E-049

番組検索のための登場人物の関係抽出

Extraction of Human Relationships for Program Retrieval

後藤淳† 関根聡‡ 八木伸行†
Jun Goto Satoshi Sekine Nobuyuki Yagi

1. まえがき

誰もが便利にデジタル放送を楽しめるように、テレビの複雑な操作を視聴者の代わりに行うTVエージェントの研究を行っている[1]。これまでに、テレビの基本操作のほかに、番組に関する質問応答機能や、番組名、出演者名、ジャンルなどのメタデータを利用した番組検索機能等を開発した。

ドラマや映画などの番組で、内容の類似する番組検索を行う場合、キーワードや既存のジャンル情報を用いるだけでは、類似する番組を探すことは難しい。番組のストーリーを考慮した検索を行うためには、その内容を示す様々な特徴量の抽出が必要である。本研究では、ドラマや映画の内容を示す特徴量の一つに、登場人物の人間関係があると考え、番組からの登場人物の関係を抽出し、その関係の類似度に基づく検索方法を検討した。

2. 人間関係の抽出

デジタル放送では、番組の紹介をテキストで記述した番組情報が付加されている。なかでも、映画やドラマなどのジャンルには、登場人物の関係などのストーリーを簡潔に記述している番組概要が含まれている。そこで、概要の各文から登場人物の個々の関係を抽出し統合することにより、番組全体の人間関係を表すデータ(関係グラフ)を生成する。関係グラフ生成の処理フローを図1に示し、以下に人間関係抽出までの処理内容を述べる。

2.1 人物表現抽出

番組概要から関係グラフを生成するには、その関係の基点となる登場人物を表す情報を特定する必要がある。番組概要では、人物を示す表現として、人名だけでなく、職業名や代名詞などを用いる場合がある(図2)。また、組織や動物を擬人化して人物のように扱うこともある。本稿では、人名(ミシェル等)などの固有表現については、機械学習を用いた抽出を行い、固有表現に含まれない一般名詞(男、女性等)や代名詞(彼、彼女等)については、人手で作成した辞書を用いて抽出する。

固有表現の抽出には、Conditional Random Field(CRF)を使用した。識別するラベルとして、関根らが提案している拡張固有表現[2]から、登場人物を表す固有表現タグ(人名、人数、称号[職業含]、生物名、組織名、GPE)を選択した。学習の素性には、語彙的な素性、意味的な素性など計57種を用いた。抽出器の性能評価のため、NHK で放送された映画 406 本の番組概要に固有表現タグを付加し、10 分割交差検定を行った。その結果、表1に示すとおり、人名で 0.956、人物表現全体で 0.909 のF値が得られた。生物名や称号の再現率が低いのは、番組概要に現れる表現が少なく、語彙的な素性の学習が足りないためと推測できる。

† NHK 放送技術研究所, NHK-STRL
‡ ニューヨーク大学, NYU

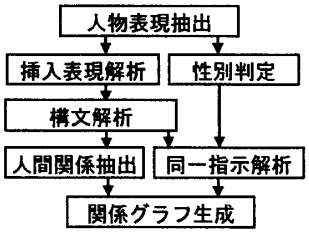

図1 関係グラフ生成の処理フロー

宜厚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。

図2 番組情報

表1 固有表現抽出結果

	Precision	Recall	F-measure
PERSON	0.957	0.956	0.956
N_PERSON	0.974	0.910	0.941
TITLE	0.932	0.659	0.772
LIVING_THING	0.923	0.429	0.586
ORGANIZATION	0.870	0.712	0.783
GPE	0.894	0.894	0.894
TOTAL	0.947	0.874	0.909

2.2 挿入表現からの人間関係抽出

番組概要には、丸括弧などを用いた挿入表現が文章内に多く出現する。挿入表現は、構文解析に悪影響を及ぼす反面、直前の情報と関係のある有効な情報が含んでいる。そのため、人に関係ある挿入表現を、正規表現によるパターンマッチングにより取得する。例えば、「長女のジョージア(ダイアン・キートン)は・・・」という表現からは、“PERSON(PERSON)” のパターンと、予め定義したヒューリスティックを用いることで、俳優“ダイアン・キートン”と役名“ジョージア”の関係が獲得できる。

2.3 構文解析結果からの人間関係抽出

人物表現抽出と構文解析の結果を用いて、人物表現の間にある構文木のノードを関係表現として抽出する。「***太郎が恋する花子は、次郎が好きだ***」という文を例に挙げる。構文解析の結果は図3のようになる。まず、人名表現抽出の結果から、人名を表すノード“太郎”、“花子”、“次郎”が得られる。次に、人名表現を基点に、構文部分木①から、太郎と花子の関係の“恋する”と、構文部分木②から、花子と次郎の関係の“好きだ”が得られる。太郎と次郎の関係は、その間のノードに人物表現の“花子”を含むため、①と②の関係から表現できる。

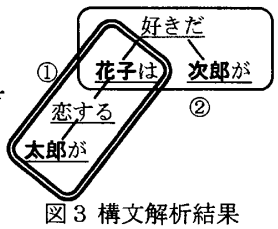

図3 構文解析結果

そのほか、人物を表すノード同士が隣り合う場合は、人物表現の同一指示解析に用いる関係が得られる。例えば、“***先生の四郎は・・***”という表現からは、“四郎＝先生”の同一指示の関係が得られる。ただし、“***四郎の先生は・・・***”は、固有表現と係り受けの順序判定により、“四郎≠先生”となる。

3.関係グラフの生成

3.1 人物表現の同一指示解析

全登場人物の人間関係を示す関係グラフを生成するためには、複数の文から導出した各々の関係を統合する必要がある。関係の統合では、複数の文に出現する異なる人物表現の照応を解析し、同一指示の表現を1つのエンティティにまとめなければならない。本稿で取り扱う同一指示解析は、関係グラフ生成が目的であるため、同一指示の候補を人物表現に限定した。これにより、一般の同一指示解析に比べ、対象とする候補を絞り込むことができる。解析に使用する素性には、語彙の一致率、係り受け関係、固有表現タグ、出現位置、代名詞の有無、人物表現の性別を用いた。この6種の素性によるルールベースの同一指示解析器を作成した。以下に、解析の素性として使用した性別の判定方法について説明する。

・性別判別

同一指示解析には、人名、代名詞などが持つ性別が重要となる場合がある。例えば、図2の番組情報で、代名詞（彼、彼女）の照応先を決めるには、ミッシェル、ブリジットや、女性、王子の性別の情報が必要である。そのため、人物表現の性別判別器を作成し、その結果を同一指示解析の素性として利用する。

人物表現のうち、人名については、機械学習を用いた性別の判別を行う。学習アルゴリズムに Support Vector Machine を用い、素性として、表層文字列、読み、文字種、文字列長、特定文字の有無を使用した。重複のない 2 万人の出演者名をデジタル放送から取得し、3種のタグ（男、女、Unknown）を人手で付加した。そのうち、1.5 万人分を学習データに、0.5 万人分を評価データに使用した。評価実験の結果、男性で 0.899、女性で 0.881 の F 値が得られた。

人名以外の人物表現の性別判定には、一般名詞 989 語（王子、女性等）と代名詞 28 語（彼、彼女等）に性別情報を加えて作成した辞書を用いる。

3.2 関係グラフと表示インタフェース

同一指示解析と関係抽出の結果を用いて、照応関係にある人物表現を1つのノードに統合し、関係グラフを生成する。関係グラフは、ノードに人物名、エッジに関係を持つラベル付グラフで表すことができる。

この関係グラフを可視化するため、Graphviz[3]を使用した表示インタフェースを開発した(図4)。エッジの数が多いときには、描画結果が煩雑となるため、便宜上、エッジ（関係）のラベルを楕円のノードとして表示している。同一指示解析の結果により統合された人物表現（例：マギー、女性、令嬢、親友、ケイト・ベッキンセール）は、固有表現の種類と、予め定義したオントロジーに基づき、リスト形式でノードに格納される。この関係グラフの描画結果は、TVエージェントにおいて番組内容を一覧するインタフェースとして利用する。

4. 関係グラフを用いた番組検索

生成した関係グラフは、ドラマや映画の内容を一意に示すメタデータと言える。そこで関係グラフの類似度を求めることにより、内容が類似する番組を探すことを検討する。

具体的には、番組の関係グラフそのものを検索キーとして使用し、キーのグラフと共通の部分グラフを持つ関係グラフを検索する。グラフの類似度は、共通部分グラフのエッジやノード数、共通部分グラフの数、などに基づき定義できるが、ここでは、エッジ数が最大となる共通部分グラフを持つ関係グラフを類似番組として抽出する。以下に検索処理の手順を示す。

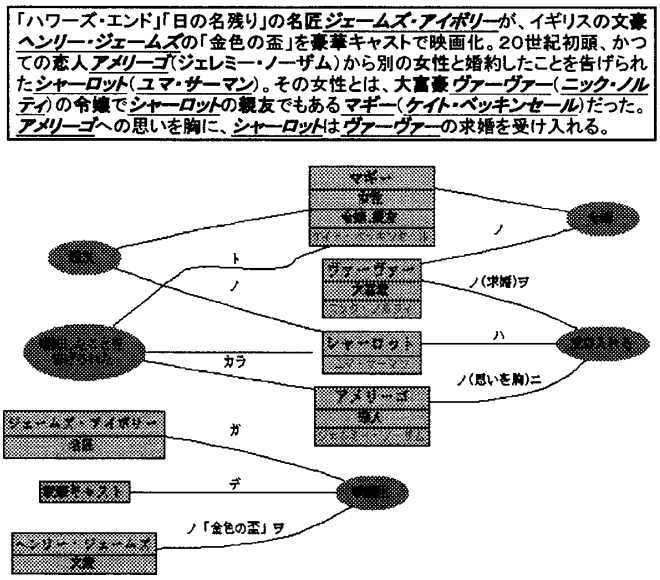

図4 関係グラフの生成

① gSpan アルゴリズム[4] により、関係グラフを一意に記述し、その部分グラフを列挙する。
② ①の処理を全ての番組に対して行い、部分グラフをデータベース（DB）に格納する。
③ 検索キーとなる関係グラフ（キーグラフ）を定める。
④ キーグラフの部分グラフを、エッジ数が多い順に選ぶ。（同数の場合はノード数の多い順に。）
⑤ ④で選んだ部分グラフがDBにあれば、その部分グラフを持つ関係グラフを類似番組として取り出す。
⑥ ⑤で結果が得られない場合、④にもどり、次の部分グラフを選ぶ。

ノード（人物）やエッジ（人間関係）の表層表現をそのままラベルに使用すると、関係グラフの種類が多くなりすぎ、共通の部分グラフを持つデータが少なくなる。上記に示した共通グラフによる検索を行うには、ノードとエッジのラベルを一定のクラスに分類し抽象化する必要がある。そのため、人物表現は、キャラクター名、職業、性別、リスト出現順序（主役度）、俳優名等の特徴をもとに、クラスタリングを行う。また、人間関係は、6つのクラス（恋愛、友人、敵対、血縁、上下、その他）に辞書ベースで分類する。分類に用いる辞書は、放送コンテンツや新聞などのコーパスから、今回開発した関係抽出手法により、人物間の関係表現を取得し作成する。

5. まとめ

本稿では、番組情報から人間関係を抽出し、関係グラフを生成する手法について述べた。また、関係グラフの類似度に基づく類似番組の検索方法を検討した。今後、関係グラフの抽象化のための関係表現辞書の構築を進める。さらに、関係グラフ生成の精度を向上させるため、人物表現のゼロ照応について検討する。

[1] Jun Goto, et al., "A Spoken Dialogue Interface for TV Operations Based on Data Collected by Using WOZ Method," IEICE Trans. INF&SYST., Vol.E87-D, No.6, 2004
[2] Satoshi Sekine, "Extended Named Entity Hierarchy," http://nlp.cs.nyu.edu/ene/
[3] John Ellson et al., "Graphviz - Open Source Graph Drawing Tools," Graph Drawing 2001, pp483-484, 2001
[4] Xifeng Yan et al., "gSpan: Graph-Based Substructure Pattern Mining," IEEE ICDM2002, 2002