離散共起表現データを用いた単語のグルーピング

内野　一†　白井　諭†　　池原　悟‡
†ＮＴＴコミュニケーション科学研究所　‡鳥取大学

１．はじめに

n-gram 統計処理を用いて、日本語テキストデータ内の単語や、固定的な言い回しを収集する手法が各所で研究されている[1][2][3]。共起表現には、いくつかの単語が連続したタイプ（連鎖共起表現）と、２種類以上の表現が文中の離れた位置に共起するタイプ（離散共起表現）があるが、自然言語の文型を収集する場合は、後者のタイプが特に有効と考えられる。しかしながら、文の単位を越えるような大きな単位で、多くの要素を持った離散共起表現を収集するには、計算量が大きくなってしまうという問題がある。本稿では、一度小さな範囲で求めた離散共起表現データを用いて表現をグルーピングすることにより、連鎖共起表現に変形し、より大きな範囲での共起表現を求める手法を提案する。

２．離散共起表現から連鎖共起表現への変形

２．１　置き換えを用いたn-gram統計処理による文型の抽出

n-gram統計処理によって、定型的表現を抽出する際、経済分野の新聞記事のように数詞や固有名詞が頻出する分野においては、長い単位を持つ表現の抽出が出来ないという問題がある。n-gram統計処理においては当然のことながら文字のレベルでの共起を集計するため、数詞のように他の値となっても文の形式にはほとんど影響しないが、表層上でのみ違いがあるような表現を、完全に別の表現として扱うためである。このような文章に対応する手法として、問題となるデータをあらかじめ別の文字列で置き換えた後に改めてn-gramの抽出を行う方法が提案されている[4]。この手法を用いることにより、数詞などによって断片化されていた連続的な表現を抽出することが出来る。しかしながら、この手法においては、語の変換において辞書を使用することを前提としているため、事前の準備が必要となる。このような制限をはずすため、２要素の離散共起表現の間に位置する表現をあらかじめ定めた基準によって自動的に同一視して、置き換えを行うことにより、連鎖共起表現に帰結させる手法を試みた。

２．２　置き換え対象となる表現

基本的には、文の構造に及ぼす影響が同一であるような表現を同一視するものとする。例えば、動詞「上がる」と「下がる」においては、語としては逆の意味を持っているが、文の構造に関しては、同一の形態をとる。このように、あくまで文型の抽出のための置き換えとして考える。

また、辞書などの語に関する情報を一切使用せず機械的に置き換えを行う上で、精度の大幅な低下を起こさないようにするため、置き換えの対象は、一語と推定できる程度の長さを持つものに限定する。また、一文字で構成される単語についても誤った置き換えを行ったときの影響が大きくなるため対象外とする。

２．３　単語のグルーピングによる文字列の置き換え

辞書情報を使用せずに語であるかどうかの判断を行うためのデータとして、機械的に判断の出来る文字コードを利用する。

（１）数値データの同一視

数値データは、算用数字、および、小数点、単位を表す漢数字によって構成されている。これらの文字コードをもった文字が連続していた場合、数値表現として置き換えを行う。

（2）離散共起表現データによる語のグルーピング

数値データに対して置き換えを行った後、n-gram統計処理を行って、2要素の離散共起表現を抽出する。抽出された要素間の文字列を原文から取りだし、以下の要件を満たしていた場合、置き換えが可能な語の候補とする。

・文字列の長さは2文字以上、6文字以内
　長い表現、および、一文字で示される助詞などの語を置き換え対象としない
・平仮名、片仮名、漢字、アルファベットを文字コードで区分し、文字列内でのコードの切り替わりは1回のみとする
　多くの単語においては、漢字＋平仮名、片仮名＋平仮名、といったように構成されており、2回以上コードの切り替わりがある場合は、単語より大きな単位であると予想される
・文字列中に記号を含まない

候補文字列の延べ数が一定以上の個数（本稿では１０回以上）となった時に、それらの文字列をグループとしてまとめ、置き換えを行う。別の2要素間から抽出された文字列は、別のグループとしてまとめる。複数のグループに共通して出現する文字列は、候補種類数が一番多いグループにまとめる。候補種類数が同じであった場合、小さな単位での置き換えを優先する立場から、平均文字列長の一番短いグループに加える。

（3）抽出グループの統合

各グループ間の関係を求めるため、置き換え後のデータを対象にして、さらに離散共起表現の抽出を行い、同様に2要素間の文字列から置き換え候補文字列を求める。この段階では、候補文字列とする条件を（2）の場合より厳しくし、文字列内でのコードの切り替えが無い場合にのみ候補とする。

すでに一度、置き換えが行なわれている文字列がさらに抽出された場合、置き換え文字列そのものの場合は、さらなる置き換えの候補とする。複数の置き換え文字列が候補として抽出されてきた場合、2つのグループが統合されることとなる。また、同時に抽出された文字列も同じグループに属するものとする。

（4）繰り返し表現の同一視

例示など、名詞の列挙による繰り返し表現は、その列挙数が変化した場合、別の表現として収集がなされる。これを一つのパターンとしてまとめあげるため、同一の置き換え文字が区切りにより列挙されている表現を抽出し、その個数によらず全体を1つの表現に置き換える。

3．実験結果

日本経済新聞社、市況速報サービス記事（8カ月分、約４２００文、１４万字）を対象に本手法を適用し、検討を行った。この記事中には、企業名や数値が頻出し、定型的表現を抽出するにはこれらの表現を同一視して処理する必要がある。記事の例を表1に示す。

表1　市況速報サービス記事例

大証修正は前日終値の水準で始まる。需給の良さから下値不安は小さいが、日経平均で８９年高値からバブル崩壊後の安値までの3分の一戻しに当たる２万２５００円近辺からは上値も重く様子見気分も強い。寄り付きの成り行き注文は、買い１５３万株、売り１２４万株。任天堂、ロームが続落。オートバクスが一時1万円台割れ。古野電、ダイダン、日理化、中外炉、シマノもさえない。半面、兼松日産農が上げ、村田製、島精機も堅調。

3．1　単語のグルーピング結果

前記手法の（2）を用いて単語をグルーピングした結果の例を以下に示す。延べ出現回数の条件としては１０回以上抽出されたものとして、グルーピングを行った。最終的には４３のグループに分かれる結果となり、企業名のグループ（例1）が１３個、相場の状態を示す語のグループ（例2）が7、株の種類を現わす語のグループ（例3）が2、ある株の状態を示す動詞のグループ（例4）が3、といった状態で、例5、6のように時間的な表現や、語の使用のされかたといった形式でも、ほぼ正しくグルーピングがなされていた。

間違ってグルーピング対象とされたものは、ほとんどが助詞＋名詞といった形式であり、その他にも長い平仮名で示された表現の一部だけが対象としてあげられているものもあった。平仮名のみで構成される語や、よく使用される助詞に関しては、一般の辞書を併用するなど特別の処理を考慮する必要がある。

グループの統合処理では、企業名を現わすほとんどのグループが統合されたが、その他のグループ間で統合されたのは、相場の状況を示すグループだけであった。グループの統合条件については、その相関の取り方などについては今後の検討課題である。

例1　第一要素　“半面、”　第２要素　“、ワキタ、”
　ローム　小野薬　青山商　村田製　島精機　日精化　日本橋
　オムロン　兼松日産農　日成ビルド

例2　第一要素　“大証修正は”　第２要素　“して始まる”
　続伸　続落　反発　反落　小反発　小幅続伸　小幅続落

例3　第一要素　“半面、”　第２要素　“の一”
　機械株　建設株　繊維株　電機株　薬品株　流通株　仕手系株　ハイテク株　仕手材料株

例4　第一要素　“ハイテク株が”　第２要素　“ほか、”
　下げた　堅調な　上げた　買われた　売られた　小安くなった　買われている

例5　第一要素　“大証修正は”　第２要素　“まで”
　後場中ごろ　前場中ごろ

例6　第一要素　“との”　第２要素　“もあり、”
　見方　空気　ムード

3．2　抽出パターン数の推移

各ステップにおける連鎖共起表現の抽出状況を表2に示す。連鎖共起表現数は抽出された連鎖共起表現の異なり数、パターン対象文数は一文全体をパターンとして抽出した文数である。パターン数はその異なり数を示している。

表2　抽出パターン数の推移

	連鎖共起表現数	パターン対象文数	パターン数	パターン当り文数
原データ	7，662	686	250	2．7
数値置き換え	7，126	992	264	3．8
単語のグルーピング	6，263	1，102	263	4．2
グループ統合	3，910	2，011	492	4．1
繰り返し表現の処理	3，639	2，332	403	5．8

原データを加工せずに文型抽出を行った場合、数詞などで分断される場合が多く、一文全体をパターンとして抽出できたのは短い特定の文がほとんどであった。

数値データの置き換えにおいては、ほとんどの記事に出現する、数値のみが異なっている文をパターンとして抽出できるようになったためであり、パターン対象文の数が増えているわりには、異なり数は増加していない。代表的な文の例を以下に示す。

例	寄り付きの成り行き注文は、買い（数値）株、売り（数値）株。	頻度　９５
	寄り付きの成り行き注文は、売り（数値）株、買い（数値）株。	頻度　３６

単語のグルーピングにおいては、第１段階でグループわけした際に１３のグループに分かれてしまい、また、この段階では、全体の半分程度の企業名しか抽出されなかったため、対象文数やパターン数に大きな変化は見られなかった。

グループの統合およびその際の企業名の抽出によって、ほとんどの企業名が抽出されたため、一文そのものをパターン化することが出来た数はほぼ倍増した。しかしながら、頻度の少ない長文をパターンとして捉えるため、１パターン当りの文の数はわずかに減少している。これは、次のように一文中での企業名の列挙数には多くのバリエーションがあり、それらを全て別パターンとして扱っているためである。この状況は、繰り返し表現の処理によって下記のような例が統合され、最終的には、対象文４２００文の半数以上の文を機械的に定型化することが出来た。

半面、（企業名）、（企業名）が下げ、（企業名）、（企業名）、（企業名）も軟調。
半面、（企業名）、（企業名）、（企業名）が下げ、（企業名）、（企業名）も軟調。

４．まとめ

本論文では、離散共起表現として抽出された２要素間の文字列を、機械的な規則によってグルーピングし置き換えを行うことにより、より長単位の離散共起表現を抽出する手法を提案した。単語のグルーピングにおいては企業名をほぼ１００％正しく自動的にまとめることが出来、また、動詞や、形容詞についても語の使用法に基づいて分類されることがわかった。企業名など一般的な辞書には掲載されない文字列を正しく分類することが可能であるため、語の属性の推定などに応用するとも可能と思われる。

この手法を市況速報記事データに適用した実験においては、半分以上の文について自動的に定型化を行うことが出来、本手法が有効であることを確認した。

今後は、一般的な辞書を併用してさらに精度を高めるなどの処理を検討し、文の範囲を超えた定型的な表現の抽出について検討していく。

参考文献

[1] 長尾，森：New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, COLING'94, pp.611-615

[2] 池原，白井，河岡：「大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法」, 情報処理学会論文誌, Vol.36, No.11,pp.2584-2596(1995)

[3] 下畑，杉尾，永田：「隣接文字の分散値を用いた定型表現の自動抽出」, 情報処理学会自然言語処理研究会報告110-11,pp71-77(1995)

[4]内野、白井、池原、新田見：「置換えを用いたn-gramによる言語表現の抽出」,電子情報通信学会技術研究報告NLC96-18,pp63-68[1996]