

自然言語における特徴的表現の抽出に関する研究

園田, 匠 / SONODA, Takumi

(出版者 / Publisher)

法政大学大学院理工学・工学研究科

(雑誌名 / Journal or Publication Title)

法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編

(巻 / Volume)

55

(開始ページ / Start Page)

1

(終了ページ / End Page)

2

(発行年 / Year)

2014-03-24

(URL)

<https://doi.org/10.15002/00010413>

自然言語における特徴的表現の抽出に関する研究

STUDIES ON EXTRACTION OF THE FEATURE EXPRESSION IN NATURAL LANGUAGE

園田 匠

Takumi SONODA

指導教員 三浦孝夫

法政大学大学院理工学研究科電気工学専攻修士課程

In natural language processing, computational linguistics are discussed well. Among others, collocation has been much discussed so far by which we expect to analyze how to obtain and enrich vocabularies. In this investigation, We propose effective process to extract collocation automatically, extraction process of conditional collocation considering asymmetry, and method of rich enhancement of NLP application. we give some experimental results to show the effectiveness of our method.

Key Words : Natural Language Processing, Collocation, Data Mining, Feature Selection

1. 問題の背景

近年、インターネットの普及に従いデジタル新聞や電子書籍、Blog や Twitter といった電子テキスト文書による情報が多く存在している。これらのテキストは多種多様な情報を保有しているが、利用者の増加に伴い情報が膨大化しており、全ての情報を人手により把握することは困難になっている。このため、これらの情報を自動的に扱うために計算機による援用が必要とされている。

計算機によるテキスト文書の自動的な処理を行う自然言語処理 (Natural Language Processing) の研究領域の一つに、計算言語学 (Computational Linguistics) という分野がある。これは計算機の観点から言語学を捉え、自然言語を論理的にモデリングすることを目的としており、文書 (文章) 中に含まれる単語やその共起の頻度、字句の相関関係や時系列等を解析し、有用な情報を獲得する研究領域である。

この中で、連語 (Collocation) が特に注目されている。連語とは特定の語の共起による自然かつ習慣的な言語表現である。例えば、英語で Strong Tea という表現は「濃いお茶」という意味をもつ。ここで Strong は語単体では「強い」を表現するが、これと同義である Powerful を用いて Powerful Tea としても「強いお茶」にしかない。このように語単位では同じ表現でも、「濃いお茶」という常套句を表現する Strong Tea のような特定の語句の強い繋がりによる共起を連語という。

連語を獲得することで、語の共起による特別な表現を示す語用論の実態の把握や、文章の係り受け構造といった記述の構造解析を行うことができる。更に連語を計算機で自動的に抽出することにより、自動推敲や自動翻訳、検索エンジンにおけるサジェスト機能等の様々な言語の高度な運用に応用することができる。

連語を計算機で自動的に抽出するにあたり、問題点として一般的な判定手法・抽出手法がないことが挙げられる。また言語は独自の形態論をもっており、文章の解析を行う上で大きく依存するため、個々の言語に独立した定式化が困難で

ある。更に抽出した共起の中で連語とみなせるものがどれだけ含まれているか、またそれをどのようにして評価を行うかといった特徴選択は十分に検討する必要がある。

本研究ではデータマイニング手法と統計手法を併用し、必要に応じた連語を抽出する手法を提案する。これにより自動的に高速かつ高精度で連語を抽出可能であることを示し、言語の高度な運用として抽出した連語から潜在的に含まれる意味を解釈することが可能であることを示す。

2. 扱う問題

(1) コロケーションのためのデータマイニング

連語 (コロケーション) を計算機で自動的に扱うことにより、語用論の実態の把握や文章の構造解析、言語の高度な運用への応用等が可能になる。しかし、これを自動的に抽出する手法は定式化がなされておらず、また連語の抽出精度については殆ど議論がなされていない。

そこで自動的に高速かつ高精度での抽出を行うため、データマイニング手法を用いる。文章に出現する単語の形態素や共起する単語間の距離、出現する頻度を限定し、連語として不適切な要素をスクリーニングすることで連語の抽出を行う。

本研究では新聞記事を対象とし、大規模なコーパスからの連語を抽出する手法を示す。また既存の辞書と比較を行い、その抽出精度を検討する。

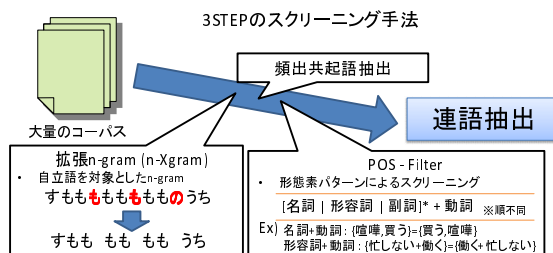


図 1 データマイニング手法による連語の抽出

(2) 統計指標を用いた日本語コーパスからのコロケーション情報抽出と精度評価

データマイニング手法を利用することで連語を自動的に抽出することが可能であるが、共起頻度によるスクリーニングでは 50%程度の精度と良好ではない [1].

そこで、抽出した共起に対し統計指標を用いた特徴選択を行う。コーパスの総語数 (総文章数) に対する共起語数のみで特徴選択を行う共起頻度と異なり、共起語の生起確率とその周辺確率を考慮して特徴選択を行うため、コーパスに依存することなく連語の強度を計ることが可能になる。

本研究ではダイス係数、自己相互情報量、T スコア、対数尤度比、条件付き確率により連語を抽出する手法を提案する。

(3) 条件付きコロケーションの抽出

連語において、構成する語が同時に発生して繋がると捉える「同時共起」に関してはこれまで盛んに議論がなされている [2]. しかし一方の語の出現がもう一方の語の出現を誘発すると捉える「条件付き共起」については殆ど議論がなされていない。

そこで共起する語群を支配語と被支配語の二つに分けて解釈し、支配語の出現を条件に被支配語が誘発して出現すると捉えた条件付き連語を検討する。例えば「馬」という支配語としたとき、被支配語として「嘶く」「食べる」「走る」など多数存在すると考えられる。一方で「嘶く」を支配語と捉えたとき、被支配語として考えられる語は「馬」のみであると予想できる。これにより、語の順序性を考慮した言語の高度な応用であるサジェスト機能やクエリ拡張などに有効であることが期待できる。

本研究では統計指標に対し条件付き確率を用いて再定義を行った条件付き統計指標として共起語の特徴選択を用い、条件付き連語を抽出する手法を提案する。また各条件付き統計指標により、抽出できる条件付き連語の傾向を検討する。

(4) 条件付き連語を用いた名詞句の文脈抽出

自然言語処理において、形態素解析や構文の解析など、文章に出現する語句を表層的に扱う研究はこれまで多くなされてきたが、語句が潜在的に持ち合わせている意味を捉え、それを抽出する議論は殆どなされていない。

そこで、条件付き連語 [3] における支配語と被支配語の関係をそれぞれ「説明される語句」と「その文脈を示す語句群」と捉え、語句の意味の抽出を行う。例えば条件付き連語により支配語として「巨人」、被支配語として「プロ野球、セリーグ、チーム」が抽出されたと仮定する。この場合は「巨人といえば、プロ野球のセリーグのチームである」といった文脈が獲得できているのではないかと考えられる。語句とその文脈を自動的に抽出することで、文書が示す内容のキーワードの抽出やラベル付けなど応用へが期待できる。

本研究ではデータマイニング手法と条件付き連語を用いて、名詞句に対しその文脈を表す語群を自動的に抽出する手法を提案する。また実験によりその抽出傾向と精度を検討する。

3. 結果

はじめにデータマイニングを用いた連語の自動的抽出について、6 か月分の新聞記事を用いて抽出を行った結果、再

現率は 70%であり、抽出できない正解は形態素解析に大きく依存することが分かる。また正解率は最大で 52%であり、中頻度での正解率が高くなっている。

統計指標を用いた抽出では、6 か月分の新聞記事と 5 種の統計指標を用いた結果、正解率は最高で 88%と共起頻度と比較して 30%の精度向上が可能となっている。また統計指標の中でも、ダイス係数、自己相互情報量、T スコアが良好であることが分かる。

次に条件付き連語の抽出では、条件付き統計指標の中でも条件付きダイス係数による抽出で 82%と高い精度での抽出が可能であると分かる。また条件付自己相互情報量は精度が下がるものの、共起後間の距離が増加しても安定した精度での抽出が可能であることが分かる。

最後に名詞句の文脈抽出について、既存の辞書で調べることが困難である固有名詞の文脈を表す語群が抽出可能であるとなっている。またその精度は頻度には依存せず、文脈をあらわす解釈語群に含まれる語の数に大きく依存することが分かる。

4. 結び

本研究では電子テキスト文書を計算機により自動的に処理することを目的として計算言語学のテーマの一つである連語に着目し、連語の自動的抽出とその精度の向上、条件付き連語の抽出、名詞句と文脈を表す語群の抽出が可能であることを示した。

データマイニングを用いた連語の抽出では、3 種の自動的スクリーニング手法を組み合わせることでこれが可能であることを示した。統計指標による連語の抽出と精度評価では、上記のスクリーニング手法に統計指標を加えることで、共起頻度と比較して高精度で抽出することが可能であることを示した。条件付き連語の抽出では、統計指標に対して条件付き確率に基づいた再定義を行い適用することで、抽出可能であることを示した。名詞句の文脈抽出では、既存の辞書で調べることが困難である固有名詞の文脈を表す語群が抽出可能であることがわかり、言語の高度な応用に有効であることを示した。

残る問題点として、言語により文章の構造が大きく変わることに対する手法の一般化を達成できなかったことが挙げられる。本研究では主に日本語を扱っているが、係り受けや修飾関係などの文章構造に依存した抽出手法ではないため、他言語への応用の可能性はあると考えられる。

参考文献

- 1) Takumi Sonoda, Takao Miura: Data Mining for Japanese Collocation, Seventh International Conference on Digital Information Management, 2012
- 2) Takumi Sonoda, Takao Miura: Mining Japanese Collocation By Statistical Indicators, 15th International Conference on Enterprise information Systems, 2013
- 3) Takumi Sonoda, Takao Miura: Conditional Collocation in Japanese, 18th Australasian Document Computing Symposium, 2013