

複数の新聞を使用した言い替え表現の自動抽出

関根 聡

ニューヨーク大学

sekine@cs.nyu.edu

概要

自然言語は柔軟であり、ひとつの事柄や出来事もいく通りもの言い方（言い替え表現）ができる。この柔軟さこそが、自然言語処理の応用を難しくしている最大の原因と捉えることもできる。そして、この問題の一番大きなボトルネックは、そのような言い替え表現を収集することの困難さにあると考えられる。本論文では、同じイベントを報道している複数の違った新聞社の記事から同じ事柄を表している言い替え表現を自動的に抽出する方法を提案する。同じ日の複数の新聞記事から同一内容の記事を発見し、同一内容の文を発見し、それらの文の構文解析結果から同一表現部分を抽出するというカスケード方式を採用した。これらの抽出過程では組織名、人名などの固有表現が同一性を判定する重要な鍵として使用されている。1ヶ月分の5つの英語の新聞を使用した実験では、1000程度の言い替え表現候補が抽出できた。

1 はじめに

自然言語は柔軟であり、ひとつの事柄や出来事もいく通りもの言い方ができる。この柔軟さこそが、自然言語処理の応用を難しくしている最大の原因と捉えることもできる。言い替え表現には、能動形や受動形といった文法的なバリエーションや名詞化、類義語や類義表現、そしてもっと複雑な文単位での言い替え表現などがある。自然言語処理応用システムが文章を理解するためには、そのように違った表現であらわされたものが実は同じ意味であると言うことを理解できる能力をなんらかの形で持たなければならない。

図1は、2001年2月21日に米国の新聞社などのWebページで報道されたハワイ沖での米国の潜水艦と日本の船舶の衝突事故において、潜水艦に搭乗した民間人が操作の邪魔になっていた可能性があると言うことを報道している記事で中心となっている文である。読みやすくするように、核ではない部分はイタリックで示してある。例を見て判るように、ひとつの出来事に対していく通りもの違った表現ができる。特にこの場合には、すべての

記事は同一の記者会見に基づいているはずであり、そこでの発言は唯一と考えられるのである。詳しく見ていると言い替えとして面白い現象が多々見られる。まず、USA Today 以外では、「<PERSON> said that he ...」という表現をしているのに対し、USA Today では断定的に表現されている。また、<PERSON> (軍人) に対しては、“a sailor”, “the sonar plotter”, “the crew member”, “a crew member”, “a crewman” とすべての新聞で違った表現が用いられている。また、「邪魔された」という表現には、「存在」という意味の動詞的表現の “were in his way”, “presence”, 名詞表現の “distraction”, そして受身形の “was destructed” という表現が2つと異なっている。同様に、「仕事を中止した」という表現には、“had to halt his task”, “was unable to finish his job”, “stopped performing that task”, “stopped performing the task”, “halted his work” といった表現が用いられている。また、4つの文では、「仕事を中止した、なぜなら民間人が邪魔したからだ。」という形式で表現されているのに対して、他の1つでは「民間人が邪魔をして、そして仕事を中止した。」という表現になっている。このように本来、インタビューされた人の1つの表現に過ぎなかったと考えられるものが非常に幅広いバリエーションを持った形で表現されており、現在あるシソーラス、類義表現辞典、類義語辞典などではカバーしきれないような奥深さがあることが伺われる。

本論文では、このような言い替え表現をコーパスから自動的に収集する方法を提案する。コーパスには例にあるような複数の新聞記事を逆に利用する。基本的なアイデアは、図1で例に挙げたような言い替え表現を同じ日の同じ事柄を報道している記事から自動的に収集するというものである。そして、その収集の際には組織名、人名、地名などの固有表現をキーとして利用する。固有表現というのは、どのような新聞であれ、ほぼバリエーションのないものであり、そのような確かな部分を手かりにバリエーションのある部分を収集するという方法は理にかなっていると考えられる。

[New York Times]

A sailor assigned to plot sonar contacts aboard the United States Navy submarine that hit and sank a Japanese fishing vessel nearly two weeks ago has told American investigators that he had to briefly halt his task because of the distraction of civilian guests.

[CNN]

But in a news conference Tuesday night, NTSB member John Hammerschmidt said that the sonar plotter on the submarine – a crew member who notes contact with other possible ships – said he was unable to finish his job plotting sonar blips because the civilians were in his way.

[USA Today]

NTSB member John Hammerschmidt also said Tuesday that the crew member responsible for tracking sonar contacts stopped performing that task within an hour of the collision because of the presence of 16 civilian guests in the submarine's control room.

[Mercury]

A crew member plotting sonar readings on a U.S. submarine that collided with a fishing vessel has told investigators he briefly stopped performing the task because he was distracted by civilian guests in the control room.

[AP]

A crewman who was plotting sonar readings also has told investigators he was distracted by civilian guests in the control room and halted his work.

図 1: 米国 5 紙による潜水艦事故原因の報道文

2 情報抽出

実は、このような言い替え表現の研究は、すでに間接的な形で情報抽出の研究の一部として行なわれてきている。情報抽出は、予め決められた形の情報を新聞記事などのテキストからテンプレート形式で抽出するという課題である。例えば、「企業合併に関する情報」というタスクの場合には、合併する企業名、合併後の企業名、合併年月日、合併企業の分野、規模などが抽出すべき情報として挙げられる。そしてこのような情報を新聞記事などから自動的に抽出するというものであり、米国では MUC (Message Understanding Conference)[6] プロジェクトを中心に、10 年来研究が行なわれてきている。現在の情報抽出の技術では、対象となる情報を含む表現をパターンとして予め用意しておき、それを新聞記事の文に適用して、マッチした場合には適切な情報を抽出するという方法がとられている。そして、最近の課題はいかにパターンを用意するかという問題になってきている。ここでいうパターンというのはひとつの事柄に対する複数の表現のことであり、言い替えれば「言い替え表現」ということである。初期の情報抽出システムでは、タスクが与えられると人間が人手でパターンを作成していた。これでは時間も手間もかかり非効率だということで、コーパスを利用した半自動的なツール [11] の作成が行なわれ、すでに、このシステムは日本語化もされている [8]。また、完全自動化の研究も行なわれており、いくつかの

研究 [5] [3] [12] では、タスクに関連した文章群から頻出の動詞やパターンを獲得し、それをパターン化しようという試みを行なっている。しかしながら、それらの手法はパターン作成の完全な自動化を行なうには欠けているものがある。彼らの手法では、頻出のパターンが複数得られることになるが、その中のいくつかは実は同一の内容を示す言い替え表現であるということは判らない。つまり、同一の意味を持つ表現もそうではないものも同じレベルで集められるだけであり、その後の処理はやはり人間が行なわなければいけない。本研究のきっかけはこの部分の自動化を目指したものである。

3 手法の概要

言い替え表現の自動抽出のシステムは 3 つのコンポーネントからなっている。

- 同じ事柄を報道する記事（同等記事：comparable articles）を同じ日に発行された複数の新聞記事から認定する。
- 同じ事柄を報道する記事から、同じ事柄を表現する文（同等文：comparable sentences）を認定する。
- 同じ事柄を認定する文から言い替え表現である部分を抽出する。

本システムでは、このようなカスケード方式を採用した。これは、直接、言い替え表現を大規模なコーパスから探すよりも効率的でかつ精度が高いと考えるためである。例えば、もし言い替え表現をコーパス全体から探すとしたら、“A Japanese car crashed in Detroit.”と“The talk between Japanese car makers and Detroit crashed.”というような表現は、同じような単語が多く使用されているために、言い替え表現として抽出されてしまうかもしれない。

本システムのキーとなる技術のひとつに固有表現がある。MUC[6]で定義された固有表現というのは、組織名、地名、人名などの固有表現や時間表現、パーセント、金額などの数値表現である。IREX[9]では、日本語を対象とした固有表現抽出の課題として、固有物名というものも導入したが、本研究の対象は英語であり、その精度もあまり高くは期待できないため、本研究では使用しなかった。固有表現というのは、違った新聞記事でもほぼまったく同一のものが使用され、バリエーションが少ないことが期待できる。言い替え表現を抽出するということは、バリエーションを探すことであり、その自動抽出にはバリエーションの少ない固有表現を利用することが非常に有効であると考えられる。この固有表現は、言い替え表現の抽出の際だけではなく、同等文の認定でも使用される。

別のキーとなる技術として、情報検索の研究で開発されたTF/IDFの技術も使用する。実は同等記事の認定は、情報検索やテキストクラスタリングに非常に似ているものである。また、TF/IDFは同等文を認定する際にも使用される。

4 実験

本節では、システムのアルゴリズムの詳細を具体的な実験の結果と共に示す。

本実験では、LDCから配布されているthe Los Angeles Times & Washington Post, the New York Times News Syndicate, Reuters News Service (General), Reuters News Service (Financial)とthe Wall Street Journalを利用した[2]。それぞれの新聞は2-3年分の記事が含まれているが、本実験では予備的なものとして、1994年の9月の記事を利用した。9月1日の記事はパラメータのチューニングに使用し、以下に示す詳細な評価には9月6日の記事を利用した。各新聞には少しだけ異なった同じ内容の記事も含まれているが、これらは、90%以上の固有表現が重なっている記事は1つだけを残すという前処理によって削除した。また、同等記事の認定は異なったニュースソース間のみで行なった。したがって、

Reutersの2つの新聞間での処理は行っていない。1日の新聞記事には80から300の記事が含まれていた。

解析に際しては、構文解析システムにApple Pie Parser[7]、Borthwickの最大エントロピー法による固有表現抽出システム[1]、依存構造解析には著者が自ら開発した決定木によるシステムを利用した。

4.1 同等記事の認定

まず、最初に同じ事柄が報道されている同等記事の認定を行なう。この課題は、情報検索やTDT (Topic Detection and Tracking) [10]で定義された類似記事検索の課題と似ている。しかし、我々の課題は、TDTの課題に比較して全く同じ事柄が報道されていることが要求されている。TDTでは、例えば日本と米国の貿易問題という課題であれば、米国側のコメントの報道も日本側のコメントの報道も含まれるが、言い替え表現の抽出においては、これは望ましいものではない。例えば、“Japan should open the market.”と“Japanese market is open.”は言い替え表現ではない。

実際のシステムでは、TDTで提案されたthe University of Massachusetts (UMASS) [4]のシステムを著者が再インプリメントし、そのパラメータを調整した。ここでは、記事間の類似度は式1で示されるように定義される。

$$ArticleSim(a_1, a_2) = \cos(W_1, W_2) \quad (1)$$

$$w_i = TF(w_i) * IDF(w_i) \quad (2)$$

$$TF(w_i) = \frac{f(w_i)}{f(w_i) + 0.5 + 1.5 * \frac{dl}{avg_dl}} \quad (3)$$

$$IDF(w_i) = \frac{\log(\frac{C+0.5}{df(w_i)})}{\log(C+1)} \quad (4)$$

ここで、 W は記事の単語ベクトルで、 w_i は単語、記事における単語 w の頻度が $f(w)$ 、記事頻度 (document frequency)が $df(w)$ 、 dl は記事の長さ (単語数)、 C は記事数、 avg_dl は記事の平均長である。プロセスに使用しない機能語などのリスト (closed class word list) は自ら作成した。ここで定義された類似度がある閾値 (パラメータ) 以上である場合に、その2つの記事が同等記事であるとした。そして、同等記事のリンクがつながっているすべての記事を集めて同等記事のグループを作成した。最終的な結果はこのパラメータにも依存するが、ここでは、網羅的に同等記事を認定することを目指すよりも、正解率を高くすることを目標にパラメータを設定した。表1に同等記事の認定の結果を示す。表には、9月6日の記事から認定された同等記事グループの数、1グループ中の平均記事数とランダムに選んだ同等記事ペアの評価結果を示す。

グループ数	20
平均記事数	3.8
評価ペア数	25
同等記事	17
類似記事	6
異なった記事	2

表 1: 同等記事の認定の実験結果

4.2 同等文の認定

次のコンポーネントは前のコンポーネントで作成された同等記事の中から同等文を認定するものである。同等記事の認定では、種々のトピックが混在している中から同等の記事を認定することであったが、今度の同等文の認定では、全体的に似通った文の集合の中から同等の文を捜し出すというより難しい課題である。例えば、トレーニングデータにある “Israel and Morocco open ties” の記事（全部で 19 文）においては、それぞれの国の名前は数多く (“Israel” は 12 文で “Morocco” は 5 文で) 言及されている。したがって、記事の検索で使用されたようなキーワードはここではあまり有効ではない場合がある。また、最終的には異なった言葉で表現されている言い替え表現を抽出することが目的であるので、固有表現以外の単語をあまり重視することはできない。

いろいろな試みをした結果、同等記事の認定で使ったものに比べてよりシンプルな TF/IDF を変形させた方法が一番有効であることが判った。また、記事頻度 (Document Frequency) にかわって、同等記事内でどれだけ文に単語が現れたかを意味する文頻度を使用した。単語頻度 (Term Frequency) は使用しなかった。これは、もしある単語が複数回使用された時には、あまりに大きな値が得られてしまうのを避けるためである。また、一般的な TF/IDF の方法と同様に機能語などは考慮に入れていない。そして、単語の種類によって異なった重みを与えた。固有表現は一般の語に比較して 2 倍の重みを与えられている。文の類似度は式 5 で与えられる。

$$SentSim(s_1, s_2) = \cos(W_1, W_2) \quad (5)$$

$$w_i = Weight(w_i) * \log\left(\frac{C}{sf(w_i)}\right) \quad (6)$$

ここで、 W は各文における単語ベクトルで、 C は 2 つの比較している記事にある文の合計数、 $sf(w)$ は、単語 w の文頻度 (sentence frequency) である。

同等記事グループ中のそれぞれの記事ペア間のすべての文ペアについて文の類似度を計算した。以下の 3 つの条件を満たした時に、2 つの文は同等文であると認定する。

	文ペアの数
同等	15 (65%)
類似	6 (26%)
異なる	2 (9%)

表 2: 同等文の認定の評価結果

1. 文の類似度がある値を越えていること。この値はチューニングデータを使用して求められた。
2. 両方の文が少なくとも 1 つ以上の共通の固有表現を有していること。
3. お互いにもっとも類似度の高い文同士であること。

表 2 に同等文の認定における評価結果を示す。「同等」というのは、それらの 2 つの文中に言い替え表現が含まれていることを示す。「類似」は、2 つの文は同じか同じような内容を表現しているが、それらの文からは言い替え表現が抽出できないことを示す。「異なる」というのはそれらの文がまったく異なっていることを示す。図 2 に、それぞれの分類に当てはまる文ペアの例を挙げる。実際、「異なる」と分類された 2 つの文は両方とも例にある文ペアであった。

4.3 言い替え表現の抽出

このようにして、言い替え表現を含む同等文が認定されたが、それらの文は完全に同等なわけではなく、文の中から言い替え表現を抽出しなければならない。例えば、挿入、形容詞などや、別の事への言及などの余分なものがそれぞれの文にあることが考えられる。それらを排除して言い替え部分だけを抽出する。

このコンポーネントでは、構文解析と依存構造解析を利用する。依存構造解析では、ヘッドとアークメントの関係から動詞とその主語、目的語などの関係や形容的な関係を解析する。そして、固有表現などの手がかりに基づいて、文の依存構造グラフを枝刈りし、言い替え表現に相当する所を抽出する。文の依存構造グラフの枝刈りは以下のように行なう。

1. それぞれの同じ固有表現にマークをつける。
2. 同じ文のグラフ内で固有表現間の距離を計算する。もし、それらの距離がある閾値以下であり、かつそれぞれの文のグラフにおける距離の差異が別の閾値以下である場合には、それぞれの固有表現の間のエッジにマークをつける。距離に上限を与えたのは、あまりに距離が離れている場合にはそれらの関係が希薄である可能性があるためであり、距

[同等]

- Under the Bush administration, the Justice Department had filed a suit supporting Sharon Taxman, declaring that her raced-based firing was illegal.
- Under the Bush administration, the Justice Department had supported a discrimination lawsuit by Sharon Taxman, a white business teacher at a Piscataway, New Jersey, high school.

—

- Gerry Adams, the militant Northern Ireland republican leader, had an unprecedented meeting Tuesday with Irish Premier Albert Reynolds in Dublin to discuss the cease-fire of the Irish Republican Army.
- Irish Prime Minister Albert Reynolds will meet Gerry Adams, president of the IRA's political wing Sinn Fein, for talks Tuesday afternoon, government sources said.

[類似]

- Delegates to the U.N. conference on population and development neared agreement Tuesday on a U.S.-backed compromise aimed at defusing religious controversy over abortion, but progress stalled at the last minute after there were objections from the Vatican.
- U.S. officials expressed hope Tuesday that they were nearing agreements over the disputed abortion and birth control sections of a United Nations plan for slowing the world's population growth.

[異なる]

- What would happen if Clinton were to lift the longstanding U.S. embargo on Cuba?
- Reversing previous U.S. policy, President Clinton has sent them to the U.S. naval base on Guatanamo on the southwestern tip of Cuba.

図 2: 文ペアの例

離の差異の制限は、お互いに違った使用がされている場合を避けるためである。

3. マークされていないエッジの内、強制的に必要なものにマークをつける。これには、動詞の主語、目的語、冠詞、to 不定詞の“to” などがある。
4. マークされたエッジの数が特定の数以上である場合に、マークがついたグラフの部分を取り出す。この制限は、あまりに小さいサブグラフの場合には、単なる複合名詞や前置詞句がついた名詞など有効な言い替え表現ではない場合があるためである。

制限に用いたすべての値は、チューニングデータを用いて求めた。

このような方法により、図 2 の「同等」に挙げた最初の文からは以下の言い替え表現が抽出される。

- Under the NE-Person administration, the NE-Organization had filed a suit supporting NE-Person
- Under the NE-Person administration, the NE-Organization had supported a lawsuit by NE-Person

この結果は、言い替え表現の自動抽出の目的を満足させる結果である。

このような方法により、1カ月の新聞記事から約 1000 の言い替え表現が抽出された。そのうち、これまでも評価で使用した 1 日分の新聞記事からは 8 つの言い替え表現が抽出されている。その内の 4 つが正しい言い替え表現であると判定された。また、同等文の認定で「異なる」と評価された文ペアからは言い替え表現は抽出されていない。

5 考察

本実験は、言い替え表現自動抽出の最初の実験であるが、実験結果からはその可能性が見い出されたと考えている。しかしながら、実験はまだ充分ではなく、これからの課題も多く残されている。

- 大規模な実験、評価

現在はまだアルゴリズムの改良やパラメータの調整中であり、大規模な評価を行なう段階にはないと考え、1 日分の評価しかしていない。現在の所、大きな問題点は文の解析精度であると考えている。例えば、前置詞句の修飾先や並列句の解析ミスがあったとしても、それだけで、重要な要素の取りこぼしにつながる可能性がある。

実験では 1ヶ月分、評価では 1 日分の新聞記事しか

使わなかったが、手元には2年分の重複した新聞記事がある。単純に計算すると約24000の言い替え表現候補が得られる計算になる。もし、これだけのデータが得られれば、例えば、それぞれの言い替え表現の信頼性の推定などが頻度を数えることによって可能になるかも知れない。また、それだけのデータが得られれば、言い替え表現のリンクも形成できる可能性がある。つまり、同じ表現を持つ言い替えペアがあれば、それらを合計して3つの表現を同じ言い替え表現と判断できる。ただし、これには危険性も伴っている。非常に一般的な表現、例えば“I take it”といったものはいろいろな表現との言い替え表現になる可能性がある(例えば、“I buy it”や“I ride it”)。しかし、このような表現同士を言い替え表現とはしたくないため、リンクの際にはこのようなことが起きるのを避ける手段を導入する必要がある。

- 単語の類似度

現在の実験では、単語は完全に一致しなければ、類似度の計算で考慮されない。しかし、シソーラスなどを利用して、単語の類似度の尺度を導入することによりより多くの言い替え表現が抽出できる可能性がある。

- 必須なものとしてでないもの

最終的に得られた言い替え表現の例を見ても判るように、その言い替え表現に必須な部分としてでない部分が存在する。この例では“Under the Bush administration”は必須ではないが、例えば、“the Justice Department”は必須である。もし、必須でない部分を削除することができれば、この言い替え表現の適用可能性が広がりより有益な結果になる可能性がある。主語や述語などの手がかりを用いるといったことが考えられるが、時に単なる形容詞や前置詞句が重要になったりする場合がある。

6 まとめ

本論文では、複数の新聞を使用して言い替え表現を自動抽出する方法を提案した。実験や評価はまだ予備的なものに過ぎないが、その実現の可能性を見出したものと考えている。今後は、実験、評価の規模を大きくしていくと共に、考察で述べたような点を実現してゆきたい。また、有益なディスカッションをしていただいたGrishman教授に感謝する。

参考文献

- [1] Andrew Borthwick, John Sterling, Eugene Agichtein and Ralph Grishman “Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition” *WVLC-98* 1998.
- [2] LDC “The North American News Text Corpus” *LDC CD-ROM* 1997.
- [3] Chikashi Nobata, Satoshi Sekine “Towards Automatic Acquisition of Patterns for Information Extraction” *ICCPOL-99* 1999.
- [4] Ron Papka, James Allan, Victor Lavrenko “UMASS Approaches to Detection and Tracking at TDT2” *DARPA: Broadcast News Workshop* 1999.
- [5] Ellen Riloff “Automatically Generating Extraction Patterns from Untagged Text” *AAAI-96* 1996.
- [6] SAIC “MUC: Homepage” <http://www.muc.saic.com/>.
- [7] Satoshi Sekine “Apple Pie Parser: Homepage” http://cs.nyu.edu/cs/project/proteus/app”.
- [8] Satoshi Sekine, Chikashi Nobata “An Information Extraction System and a Customization Tool” *JSPS-Hitachi Workshop* 1998.
- [9] Satoshi Sekine, Hitoshi Isahara “IREX: IR and IE Evaluation project in Japanese” *LREC-00* 2000.
- [10] Charles Wayne “Topic Detection & Tracking: A Case Study in Corpus Creation & Evaluation Methodologies” *LREC-98* 1998.
- [11] Roman Yangarber, Ralph Grishman “Rapid Customization of Information Extraction Systems” *In the proceedings of the Symposium on Advanced Information Processing and Analysis* 1997.
- [12] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, Silja Huttunen “Automatic Acquisition of Domain Knowledge for Information Extraction” *COLING-00* 2000.