2G2-04

照応解析を利用した放送番組からの登場人物の相関図生成

Generation of Correlation Charts from TV programs based on Anaphora Resolution

後藤 淳[*1*2*3] Jun Goto　八木伸行[*1] Nobuyuki Yagi　相澤彰子[*3] Akiko Aizawa　関根 聡[*4] Satoshi Sekine

[*1]NHK 放送技術研究所 NHK Science and Technical Research Laboratories
[*2]総合研究大学院大学 Graduate School of Advanced Studies
[*3]国立情報学研究所 National Institute of Informatics
[*4]ニューヨーク大学 New York University

This paper describes a method to create correlation charts based on anaphora resolution. The method extracts expressions that describe attributes of individuals and relations between them from plain texts such as TV program abstracts for dramas and movies in Electronic Program Guides (EPG), and integrates the extracted relations using the results of anaphora resolution to create human correlation charts. Evaluation reveals that the charts created by the method cover 86.7% of attributes of individuals and 71.7% of relations compared with charts made by a human subject.

1. はじめに

デジタル放送開始に伴い、EPG(Electronic Program Guides)により、番組本編の映像・音声と共に、番組の内容に関する番組概要などのメタデータが得られるようになった。中でも映画やドラマなどのジャンルでは、背景、あらすじ、登場人物の紹介などが含まれており、視聴者は番組を探す際、番組概要を表示させ放送番組の具体的な内容を確認することができる。しかし、見たい番組を選ぶためにザッピングしている場合は、ユーザは内容を知りたい要求はあるものの、短時間で文章を読むことは難しいため、よほど興味のある番組でない限り、番組概要を読まないのではないだろうか。

一方、雑誌やインターネットの記事では、映画やドラマなどの内容を紹介するとき、人物の関係を示した図（人物相関図）を本文と共に掲載している。情報の可視化は、読者に文章に対する興味を引かせるのと同時に、文章の内容把握に役立つためであると考えられる[1][2]。放送コンテンツのインタフェースでも、概要と共に相関図を表示すれば、視聴者に番組内容を短時間で把握させることができ、リラックスした状態やザッピング時の情報提示への応用を期待できる。そこで、番組それぞれに付与されるEPGの番組紹介文（番組概要）を用いて、自動で人物相関図を生成する手法について検討する。

テキストからの人間関係生成の関連研究として、松尾らは、Web に出現する研究者の共起頻度から関係の強さを求め、人間関係ネットワークを生成している[3]。また馬場らは、小説のテキストデータから登場人物が出現する“場面”という概念を用いて、場面に共起する人物には関係があると捉え、相関図を生成している[4]。これらの研究では、インターネットや小説などの人名が頻出する大規模データを対象にするため、大量の人物表現が得られ、各記事や場面ごとの再現率が必ずしも高い必要はない。そのため、これらの先行研究では、人物表現の照応の解決にテキストの類似度のみを使用しており、代名詞や一般名詞が固有名詞を示すような場合は考慮していない。さらに、人物間の関係も人物表現の共起頻度による関係の強さを用いており、その関係の詳細な種類までは対象としていない。

連絡先：後藤淳，NHK 放送技術研究所，〒東京都世田谷区砧 1-10-11，goto.j-fw@nhk.or.jp

本研究では、照応解析を用いて、表層表現の類似度だけでは取り扱うことができなかった代名詞や一般名詞が指す人物を同定することにより、番組概要などの少量の文章や人物名の出現頻度が少ない場合でも、それぞれの人物関係の抽出し、ストーリ上の人間関係の構造を取得する。

2. 照応解析に基づく相関図生成

本稿では、自然言語から人物とその間にある関係を抽出した上で、（人、関係、人）の三つ組を人物の照応解析結果により結合するアプローチをとる。相関図生成の流れを図1に示す。以下に各処理内容について説明する。

1. 人物表現抽出
番組概要から、登場人物を示す表現を抽出する。
2. 人物表現間の関係抽出
抽出した人物表現と構文解析の結果を用いて、人物表現間にある関係を抽出する。
3. 人物表現の照応解析
文章中の異なる場所に出現する人物表現の照応関係を抽出する。
4. 相関図の生成
照応解析結果を用いて、人物表現を統合し、ネットワーク状の人間関係相関図を生成する。

図 1　処理の流れ

2.1 人物表現抽出

番組概要から相関図を生成するには、まず相関図のノードとなる人物を文章中から抽出する必要がある。人物を示す表現は、人名、職業名、男や女などの一般名詞、代名詞などが含まれる。図2に EPG の番組概要の例を示す。太字が人物表現として抽出するものである。以下に、本稿で取り扱う人物表現の種類について述べる。

(1) 固有表現

固有表現とは、人物名などの固有名詞に数値表現を加えたものである。関根らは、固有表現の定義を階層化した拡張固有表現（Extended Named Entity, ENE）[1]を提案している。本稿では、ENE から登場人物として有用なタグを選択し使用する。

登場人物としては、人名の抽出が最も重要であるが、映画の概要では、人名の代わりに、教授などの職業・肩書きを人物表現として利用することがある。また、生物名や組織等を擬人化して扱う場合がある。図2の例では、王子や首相を人物を示

[1] http://nlp.cs.nyu.edu/ene/

す表現にそのまま用いている。固有表現は、その種類が多様であることや複数の形態素を統合した表現が多いため、表現の抽出には、機械学習を用いることとする。

(2) 一般名詞

一般名詞のなかにも人物を示す名詞が含まれている。番組概要では、名前には言及せず、抽象的に男、女などの一般名詞(一般人物表現)を用いる場合も多い。また、令嬢、一人娘、敵などのように人物を示すと同時に、他の人物との関係を示す名詞(関係人物表現)がある。これらの語彙は関係を抽出する際にも有用となる場合があるため、一般人物表現とは分けて抽出を行う。一般人物表現や関係人物表現は、複合語などの問題もあるため、固有表現抽出と同様に機械学習を利用する。

(3) 代名詞

通常、日本語の文章では、彼、彼女、彼ら、等の人称代名詞はあまり使われない。一度使われた固有名詞は、以後の文で代名詞に置き換わらずに使用されるか、省略されることが多い。しかし、映画などの紹介文では人称代名詞がよく使用されている。NHK の映画 378 本の EPG 文を調べたところ、158 本で用いられている。映画の紹介文では、限られた文字数で、1本毎にストーリのエッセンスを紹介するため、文字数が多い人名の連続した使用を避け、代名詞を使うからと考えられる。そのため、映画などの紹介文章では、人称代名詞の把握は、人間関係の構造を得るために非常に重要である。代名詞の抽出は、その種類が限られているため、予め作成した辞書に基づき行う。

B・バルドーが、フランス**首相**の**一人娘ブリジット**という**令嬢**役にふんしたラブ・コメディー。**官房長**の**ミシェル**に恋焦がれる**ブリジット**は、初めは相手にされなかったものの、ちょっとした作戦が功を奏し、見事**ミシェル**と結婚。しかし、**彼**の派手な女性関係にうんざりした**彼女**は、**国賓**の**王子**との浮気を宣言する。**王子**役はアカデミー主演男優賞に4度ノミネートされたフランスの**2枚目スター**、**シャルル・ボワイエ**。

図2 番組概要の例

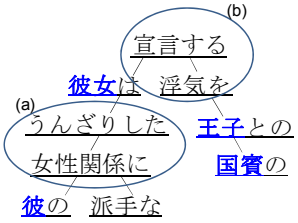

図3 関係抽出例

2.2 人物表現間の関係抽出

(1) 構文情報に基づく人間関係の抽出

番組の紹介文では、限られた文章で登場人物や内容を紹介するため、1文に複数の人物表現が現れる。そこで、人間関係の抽出は、1文の構文解析結果(構文木)から関係を抽出する。人物表現抽出で得られた固有表現と構文木との対応を取り、構文木上の人物表現を含む2つの節の間にある表現を関係として取得する。

例えば、図2の「彼の派手な女性関係にうんざりした彼女は、国賓の王子との浮気を宣言する。」という文を構文解析すると図3のようになる。この文で人物表現は、{彼、彼女、王子、国賓}である。そのため、”彼の”の節と”彼女は”の節の間にある表現(a)と、”彼女は”と”王子との”の間にある表現(b)が関係表現として抽出される。

(2) 単文への分割

関係抽出の精度は構文解析に依存するため、できる限り構文解析結果を向上させることが重要である。このため、構文解析が誤った結果を出しやすい複文を単文へ分割することによって、構文解析の精度向上を図った。ここでは、金らの分割処理に基づいたルール[5]を用いた。

(3) ゼロ主語の補完

日本語では、話題の中心になっている物や人は、次の文では省略されることがある。本稿では、2つの人物表現の間にある関係を、構文解析結果(構文木)から抽出する方針をとるため、人物名が省略されると関係を抽出することができない。

そこで、主語若しくは主題がない文については、以前の文から主語の補完を行う。主語補完には、

「主題(ハ) ＞ 主語(ガ格) ＞ 間接目的(ニ格)＞直接目的(ヲ格)＞ その他」

のセンタリング理論に基づく表層格を利用した優先度[6]を利用する。

2.3 人物表現間の照応関係抽出

多くの場合、抽出された人物表現は、表層上は異なる表現であっても、同一のエンティティを示している場合がある。また逆に同一の表現であっても異なるエンティティを指す場合がある。そこで、照応解析を用いて同一のエンティティを示す人物表現を獲得する。本研究では、照応解析を行う際に、人物表現抽出により抽出された人物表現のみを対象とする。これにより、名詞全体の集合から先行詞を選ぶ問題に比べ、大幅に候補の対象を削減することができ、精度の向上が期待できる。

また、代名詞などの照応詞の照応先(先行詞)を決める際、前方照応を対象とする。つまり、後方照応は考慮せず、照応先は、その出現位置より前に在ると仮定する。後述する条件に合致し、代名詞の最も直近の文に現れる先行詞を照応先として取得する。

(1) 人物表現間の照応解析

以下に照応詞となる人物表現及び先行詞となる人物表現について説明する。

(a) 照応詞となる人物表現

- **代名詞**

番組概要には、人物名の代わりに代名詞(彼、彼女、彼ら)が用いられる。代名詞は、ほとんどの場合、他の人物表現を参照している。そのため、代名詞は照応詞として選択され、後述する条件(性別、単数複数、格助詞)などの情報を用いて、代名詞より前の人物表現を参照する。

- **連体詞＋一般人物表現**

”その”などの連体詞に続く一般人物表現は、代名詞などと同様の処理を行う。例えば、”その男"は、性別情報と単数複数情報に基づき、代名詞の”彼”と同様の処理を行う。

ただし、指示代名詞に続く名詞が、関係人物表現(妹、父等)である場合は、例外とする。例えば、”トムとその妹”では、その妹”は、妹のオブジェクトを参照するのではなく、"その”がトムを参照している。

- **人数表現**

番組概要には、人数の表現で複数の人物を指す場合がある。例えば、「資金を集めるため、2人はバンドを結成するが…。」という文では、前の文で出現しているキャラクタの2人を指している。このときは、2 人の照応先は複数の人物表現か、単数の人物表現を人数分照応する。

(b) 先行詞のなる人物表現

- **性別の一致**

彼や彼女、その男などの照応詞が性別情報を持っている場合、その照応先は性別に矛盾がないものを選ぶ必要がある。図2の例では、彼や彼女がミッシェルを指すのかブリジットを指すのかは、性別情報を考慮しない限り解決できない。

そのため、人物表現の性別を特定する手段を考える。日本人の名前(First Name)は表層やその読みから性別を判定することが比較的容易である。また、英語圏でも性別によって使われる First Name はある程度決まっている。そこで、機械学習で人名と性別の関係を学習することにより性別の判定を行う。

- **単数複数の一致**

単数の照応詞は、単数の属性を持つ先行詞を照応する。例えば、“彼女”が、複数の属性を持つ”姉妹”を照応することはできない。同様に、彼ら、その男達などの複数の属性を持つ照応詞は、複数を示すグループや組織などの名詞を照応するか、若しくは照応先が複数とする。

- **格助詞による優先順位**

ゼロ主語の補完の際に用いた表層格に基づく優先度を用いて、照応先を決定する。対象の文で主格が省略されているときは、ゼロ主語と扱い、さらに前の文から先行詞を取得する。

(2) 人物表現間の関係に基づく照応

- **同一節、ノ格**

人物表現 A がノ格で他の人物表現 B を修飾しているとき、若しくは同一文節にあるとき、A と B は同一指示とする。ただし、関係人物表現がノ格係り先であった場合は例外として照応しない。例えば、”弟のトム”では、弟＝トムである

- **表層文字列**

同順序で現れる文字の割合・数により、照応とするかを判定する。ただし、他の条件で整合性が合わない場合、表層が一致しても照応とはしない。例えば、”若手の刑事 A さん”と”ベテランの刑事 B さん”は、称号や人名に矛盾が生じるため、”刑事”の文字列は一致するが、同一指示とはしない。

- **特定の関係表現**

関係抽出で得た人物表現のペアで予め定義した関係にあるものは同一指示として取り扱う。例えば、“ A は B である”等の is-a 関係は、同一指示として取り扱う。そのほか、"演じる”、”ふんする”などの俳優と役名との関係についても広い意味での同一指示として取得する。

2.4 人間関係のグラフ構造取得と相関図生成

人物表現と関係抽出により得られた(人、関係、人)の三つ組を照応解析の結果により統合する。照応関係にある人物のペアを同一のノードにまとめ上げ、人物表現間の関係をエッジとしてノードを結ぶことにより、線の関係だった人間関係を、ネットワーク状のグラフ構造にすることができる。得られたグラフ構造を可視化するため、人物表現をノード、関係をエッジとする相関図を描画する。

3. 人物相関図生成システム

3.1 システム構成

実装したシステムの特徴を下記に示す。

- 固有表現抽出、照応解析、関係抽出の開発データとして、番組概要（映画 207 本、連続ドラマ 8 シリーズ）に対し、ENE から人物表現タグ 6 種と、一般人物表現(GENERAL_NOUN)、関係人物表現(RELATIONSHIP)のタグを付与したものを用いた。
- 形態素解析器に Chasen[2]、構文解析器に Cabocha[3] を用いた。
- 人物表現抽出では、学習に Conditional Random Fields (CRF)を用い、開発データからモデルを作成した。学習の素性には、形態素の、読み、品詞、文字種、EDR の概念など 57 種を用いた。
- 照応解析器は、2.3 で説明した条件に基づき、ルールベースで作成した。
- 性別の自動判定には、Support Vector Machine による判別器を用いた。学習には、放送から取得した 2 万人の人名に対し性別(Male, Female, Unknown)を付与したデータを用いた。
- 人間関係を表すグラフの形式には、Graphviz[4] のフォーマット dot を利用し、独自の相関図表示ソフトウエアを開発した。

3.2 処理例

図2の番組概要から、相関図が生成される過程を説明する。まず、番組概要から人物表現抽出に基づき、以下の人物表現の集合 P が抽出される。

P={B・バルドー,首相,一人娘,ブリジット,令嬢,官房長,ミシェル,ブリジット,ミシェル,彼,彼女国賓,王子,王子,2枚目スター,シャルル・ボワイエ}

各文を構文解析し、構文木上で人物表現の間にある関係を抽出した結果を表1に示す。RELATION は、PERSON1 と PERSON2 に対して抽出された関係表現である。

表２は、照応解析の結果、人物表現の集合 P から得られた同一指示の関係にある表現のペアを示している。

表 1 の人物間の関係と表 2 の照応関係から得られた人物相関図を図4に示す。自動生成した相関図では、人物の性別により、ノードの色(女性:桃、男性:青)を変更している。また、“演じる”や “役の”という関係で照応と判断されている B・バルドーやシャルル・ボワイエなどの俳優名は登場人物名と区別するため、赤字で示している。照応解析の動作を例示するため、代名詞(彼、彼女)もノードに表示する。

4. 番組概要からの人物相関図の生成実験

システムで自動生成した相関図の性能を測るため、NHK で実際に放送された、開発データに含まれていない映画 79 本を対象に評価を行った。このデータに対して、人物表現のタグ、照応関係、相関図を人手で作成した。なお、評価用の相関図

表1 人物表現間の関係抽出

Person1	Person2	Relation
令嬢	B・バルドー	ふんした
首相	ブリジット	一人娘
ミシェル	ブリジット	恋焦がれる
ブリジット	ミシェル	結婚
彼	彼女	女性関係にうんざり
彼女	王子	浮気を宣言
王子	シャルル・ボワイエ	役は

表2 照応解析結果

人物表現ペア		人物表現ペア	
B・バルドー	令嬢	ブリジット	彼女
一人娘	ブリジット	ミッシェル	ミッシェル
ブリジット	令嬢	王子	王子
官房長	ミシェル	2枚目スター	シャルル・ボワイエ
ブリジット	ブリジット	国賓	王子
ミシェル	彼		

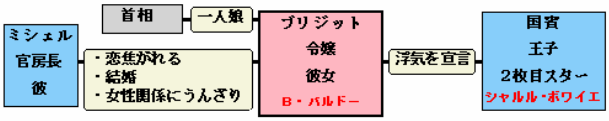

図4 相関図生成結果

[2] http://chasen.naist.jp/hiki/ChaSen/
[3] http://chasen.org/~taku/software/cabocha/
[4] http://www.graphviz.org/

は、概要の語彙のみを用いて作成した。相関図の評価は、人物表現抽出の結果が与えられた条件で行った。人物表現抽出の評価は別に実施した。

(1) 相関図評価

人手で作成した相関図 C_H と自動作成した相関図 C_S の比較により評価を行った。C_H と C_S の両相関図のノードに含まれる人物表現の一致数と、C_H の人物表現の総数との比をノード一致率と定義する。同様に両相関図のエッジに含まれる関係表現の一致数と、C_H における関係数の総数との比をエッジ一致率と定義する。表3に実験結果を示す。ノードの一致率は照応解析の精度に影響を受け、エッジの一致率は、構文解析の精度に影響を受けている。

実験の結果、ノードの一致率は 0.867 であった。人物に限定した照応解析が比較的上手く動作したためであると考えられる。各処理についての評価結果を表4に示す。特定関係に基づく照応の再現率が低い。これは、開発データの規模が十分でないため、照応とすべき関係を網羅できていなかったためと考えられる。また、連体詞の照応で、センタリング理論による格の優先度よりも、意味の近い先行詞を取るべきである事例があった。「**エディは、プロモーターから新人ボクサーを売り出すための依頼を受ける。その選手は・・・・**」という文章で、現行システムは、照応詞"その選手"は、前文の主題である"エディ"を照応先としてしまう。"ボクサー"を照応先に優先したい場合、「選手＝スポーツを行う人」「ボクシング∈スポーツ」「ボクサー＝ボクシングを行う人」などの事前知識が必要である。今後、人物表現のなかでも、より細かな意味を捕捉できる仕組みを検討する必要がある。

人物間の関係のエッジの一致率は 0.717 であった。本システムでは、構文解析の精度向上のために簡易な複文の分割処理を導入しているが、まだ網羅性は十分ではない。そのため、失敗例では、複文分割の対象から漏れた並列構造の文で、構文解析結果が誤っているものが多数含まれていた。そのほか、関係抽出において以下の問題点が残されている。

- "2 人は愛し合う"等の人数表現が動作主の場合、その動作を 2 人の関係として抽出できない。
- "恋人は・・"のような関係人物表現の対象が明示的に示されない時、現状では所有格の補完をしていないため、誰の恋人かを判定できず関係を獲得できない。
- 目的語がない場合も前文から補完を行っていないため、1 文に人物表現が1人しか現れず関係を抽出できない。

エッジの一致率を向上させるには、これらの問題を解決する処理を検討する必要がある。

(2) 人物表現抽出評価

システムをトータルで動作させるためには、人物表現の抽出の精度が重要となる。本システムでは、代名詞以外については、機械学習で人物表現抽出を行っている。人物表現抽出の結果を表 5 に示す。

結果では、一般人物表現（GENERAL_NOUN）と称号（TITLE）に、多くの誤分類があった。これは前後に現れる単語や構成される単語が類似しているためと考えられる。両表現は人間でも分類に迷う場合があるため、本タスクではタグの統合を検討する。また組織名や生物名の精度が低いのは、学習データの規模が十分でないことや、EDR の分類とテストデータの文中に出現した表現が一致しないものがあり、意味素性の効果があまり得られなかったためと考えられる。

固有表現抽出において、称号、生物、地理的政治的エンティティ（GPE)は、学習データを増やしたり、静的な知識を用意することにより精度を比較的容易に向上させられると考えられる。しかし、人名については、既に精度が高い上に、地名や組織名（ORGANIZATION）と表層上は似ていることがあるため、取り違えることが多く、本稿で用いた前後の形態素の素性だけでは、改善することは難しい。そのため、格フレーム[7]などを利用し、人名が取りうる動詞とその格情報などの知識を素性として導入することが必要である。

表3 相関図生成結果

	一致数	総数	一致率
ノード	442	509	0.867
エッジ	147	205	0.717

表4 照応解析結果

	再現率	適合率	F 値
照応詞	0.878	0.818	0.847
ノ格、同節	0.988	0.976	0.982
表層文字列	0.929	0.987	0.957
特定関係	0.645	0.909	0.755

表5 人物表現抽出の結果

	再現率	適合率	F 値
PERSON	0.910	0.928	0.919
TITLE	0.798	0.736	0.766
ORGANIZATION	0.533	0.980	0.690
GPE	0.937	0.843	0.887
LIVING_THING	0.833	0.714	0.769
N_PERSON	1.000	1.000	1.000
GENERAL_NOUN	0.759	0.787	0.773
RELATIONSHIP	0.869	0.935	0.901
TOTAL	0.827	0.935	0.878

5. まとめ

本稿では、自然言語で記述された番組概要から相関図を自動生成する手法について述べた。その評価実験結果、人手で作成した相関図と比較して人物表現の一致率 0.867、人物間の関係の一致率 0.717 が得られた。今後は、大規模なコーパスから作成した事前知識を用いることで、意味的な尺度を取り入れ精度向上を図る予定である。

また、人間関係は、連続ドラマ等では話が進むにつれて変化する場合があるため、複数回の番組内容と統合し、動的な関係の変化を捕捉する仕組みについても検討する。

REFERENCES

[1] J. Rankin, K. Harwood and P. Mirenda: "Influence of graphic symbol use on reading comprehension," Augmentative & Alternative Communication, Vol.10, No. 4, pp.269-281 (1994).

[2] T. Kato, M. Matsushita and Y. Kando: "MuST: A Workshop on Multimodal Summarization for Trend Information," Proc. of NTCIR-5, pp.556-563 (2005).

[3] Y. Matsuo and Y. Yasuda: "An Analysis of Researcher Network Evolution on the Web," Proc. of ANDI'05 (2005).

[4] 馬場, 藤井:"小説テキストを対象とした人物情報の抽出と体系化", 言語処理学会第 14 回年次大会, pp.574-577 (2007).

[5] 金, 江原:"日英機械翻訳のための日本語長文自動短文分割と主語の補完", 情報処理学会論文誌, Vol.35, No.6, pp.1018-1028 (1994).

[6] M. Walker, M. Iida, and S. Cote: "Japanese discourse and the process of centering, Computational Linguistics," Vol. 20, No. 2, pp.193–233 (1994).

[7] 河原, 黒橋:"高性能計算環境を用いた Web からの大規模格フレーム構築", 情報処理学会 自然言語処理研究会 171-12, pp.67-73 (2006).