主題構造解析による新聞記事からの気象情報の抽出と応用

大村 高史† 田村 直良††

†横浜国立大学大学院 工学研究科 電子情報工学専攻
††横浜国立大学大学院 環境情報研究院
{takashi,tam}@tamlab.eis.ynu.ac.jp

1 はじめに

近年、電子化された文書は増加の一途をたどっており、WWWの普及と相まって、これらの文書を手軽に入手することも可能になってきた。しかしながら、膨大な文書の中から欲しい情報を探したい、あるいは文書の集合を分析して傾向をつかみたいといった、利用者のさまざまな要求に十分対応できる情報アクセス手段はまだ乏しいのが現状である。

現在、我々は容易に入手可能であり、多方面に渡り事実の提示や物事の解説が多く含まれる新聞記事を情報源として、気象庁が発信している主として数値的な情報からでは得ることができない、社会生活に密着した気象に関わる知識を獲得することを目標として研究を行っている。気象情報は我々の生活と非常に深い関係を持っており、このシステムの実現は、我々の生活にとって非常に高い利益をもたらすと考えられる。その一端として、本研究では、文章から知識発見するために必要なデータベースの構築を目指し、データベースに蓄積するための事例の抽出方法について検討する。

事例 (気象記事) の抽出は、基本的には新聞記事から気象に関する字句表現と、その気象の事象が発生した地域の名称抽出に基づく。これは、それぞれ単独に固有名詞の抽出をするのでは全く意味がなく、気象情報とそれが発生した場所の組として情報を抽出する必要がある。単独の固有名詞の抽出に関する研究は盛んに行われているが、このように複数の語句を関連付けて抽出する研究はまだあまり注目されていない。

本研究で取り扱う２つの対象は、同一文内に定型的に出現するものではなく、その上、必ずしも同一文内に含まれているとも限らない。この場合、表層上のパターンマッチによる抽出手法は非常に困難であり、文章の内容を理解した上での抽出が必要不可欠である。そこで本研究では、文章中の話題の展開を把握することにより、語句の関連性を重視した抽出手法を提案する。

2 主題構造解析

文章は文を累積したものであるが、単なる集合体ではなく、時間的・線条的な累加・連続として成立する流れであり、文の継起的連続というところに、文章の特性がある [2]。

本研究では、一文にはその文において中心の話題となる語句 (主題) が存在すると仮定する。そして、文間の主題の連鎖関係を解析することで、文章全体の話題の展開を把握することができると考えている。この解析を以降では主題構造解析と呼ぶ。以下に、主題構造解析について説明する。

2.1 主題の抽出

主題構造解析をするために、トピックと主題の抽出を行う。トピックと主題の定義を以下に示す。

- トピック：文章は、その全体の話題を表すようなトピックを持っていると仮定する。本研究では、新聞記事のトピックをそのタイトル中に出現する「名詞句」及び「名詞」と定義する。

- 主題と題述 [2]：各文は、主題構造を持つと仮定し、主題と題述とから構成されているとする。具体的には、係助詞「は」の直前に出現する「名詞句」もしくは「名詞」を主題と定義し、文の主題以外の残りの「名詞句」及び「名詞」を、題述と定義する。

主題となりうる語句が複数存在する場合は、一文中でより文末に近い語句を修飾している主題候補を抽出する。ただし、同じ語句を修飾している場合は、より文末の近くにあるものを抽出する。

2.2 主題の連鎖関係の種類

記事中の各文間が、下記の条件の、５種類の連鎖関係のうちで少なくとも１つを満たすものとし、何らかの結束性を持っているとする。

A 主題維持：直前の文の主題と同一か、基準以上の類似度のある主題を持つ場合。

B 主題変化：直前の文の題述のいずれかと同一か、基準以上の類似度のある主題を持つ場合。

C 主題回復：最も近い主題変化の直前の主題と同一か、基準以上の類似度のある主題を持つ場合。

D トピックの導入：文章のトピックと同一か、基準以上の類似度のある主題を持つ場合。

E 主題派生：上記のいずれにも該当しない場合。この場合、直前の文やトピックとは関連性の低い文となる。

ただし、基準以上の類似度とは、一方の語句が他方の部分文字列になっている場合とする。

2.3 主題の連鎖関係の決定

主題の連鎖関係を決定するルールを示す。

- ルール１：原則として、結束関係の強さはＡ＞Ｂ＞Ｃ＞Ｄ＞Ｅとし、可能な限り結束性の高い連鎖を採用する。ただし、主題を抽出する際、主題が省略されている文に関しては、省略 (ellipsis) により結束構造 (cohesion)[2] があるものとして、主題の維持と見なす。
- ルール２：第１文に関しては、前文が存在しないため、「トピックの導入」が採用できる場合にそれを採用する。ただし、第１文に主題が存在しない場合に関しては、「トピックの導入」を採用し、主題は定めない。そして、第２文との主題の連鎖関係は、第２文の主題が何であっても「主題の維持」とする。
- ルール３：ルール１にもとづき、連鎖関係を決定した結果、Ｄ「トピックの導入」とＥ「主題の派生」のどちらかになった場合、連鎖関係の修正処理を試みる。これは、主題を採用する際にあらかじめ抽出しておいた主題候補 (主題を修飾する名詞句及び名詞) をもとに行う。主題候補を主題と仮定し、再度ルール１にもとづき連鎖関係を決定し、その結果がＡ～Ｃの連鎖関係となった場合にのみ、この主題と連鎖関係を採用する。ただし、主題候補が複数存在する場合は、主題の直前に存在するものから順に修正処理を試みる。

3 気象情報の抽出

気象に関する事項の抽出のために気象語句のタグ付けを行い、その後タグ付けされた記事を対象に気象に関して記事の分類を行う。

3.1 気象語句のタグ付け

本研究では、気象語句を、分類語彙表 [3] に掲載されている気象に関連の深い項目 (1.514 天災、1.5150 気象、1.5151 風、1.5152 雲、1.5153 雨・雪、1.5154 天気、1.5155 波・潮) に含まれる名詞とする。動詞を含む他品詞に関しては、気象を示すものを分類することが難しく、又高い精度が望めないことが予想されるため省略する。

タグ付け手法としては、前述の気象語句をリストとして用意しておき、記事毎に表層上のパターンマッチング手法により気象語句にタグ付けを行う。パターンマッチングでは、文章の任意の位置からマッチングを行うと、誤ってタグ付けを行う確率が高くなるため、形態素解析ツール茶筌 [4] の結果で得られる形態素の境界をもとに、完全一致法によってマッチングを行う。気象語句が、二つの形態素にまたがる場合に関しては、その区間による完全一致でマッチングを行う。

3.2 気象記事の分類

気象語句のタグ付けを行った記事を対象に、気象と関連の深い記事と気象と関連の浅い記事を機械学習システムＣ 4.5[1] を用いて分類する。気象記事をデータとして蓄積する際に分類を行い、蓄積するデータに分類項目を付け加えることによって、データ利用者が用途によって利用しやすくなることが期待される。

4 地名の抽出

4.1 地名データベース

電子化された郵便番号簿をもとに地名を階層化し、PERL の GDBM を用いて地名データベースを作成する。階層化することで、地名を単に名称だけでなく包含関係も考慮した系統的な情報の蓄積を実現することができる。

気象に関する記事をデータとして蓄積することを考慮すると、せまい地域に関してデータを蓄積することは、あまり有効ではない。そこで、第１階層：「国名」、第２階層：「地方名」、第３階層：「都道府県名」、第４階層：「市区町村名」として階層化し、第４階層よりせまい地域に関しては階層化しない。

4.2 地名のタグ付け

気象語句のタグ付けと同様に形態素解析ツール茶筌の結果で得られる境界をもとに、地名データベースに存在する地名とのマッチングを行う。また、タグ付けされた後の処理を効率化させるために、階層毎に異なるタグを使用する。

4.3 地名の抽出

タグ付けされた地名には、気象語句と全く関連の無い地名も多く含んでおり、気象語句と関連性の高い地名のみを抽出する必要がある。

手法としては、まずあらかじめ新聞記事に対して主題構造解析を行い、その結果をもとに、地名を抽出するにあたって関連性が高いと考えられる文の集合を抽出する。この文集合を以下では、related sentence とする。

- related sentence の抽出方法

  related sentence の抽出方法としては、気象語句と地名の関連性を重視し、以下の条件をもとに抽出を行う。ただし、ここではある文Sとその前文との連鎖関係によって条件を設定している。また、本研究では、離れた文に出現する気象語句と関連した地名をいかにして抽出するか、いかに気象語句と関連の低い地名を削除するかということを目的としている。そのため、一文中で気象語句と地名の両方が出現している場合は、その2つの語句は関連性が高いということを前提としている。

  1. 前文との連鎖関係が「主題の維持」の場合、文Sと前文は同一の内容に関して述べられていると考えられ、文Sと前文の中で出現する気象語句と地名は関連性が高いと考えられるため、文Sと前文は related sentence であるとする。
  2. 前文との連鎖関係が「主題の変化」の場合、前文の題述である語句、つまり文Sの主題である語句が気象語句である場合にのみ、文Sと前文を related sentence とする。これは前文の内容に気象語句が出現し、その気象語句に関して文Sが述べられていると考えられるためである。
  3. 連鎖関係が「主題の回復」の場合、文Sと主題の回復元である文の関係は、主題の維持の場合と同様であるため、related sentence であるとする。
  4. 連鎖関係が「トピックの導入」の場合、タイトル中に出現する語句を主題としている文Sとタイトルの内容は非常に密接したものであると考えられるため、同様に文Sとタイトルは related sentence とする。
  5. 前文との連鎖関係が「主題の派生」の場合、文Sは他のいかなる文とも related sentence を構成しない。

  主題構造解析をもとにした related sentence の抽出例のイメージを図1で示す。

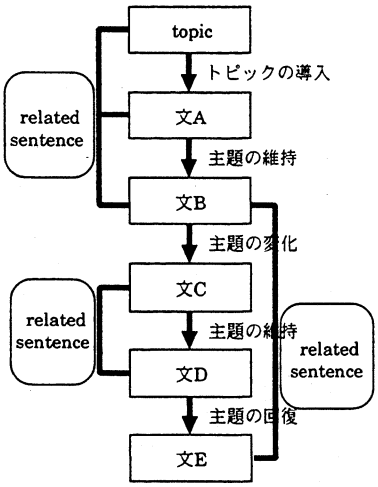

図 1: related sentence の抽出例

- 地名の抽出方法

  作成された related sentence の中から地名を抽出する。この作成された related sentence は、主題構造解析の結果をもとに得られたものであり、この中に出現する気象語句と地名は関連性が高いと考えられる文の集合である。そのため、related sentence 内に気象語句と地名の両方が出現した場合は、その地名を全て抽出する。

4.4 地名のあいまいさの解消

本研究における第4階層の地名には、いくつかつづりが等しいものが存在する。そのようなあいまいな地名に対して、地名の同定を行う。手法を以下に示す。

地名の抽出において文章を左から右へ走査する際、都道府県名が出現すれば配列にスタックする。もし文章中の地名にあいまいさがなければ、この配列を参照せずに同定できる。あいまいな地名が出現した時には、それと同定するべき地名候補の上位階層地名がスタックに格納された地名と比較され、等しい地名が選択される。

5 実験と評価

地名の抽出の際、最も重要なことは、主題構造解析の結果を用いることによって気象語句と関連性の高い地名を正確に抽出できているかどうかである。これに関して、「A：本論文で提案したシステムによって地名を抽出する方式」と「B：一記事中に気象語句と地名が共起した場合に限り、その地名を気

象語句と関連のある地名として抽出する方式」の２つの方式を比較して検討する。その結果を表1で示す。ただし、実験では９３年度日本経済新聞からランダムに選出した５６記事に対して地名の抽出を行い、その結果に対して検討を行う。また、評価では再現率、適合率、Ｆ値を求める。Ｆ値($0 < F < 1$)とは、再現率Ｒと適合率Ｐのトータルでの精度を表す指標である。

抽出手法	再現率	適合率	Ｆ値
Ａ方式	36.8	58.3	0.451
Ｂ方式	68.3	50.0	0.577

表 1: 実験の評価１

表1の結果から、Ｂ方式では、離れた文で気象語句とそれと関連性の高い地名が出現した場合、地名を抽出することは不可能である。そのため、この手法の再現率は極めて低い値になっている。つまり、このことから記事中には気象語句とは離れているが、関連性の高い地名が多く存在することが分かる。

またＢ方式では、Ａ方式と比較すると再現率は良いが、適合率はやや低くなっている。これは、気象語句とは離れた文に存在する、気象語句と関連性の高い地名は抽出することができているが、その反面、関連性の低い地名に対しても抽出を行ってしまっているためであると考えられる。このような場合の多くは、気象の意味として用いられていない語句に気象のタグ付けを行ってしまっている場合である。

ここで実験として、本システムよって離れた語句を抽出する際の抽出精度について、地名と対応する気象が全ての気象語句とする場合と、気象の意味として用いられている気象語句に限った場合とに分けて、それぞれに対して本システムによって地名を抽出し、その結果を表2に示し比較検討する。ただし、気象の意味として用いられているかどうかについては人手の判断によるものである。

地名抽出の対象	正解率
全ての気象語句	44.1
気象の意味として用いられている気象語句	73.3

表 2: 実験の評価２

表2の結果から、気象語句と関連のある離れた地名を抽出する際、気象語句が気象の意味として用いられている場合については、約７割の正解率が得られた。しかし、気象語句が気象の意味として用いられていない場合が多くあるために全体としてはあまり良い結果が得られなかった。このことから、主題構造解析自体は有効な手段であるが、格構造解析等により気象語句が気象の意味として用いられているかどうかを解析する手段が新たに必要であることが分かった。

図2に９３年１０月３０日の日本経済新聞に掲載されている記事をもとに、地名の抽出例を示す。

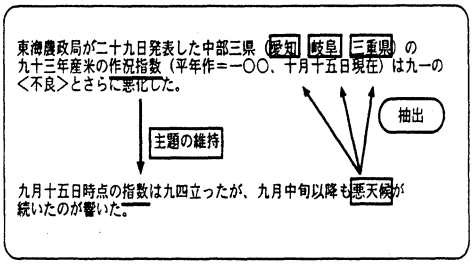

図 2: 地名の抽出例

6 まとめと今後の展望

本研究では、新聞記事を対象に主題構造解析をすることによって話題の展開を把握することにより、気象語句と関連性の高い地名を抽出する手法を試みた。その結果、パターンマッチング手法や一文中の共起情報からの抽出ではなすことができなかった、離れた文からの関連語句の抽出をすることが可能となった。

また、主題構造解析における一文の単文化や主題の抽出精度向上、気象語句の抽出における抽出精度向上など検討すべき問題も多く残されている。

参考文献

[1] J.R. キンラン. ＡＩによるデータ解析. トッパン, 1985.

[2] M. A. K.Halliday. An introduction to functional grammar second edition. くろしお出版, 2001.

[3] 国立国語研究所. 分類語彙表. 秀英出版, 1993.

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム「茶筌」version2.0 使用説明書 第二版. NAIST Technical Report, 奈良先端科学技術大学院大学 松本研究室, 1999.