理工学専攻修士論文要旨

提 出 年 度： 2005
提　出　日： 2006/02/27
専 修 コ ー ス： 知能情報コース
学 生 番 号： 35604005
学 生 氏 名： 安部　建助
研究指導教員： 原田実教授

（論文題目）
固有表現抽出・分類システムの開発研究

（内容の要旨）
昨今，インターネットや新聞などの，いわゆる「知識の宝庫」と呼ばれるものには，質問応答などの自然言語処理を行う上で有効な情報が大量に含まれている．しかし，自然言語処理を行う上で「固有表現」は，その解析精度を落とす要因となり，「固有表現」の語意を正しく捉えることが非常に有意義であると考えられる．そこで本研究では，本研究室が使用している電子化辞書 EDR に，1)語意の定まった固有表現一覧から固有表現を登録する，2)固有表現と思われるものを含む文中から，文脈を考慮してその語意を推定し登録する，という 2 手法で未収録の固有表現の語意を EDR に登録し，より高精度な辞書の構築を目指す．

これまでの固有表現処理では，予め用意した語の並びのパターンを表層的に日本語文章に適用し，そのパターンに適合した文字列を固有表現として抽出する手法が多くとられてきた．しかし，このような手法では文章の意味理解を行っていないために，語の意味レベルでの解釈を必要とする判定は困難となっている．この問題に対し本研究の 2)では，意味解析システム「Sage」によって日本語文章を意味解析した意味グラフを入力とすることで，語を意味レベルで解釈したものを基に固有表現の語意推定処理を行うことを可能にする．

本研究の具体的な処理として 1)については，インターネットや書籍などから固有表現の一覧情報を獲得し，その一覧を EDR 日本語単語辞書・概念辞書に追加するための情報を持たせたレコード情報に加工して登録することを行う．2)については，意味グラフを入力として，

ⅰ)固有表現候補抽出
ⅱ)共起関係詞＆受け側主辞獲得
iii)全文検索システム Namazu を用いての毎日新聞検索
iv)EDR 共起辞書検索
ⅴ)上位概念集計からの語意推定

という 5 ステップで行う．ⅰ)では，Sage が解析した結果，EDR 概念辞書に未収録で語意が不明な語でありかつそのリファレントが存在する語を固有表現候補（以降，未知語とする）として抽出する．ⅱ)では，ⅰ)で抽出された未知語 frame の共起関係詞とその受け側 frame の主辞を獲得する．iii)では，当該未知語表記を含む記事を，毎日新聞過去 4 年分から最大 50 個検索し，獲得した記事の意味グラフ内で，ⅱ)と同様に当該未知語 frame の共起関係詞とその受け側主辞を獲得する．iv)では，ⅱ)，iii)で獲得した共起関係詞と受け側主辞の語意を用いて EDR 共起辞書を引き，その共起事例の中から語意を持つ（EDR 概念辞書に既収録の）係り側を獲得する．最後にⅴ)では，iv)で獲得した複数の係り側がそれぞれ，予めこちらが定めた固有表現の語意推定のための 17 個の概念候補の内のどれを上位に持つかを探索，集計する．そして，その中で最も多かった概念候補を当該未知語に与える語意の妥当な上位概念として EDR 日本語単語辞書・概念辞書に登録することを行う．

上記の手法を行うことにより，1)では湖や株式会社など，合計 7623 語の固有表現を登録した．2)ではインターネットから抽出してきた 41 文章の記事に対して処理を行い，64%の精度で固有表現に対し妥当な語意を推定することができた．