様式 C-19

科学研究費補助金研究成果報告書

平成　22 年　6 月　10 日現在

研究種目：　基盤研究（C）
研究期間：　2007　～ 2009
課題番号：　19500135
研究課題名（和文）　専門分野テキストコーパスからの体系化された用語抽出

研究課題名（英文）　Structurized Term Extraction from Academic Text Corpora

研究代表者
小山　照夫　（KOYAMA TERUO）
国立情報学研究所・情報社会相関研究系・教授
研究者番号：80124410

研究成果の概要(和文):本研究ではまず､専門分野テキストコーパスに出現する複合語用語を、高い精度で網羅的に抽出する手法を確立した。日本語名詞形態素の内で、複合語構成上制約のあるものを整理し、また、形態素解析誤りの影響を受けやすい部分からの候補抽出を避けることによって、一定の抽出精度を保ちながら、コーパス内出現頻度の低いものまで網羅的に用語を抽出することが可能となる。本研究ではまた、複合語の入れ子関係を用いた用語の階層的構造化と、部分研究領域に強く関連する形態素を選択し、これらの形態素を要素として含む複合語を選択することにより、部分研究領域に関連づけた用語体系化が可能となる事を明らかとした。

研究成果の概要（英文）: In this study, we established a method for comprehensive term extraction from domain text corpora with high precision. The method is based on basically two new principles. One is the reconsideration and modification of Japanese morpheme classification, and another is the evaluating the certainty of composite boarders. We also have developed methods to structurize terms from two points of view, namely, nesting relations between composites, and the term relationships to various research subdomains.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,200,000	360,000	1,560,000
2008 年度	1,000,000	300,000	1,300,000
2009 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：用語抽出、用語体系化、自然言語処理、形態素解析

１．研究開始当初の背景

(1) 専門分野の文献を高度活用するためには文献で用いられる用語を体系化して整理する必要性があるが、人手で用語を抽出する方法では十分な整備を行うことが難しい状況である。

(2) 一方、自然言語処理の応用としての用語

抽出については、様々な研究の成果から、完全な自動抽出は困難であるとしても、用語集整備の手段として活用できる可能性は高いと判断できる状況であった。

(3) また、研究開始までの我々の研究から、複合語の入れ子関係が用語間の関係整理に利用可能であること、また、特定領域についていくつかの部分研究領域を考えることができる場合、部分領域を代表する形態素との共起関係を用いることにより、用語候補を部分領域に関連付けて整理できる可能性のあることも予想された。

(4) このような状況から、完全な自動化は困難であるとしても、分野テキスト集合から高い精度で網羅的に用語候補を抽出し、また、抽出された用語の相互関係や利用される文脈などの視点から構造化の示唆を与えることのできるシステムが実現できるならその価値は大きいと考えられた。

２．研究の目的

本研究では以下の項目を目的とする

(1) 日本語研究文献コーパスから、複合語用語を、一定以上の抽出精度を確保しながら、網羅的に抽出する方法論を確立する。

(2) 特定研究文献コーパスについて、共起関係に基づき、領域に含まれる部分研究領域を同定するとともに、抽出された用語候補を部分領域に関連付けることにより、用語候補を体系的に分類する方法を確立する。

(3) 抽出された用語候補について、複合語の入れ子関係を利用して候補間の関係を推定し、用語候補を階層的に体系化する方法を確立する。

３．研究の方法

(1) 用語抽出：日本語では用語の多くは複合語の形を取り、また、複合語として現れる語の多くは何らかの研究分野の用語となっている場合が多い。これに対して句構造が用語として用いられることはむしろ稀である。日本語の複合語は名詞性の形態素連接の形をとると考えられるところから、日本語テキストからの用語抽出において網羅性の高い用語抽出を行うことは容易なはずである。しかし、実際には分かち書きを行わない日本語テキストでは、形態素解析にある程度の誤りを生じることは避けられず、また、現在存在する形態素辞書では形態素分類が、複合語抽出を行う上で必ずしも十分ではないという問題がある。結果として単純に名詞系の形態素連接を取りだしたのでは、精度の低い結果しか得られないという問題がある。

精度の低い候補集合から用語として成立する可能性の高いものだけを選択する方法として、従来は統計的指標を用いる方法が提案されてきた。しかしながら統計的指標が有効性を持つためには、候補がコーパス内で一定程度以上の出現頻度を持つことが要求される。実際にコーパス内の用語性が認められる文字列を調べてみると。頻度の低いもの、極端にいえばコーパス内に一度しか出現しないものであっても用語性の高いものが認められるのみならず、頻度の低いものほど、種類が多いという傾向がみられる。このような状況の下で、統計的方法だけに基づいて、精度を確保しつつ網羅的な用語候補抽出を行うことは困難であると考えられる。

本研究では、網羅的な用語候補抽出を目指すという観点から、統計的指標に全面的に依存するのではなく、用語候補の外形的構造と、その前後に対する接続関係を手掛かりにした用語抽出を試みる。

日本語複合語抽出にあたって、不適切な候補を抽出する要因を検討した結果、日本語形態素の分類に不十分なところがあることと、形態素解析の誤りにより不適切な候補が取りだされることの二点が主要な問題となっていることが明らかとなった。そこで用語候補抽出の基本的方針として、形態素辞書の上では接尾辞、接頭辞、副詞可能名詞などの名詞系形態素と分類されているものの内、問題があると考えられるものについて、特別な扱いを必要とするものをリストの形で管理することにより、少なくとも頻度の大きい形態素については、その影響を緩和することを試みる。また、本来名詞的使用が考えにくい動詞連用形についても、それらをリストとして管理することにより、用語候補の構成要素としては認めないこととする。

形態素誤りの影響としては、候補形態素列に含めることが不適切な名詞系形態素と判定される文字列が出現することと、候補形態素列の前後境界が不適切なものを抽出してしまうことが挙げられる。これについては特に影響の著しい、ひらがな一文字の名詞形態素を候補の構成要素としないこと、また、候補として得られた形態素列の前後にどのような形態素が出現するかを検査し、特定の場合には形態素列を候補とはしないという抽出方法を試みた。

従来の用語抽出研究では、用語候補の頻度を重視する観点から、他の候補の中に入れ子になった形態素列も候補としての可能性を調べる方法を採用している。しかし、コーパス中に一度も独立した形で出現しないものを候補とすることには疑問が残る、今回の方法では、独立して最長の候補形態素列として少なくとも一度はコーパス中に出現する物のみを抽出することとしている。

(2) 部分研究領域との関係：本研究に先だって行った、コーパスに特有な名詞形態素とサ変名詞形態素の共起に基づく部分研究領域推定を試みる。また、推定された部分研究領域に特徴的な文書集合を決定し、その集合に特異的に出現する用語候補を選択する方法の効果を確認する。

この方法は有効であると期待できるが、一方で統計的指標に依存する関係から、コーパス内生起頻度の低い候補の判定については有効性に疑問が残る。そこで、同様の統計的指標の適用ではあるが、形態素のレベルでその部分領域との関連性を求める方法を新しく検討した。形態素と部分領域との関係も、実際には統計的指標を用いなければ評価することは困難だが、用語候補と比較すると出現頻度として圧倒的に多数であることが期待できる形態素レベルで関連性が明らかになるならば、その信頼性は低頻度の用語候補と比較して高いことが期待できる。

形態素と部分研究領域との関連の強さを評価するためには、これまでの研究と同様に、サ変名詞出現パタンによって文書を分類する方法も考えられるが、今回は部分領域に特有と考えられる文字列を人間によって指定させ、指定された文字列と各形態素との共起傾向を評価指標として用いることを試みた。

(3) 入れ子関係に基づく用語候補間の階層関係整理：特定の複合語用語が他の用語を入れ子として含む場合、その両者の間には意味的関係があると考えられる。ここで、入れ子として含まれる側が含む側の先頭部分(tail)になっているか、あるいは最終部分(head)になっているかで意味関係は変わってくるし、また、付加された部分がどのような性質を持つかでも関係は変わってくることが予想される。そこでこの二つの観点から、入れ子関係にある二つの用語候補の間の関係整理を試みる。

４．研究成果

(1) 用語抽出：今回提案する手法を、NTCIR-I 学会発表抄録データに適用した。学会抄録データに収録されているデータの内、情報処理学会研究会抄録、約 27,000 件を対象としている。各抄録はタイトルを加えた平均文字数 290 文字、文字数の標準偏差は 74.7 文字であった。

この文書集合に今回提案する用語抽出手法を適用した。候補の前後接続関係をチェックしない方法と、前後接続関係をチェックする方法とを適用した場合を比較する。結果は抽出候補数と、抽出された候補集合からランダムに 500 候補を取りだし、用語となる可能性があるかどうかを人手で判定し、抽出精度を評価した。結果は次の通りである。

前後検査あり、頻度制限なし
　抽出数：130,876　　精度：84.6%
前後検査なし、頻度２以上
　抽出数：　46,609　　精度：85.8%

候補前後の検査を行わない場合、全候補を抽出すると精度が低下する傾向がみられる。候補をコーパス内生起頻度２以上のものに限定すると、精度は向上するが、一方で抽出可能な候補数は当然ながら減少する。精度評価の観点からは、候補前後の検査を行わず、生起頻度を２以上とした場合と、候補の前後検査を行った上で全候補を抽出した場合の精度はほぼ同等である。一方で抽出候補数は約 2.8 倍に増加している。この結果から、抽出精度を確保しながら、網羅的に複合語用語を抽出する方法論が確立できたと結論できる。

海外の研究では、Daille 等がフランス語用語の抽出に、外形的基準を主として用いた試みを行っている。言語や対象分野の相違から単純な比較はできないが、網羅性や精度については同等以上の結果が得られていると考えられる。その他の海外の研究では、統計的指標の利用が主となっているが、精度向上は認められるものの、網羅性という視点からは今回我々が採用した手法が有利であると考えられる。

今回の結果は、用語抽出結果をそのまま用語集の見出しにできるわけではないが、全て人手でコーパスから用語を選び出すことと比較すれば、大幅な用語選択の効率化が図れると判断できる。

(2) 部分研究領域との関係：我々の先行研究で主要名詞形態素と、主要サ変名詞形態素の共起傾向を用いたクラスタ分析により、主要な部分研究領域のいくつかを同定し、用語分類が可能になることを示してきたが、情報処理学会抄録コーパスについても、同様の結果が得られることが確認された。このことからこの手法は分野を限定せず、広く適用可能であることが示された。

一方、形態素レベルでの部分領域を特徴づける要素の同定では、情報処理分野の文献を他の分野の文献と比較することにより、予め分類すべき形態素を情報処理分野に特有なものに限定することにより、想定する部分領域に特に関連の強い形態素を効果的に同定できるようになった。この手法を適用することにより、指定する研究領域に関連する用語を幅広く収集できることが明らかとなった。

部分研究領域に関連付けた用語分類の試みはほとんど行われてきていない。しかしながら、用語の体系化という視点からは有効な

手法と考えられるところから、今後さらに発展が期待される。

(3) 用語の階層的整理：二つの複合語用語候補が入れ子関係にある場合、入れ子となっている候補が入れ子を含む候補の先頭部分に一致すれば、入れ子を含む候補は入れ子となる候補の下位語として位置づけられ、逆に末尾部分に一致するなら関連語として位置づけられる。今回は、入れ子を含む候補の部分列として最長のものだけを検討した。最長の部分列となる候補が、先頭にも末尾にもならない場合も存在するが、このような場合のほとんどでは入れ子となる候補に、入れ子として含んでいる候補の末尾部分を含めても用語として成立するため、このような形で拡張した候補を用いて階層的整理を行う。

実際に整理を行った結果では、上位－下位語関係や関連語関係は、ほぼ想定通り成立していることが明らかとなった。しかし一方で、一概に下位語と言っても、さまざまな階層の様相が存在することも明らかとなった。このような多様な関係を整理しないままでまとめてすべて示したのでは、必ずしも理解しやすい体系化にはならない可能性がある。入れ子となる候補と入れ子を含む候補の差分部分の形態素の種類に応じて整理すると、やや見やすい関係が得られるが、未知語などが入ってくると必ずしも明快な分類ができなくなることもある。

入れ子関係に基づく関係整理では、一定の関係性が明らかになってはいるが、十分な体系化ができたとは言い難い。今後は形態素の分類見直しや、複合語内部での形態素間係り関係の解析などを行い、より整理された形での階層関係整理を行う必要があると考えられる。

全体としてみるなら、複合語用語抽出に関して、網羅性が高く、かつ抽出精度を落とさない方法を確立したこと、形態素レベルで部分研究領域との関連を利用することにより、抽出された用語の分類を行う可能性を示したことが今回の研究の成果であるといえる。一方、入れ子関係に基づく階層的整理では、整理の視点をより明確にする必要があり、そのためには形態素分類のより詳細な検討が必要となるという課題が明らかになった。また、例えば用語集編纂などの目的に応用しようとする場合には、完全自動化には無理があるところから、人間の作業を支援する環境の整備も重要な課題となることが予想される。

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) 小山照夫、日本語テキストからの複合語用語抽出、情報知識学会誌、査読無、vol.19、No.4、pp.306-315、2010.

〔学会発表〕（計　4件）
(1) 小山照夫、竹内孔一、日本語複合語用語の入れ子関係に基づく階層的体系化、信学技報、vol.107, no.158, NLC2007-9, pp.49-54, 電子情報通信学会、2007.7.24、徳島大学.

(2) 小山照夫、竹内孔一、用語クラスタリングに基づく部分研究領域推定と用語分類、情処研報 vol.2008, 2008-NL-183, pp.87-92, 情報処理学会、2008.1.22、国立情報学研究所.

(3) 小山照夫、竹内孔一、形態素出現パタンに基づく文書集合類似性評価、情処研報 vol.2008, 2008-NL-188, pp.51-56,情報処理学会、 2008.11.26、九州大学.

(4) 小山 照夫、竹内 孔一、候補の接続関係を考慮した複合語用語抽出、情報処理学会研究報告、SIGNL-193, pp. 13/1-6, 情報処理学会、2009.9.29、京都大学.

〔その他〕
ホームページ等
http://research.nii.ac.jp/~koyama/official/tmrec/

６．研究組織
(1)研究代表者
小山　照夫（KOYAMA TERUO）
国立情報学研究所・情報社会相関研究系・教授
研究者番号：80124410

(3)連携研究者
竹内　孔一（TAKEUCHI KOICHI）
岡山大学大学院・自然科学研究科・講師
研究者番号：80311174