格関係の比較を用いた複数テキスト間の重複・差分の検出

成松 深[†]　河原 大輔[‡]　黒橋 禎夫[††]　西田 豊明[††]

[†]東京大学工学部　[‡]京都大学大学院情報学研究科　[††]東京大学大学院情報理工学系研究科

{narimatu, kawahara, kuro, nishida}@kc.t.u-tokyo.ac.jp

1. はじめに

言語は情報を伝達する上で最も有効な手段であるが、同じ内容を様々に表現できるという自由度をもっている。これは、微妙なニュアンスを伝えたり表現を豊かにするという言語の良い性質でもあるが、計算機による言語処理・情報伝達という面では大きな障害である。

本研究では複数の関連するテキストを対象としてこの問題を扱う。たとえば一方のテキストで「2人組は刃物を持っていた」、もう一方で「犯人らは刃物を持っていた」となっていれば、ここでは「2人組」と「犯人ら」がほぼ同じ意味で使われていることがわかる。このような問題を正しく認識できれば、テキスト間の重複と差分を判別し、それに基づいて、複数文章要約で重複情報を重要と考えたり、関連文章を提示する際に差分情報を強調することができる。

本研究では、このような「ほぼ同じ意味で使われている」ことを格関係に基づき判断する[*1]。すなわち、上の例では「～が刃物を持つ」が同じことから「2人組」「犯人ら」が同じだろうと推測するのである。

このような処理は表層的な構文解析だけではできない。文章中には多数の省略があり、また、提題助詞や連体修飾によって表層格がかくれることも多いからである。このような格・省略解析は、大規模コーパスから格フレームを自動構築する河原らの方法[4,5]によって、徐々に可能となってきており、本研究ではその成果を利用する。

本システムの入力は、互いに関連する2つの文章とする。このうち重複・差分を判別したい文章を判別対象文章、その比較対象を比較先文章とよぶ。入力文章は、JUMAN・KNPによって、形態素・構文・格・省略解析を行う。重複・差分の判別単位は文節である。重複・差分の判別には形態素の比較および用言の格情報の比較という2つの手法を用いる。

[*1] 本稿では「刃物を持つ」のような用言と項の関係を格関係、「刃物」をその格要素と呼ぶ。

2. 重複情報と差分情報

本研究での「重複情報」とは、その文節が持つ内容が比較先の文章に存在している文節のことを示す。

(1) 同表現で同じ意味を表す場合

「ガードマンの」⇔「ガードマンの」
:文節内の形態素の完全一致

「奪い」⇔「奪って」
:活用が異なっている同じ用言

(2) 異表現で同じ意味を表す場合(表現のずれ)

「集金作業を」⇔「集金の」「作業を」
:文章区切り等の構造的ずれ

「ピストル」⇔「拳銃」
:文脈独立の同義表現

「2人組は」⇔「男らは」
:文脈に依存する同義表現

「小平市で」⇔「同市では」
:繰り返しを避けるための代用表現

一方、本研究での「差分情報」とは、その文節の内容が比較先の文章に存在しない文節のことを示す。

(1) 追加:比較先の文章には全く登場しない内容が新たに現れた場合

「犯人は徒歩で逃走した」⇔「犯人は逃走した」
この場合は「徒歩で」が追加情報となる。

(2) 相違:比較先の文章で触れられた内容について、それを取り消して異なる内容が現れた場合

「犯人は徒歩で逃走した」
⇔「犯人は車で逃走した」

「2000人が集会に参加した」
⇔「3000人が集会に参加した」
この場合は「車で」「3000人が」に対してそれぞれ「徒歩で」「2000人が」が相違情報となる。

(3) 詳細度のずれ:比較先の文章で触れられた内容と詳細度の異なる内容が現れた場合

「犯人はセダンで逃走した」
⇔「犯人は車で逃走した」

「2017人が集会に参加した」
⇔「約2000人が集会に参加した」

この場合は「車で」「約2000人が」に対してそれぞれ「セダンで」「2017人が」が詳細化情報となり、「セダンが」「2017人が」に対してそれぞれ「車で」「約2000人が」が一般化情報となる。

3. 形態素の比較による重複情報の判別

判別文節中の自立語形態素について、比較先文章の出現形態素リストと比較する。文節中の自立語形態素全てが比較先文章の出現形態素リスト上で一定の距離(現在は5文節)以内に出現している場合に、この文節を重複情報と判別する。形態素一致は、語の一致でも、シソーラスにおける一致(NTT 日本語語彙大系[6]の意味素性が同じ)でもよいとした。

例えば、「集金作業を」という文節では、自立語形態素「集金」「作業」が比較先文章の形態素リストで5文節以内にあれば重複情報とみなす。

4. 格情報の比較による重複情報の判別

形態素の比較による重複情報の判別では「表現のずれ」、特に文脈に依存する同義表現の判別はできない。そこで、格情報の比較を用いて重複情報の判別を行う。この処理では差分情報のうち「相違」「詳細度のずれ」を重複情報と誤解析する可能性があるが、実際の関連文章中では「相違」「詳細度のずれ」の確率は低く、本稿では対象外とする。

4.1 格関係の比較による体言の同義推定

判別対象文章の格情報リスト上にあるそれぞれの用言について、比較先文章の格情報リスト上にあるものと比較していく。この際、

> ある用言が比較先文章の格情報リスト上にもある場合にそれぞれの格情報を比較する。そして1つ以上の格関係が一致している場合、おなじ表層格を持つ格要素同士を同義語と推定する。

格要素の一致は、複合名詞は一語とした上で、文字列としての完全一致・部分一致、シソーラスにおける一致のいずれでもよいとした。

例えば判別対象文章中に「逃走する 2人組:ガ 乗用車:デ」、比較先文章中に「逃走する 犯人ら:ガ 乗用車:デ」という格情報があれば、に用言「逃走する」が同じで格関係「乗用車:デ 逃走する」が一致するのでガ格の「2人組」と「犯人ら」が同義語と推定さ..れ、この句を含む文節は重複情報とされる。

同義語関係にあるとされる句(形態素列)は同義語リストに記載され、格情報の比較(全体で2回行う)を以降でおこなう際にも用いられる。

4.2 格関係の比較による用言の重複推定

判別を行おうとする文章の格情報リスト上にあるそれぞれの用言の格情報について、比較先文章の格情報リスト上のものと比較していく。このとき、

> 互いに異なる用言の、それらの持つ格要素と表層格の組が2つ以上一致している場合にそれらの用言が同じ意味を表していると推定し、重複情報と判別する。

格要素の一致は完全一致、部分一致、シソーラスにおける一致のいずれでもよいとした。

たとえば、「乗用車を逃走に使う」と「乗用車を逃走に用いる」の場合、異なる用言「使う」「用いる」について「逃走:ニ」「乗用車:ヲ」が一致しているため「使う」と「用いる」が同じ意味を表しているとみなし、重複情報とされる。

5. 出力と具体例

それぞれの文節にはその判別結果によって 4.1 節又は 4.2 節の条件に相当すれば「重複」、そうでなければ「差分」の属性がタグ情報として与えられる。ただし、形式的な文節(接続詞・副詞・指示詞・形式名詞・副詞的名詞および汎用度の高い動詞(「ある」「する」等)を含む文節)はこの処理の対象外とした。

本システムは比較先文章をユーザがすでに読んだ文章、判別対象文章をユーザが新たに読む文章であると想定し、ユーザの新規情報の獲得効率を高める目的で差分情報の強調表示を行う。ただし、1文中で差分情報とされた文節の割合が一定値(現在は0.65)を超えている場合は、その文全体を差分情報として表示している。

システムの処理の具体例を図2にあげる。

6. 実験と考察

6.1 実験データ

互いに関連する新聞記事を10組用いてシステムの評価実験を行った。1つの組は2文章からなり、一方の文章を重複・差分を判別する文章、他方をその比較先の文章とし、その方向の処理と逆方向の処理の

[判別対象文章]
〔原文〕
JA花小金井支店で、現金輸送車が拳銃を持った2人組に襲われました。犯人らは現金を奪うと青梅街道を車で逃走、けん銃を持ったまま行方が分かっていません。3日午後3時40分頃、小平市花小金井のJA花小金井支店で、集金作業をしていたガードマンに、2人組の男が後ろから近づきました。男らは片言の日本語で「金を全部出せ」と言うと、いきなりけん銃を発砲しました。男らは輸送車に積み込まれたおよそ1億3200万円を奪って逃走しています。

〔格情報〕
持つ　2人組:ガ 拳銃:ヲ
襲う　JA花小金井支店:デ 現金輸送車:ガ 2人組:ニ
奪う　現金:ヲ 2人組:ガ
逃走する　犯人ら:ガ 青梅街道:ヲ 車:デ
持つ　けん銃:ヲ 犯人ら:ガ
分かる　犯人ら:ガ2 行方:ガ
する　ガードマン:ガ 集金作業:ヲ
近づく　40分頃:時間 JA花小金井支店:デ ガードマン:ニ 男:ガ 後ろ:カラ
出す　男ら:ガ 2人組:ニ 金:ヲ 全部:無格
言う　男ら:ガ 日本語:デ 出せ:ト
発砲する　男ら:ガ けん銃:ヲ
積み込む　1億3200万円:ガ 輸送車:ニ
奪う　2人組:ガ 1億3200万円:ヲ
逃走する　男ら:ガ

[比較先文章]
〔原文〕
東京・小平市で3日午後、現金輸送車が2人組の男に襲われ、現金などおよそ1億700万円が奪われた。午後3時35分ごろ、小平市のJA東京むさしの花小金井支店前で、ガードマンが集金を終えて現金輸送車に戻ろうとしたところ、ピストルを持った2人組の男に「金を全部出せ」などと片言の日本語で脅された。2人組は、ガードマン2人を車の荷台部分に押し込み、ピストル3発を発射し、車の中にあった現金およそ1億700万円を奪って乗用車で逃走した。

〔格情報〕
襲う　東京・小平市:デ 午後:時間 現金輸送車:ガ 男:ニ
奪う　男:ニ 1億700万円:ガ
終える　ガードマン:ガ 集金:ヲ
戻る　35分ごろ:時間 花小金井支店前:デ 現金輸送車:ニ ガードマン:ガ
持つ　2人組:ガ ピストル:ヲ
出す　男:ガ 1億700万円:ニ 金:ヲ 全部:無格
脅す　男:ニ 出せ:ト 日本語:デ
押し込む　2人組:ガ ガードマン2人:ヲ 荷台部分:ニ
発射する　ピストル3発:ヲ 2人組:ガ
ある　現金:ガ 中:ニ
奪う　2人組:ガ 1億700万円:ヲ
逃走する　2人組:ガ 乗用車:デ

(注)下線は双方に出現する用言

[同義と推定されたもの]
ピストル＝拳銃＝けん銃
東京・小平市＝JA花小金井支店
男＝2人組＝犯人ら＝男ら
1億700万円＝現金　乗用車＝車
言う => 脅す 発砲 => 発射

[結果]　:差分情報が強調表示されている
JA花小金井支店で、現金輸送車が拳銃を持った2人組に襲われました。犯人らは現金を奪うと【青梅街道を】車で逃走、けん銃を持ったまま【行方が分かっていません。】3日午後3時【40分頃、】小平市花小金井のJA花小金井支店で、【集金作業を】していたガードマンに、2人組の男が【後ろから近づきました。】男らは片言の日本語で「金を全部出せ」と言うと、いきなりけん銃を発砲しました。男らは輸送車に【積み込まれたおよそ1億3200万円を】奪って逃走しています。

図2 処理の具体例

双方で評価を行った。試行には前もって筆者による「正解」を文節ごとに作成しておき、システムの出力と比較した。正解の分布は表1のようになった。

6. 2 実験結果

正解が「同表現重複」「形式的文節」のものを除いた文節を対象として、「異表現重複」文節(表現のずれが起こっている文節)についてその適合率・再現率を「形態素の比較のみによる重複・差分の判別」と「形態素の比較と格関係の比較双方による重複・差分の判別」の2つの場合で調べた。結果を表2に示す。

表2のように、格情報の比較を用いることで再現率が2.1ポイント上昇しているものの59.2%にとどまっている。表現のずれを解析できなかった40文節中、用言は15文節、体言は25文節であった(解析に成功した文節は全て体言であった)。

表2に示したように、格関係を調べることで新たに(形態素の比較に加えて)見つかったものは2例であるが、形態素の比較でシソーラス等で発見されたもので、格関係によっても発見されたものもある。そこで、格情報の比較を用いた判別により、異なる単語で同じ意味を表すと推定された格要素(体言)および用言全体についてその正誤を調べた結果を表3に示す。

表1 文節ごとの正解の内訳

同表現による重複(同表現重複)	475 文節	26.7%
異表現による重複(異表現重複)	102 文節	5.7%
差分情報	958 文節	53.8%
形式的文節	247 文節	13.9%

表2「異語重複」文節の再現率

	適合率	再現率
形態素の比較のみ	51.9% (56/108)	57.1% (56/98)
形態素の比較＋格関係の比較	52.7% (58/110)	59.2% (58/98)

表3 格関係比較による同義推定の内訳

同じ意味を表していたもの	体言 21
同じ意味を表さないもの	体言 10,用言 4

6.3 考察

評価実験から、格関係の比較を用いた判別は複数文章間の「表現のずれ」の解決に寄与しているといえるが、解決できていない箇所も多く残っているため、さらに種々の解析を精巧化する必要がある。

解析失敗のうち約3割は格・省略解析のミスによるものであった。それ以外の本研究のアルゴリズム固有の問題について失敗例を以下に示し考察する。

表現のずれによる重複情報であることを検出できなかったものには、以下のような例がある。この例では「セスナ機」と「小型機」の表現のずれを判別できなかった。

(例1) 墜落する セスナ機:ガ (白岩山の)西側斜面:ニ
⇔ 墜落する 小型機:ガ 山中:ニ

また、差分情報を重複と誤解析した例には次のようなものがある。

(例2) 出航する はまな:ガ 情報収集目的:デ 9日:ニ
(「はまな」は護衛艦2隻とともに情報収集目的で9日に出航したが…)
⇔ 出航する 3隻:ガ 9日:ニ
(情報収集のため9日に出航した3隻のうちの補給艦「はまな」が…)
この場合「はまな」と「3隻」を同義語と判別してしまう。

(例3) 反対する 漁民ら:ガ 干拓工事:ニ
⇔ 賛成する 漁民ら:ガ 干拓工事:ニ
この場合、「反対する」と「賛成する」を同じ意味と推定して重複情報としてしまう。

これらの問題の解決法として、「白岩山の西側斜面」「3隻のうちの補給艦『はまな』」のように格要素をより大きな単位で参照することが考えられる。また、用言の重複判定には対義・類義辞書などの別の知識を併用して判別を行ったほうがよいと考えられる。

また別の問題として、現在は処理の単位を文節としているが、そのために次のような場合にすべて重複情報となってしまう。

(例4)「A社とB社はXの問題を抱え、A社は倒産した。」
⇔ 「A社とB社はXの問題を抱え、A社に続いてB社も倒産した。」

格関係の比較をより重要視して、文節ではなく格関係の単位で差分情報と扱う必要がある。

7. おわりに

同じ話題について書かれた複数文章を対象とした重複・差分情報の判別手法について提案した。

用言を中心とする格情報の比較を用いて重複・差分情報の判別をすることにより、単語の比較のみでは解決できない「表現のずれ」の解決を目指した。

システムの評価では、提案手法である格関係の比較を用いた重複・差分の判別が表現のずれの解決に寄与していることがわかったが、その効用は低くさらに種々の解析を精巧化する必要がある。

参考文献

[1] 上田良寛,小山剛弘. 共通意味断片の抽出による複数文書要約. 言語処理学会第6回年次大会 発表論文集, pp.360-363, 2000.

[2] 山本和英,増山繁,内藤昭三. 関連テキストを利用した重複表現削減による要約. 電子情報通信学会論文誌, '96/11, Vol.J79-D- II, No.11, pp.1968-1972, 1996

[3] 黒橋禎夫,長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol.1, No.1, 1994

[4] 河原大輔,黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol.9, No.1, pp.3-19, 2002

[5] 河原大輔,黒橋禎夫. 自動構築された格フレーム辞書に基づく省略解析. 言語処理学会 第7回年次大会, pp.498-501, 2001

[6] NTTコミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997