3P1-10in2

Web 上の人物の概要文の作成

Generating Summary Sentences for People on the Web

村上 晴美[*1] 小西 利宗[*1] 浦 芳伸[*2]
Harumi Murakami Toshimune Konishi Yoshinobu Ura

[*1] 大阪市立大学大学院創造都市研究科
Graduate School for Creative Cities, Osaka City University

[*2] winspire
winspire

To help users understand and select people on the web, we developed a method of generating summary sentences for the results of web people search. We extracted attribute information about people (furigana reading for person name, birth date, death date, place of birth, vocation, organization, and position) from the HTML files of person clusters that were manually classified into individuals and generated summary sentences whose style resembles the first sentence of Wikipedia. We evaluated the method using 20 queries (person names) * 50 web search results.

1. はじめに

著者らの研究室では, Web 上の人物の要約に関する研究を行っている[上田 07, 村上 09]. これまでに, Web 人名検索結果を同姓同名人物に分離して, 表[上田 07]や地図[Murakami 09]に表示するインタフェースを試作し, 人物を人間が識別するために都道府県やキーワード[上田 07], 職業関連情報[上田 09]を抽出する手法を考案した. また, 人物関連情報を履歴書形式で出力する手法を提案し[上田 10], 人物の履歴情報を地図上に表示するプロトタイプを試作した[Murakami 14].

本稿では, Web 上の人物の概要文を作成する. Web 人物検索結果の HTML 文書から, 氏名のよみ, 生年月日, 没年月日, 出身地, 職業, 所属と役職を抽出し, Wikipedia の第一文風の概要文を生成する.

2. Wikipedia 風の概要文の作成

先行研究[佐藤 05]による 20 の日本人氏名を用いて, Google Web APIs で 50 件の検索を行い, 検索結果を同姓同名人物に人手で分離した. データセット中には Wikipedia のページがある人物が 14 人物存在した. 実在人物が 12 人, 架空人物が 2 人であり, 実在人物の中生存人物が 8 人であった. この実在人物 12 人の第一文の先頭のフォーマットは概ね以下のとおりであることを確認した.

氏 名(氏のよみ 名のよみ、生年月日 - 没年月日)は、地域等の職業等。

地域等は 8 人が「日本」であり, 2 人が「東京都出身」と「福島県出身」, 1 人が「アメリカ合衆国を拠点に活躍する大阪府枚方市出身の日本」であり, 1 人は存在しなかった。職業等は, 1-5 の職業(所属等を含む)が列挙されており, 平均は 1.9, 内訳は職業が 16, 所属＋役職が 6, 学位が 1 であった. なお, 12 人物中 11 人物に第二文以降があり, 第二文の内容は所属＋役職が 9, 他が 5(本名, 旧姓, 出身地, 血液型, 業績)であった.

上記の分析により, 本研究では, 氏のよみ, 名のよみ, 生年月日, 没年月日, 出身地, 職業, 所属と役職を抽出対象とすることにした.

本研究のアプローチは, 属性情報抽出と概要文生成の 2 段階で構成される.

連絡先:村上 晴美, 大阪市立大学大学院創造都市研究科, 大阪市住吉区杉本 3-3-138, harumi@media.osaka-cu.ac.jp

3. 属性情報抽出

3.1 氏名のよみ

氏と名で分けて処理を行い後で結合する. 前処理として文書のカタカナをひらがなに, 大文字を小文字に変換する.

「人名漢字辞典-読み方検索(http://kanji.reader.bz/)」を用いてよみ候補を作成し, 小文字のローマ字表記を加える. (例)伊庭⇒いば, iba;幸人⇒ゆきと, ゆきひと, yukito, yukihito

よみ候補で文書を検索し, 一致したものを抽出する. 複数ある場合は前方(Web の上位⇒文書の上方)を優先する.

なお, 複数ある場合は以下同様(前方優先)である.

3.2 生年月日と没年月日

「生まれ」や「年月日」等に着目した正規表現を用いて西暦の生年月日及び没年月日を抽出する.

3.3 出身地

人物のプロフィールが含まれやすい氏名の前後 100 文字の文字列を取得する. その中から「出身」の前後 10 文字ずつ, 「生まれ」の前 10 文字にある「都道府県」を都道府県辞書を用いて抽出した. その際「出身者」等の表記のある箇所を除外した.

3.4 職業

Wikipedia の職業一覧ページの最も長い職業名を参考にして, 氏名の前後の 20 文字ずつを取得する. その中から「師」「士」等や, 最後が「ー」で終わる 4 文字以上のカタカナが含まれている文字列を抽出し, 形態素解析をかけ, 連続した名詞を結合する. 結合された文字列の内, 最後の文字が「師」「士」「ー」等となっている文字列を職業として抽出する.

3.5 所属と役職

所属と役職に分けて抽出し, 後で結合する.

(1) 所属

組織を表す文字列は非常に多いため上位 5 件の文書を用いる.

「センター」「営業所」「病院」「大学」等の文字列が含まれている行を抽出し, 形態素解析をかけ, 連続した名詞を結合する. この際, 経歴を表す「入学」「卒業」等を除外する. 結合された文

字列の内，語尾が抽出に用いた文字列になっているものを所属として抽出する．

また，「株式会社」「クリニック」「スタジオ」等の固有名詞が連続すると考えられる文字列では，連続した記号以外の形態素を結合する．結合された文字列の内，文字列の先頭もしくは語尾が抽出に用いた語となっているものを所属として抽出する．

(2) 役職

所属を抽出できた場合のみ，抽出された所属名を用いて役職を抽出する．

上位 5 件に出現する所属名の後方 50 文字から，「社長」「所長」「院長」「取締役」「幹事」等が含まれている文字列を抽出し，前方に連続した名詞を結合する．

4. 概要文生成

以下に，概要文生成のアルゴリズムを示す．人物の属性情報を与えると，「氏 名（氏のよみ 名のよみ、生年月日 - 没年月日）は、出身地の職業。所属役職。」という概要文を生成する．

function Generate-Summary(person-attribute-information)
returns a summary
  Name←氏 名
  Yomi←氏のよみ 名のよみ
  Birth←生年月日
  Death←没年月日
  Place←出身地
  Job←職業
  Org←所属
  Post←役職
  if Place is empty **then** Place←“日本の”
  else Place←Place + “出身の”
  end if
  if Job is empty **then** Job←“人物。”
  else Job←Job + “。”
  end if
  if Org is not empty **then**
   if Post is empty **then** Post← “所属。”
   else Post ← Post + “。”
   end if
  end if
  Summary←Name + “（” + Yomi + “、” + Birth + “ -” + Death + “）は、” + Place + Job + Org + Post
  return Summary

5. 出力例

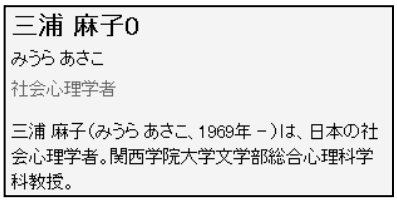

図 1：出力例

図 1 に，人物「三浦麻子 0」のクラスタ（HTML32 個）を入力とした Google のナレッジグラフ風の出力結果を示す．本研究の手法で出力される氏名のよみ，職業，概要文から構成される．

6. 実験

Google Web APIs で 20 氏名×50 件の検索を行い，同姓同名人物に人手で分離したデータセットには，80 人物が存在した．

概要文は 100%(80/80)生成できた．ただし，15%にあたる 12 人物は「氏名は、日本の人物。」というほぼ無意味な結果である．この中 92%(11/12)にあたる 11 人物ではページ数が 1 であった．

表 1 に属性情報抽出の結果を示す．氏名のよみ，生年月日，没年月日，出身地の適合率は 80%以上と比較的良好であったが，職業と所属と役職の適合率は 60%台であり改良の余地が大きい．再現率も職業，所属と役職では低く改善の必要がある．

表 1：属性情報抽出結果

氏名のよみ		生年月日		没年月日	
適合率	再現率	適合率	再現率	適合率	再現率
100% (37/37)	95% (35/37)	82% (14/17)	70% (14/20)	100% (3/3)	75% (3/4)
出身地		職業		所属	
適合率	再現率	適合率	再現率	適合率	再現率
93% (13/14)	93% (13/14)	67% (16/24)	24% (16/67)	62% (39/63)	59% (39/66)
役職					
適合率	再現率				
66% (19/29)	38% (19/50)				

7. おわりに

Web 上の人物の理解と選択を支援するために，Web 人物検索結果の HTML 文書から，氏名のよみ，生年月日，没年月日，出身地，職業，所属と役職を抽出し，Wikipedia 風の概要文を作成する実験を行った．

参考文献

[上田 07] 上田 洋，村上 晴美：Web 上の同姓同名人物を分離して人物属性情報を表示するシステム，2007 年度人工知能学会全国大会（第 21 回）論文集 (2007)

[村上 09] 村上 晴美，上田 洋：Web 人名検索結果の要約と可視化を目指して：2009 年度人工知能学会全国大会（第 23 回）論文集 (2009)

[Murakami 09] Murakami, H., Takamori, Y., Ueda, H., Tatsumi, S.: Assigning Location Information to Display Individuals on a Map for Web People Search Results, In *AIRS 2009*, Springer-Verlag, pp.26-37 (2009)

[上田 09] 上田 洋，村上 晴美，辰巳 昭治：Web 上の同姓同名人物識別のための職業関連情報の抽出，システム制御情報学会論文誌，Vol.22，No.6，pp.229-240 (2009)

[上田 10] 上田 洋，村上 晴美，辰巳 昭治：Web 上の人物理解のための履歴書作成，人工知能学会論文誌，Vol.25，No.1，pp.144-156 (2010)

[Murakami 14] Murakami, H., Tang, C., Wang, S., Ueda, H.: Vitae and Map Display System for People on the Web, In *IEA/AIE 2014*, Springer-Verlag, pp.348-359 (2014)

[佐藤 05] 佐藤 進也，風間 一洋，福田 健介，村上 健一郎：実世界指向 Web マイニングによる同姓同名人物の分離；情報処理学会論文誌：データベース，Vol.46, pp.26-36 (2005)