7K-05

同じ出来事についての記事からの共通点と差異の抽出

皿海　宏明†　　湯本　高行‡　　新居　学‡　　上浦　尚武‡

†兵庫県立大学工学部　　‡兵庫県立大学大学院工学研究科

1. 背景・目的

同じ出来事について複数の記事があったとき，発信者ごとの意見や発信者に左右されない情報を知るためには記事の比較が必要となる．しかし，記事を見比べ，記事ごとの差異もしくは共通の内容を人手で詳細に発見することには過大な負荷がかかる．また書き手の違いによる表現のずれは単純な比較を困難にする．そこで本研究では文章間の表現のずれを吸収する意味的な比較の手法を開発する．

2. 重複部分の定義

同じ出来事を表す２つの文章のどのような部分を重複部分とするかを定義する．述語項構造の重複を重複部分とし，述語，項単体での意味的重複は重複部分としない．表１に例を示す．下線は重複部分を示す．項が重複かどうかについては，関係性と表現の観点から分類して定義する．表２に対応と例を示す．例は対応関係に続けて対応する文を（例文｜例文）と示す．下線は重複部分を表す．

表１　重複部分の例

	文１	文２
述語，項単体	男は逃走した	男は連続通り魔事件の犯人とみられている
述語項構造	男が男性の首を刺して逃走した	男性が男に首を刺されて死亡した

表２　語の関係性と表現の対応関係

	数値・日時・地域	単体	並列
類義	全て重複（１００円｜百円）	全て重複（危険ドラッグ｜危険ハーブ）	全て重複（背中と首｜背中や首）
省略	全て重複（１０万円｜１０万１０５５円）	全て重複（省エネルギー｜省エネ）	全て重複（首など｜背中や首）
含意	全て重複（兵庫県｜姫路市）	重複なし（鳥｜インコ）	重複なし（肢体に｜腕や足に）
修飾	全て重複（約１００円｜１００円）	修飾以外重複（チョコ工場｜工場）	修飾以外重複（工場｜チョコ工場や飴工場）

Extraction of Common and Different Points from Articles about Same Event

†Hiroaki Saragai, School of Engineering, University of Hyogo

‡Takayuki Yumoto, Manabu Nii and Naotake Kamiura, Graduate School of Engineering, University of Hyogo

3. 重複部分の発見

重複部分の発見では語の同定によって，候補とした語より重複語組を作成する．重複の確からしさの値として重複語組に対して対応スコアを設定する．

前処理として文章から語を抽出する．まず，記事に含まれる【文】，（文），=文=をフィルタリングにより取り除き，形態素解析器 JUMAN と日本語構文・格・照応解析器 KNP を用いて記事を解析する．解析結果より名詞，動詞，形容詞，副詞を含む連続する形態素を語として抽出する．ただし，文中によく出現し対応関係の取りにくい頻出語「する」「こと」については単体で抽出する．また文節をまたぐ場合は文節で分ける．

3.1. 文脈独立な同定

文脈独立な同定では語単体の表層的および意味的な観点から語を比較し，重複語組を発見する．具体的には以下の手順で同定を行う．

(1) 数値表現による語の同定
(2) 完全一致による語の同定
(3) 編集距離による語の同定
(4) 部分一致による語の同定
(5) 類語辞典による語の同定

まず数値表現による語の同定では，数値表現規格化器 normalizeNumexp[1]を新聞記事の表現により対応できるように拡張したものを用いて数量表現・時間表現を抽出し規格化を行う．規格化された数値表現が一致や省略で同一のものと見なせるかどうかを判定する．同定した組を重複語組とし，対応スコアを１とする．

完全一致による語の同定では，語の形態素の基本形が全て一致したものを同定する．同定した組を重複語組とし，対応スコアを１とする．ただし，頻出語を含む場合，間違った組を対応させる場合が多くなるため対応スコアを 0.4 とする．

編集距離による語の同定では，語の合計文字数で割った Damerau-Levenshtein 距離を１から引いた値を対応スコアとして重複語組を同定する．ただし，いずれかが３文字未満，もしくは対応スコアが 0.66 未満のものは同定しない．

部分一致による語の同定では，次の式で表されるコサイン距離を対応スコアとする重複語組を同定する．ただし，いずれかの形態素の数が２つ未満，もしくは対応スコアが 0.66 未満のものは同定しない．

$$対応スコア = \frac{一致した形態素数}{\sqrt{片方の形態素数 \times もう片方の形態素数}} \quad (1)$$

類語辞典による語の同定では，語の品詞が同じとき，名詞であれば Wikipedia 辞書，他は日本語ワードネットおよび ALAGIN Forum で公開されている含意関係辞書を用いて関連があれば同定する．同定した組を重複語組とし，対応スコアを 0.8 とする．

3.2. 文脈依存な重複語組の絞込み

次の手順で重複語組を位置や文脈，対応スコアから確からしいものに絞り込む．

(1) 1 対 1 対応による語の選定
(2) 係り受けによる語の選定

1 対 1 対応による選定では，共通の語を含む重複語組の対応スコアに 0.1 以上の開きがあれば明確な差があるとして，対応スコアの低いものを削除する．そうでない場合は共通の語の近傍にあるスコア 0.8 以上の最大 4 つの重複語組を基準組とする．次に共通の語を含まない側の文章を見て，それぞれの組の共通でない語と選択した基準組との符号つきの位置の差をそれぞれ求める．この位置の差を共通でない語ごとに合算し相対位置とする．同様に求めた共通の語の相対位置との差の絶対値より相対距離を求め，遠いものを削除する．例を図 1 に示す．例では重複語組 A の相対距離が 0，B が 10 となるので，遠いほうである B を削除する．

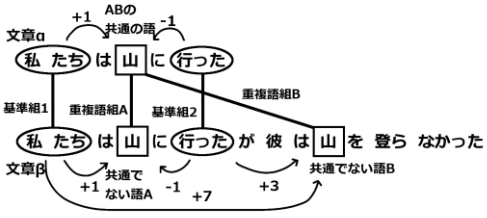

図 1　1 対 1 対応の例

係り受けによる選定では，他の重複語組の両方ともの語と間接的に係り受け関係にない重複語組が動詞を含めば，不自然であるとして削除する．例を図 2 に示す．例では重複語組と係り受け関係にある重複語組がないため重複語組を削除する．

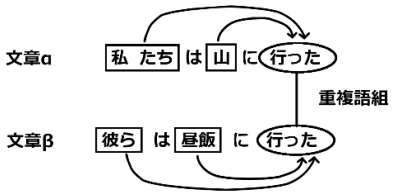

図 2　グループ化の例

3.3. 文脈依存な同定

文脈依存な同定では次の手順で前述の重複語組を基点とした重複語組の語の文脈を考慮した同定を行う．

(1) 係り受けによる語の同定
(2) 表層格フレームによる語の同定

まず係り受けによる語の同定では，2 つ以上の重複語組から係っている語の組み合わせを同定する．同定した組を重複語組とし，係る重複語組の数が 2 ならば対応スコアを 0.8,それ以上なら 1 とする．

表層格フレームによる語の同定では対応する表層格フレームの組から重複語組を発見する，対応スコア 0.6 以上の重複語組が述語または項で 2 つ以上ある表層格フレーム組を対象とする．これらの組より一致する格の項または述語どうしを同定する．同定した組を重複語組とし，対応スコアを 0.8 とする．例を図 3 に示す．例では表層格フレームα，βはガ格，述語が対応スコア 0.6 以上の重複語組であるので表層格フレーム組とする．この組では 2 格が一致しているので 2 格の語を対応スコア 0.8 の重複語組とする．

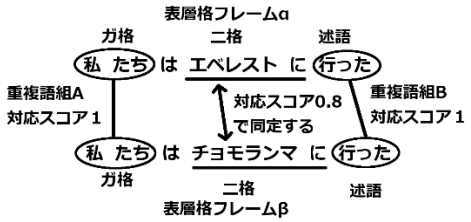

図 3　表層格フレームによる同定の例

3.4. 重複部分の決定

ここまで発見した重複語組を文章の重複部分とし，重複語組に含まれない語を差異部分とする．またこの重複部分と差異部分を色分けをして表示するアプリケーションを開発した．

4. 実験

評価には「CD-毎日新聞データ集 2014 年版」と 2014 年版の「読売新聞記事データ」の同じ出来事を示したリード 50 組を用いた．

手法より重複部分を求めた結果と，人手により求めた正解をそれぞれ，インデクシングした語が完全に重複部分に含まれるものを数え，提案手法の結果での数，正解での数，両者に含まれる数より適合率，再現率，F 値を求め評価指標とする．またベースラインとして，文章全てを重複とした場合と完全一致のみ用いた場合について評価値を算出したものを用いた．結果を表 3 に示す．表 3 より提案手法では完全一致より適合率は低いが再現率は高い，また全て重複と比べても F 値が高いことより，表現にずれのある語の同定が適切に行えていると考える．

表3　実験結果

	適合率	再現率	F 値
提案手法	0.868	0.788	0.820
全て重複	0.612	1.000	0.744
完全一致	0.878	0.642	0.732

謝辞　本研究の一部は，平成 27 年度科研費若手研究(B)「情報の詳細関係に基づく Web ページの組織化」（課題番号：24700097）によるものである．

参考文献

[1]normalizeNumexp http://www.cl.ecei.tohoku.ac.jp/~katsum a/software/normalizeNumexp/