Personal Information Extraction from Korean Obituaries

Kyoung-Soo HAN†a, Member

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-

References

1. Ravichandran and Hovy proposed a pattern learning method that uses a training set that is constructed using question-answer pairs in order to utilize the extraction result in a question answering system [2].
2. Yarowsky extracted biographical facts from free text using various models that rely on contextual patterns, relative attributes, and awards [1].
3. Chen et al. classified Web pages with contextual patterns and the types of family relationships, the present format of obituaries [3].

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

Kyoung-Soo HAN†

Copyright © 2013 The Institute of Electronics, Information and Communication Engineers
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 from a set of candidate words \( W = \{w_1, w_2, \ldots, w_n\} \) 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 \( T = \{t_1, t_2, \ldots, t_m\} \).

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

- **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., 조교수 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.

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

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 (프소) Inc. and [드화] Electronics, are compiled from the affiliation names. The occupation dictionary is explained in the following section.

\[
Score_{aff}(w_i) = \begin{cases} 
-1 & \text{if } w_i \text{ starts with } \text{PREF} \\
-1 & \text{if } w_i \text{ ends with } \text{SUFFIX} \\
-1 & \text{if } w_i \text{ ends with } \text{OCCUP} \\
0 & \text{otherwise} 
\end{cases} \tag{3}
\]

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.

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

where \( \text{Length}(w_i) \) 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: 김동 Gil-Dong, 강동 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, 임원 교수 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 unknown, 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
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 김인호, 서울전자 Seoul Electronics. 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-

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Total</th>
<th>Correct</th>
<th>Partially Correct</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Name</td>
<td>1678</td>
<td>1616</td>
<td>-</td>
<td>0.9631</td>
<td>0.9717</td>
</tr>
<tr>
<td>Affiliation</td>
<td>1317</td>
<td>1064</td>
<td>173</td>
<td>0.9593</td>
<td>0.9508</td>
</tr>
<tr>
<td>Position</td>
<td>1499</td>
<td>1371</td>
<td>47</td>
<td>0.9460</td>
<td>0.9686</td>
</tr>
<tr>
<td>Currency</td>
<td>1572</td>
<td>1501</td>
<td>-</td>
<td>0.9548</td>
<td>0.9684</td>
</tr>
<tr>
<td>Relationship</td>
<td>1533</td>
<td>1449</td>
<td>-</td>
<td>0.9452</td>
<td>0.9596</td>
</tr>
<tr>
<td>Age</td>
<td>6</td>
<td>6</td>
<td>-</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Gender</td>
<td>684</td>
<td>665</td>
<td>-</td>
<td>0.9722</td>
<td>0.9881</td>
</tr>
<tr>
<td>Death Place</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Death Cause</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>Death Date</td>
<td>145</td>
<td>142</td>
<td>-</td>
<td>0.9793</td>
<td>1.0000</td>
</tr>
<tr>
<td>Death Time</td>
<td>141</td>
<td>138</td>
<td>-</td>
<td>0.9787</td>
<td>1.0000</td>
</tr>
<tr>
<td>ALL</td>
<td>1599</td>
<td>1280</td>
<td>187</td>
<td>0.9174</td>
<td>0.8821</td>
</tr>
</tbody>
</table>

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


1http://people.search.naver.com/