日本語テキストから用語を網羅的に取り出します

候補の文中位置を考慮した日本語テキストからの用語抽出

Term extraction based on the connective relations of candidates

発表者：小山　照夫（情報社会相関研究系）
by:Teruo KOYAMA (Information and Society Research Division)

何がわかる？

日本語文中に出現する複合語用語候補について、その前後に位置する形態素の性質によって、境界の信頼性が変わる性質を考慮することにより、精度を低下させることなく、より網羅的な用語抽出ができるようになります。

どんな研究？

学術研究文書を高度利用するためには、文書を記述するために用いられる用語を整理することが重要です。この研究では文書集合から網羅的に用語候補を取り出すことを目的としています。

研究の目的

特定分野の学術文書集合では、どのような用語が用いられているのでしょうか。

特定学会の研究抄録など、当該研究分野の文書集合が与えられたとき、文書集合中にどのような用語が出現しているのか、可能性の高い複合語の集合を網羅的に抽出するための方法論を明かにします。

研究の方法

日本語では用語の多くは複合語の形を取ります。また複合語は多くの場合特定分野の用語となっています。したがって日本語では複合語を正しく取り出すことができれば、高い精度で網羅的に用語が抽出できると期待されます。

日本語の複合語は名詞的形態素の連接という形を取ることから、複合語を正しく取り出すことは基本的には容易なはずです。しかし実際には、名詞系形態素の分類が必ずしも十分ではないこと、また形態素解析にはある程度誤りが起きることから、いくつかの処理をして抽出精度を高める必要があります。

名詞系形態素の分類については、複合語の要素となることができないもの、複合語の先頭また、末尾にくることができないものを、リストの形で管理します。

また、形態素解析誤りの影響を緩和するため抽出された複合語候補の前後に位置する形態素に制約を設け、複合語の境界が確実なものだけを選び出します。

以上の工夫をすることにより、用語とならない形態素列の抽出を抑えながら、しかも網羅的にテキストコーパスから用語を抽出することが可能となります。

提案する手法の有効性を確認するため、情報処理学会研究発表抄録から用語を抽出した結果を示します。

ここでは候補の前後接続関係を調べることの効果を評価するため、比較として、候補の前後接続関係の検査を省略するかわりに、候補のコーパス内出現頻度を2以上に限定した時の結果を併せて示します。

候補の前後を確認することにより、抽出精度をほとんど低下させることなく、抽出候補数を3倍近くに増やすことができます。

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%

NII　連絡先：　小山　照夫(Teruo KOYAMA)／国立情報学研究所　情報社会相関研究系　教授
TEL: 03-4212-2509　FAX: 03-3556-1916　Email: t_koyama@nii.ac.jp