Generating Wikipedia-Like Biographical Sentences from Web People Search Results

Harumi Murakami, Toshimune Konishi, Yoshinobu Ura

Generating Wikipedia-Like Biographical Sentences from Web People Search Results

Harumi Murakami
Graduate School for Creative Cites
Osaka City University
Osaka, Japan
Email: harumi@media.osaka-cu.ac.jp

Toshimune Konishi
Graduate School for Creative Cites
Osaka City University
Osaka, Japan

Yoshinobu Ura
winspire
Wakayama, Japan

***Abstract*—To help users understand and select people on the web, we developed a method of generating biographical sentences for the results of web people search. We extracted attribute information about people (birth date, death date, birthplace, vocation, organization, and position) from the HTML files of person clusters that were manually classified into individuals and generated biographical sentences whose style resembles the first sentence of Wikipedia. We evaluated the method using 20 queries (person names) $\times$ 50 web search results. Our experimental results show that it is easy to generate Wikipedia-like biographical sentences of people using (1) a reasonable Wikipedia-like template and (2) a simple generating algorithm that combines extracted attribute information and predefined words.**

***Keywords*—biographical sentence, biographical summary, attribute information, web people search**

I. Introduction

The popularity of web people searches continues to rise as the number of people increases about whom the web can provide information. Through people search, users must distinguish among different people and select appropriate search results. If the list is merely "person 1, person 2, and so on," users have difficulty determining which person they should select. Appropriate labels shown with people should help users select the person they want. Research exists that assigns labels to people. For example, Wan et al. [1] separated web people search results and assigned titles to person clusters. Ueda et al. [2] assigned vocation-related information to person clusters. However, what kind of information is useful to understand or select a person? One answer might be "Google's knowledge graph," which is displayed on the right side on a keyword search by a person's name. Typically, if the searched person has a Wikipedia page, its 1st sentence is displayed.

In this paper, we present a method of generating Wikipedia-like biographical sentences for the results of web people search. We extract attribute information about people from the HTML files of person clusters that were manually classified into individuals and generate biographical sentences whose style resembles the first sentence of Wikipedia.

II. Approach

We extract attribute information about Japanese people (birth date, death date, birthplace, vocation, organization and position) and generate biographical sentences for them. When no attribute information is extracted, predefined words are used to fill the gap to form sentences.

A. Extracting Attribute Information

1) Birth and Death Dates: We extract one birth date and one death date using regular expressions that contain such terms as "born" or "year, month, date." We choose the most frequent candidates. If the most frequent candidates exceed one, a candidate is selected from the highest ranked web page. If more than one candidate remains, a candidate is selected that is found in the upper most part of the page. This "frequency first, high rank page second, upper part third" strategy, which selects one answer from the candidates, is common for the following extracting algorithms.

2) Birthplace: We obtain 100 characters before and after the person names since person profiles tend to be included. From these characters, we obtain 10 characters before and after "come from" and 10 characters before "born". By exact-matching prefecture names with these strings, prefecture names are extracted as birthplaces.

3) Vocation: The vocation list page of Wikipedia contains Japanese vocation names. Since the longest vocation name has around 20 characters, we obtain the 20 characters before and after person names. Character strings that contain such suffixes as "er" or "et" (usually used for vocations) are selected. We perform morphological analysis and concatenate the forward continuous nouns with the suffixes. Concatenated strings, which end with the suffixes used for the selection, become vocation candidates.

4) Organization: Since many organization candidates are contained on web pages, we only use the top five for extraction. Lines that contain such terms as "center," "hospital," and "university" etc. are extracted and analyzed morphologically, and continuous nouns are concatenated with the terms. Concatenated strings that end with the terms used for the selection become organization candidates. In addition, lines that contain such terms as "Inc.," "clinic," or "studio," are extracted and analyzed morphologically, and except for symbols, morphemes are concatenated with the terms. Concatenated strings that begin or end with the terms used for the selection become organization candidates.

5) Position: When only an organization is extracted, an associated position will be extracted using the extracted organization. From the 50 characters before and after the above organizations, we select such terms as "president," "professor," and " executive" and analyze them morphologically and concatenate the forward continuous morphemes with the terms.

Concatenated strings that end with the terms used for the selection become position candidates.

B. Generating Biographical Sentences

We generate biographical sentences with a template: "Name (Birth - Death) is a Birthplace Vocation Organization Position" from the extracted attribute information. When no attribute information is extracted, such words as "Japanese," "from," "person," and "belong to" are used to fill the gaps to form natural sentences. The generating algorithm is described below.

function Generate-Japanese-Summary(person-attribute-information)
returns a summary

if Birthplace is empty **then** Birthplace ← "Japanese"
else Birthplace ← Birthplace + "from"
end if
if Vocation is empty **then** Vocation ← "person."
else Vocation ← Vocation + "."
end if
if Organization is not empty **then**
  if Position is empty **then** Position ← "belongs to."
  else Position ← Position + "."
  end if
end if
Summary ← Name + "(" + Birth "-" + Death + ") is a"
+ Birthplace + Vocation + Organization + Position
return Summary

C. Example

> Asako Miura (1968 -) is a Japanese social psychologist. Professor of Department of Psychology, Faculty of Letters, Kwansei Gakuin University.

Fig. 1. Example of generated biographical sentences

Figure 1 shows an example of biographical sentences generated for Asako Miura whose birth date, vocation, organization, and position were extracted. Since her birthplace was not extracted, "Japanese" was added.

III. Experiment

As queries, we used 20 Japanese person names that were used in related work [3]. 50 web pages (HTML files) were obtained for all 20 queries from web searches (i.e., 20 × 50 = 1,000 HTML files). We manually classified these web pages into different people. 80 people existed.

For the attribute information extraction, we checked all of the extracted information for the 80 people using the following evaluation measures:

$$Precision = \frac{\text{correct answers by method}}{\text{people to whom information was assigned by method}} \quad (1)$$

$$Recall = \frac{\text{correct answers by method}}{\text{people to whom information was assigned manually}} \quad (2)$$

Table I shows the results of the attribute information extraction. The precision of birth date, death date, and birthplace exceeded 80% and seems reasonable. However, the recall of vocation, organization, and position was low and must be improved.

TABLE I. Results of extracting attribute information

Birth date		Death date		Birthplace	
Precision	Recall	Precision	Recall	Precision	Recall
82% (14/17)	70% (14/20)	100% (3/3)	75% (3/4)	93% (13/14)	93% (13/14)
Vocation		Organization		Position	
Precision	Recall	Precision	Recall	Precision	Recall
67% (16/24)	24% (16/67)	62% (39/63)	59% (39/66)	66% (19/29)	38% (19/50)

For biographical sentences, 100% (80/80) were generated; in other words, it was easy to generate Wikipedia-like biographical sentences of people. However, for 12 people (15%), those sentences were just default sentences: "person name is a Japanese person." In other words, no attribute information was obtained. For these 12 people, since only one page exists, and there is insufficient information for them.

IV. Related Work

Schiffman et al. [4] produced biographical summaries from newspaper articles using linguistic knowledge with corpus statistics. Basically their approach selects suitable sentences from a large corpus without producing Wikipedia-like sentences. We analyzed the first sentence of Wikipedia articles for people and presented an easy-to-generate template and an associated robust generation algorithm specialized to people. Some research (e.g., [5]) automatically generated general Wikipedia articles. We focus on people on the web and presented a method. We believe that our work's main contribution is to generate Wikipedia-like biographical sentences for people on the web. To the best of our knowledge, this is the first research of its kind.

V. Summary

We developed a method of generating biographical sentences for the results of web people search. We extracted attribute information about people (birth date, death date, birthplace, vocation, organization, and position) and generated biographical sentences whose style resembles the first sentence of Wikipedia. We evaluated the method using 20 queries (person names) × 50 web search results.

Acknowledgements.: This work was supported by JSPS KAKENHI Grant Number 25330385, 16K00440.

References

[1] X. Wan et al., "Person Resolution in Person Search Results: WebHawk," in: *Proc. CIKM 2005*, 2005, pp. 163–170.

[2] H. Ueda et al., "Assigning Vocation-Related Information to Person Clusters for Web People Search Results," in: *Proc. GCIS 2009*, 2009, vol. 4, pp. 248–253.

[3] S. Sato et al., "Distinguishing between People on the Web with the Same First and Last Name by Real-world Oriented Web Mining," *IPSJ Transactions on Databases*, vol. 46, no. 8, pp. 26–36, 2005.

[4] B. Schiffman et al., "Producing Biographical Summaries: Combining Linguistic Knowledge with Corpus Statistics," in: *ACL 01*, 2001, pp. 458–465.

[5] C. Super and R. Barzilay, "Automatically Generating Wikipedia Articles: A Structure-Aware Approach," in: *ACL 09*, 2009, pp. 208–216.