新聞記事における顔領域と名前との自動対応づけ

山田 剛一　　杉山 一成　　与那嶺 靖典　　中川 裕志
横浜国立大学 工学部
{aron,ksugi,yasunet,nakagawa}@naklab.dnj.ynu.ac.jp

1 はじめに

マルチメディアを構成する各メディアの内容の間には、例えば顔画像と人名が同一人物を表しているといった、意味的な関係があるのが普通である。しかし、マルチメディアの内容検索を行なう研究の多くは単一のメディアの情報により検索を実現しており、メディア間の意味的な関係を利用していない。マルチメディアを統合されたメディアの意味によって検索するには、メディア間の関係を解析し利用することが必要である。

本研究では、マルチメディアコンテンツとしてWeb上で公開されている写真ニュースを利用し、その記事中の人名と写真中の人物の対応関係を、各メディアの特徴を学習することにより判断するシステムを構築した。このシステムを用いることにより、例えば人物を検索するとテキストに書かれた情報だけでなく写っている写真も出力するような、各メディアの意味内容の連係を生かした検索システムが実現できる。

2 各メディアの解析結果を統合するためのアーキテクチャ

写真ニュースは、記事本文と写真画像からなる。図1に示すような記事中の人名と写真中の人物を対応づけるために必要となる最初の処理は、記事本文からの人名の抽出と、写真画像における人物の認識である。しかし、記事本文に現れる人名に対応する人物が写真に必ず写っているわけではないし、写真には記事には現れない人物まで写っていることも多い。両メディアに共通して現れる人物だけを選択する仕組みが必要である。

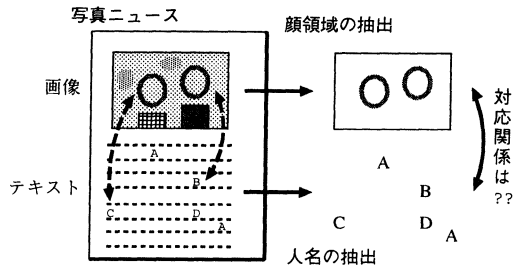

図 1: 人名と顔領域の抽出

そこで本研究では、「両メディアに共通して現れる人物とそうでない人物とでは、記事中の人名の現れ方や、写真中の顔の現れ方に違いがある」という仮定に基づき、各メディアにおけるその現れ方の違いを学習するという方法をとる。

システムの構成としては、メディアごとに処理内容が異なることから、記事本文を処理する言語モジュール、写真画像を処理する画像モジュールがある。これらのモジュールは重みつきの候補をそれぞれ出力するので、それらから最終的な結果を導く統合モジュールも必要である。統合モジュールは各モジュールによって出力される候補をまとめ、それらの各モジュールによる重みから全体としての重みを計算し、出力する。

言語、画像の各モジュールにおいて両メディアに共通して現れる人物の人数を推定するわけであるが、同じ人数を推定するのであるから、片方のメディアの判断は、もう片方のメディアが何人と判断したかによって影響を受ける。そこで、本システムでは次のような方法をとっている。まず、実際に記事と写真に共通して現れる人物の人数について、あらかじめ数通り仮説をたてる。次にその各仮説の下に両モジュールが結果を出力し、その結果の組み合わせの中から最も確からしい解を選択するという方法である。今回は、両メディアに共通して現れる人物の人数が{1人,2人,3人以上}という3つの仮説を用いた。

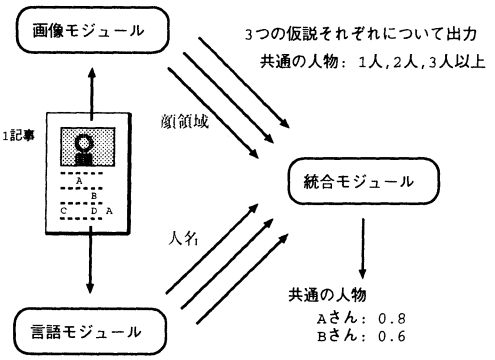

図 2: システム構成の概観

3 テキストからの人名候補の抽出

言語モジュールは、記事中に現れる多くの人名の中から、写真に現れている人物の人名を抽出し重みつきで出力する。

人名が写真に写っている人物を指している場合には、その人物は記事の内容において重要な人物であることが多いので、その人名の周囲の言語表現には何か特徴があると考えられる。本手法ではこの考えに基づき、人名の前後の言語表現によって、その人名が写真中の人物であるか否かを判断する。その言語表現の特徴は、機械学習によって獲得する。

3.1 属性として用いる言語表現の特徴

本システムでは、まず記事本文を日本語形態素解析システム JUMAN version 3.6 [1] により形態素解析し、そ

の結果から機械学習で用いる各属性値を生成する。用いる属性は以下に述べるものである。

人名を主辞とする複合名詞の情報

新聞記事では、人物を導入するときに肩書きや年齢、生年などの情報を人名の前後に配置することが多い。それらが存在するかどうかを調べ、その有無を属性としている。なお、これらの情報を表していると解析された形態素は、人名の周囲の形態素列(後述)には含まれない。

人名の構文上の位置

注目している人名の格は何か、対応する述語は何かといった情報は構文解析・意味解析をしなければ得られないものであるが、意味を反映しているので重要な情報である。そこで、構文解析を必要としない簡単な解析によって、これらの属性を近似的に生成することにした。

まず、格を解釈する代わりに、後続する助詞を属性とすることにした。当然、取り立て助詞が用いられている場合などがあるので格とは直接対応しないが、人名の構文上の位置を反映する属性として有効である。

述語については、構文解析を行わないので、人名の後ろにあり人名に最も近い用言を対応する述語として選択している。

人名の周囲の語の品詞情報

人名の周辺の言語表現の特徴をつかむため、人名の前後 n 形態素の品詞情報を、相対位置の情報も含めて使用することにした。品詞情報としてはJUMANの品詞体系に基づき、品詞と品詞細分類の2つを属性としている。ただし、用言については品詞だけとした。なお、評価の際には $n = 5$ としている。

人名の出現位置と出現頻度

人名の出現位置は、文書の構造を近似的に反映するものとして重要である。人名の出現位置がタイトル内か、あるいは本文なら何文目、何段落目であるかを属性として用意した。また、記事の先頭から何番目に現れた人名であるか、および記事中の出現頻度といった統計情報も属性としている。

3.2 機械学習による決定木の生成

これまでに見てきた各属性の属性値から、分類モデルを決定木で表現するタイプの学習プログラムであるC4.5(Release 8) [2] およびC5.0 [3] を用いて決定木を生成する。クラスは、{ 人名の表す人物が画像中に存在, 非存在 } の2クラスである。

各クラスに属する事例数のバランスが違う複数の学習データを用意し、そのそれぞれから生成される決定木の判断を総合して重みづけする手法 [4] を用いることにより、重みつきの出力としている。これにより、言語 / 画像の各モジュールからの候補を統合モジュールで選択する際の自由度を確保している。

4 画像からの顔領域候補の抽出

画像モジュールは、テキストに現れる人物の顔領域を画像中から抽出する。言語モジュールと同様の学習手法を用い、領域の特徴量を学習し、判断している。

画像中の人物とテキスト中の人名の対応づけを行なうためには、顔の認識を行うだけでなく、テキスト中に現れるような顔を選択して抽出することが必要である。このため、画像の特定の部分が目的の顔であるか否かを判断するための材料として、顔か否かを反映する特徴量とともに、テキストに人名が現れた顔か否かを反映する特徴量が必要となる。

また、1枚の画像中における顔の位置関係、面積といった特徴が写っている人数によってそれぞれ違う(図 3)ので、両メディアに共通して現れる人物の人数が1人、2人、3人以上の3通りの場合に分けて学習 / 判断を行なった。

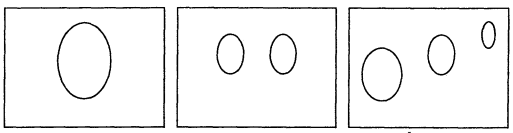

図 3: 人数別の特徴

4.1 顔領域候補の抽出

肌色分布が (R, G, B) 空間でガウス分布に従う [5] と仮定し、*Mahalanobis* 距離を導入した。85人分の画像から頬の部分の肌色 [5] を 5×5 のブロックで抽出し、これらのデータから濃度平均値 $\boldsymbol{M}$、分散共分散行列 $\boldsymbol{V}$、輝度 $\boldsymbol{I}$ を求め、式(1)の右辺で定義される *Mahalanobis* 距離を求めた。ここで、

$$d^2 > (\boldsymbol{I} - \boldsymbol{M})^T \boldsymbol{V}^{-1} (\boldsymbol{I} - \boldsymbol{M}) \qquad (1)$$

を満たす画素を肌色候補画素として、顔領域候補を抽出した。閾値 d の値は実験により定めた。

4.2 抽出領域の特徴パラメータ

形状のパラメータとしては、最大領域に対する面積比、縦横比、矩形度、楕円度、円形度を、色のパラメータには、R, G, B, 輝度 Y のそれぞれの平均値を使用した。構図を反映するパラメータには、以下の8つを用意した。

- 画像全体の縦横比
- 領域の重心の x, y 座標
- 画像の中心と領域の重心との距離
  画像の対角線の半分の長さで正規化した値を用いている。
- 画像の中心から何番目に近い領域か
- 画像の上端中心から領域の重心までの距離
  画像の上端中心から左下端（または右下端）までの距離で正規化した値を用いている。
- 画像の上端の中心から何番目に近い領域か
- 画像を 3×3 に9分割した時、領域の重心がどこに含まれるか

4.3 機械学習による決定木の生成

これまでに見てきた各属性の属性値から、言語モジュールと同様にC4.5/C5.0を用いて決定木を生成している。クラスは、{人名が記事本文に現れる人物の顔領域, そうでない領域}の2クラスである。

5 人名と顔画像の対応づけ

言語モジュールと画像モジュールが出力する対応づけの各候補を総合的に評価し、システム全体としての結果を導く統合モジュールについて述べる。

5.1 入力

言語、画像の各モジュールは、両メディアに共通して現れる人物の人数が{1人,2人,3人以上}の各仮説の下に動作するので、1つのモジュールの出力が3種類、両モジュールあわせて6種類の出力となる。1つの記事において、両メディアに共通して現れる人物は1人とは限らない。そこで、言語、画像の各モジュールは、1つの記事に対し重みつきの候補を複数出力する。これを重みの高い順に並べると、図4の関数として表現することができる。ただし言語モジュールの出力は、実際には単なる重みの列ではなく、人名と重みの組である。

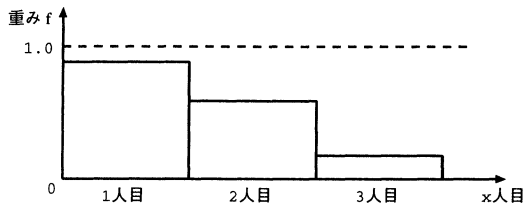

図 4: 1つの仮説における各モジュールの出力

両モジュールの出力の重みを以下の式で表す。

$$(\text{言語モジュールの出力}) = f_{lang}(n,x) \quad (2)$$

$$(\text{画像モジュールの出力}) = f_{image}(m,y) \quad (3)$$

ただし、n,mは人数(仮説)である。x,yは候補の順位(重みの高い順)で、例えば$f_{lang}(n,2)$は2番目に重みの高い人物の重みを表す。

5.2 組み合わせの評価

各メディアで3通りの仮説をたてて結果を出しているので、3×3通りの結果の組み合わせが存在する。その中から最も確かな組み合わせを選択し、システムの結果とする。各組み合わせを評価するために、結果の近さという概念を導入する。

メディア間の近さ

メディア間の近さf_mは、言語、画像の両メディアの出力する結果の近さを評価する度合いである。次式で定義される。

$$f_m(n,m) = \sum_{z=1}^{M} \frac{|f_{lang}(n,z) - f_{image}(m,z)|}{f_{lang}(n,z) + f_{image}(m,z)} \quad (4)$$

ただし、候補を多く出力したメディアの出力人数をMとする。

この式によれば、メディア間で同一順位zの候補の重みの値が近いほど、全体の値も小さくなる。つまり、$f_m(n,m)$が小さいほど、画像メディアの出力と言語メディアの出力は近い。

メディアと仮説との近さ

仮説を立てて求めた結果であるf_{lang}やf_{image}が仮説と一致しなければ、その仮説は妥当でなかったと判断すべきである。そこで、仮説と結果の近さを評価する。まず、両メディアに共通して現れる人物の人数がnであるという仮説は以下の関数f_aとして記述することができる。

$$f_a(n,x) = \begin{cases} 1 & (x \le n) \\ 0 & (x > n) \end{cases} \quad (5)$$

ただし、xは候補の順位である。

言語と画像の各モジュールごとに仮説を立てているので、言語モジュールにおける出力と仮説との近さf_{al}、画像モジュールにおける出力と仮説との近さf_{ai}の双方を求める。前述のメディア間の近さと同様、f_{al}, f_{ai}はそれぞれ次のような式で表せる。

$$f_{al}(n) = \sum_{z=1}^{N} \frac{|f_{lang}(n,z) - f_a(n,z)|}{f_{lang}(n,z) + f_a(n,z)} \quad (6)$$

$$f_{ai}(m) = \sum_{z=1}^{N} \frac{|f_{image}(m,z) - f_a(m,z)|}{f_{image}(m,z) + f_a(m,z)} \quad (7)$$

ただし、仮説の人数と候補の数のうち大きい方をNとする。なお、仮説が「3人以上」の場合には、順位zが4以上の候補については仮説との相違が最低でも$z/10$はあるとする。これは「3人以上」という仮説が3人よりも少し多い程度を想定していることの反映である。

仮説の一致度

言語、画像の各モジュールの間で仮説が一致していなければ、なんらかの矛盾が生じている。しかし、それぞれのモジュールの処理は完璧ではないため、矛盾した状況も排除せず、重みを小さくすることで対処している。この仮説の組合せの重み$D(n,m)$として、表1を用いる。

表 1: 仮説の一致度

		言語の仮説		
		1人	2人	3人以上
画像の仮説	1人	1.0	0.7	0.5
	2人	0.6	1.0	0.5
	3人以上	0.3	0.4	0.8

近さの統合

上記で求めた2種の近さと一致度を元に、各組み合わせについて最終的な近さ$f(n,m)$を求める。この値が小

さいほど、一致しているということになる。

$$f(n,m) = \{f_m(n,m)+1\}\{f_{al}(n)+1\}\{f_{ai}(m)+1\} \times \frac{1}{D(n,m)} \quad (8)$$

5.3 出力の統合

統合された近さ $f(n,m)$ の最も小さい組み合わせが求まると、言語、画像の各モジュールを代表する出力が定まる。この2つの出力を統合し、システム全体の出力とする。この統合では各候補(順位 z)ごとに、以下の2式にしたがって重みを求める。式(9)は両メディアにおいて、人数に関して同じ結果が導けることを重視しており、一方、式(10)はどちらかのメディアが重みを高くつければ、システムとしても高い重みを出力する。

$$\forall z, f_{union}(n,m,z) = f_{lang}(n,z) \times f_{image}(m,z) \quad (9)$$

$$\forall z, f_{union}(n,m,z) = 1 - \{1 - f_{lang}(n,z)\}\{1 - f_{image}(m,z)\} \quad (10)$$

各モジュールの出力と同様、システム全体の出力も図4のような重みつきの出力である。言語モジュールの出力から各順位の人物の重みと人名との対応がついているので、「A さん: 重み 0.8 で画像にも出現」「B さん: 重み 0.4 で画像にも出現」といった出力となる。

6 評価

実際の写真ニュースを解析し、メディアの統合の効果を評価した。その効果を確認するため、言語モジュールおよび画像モジュール単体での評価も行った。

評価に用いたデータは、毎日新聞社が公開しているWebページ「AULOS」[6]の写真ニュースである。写真ニュースの性格上、長い記事は少なく、また通常、写真にはキャプションはついていない。画像のピクセル数はまちまちであるが、250 × 200 程度のものが多い。

今回の評価では1997年5月と6月の記事のうち、写真がカラーで、記事本文に画像中の人物の人名を含む228記事を使用した。言語モジュールは4 fold、画像モジュールは3 foldの交差検定を行った。

記事中の人名が画像に現れるかを正しく判断できるか評価するため、両メディアに共通して現れる人物の再現率 / 適合率を求めた。再現率 / 適合率は次の式で定義した。

$$\text{再現率} = \frac{\text{正しい出力の重みの総和}}{\text{両メディアに現れる人物の数}} \quad (11)$$

$$\text{適合率} = \frac{\text{正しい出力の重みの総和}}{\text{すべての出力の重みの総和}} \quad (12)$$

各モジュールの評価結果は、表2,3のようになった。

統合モジュールには式(9)、式(10)による違いがあるが、式(10)を用いた場合には、統合により再現率 / 適合率とも向上している。式(10)で再現率が向上するのは、片方のモジュールで取り損ねてしまう人物をお互いに補い合っているからであり、適合率も向上するということはこのプロセスがかなり有効に機能していることを示している。

表2: 各出力の評価(C4.5 を使用)

出力元		再現率	適合率
言語モジュール		0.80	0.53
画像モジュール		0.54	0.35
統合モジュール	式(9)	0.68	0.64
統合モジュール	式(10)	0.87	0.55

表3: 各出力の評価(C5.0 を使用)

出力元		再現率	適合率
言語モジュール		0.81	0.61
画像モジュール		0.56	0.21
統合モジュール	式(9)	0.70	0.65
統合モジュール	式(10)	0.86	0.62

一方、式(9)を用いた場合には、再現率が落ちるものの、式(10)よりも高い適合率となった。これは、式(9)では片方のメディアのみに現れるノイズを無視するように統合するためである。

7 おわりに

テキスト中の人名と画像中の人物との対応づけを行うシステムを構築した。言語モジュール、画像モジュールの各モジュール単体での解析結果を統合することにより、対応づけを実現した。今後は対応づけの精度向上を図るとともに、対応づけした人物に対して各メディアにおける特徴量を付随させ、マルチメディアにおける人物検索システムを構築する予定である。

謝辞 充実した写真ニュースをWeb上で発信している毎日新聞社に敬意を表すると同時に感謝いたします。

参考文献

[1] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.6. 京都大学大学院情報学研究科, Dec 1998.

[2] J.Ross Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc., 1993. (古川康一 監訳. AI によるデータ解析. トッパン, 1995).

[3] RuleQuest Research Pty Ltd. *See5 / C5.0*, 1998. http://www.rulequest.com/.

[4] 山田剛一, 杉山一成, 与那嶺靖典, 中川裕志. 新聞記事における顔写真と文書表現との自動対応づけ. 第4回知能情報メディアシンポジウム論文集, pp. 39–46. 知能情報メディア時限研究専門委員会 電子情報通信学会 , Dec 1998.

[5] Shin'ichi Satoh, Yuichi Nakamura, and Takeo Kanade. Name-it: Naming and detecting faces in video by the integration of image and natural language processing. *Proc. of International Joint Conference on Artificial Intelligence*, pp. 1488–1493, 1997.

[6] 毎日新聞社. 毎日新聞 AULOS 写真ニュース. http:// aulos.mainichi.co.jp/, 1997.