様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　４月　４日現在

機関番号：１３３０１
研究種目：基盤研究（Ｃ）
研究期間：2010～2012
課題番号：22500126
研究課題名（和文）　一度しか出現しない単語の意味推定とその応用に関する研究

研究課題名（英文）　Research of Semantic Category Prediction of Extremely Rare Words and It’s Application

研究代表者
佐藤　賢二（SATOU KENJI）
金沢大学・電子情報学系・教授
研究者番号：10215783

研究成果の概要（和文）：非常に稀な言葉や新しく生まれた言葉は辞書にも登録されておらず、出現する文書も限られるため、コンピュータを用いて言葉の意味を推定するのが難しい。しかし、人間同士の会話では、そのような言葉でもわずか１～２文の情報から手掛かりを得て、自然に意味推定を行っている。本研究ではこのような言葉に対しても、構文や修飾などの情報を用いることで意味推定が可能であることを明らかにした。

研究成果の概要（英文）：Since an extremely rare word or a newly coined word is not in any dictionary and occurs only in a few sentences, it is difficult to computationally infer the semantic category of such a word. However, in humans’ daily communication, such an inference is naturally done from only a few sentences. In this study, it is shown that by using various information about sentence structure and modifier-modificand relationships, a semantic category of an extremely rare word could be inferred computationally.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
平成２２年度	1,300,000	390,000	1,690,000
平成２３年度	900,000	270,000	1,170,000
平成２４年度	500,000	150,000	650,000
年度			
年度			
総　計	2,700,000	810,000	3,510,000

研究分野：知能情報学
科研費の分科・細目：情報学・知能情報学
キーワード：人工知能，自然言語処理，画像、文章、音声等認識

１．研究開始当初の背景

現在の情報処理技術では、原則としてコンピュータは単語の意味を理解しない。例えば、インターネット検索で自分の思い通りの検索結果が返ってきたとしても、それは入力したキーワードの出現頻度や Web ページ間のリンク情報によりたまたまそういう結果が得られただけで、検索プログラムが言葉の意味を理解しているわけではない。Google が公開している N-gram データが端的に示すように、何十万何百万という単語の統計情報を収集することにより、より知的で高精度なテキスト検索やテキスト分類が可能にはなるが、現在の技術から一歩踏み込むためには、単語の意味を考慮した新しいテキスト処理技術を模索する必要がある。

一般的なテキスト処理で単語の意味を取り扱う場合、大きく分けると辞書を利用するアプローチと、統計情報のみを利用するアプ

ローチがある。前者の場合、専門家が作成した詳細な意味情報が利用できる反面、辞書登録されている単語の数が限られ、新しく出現した単語に即応できないなどの問題がある。後者の場合、カバーできる単語の数は飛躍的に増えるが、実際には統計情報から意味内容が推測できるのは、多くのテキストに出現する高頻度な単語に限られる。例えば、orange と grape はどちらも似たような形容詞の修飾を受け、似たような動詞の主語や目的語になるため、大量のテキストから頻度情報を計算すれば、２つの単語が同様の文脈で使われており、それゆえ意味的類似性が高いことが容易に分かる。しかしながら、十分な数のテキストに出現しない単語、例えば新しく品種改良されて命名されたばかりの果物の名前に対しては、このアプローチは全く通用しない。そして、単語の出現頻度がべき乗則（Zipf の法則）に従う以上、このような低頻度の単語の方が圧倒的に種類が多い。

一方、人間が行っている自然言語理解においては、当該の単語（意味推定の対象となる特定の未知語）が出現している文が２～３個しかなくても、容易に意味の推定が成立している。そこでは周辺の文脈情報を元にした多彩な推論が複合的に行われており、統計情報や意味は未知語よりもむしろ手がかりとなる周辺の単語にこそある。つまり、ある程度頻度が高く意味が確立しているような周辺の単語の情報を複合的に用いれば、１度しか出現しない単語（自分自身の統計情報が全く利用できない単語）に対しても、意味推定が可能になると考えられる。

以上の背景から、本研究では人間が行っている未知語の意味推定をコンピュータ上で模倣することにより、極端な低頻度の単語に対しても高い精度で意味推定を行う手法を着想するに至った。

２．研究の目的

一般的なテキスト処理で単語の意味を取り扱う場合、大きく分けると辞書を利用するアプローチと、統計情報のみを利用するアプローチがある。しかしながら、十分な数のテキストに出現しない単語、例えば新しく品種改良されて命名されたばかりの果物の名前に対しては、これらのアプローチは全く通用しない。そして、単語の出現頻度がべき乗則に従う以上、このような低頻度の単語の方が圧倒的に種類が多い。

一方、人間が行っている自然言語理解においては、当該の単語（意味推定の対象となる特定の未知語）が出現している文が２～３個しかなくても、容易に意味の推定が成立している。そこでは周辺の文脈情報を元にした多彩な推論が複合的に行われており、統計情報や意味は未知語よりもむしろ手がかりとなる周辺の単語にこそある。つまり、ある程度頻度が高く意味が確立しているような周辺の単語の情報を複合的に用いれば、１度しか出現しない単語（自分自身の統計情報が全く利用できない単語）に対しても、意味推定が可能になると考えられる。

以上の背景から、本研究では人間が行っている未知語の意味推定をコンピュータ上で模倣することにより、極端に出現頻度の低い単語、究極的には一度しか出現しない単語の意味を、高精度に推測する手法を開発することを目指した。

３．研究の方法

本研究ではまず、意味推定の基礎として頻度の高い単語の意味を確立する作業を行った。ここで、単語の意味とは直観的には意味カテゴリ（例えば orange なら果物、nurse なら職業、など）を指す。ここでは主に名詞、形容詞、動詞を対象として、統計情報を元にクラスタリングを行い、高頻度単語の意味カテゴリを構築した。計算の元になるテキストコーパスは、一般的なものと分野固有のものを用意し、前者については Google Web 1T 5-gram と Wikipedia、後者については Medline を選択した。これに関連して、大規模なテキストデータから自動的に意味カテゴリを構築するための基礎技術として、引力に基づく新しいクラスタリングアルゴリズムを開発した。

次に、構築した意味カテゴリの精度を評価するために、２つの文の類似度を計算する予備的な実験を行った。実験では、文を主語・述語・目的語の３項組として単純化し、２つの文に含まれる主語同士・述語同士・目的語同士の類似度を計算した上で、サポートベクターマシンを用いて類似文と非類似文の分類を行った。この分類実験には Microsoft Research Paraphrase Corpus を用いた。

次に、構築した高頻度単語の意味情報と、未知語が出現するテキストの情報から、未知語の意味カテゴリの候補を絞り込む手法を開発することを試みた。まず、未知語を含む文の構文解析結果を元に、主語・述語・目的語の関係から未知語の意味カテゴリがどの程度推定可能かを調べた。対象としては Medline から抽出した生物医学文献アブストラクトを用い、主語または目的語となる名詞句の意味カテゴリ情報として Genia Corpus を用いることにより、学習および予測のための特徴ベクトルを生成した。学習器としてサポートベクターマシンおよびナイーブベイズを用いて実験を行った。

最後に、未知語を修飾している名詞句の構造（形容詞の情報）を用いた意味カテゴリの

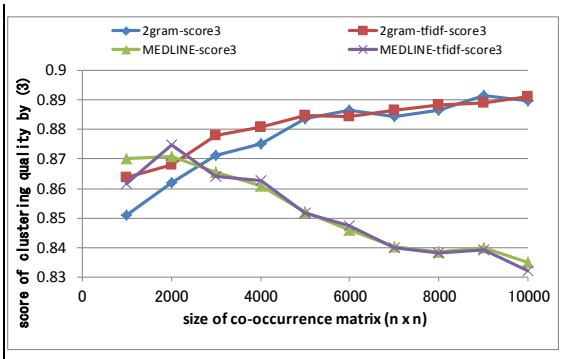
絞り込みに関する研究を行った。ここでは、修飾語である個々の形容詞が、被修飾語である名詞の意味カテゴリをどの程度限定できるかを数値化することにより、意味推定に役立つ形容詞とそうでない形容詞を明らかにした。実験の際、2011 年に BioMed Central で発表された約１万本の論文のフルテキスト情報をテキストコーパスとして用い、約 100 万個のセンテンスから名詞句を抽出した後、形容詞＋名詞の２ワードから成る名詞句に絞り込み、修飾・被修飾の関係についてクラスタリングを行うことで、意味的に類似した名詞のクラスタが得られることを確認した。さらに、修飾する（あるいは修飾を受ける）単語の種類数と頻度に基づいたスコアを用いることにより、形容詞の意味的決定能を自然に数量化できることを明らかにした。

４．研究成果

構築した意味カテゴリの精度を評価するために、２つの文の類似度を計算する予備的な実験を行った結果、コーパスから自動構築した単語の意味カテゴリと対象文の主語・述語・目的語しか用いていないにも関わらず、辞書等の情報を用いた従来手法に匹敵する精度が得られることが分かった（表１）。その一方で、Medline コーパスのみを用いて構築した意味カテゴリよりも、Google Web 1T 5-gram を用いて構築した意味カテゴリの方が、特に低頻度単語において信頼性が向上することが分かった（図１）。与えられた文と意味的に類似した文を高速検索するシステムのプロトタイプを作成し、実用性について検証した。また、意味カテゴリの構築に関連して開発した新しいクラスタリングアルゴリズムでは、物理現象を模倣してサンプル間に引力を定義し、サンプルの移動を繰り返すことにより、従来よりも高精度かつロバストにクラスタリングを行えることが分かった。

（表１）

	Acc	Prec.	Rec.	F.
MEDLINE (idf 処理無し)	67.8%	69.6%	91.7%	79.1%
Wikipedia (idf 処理無し)	68.7%	70.8%	89.9%	79.2%
Wikipedia (idf 処理あり)	70.0%	72.0%	89.8%	79.9%
従来手法	70.3%	69.6%	97.7%	81.3%



（図１）

次に、構築した意味カテゴリと、未知語を含む文の構文解析結果を元に、主語・述語・目的語の関係から未知語の意味カテゴリがどの程度推定可能かを調べた結果を示す（表２、３）。実験の結果、この問題に対しては学習器としてサポートベクターマシンを用いた方が高い予測精度が得られること、９つの意味カテゴリから１つを予測する問題で約６５％の精度が得られること、主語の予測よりも目的語の予測の方がやや高い精度が得られること、動詞の情報が重要であること、などが分かった。今回は動詞の意味カテゴリや単語の意味カテゴリ階層の情報を用いていないこと、名詞に関しては主語や目的語の名詞句の意味的主辞となる単語しか用いていないことなどを踏まえると、この予測精度は十分高いと考えられる。

（表２．主語の意味カテゴリの正解率）

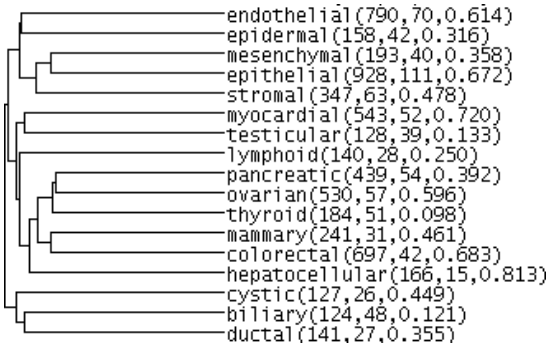
使用した特徴	NB	SVM
目的語、目的語の意味	50.0%	50.3%
述語、目的語、目的語の意味	55.9%	62.6%
全ての特徴（主語関係以外）	56.1%	61.9%

（表３．目的語の意味カテゴリの正解率）

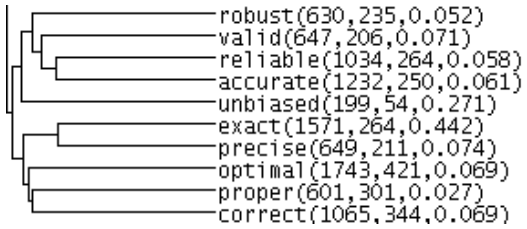
使用した特徴	NB	SVM
主語、主語の意味	51.8%	48.8%
述語、主語、主語の意味	58.2%	64.6%
全ての特徴（目的語関係以外）	57.7%	63.8%

最後に、名詞句に含まれる形容詞の情報を用いた名詞の意味カテゴリの絞り込みに関する実験結果の一部を示す。図２に示す形容詞の大部分は、それが修飾する名詞の意味カテゴリが非常に限定されるため、例えば colorectal ならそれが修飾する名詞の意味カテゴリはがん関係であることが高い精度で分かる。その一方で、図３に示す形容詞の大部分は、非常に多種類の名詞を修飾するため、殆ど意味カテゴリの決定能を持たない。このように、どの形容詞がどの程度の意味的決定能を持つかを数量化することに加えて、修飾される名詞がどのような意味カテゴリのバリエーションを持ち得るかについても明ら

かになった。今後はこれらの結果を総合し、実用的な意味推定システムの開発を目指す。



（図２）



（図３）

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Tran,V.A., Clemente,J.C., Nguyen,D.T., Li,J., Dang,X.T., Le,T.T.K., Nguyen,T.L.A., Saethang,T., Kubo,M., Yamada,Y., Satou,K., (IMPACT: A Novel Clustering Algorithm based on Attraction), Journal of Computers, 査読有, Vol.7, No.3, 2012, 653-665 DOI:10.4304/jcp.7.3.653-665

〔学会発表〕（計３件）

① Satou, K. , (ANALYSIS OF NOUN PHRASES EXTRACTED FROM BIOMEDICAL TEXTS FOR SEMANTIC CATEGORY PREDICTION), The Pacific Symposium on Biocomputing 2013 (PSB2013), 2013.1.5, Fairmont Orchid (Hawaii)

② 武田徳明, 佐藤賢二, (構文情報を用いた名詞の意味推定), 平成 24 年度電気関係学会北陸支部連合大会, F2.情報処理・その他, F-118, 2012 年 9 月 2 日, 富山県立大学（富山県）

③ Satou,K., (SEMANTIC CLUSTERING OF BIOMEDICAL WORDS USING GOOGLE WEB 1T 5-GRAM), The Ninth Asia-Pacific Bioinformatics Conference (APBC2011), APBC2011 Conference, P285, 2011.1.12, Songdo Convensia (Incheon)

６．研究組織
(1)研究代表者

佐藤　賢二（SATOU KENJI）
金沢大学・電子情報学系・教授
研究者番号：10215783