Wikipedia エントリを用いたニュース字幕中の人物の名寄せ
Person Identification in Broadcast News Caption using Wikipedia Entries

鈴木　規之† Noriyuki SUZUKI
奥岡　知樹†† Tomoki OKUOKA
高橋　友和††† Tomokazu TAKAHASHI
井手　一郎††,†††† Ichiro IDE
出口　大輔†† Daisuke DEGUCHI
村瀬洋†† Hiroshi MURASE

†名古屋大学 工学部　〒464-8603 愛知県名古屋市千種区不老町
††名古屋大学大学院 情報科学研究科　〒464-8601 愛知県名古屋市千種区不老町
†††岐阜聖徳学園大学 経済情報学部　〒500-8288 岐阜県岐阜市中鶉 1-38
††††国立情報学研究所　〒101-8430 東京都千代田区一ツ橋 2-1-2

1. はじめに

ニュース映像には、様々な人物名称が出現する。一般的に知名度の高い人物は、フルネームで呼ばれることは少なく、しばしば苗字と役職名で表される。このような人物名称は、外部知識がなければ名寄せが難しい。小笠原ら[1]は外部情報としてニュース映像中の顔画像を利用した。

本講演では、ニュース字幕（Closed-Caption）中に出現した人物名称を Wikipedia 中のその人物に関するエントリに対応付けることにより、名寄せする手法を紹介する。

2. 提案手法

2.1 提案手法の概要

人物の特定に必要な外部情報として、Wikipedia 日本語版[1]を利用する。Wikipedia 日本語版の人物に関するエントリには、略歴などの人物に関する豊富な記述がある。これらの情報を利用することにより、ニュース字幕中の人物を特定する。

2.2 人物に関する Wikipedia エントリの抽出

Wikipedia 記事は、記事毎にカテゴリ分けされている。本研究では、まず、人物に関するカテゴリを利用し、人物について記述したエントリを抽出する。表 1 は抽出に用いたカテゴリの例である。以下ではニュース字幕中に出現した役職名から、Wikipedia でカテゴリ名となっているものを人手で選んだ。12 種の役職名を用いてカテゴリを抽出し、Wikipedia 日本語版全体（2009/11/24 時点の記事）から、人物に関するエントリ約 25 万件を抽出した。

表 1　抽出に用いたカテゴリの例

2.3 ニュース字幕と Wikipedia エントリの照合

ニュース字幕から人物名称を抽出するために、4〜5 文のニュース字幕を一つの文章とし、それに対し、形態素解析（形態素解析器は茶筌[2]を使用）を行い、「名詞（列）＋役職名」という条件を満たす名詞列を得る。以下では役職名（20 種）は人手で与えた。これらの人物名称と、Wikipedia 中の人物に関するエントリ及びカテゴリ情報を照合し、名寄せ候補となるエントリを抽出する。照合元のニュース字幕から作られる単語ベクトルにおいて、照合中の人物名称の頻度が 2 倍になるように加重して強調し、2.2 で抽出した全てのエントリに対して、単語ベクトル間の内積による類似度を求める。類似度比較の結果、類似度が最大のものを名寄せ対象のエントリとする。

3. 評価実験

3.1 実験条件

ニュース映像として、「NHK ニュース 7」2001/3/16〜2001/8/29（100 日分）より、人物名称を抽出、そのうち 2.2 で抽出した人物に関するエントリの中に正解が含まれる人物名称（56 件）を人手で抽出し、提案手法により名寄せを行った。最後に正解を人手で与え、正解率を求めた。

3.2 結果と考察

本実験では、56 件の人物名称に対し 31 件を正しく名寄せ出来た。よって正解率は約 55%であった。

名寄せ誤りの理由として、エントリによって記述量が大きく違い、特徴ベクトルが抽出しづらいことが挙げられる。この問題に対処するためには、各エントリから抽出する特徴ベクトルの大きさをできるだけ揃える必要がある。

4. むすび

ニュース字幕中に出現する人物名称を、Wikipedia 日本語版の人物に関するエントリを用いて名寄せする手法を提案した。

今後は、人物に関するカテゴリと、カテゴリに属する役職名を厳選することで、より多くの人物に対応することのほか、類似度計算方法を工夫することで、より高精度な名寄せを行うことができると考えている。

5. 謝辞

本研究は、科研費によった。

6. 参考文献

[1] 小笠原, 高橋, 井手, 村瀬, “ニュース映像アーカイブにおける登場人物の顔照合を利用した名寄せ”, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, 2006-264 pp.55-60 ,March 2007

[2] Wikipedia 日本語版　*http://ja.wikipedia.org/*

[3] 形態素解析器[茶筌] *http://chasen-legacy.sourceforge.jp/*