, and the occupation dictionary had about 650 entries.

The correct answer that should be extracted from the test set was constructed manually, and it was used to carry

[†]http://kostat.go.kr/portal/korea/kor_nw/2/2/1/index.board?bmode=read&bSeq=&aSeq=46672

^{††}<http://kostat.go.kr/kssc/main/MainAction.do?method=sub&catgrp=kssc&catid1=kssc02>

^{†††}http://news.chosun.com/svc/list_in/list.html?catid=3N3

Table 1 Evaluation results for 49 obituary web pages.

	Total	Correct	Partially correct	Recall	Precision
Person Name	1678	1616	-	0.9631	0.9717
Affiliation	1317	1064	173	0.9393	0.9508
Position	1499	1371	47	0.9460	0.9686
Currency	1572	1501	-	0.9548	0.9684
Relationship	1533	1449	-	0.9452	0.9596
Age	6	6	-	1.0000	1.0000
Gender	684	665	-	0.9722	0.9881
Death Place	0	0	-	N/A	N/A
Death Cause	1	1	-	1.0000	1.0000
Death Date	145	142	-	0.9793	1.0000
Death Time	141	138	-	0.9787	1.0000
ALL	1599	1280	187	0.9174	0.8821

out the evaluation automatically and very strictly. An extraction was judged to be correct only if it was matched exactly to the correct answer. If an extraction included the whole answer expression, it was judged to be partially correct. If an extraction result included only some of the words from the correct answer, then it was judged to be incorrect. For example, if the correct answer was *Samsung Electronics Service Yongsan Branch* for affiliation, then the extraction results of *representative of Samsung Electronics Service Yongsan Branch* and *Samsung Electronics Service* were judged to be partially correct and incorrect respectively. Each attribute was evaluated in connection with a specific person. In the sample obituary in Fig. 1, this study did not judge the affiliation 서울전자 *Seoul Electronics* alone, but judged the affiliation in association with the personal name, <김인호, 서울전자>. The extractions of the nameless *null persons* were not evaluated.

The extraction performance was measured for each attribute based on recall and precision. If N_k is the total number of an attribute A_k in the test obituaries, C_k is the number of correct ones, and I_k is the number of incorrect ones, then the recall is C_k/N_k , and the precision is $C_k/(C_k + I_k)$.

Table 1 presents the evaluation results. Our method achieved high performances for each of the attributes in terms of recall and precision. The performances were higher for the finite attributes, such as age, gender, date of death, and time of death, than for the infinite ones. Although the gender was estimated using the values from other attributes, the performance was still good. The ALL row in Table 1 shows the performances, when the result was judged correct, only if all of the attributes for a given person were correct. As shown in the table, the proposed method extracted the personal information thoroughly and precisely.

Most of the errors were caused by failures in extracting a personal name, which is a key attribute for this task. One of the reasons why the personal name was not extracted correctly is that some obituary pages were written with ambiguous format, so the extraction patterns failed to match. In some cases, the personal name that was extracted was not correct because the family name was not restored.

In order to compare the extraction result with an exist-

ing dictionary of personal names, this study compared the extraction results with the results from the Naver People Search[†] using 26 personal names from the most recent obituary page in the test set. Naver is a well-known Web search engine in Korea, and the Naver People Search provides biographical information for executive level leaders that are politicians, businessmen, professors, and so on. Out of the 26 names in the most recent obituary page, 8 of the names were names of executives. Although the proposed method extracted the personal information for all of the 26 names, the Naver People Search showed biographical information with no family relationship for only 3 of the names. Moreover, one of the occupations that was retrieved from Naver was out of date. This experimental result implies that the proposed method will be useful for expanding the existing biographical dictionaries by extracting new personal information that is not in the dictionaries.

4. Conclusion

This study proposes a method for extracting personal information, such as personal names, affiliations, and relationships, from Korean obituary Web pages. Experimental results show that the proposed method extracts the personal information thoroughly and precisely. The extraction results will be useful for expanding the existing biographical dictionaries and for acting as a seed dictionary for a named entity recognizer. My next study will attempt to identify the same person from additional obituary sources and integrate the information from the multiple sources. The extraction of personal events in obituaries is worthy of future study.

References

- [1] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó, "WePS-3 evaluation campaign: Overview of the web people search clustering and attribute extraction tasks," Proc. Conference on Multilingual and Multimodal Information Access Evaluation, 2010.
- [2] D. Ravichandran and E. Hovy, "Learning surface text patterns for a question answering system," Proc. 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Stroudsburg, PA, USA, pp.41–47, 2002.
- [3] N. Garera and D. Yarowsky, "Structural, transitive and latent models for biographic fact extraction," Proc. 12th Conference of the European Chapter of the ACL (EACL 2009), p.300, 2009.
- [4] Y. Chen, S.Y.M. Lee, and C.R. Huang, "A robust web personal name information extraction system," Expert Systems with Applications, vol.39, no.3, pp.2690–2699, 2012.
- [5] P. Andrzejewicz and B. Nachyla, "Automatic extraction of profiles from web pages," in Intelligent Tools for Building a Scientific Information Platform, ed. R. Bembenik, L. Skonieczny, H. Rybinski, M. Kryszkiewicz, and M. Niezgodka, Studies in Computational Intelligence, vol.467, pp.415–431, Springer Berlin Heidelberg, 2013.
- [6] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.K. Ng, and R.D. Smith, "Conceptual-model-based data extraction from multiple-record web pages," Data & Knowledge Engineering, vol.31, no.3, pp.227–251, 1999.

[†]<http://people.search.naver.com/>