様式 C-19

科学研究費補助金研究成果報告書

平成２１年　４月１０日現在

研究種目：基盤研究（Ｂ）
研究期間：2006～2008
課題番号：18300051
研究課題名（和文）　語構成を考慮した多言語の語彙および用語解析システムの研究開発

研究課題名（英文）　Research and Development of Multi-lingual Lexical Analysis System based on Word Construction

研究代表者
松本　裕治　（MATSUMOTO　YUJI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：10211575

研究成果の概要：日本語および中国語の複合語の語構成分類を統語および意味関係に基づいて行い，内部構造情報の記述法を考案した．また，複合語の内部構造記述のアノテーションを行うためのインタフェースと検索機能を備えた辞書管理システムを開発した．従来開発してきた日本語および中国語の辞書項目の拡張を行うとともに，新たに，複合語情報を含む英語辞書を開発し，それぞれ研究目的に自由に使える辞書として公開した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	5,000,000	0	5,000,000
２００７年度	4,800,000	1,440,000	6,240,000
２００８年度	4,500,000	1,350,000	5,850,000
年度			
年度			
総　計	14,300,000	2,790,000	17,090,000

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：自然言語処理，語彙解析，辞書，多言語処理，形態素解析

１．研究開始当初の背景

コンピュータによる言語の解析においては，単語の認定と品詞の同定が最も基本的な処理である．しかし，日本語や中国語などでは，文の表記において単語を区切るためのデリミタを用いないため，単語への分かち書きの曖昧性が問題である．例えば，中国語については，ACL(Association for Computational Linguistics)の中国語処理部会が開催している SIGHAN Workshop の中で，単語分かち書きを共通のタスクとするコンテストが行われていた．2005 年 10 月の会議では，４種類の中国語コーパスについて，訓練データとテストデータが準備され，共通のデータでのシステム比較が行われた．これらのコーパス間には分かち書きの認定基準に隔たりがあり，ある訓練データで学習したシステムを用いて他の種類のコーパスを解析した結果は，同種のテストデータの解析結果に比べて著しく精度が劣ることが報告された．このようなことが生じる最大の原因は，単語分かち書きの基準がコーパス毎に異なることにあった．この問題は，中国語や日本語だけの問題ではなく，英語においても，連語，熟語，固有名詞，専門用語のように複数の語がつなが

って初めて意味を持つ一連の表現（複合語）があり，また，屈折や派生のように，語の接頭，接尾などの断片を解析して初めて語の役割や意味が明らかになる表現がある．文の分かち書きの問題は，種々の基準が雑多に存在するのが困難さの要因ではなく，それぞれの応用や目的によって単語（あるいは用語）の構造のどの部分までを基本要素として捉えるかという考え方が異なることにその要因がある．

よって，問題の本質は，分かち書き標準となる単一基準を考えるのではなく，語や用語がもつ語構成とその分類を行い，語構成の中で，どのような構成までをまとめ上げや区切りの単位と考えるかを明確に指定できる構造として解析しておくことによって，分かち書きを考えるということにある．例えば，複数の語がまとまることによって統語的な振る舞いを異にする場合（“ with respect to ”のような英語の前置詞相当表現や，「に関する」のような日本語の助詞相当表現など）を連語としてまとめるという視点もあるし，構成要素となっている単語からは全体の意味が取れない複合表現をまとめあげるという視点もある．このような基準をわかりやすく特定できるような構造を持った辞書，用語データベースを構築する手法や基本技術を蓄積することが重要であると考えた．

２．研究の目的

これまであまり明確に規定されてこなかった分かち書きや用語の切り出しの問題を，上で記述した考え方にしたがって解決するために，次のような機能を備えた辞書および用語解析システムを研究・開発することを目標とした．

(1)屈折・派生などによる語変形，複合語の構成に関する分類と分類基準の策定
(2)複合語や連語の語構成に関する明確な情報を記述した辞書の開発
(3)辞書に登録されていない複合語，専門用語，固有名詞などの多単語からなる表現を文中から特定する手法の開発とその実装

(1)については，屈折や派生などの語変形に関しては言語毎に既存の研究があり，辞書や事例を調査することにより網羅的な分類と記述が可能である．複合語については，辞書に登録すべき語には様々な観点からの分類が可能である．例えば，英語の“ in short ”，“ at last ”などは，２つの語がまとまって一つの副詞のように機能し，統語的な観点からの連語である．寝台車の意味の ”sleeping car ” は，意味的な連語であり，構成的に意味を取ることが難しく辞書に登録すべき語である．その他，縮約，句動詞，熟語，連濁現象のように読みの変化を伴う連結現象，など，複合語の構造についての網羅的な分類を行うことが重要である．分類された語構成に従って，(2)では，様々な視点から複合語として登録すべき語を大規模なテキストデータから抽出し，各語の語構成の明確な記述を行った代表的な辞書を作成する．このために，語構成解析や解析結果の確認・修正を行うためのツールを同時に開発する．(3)では，複合表現となっている固有表現などの候補を自動抽出し，前後文脈より，それらが真に複合的表現であるか，また，どのような構造をもった表現であるかを自動解析，あるいは，解析支援する手法の開発を目的とする．この処理が必要なのは，どのように辞書を充実させようとも，すべての可能な複合表現を網羅することはできないからである．単純な例として，例えば，「日本人」を「日本」という国名と「人(じん)」という接尾辞から構成される複合語として辞書に登録することを仮定する．一旦これを行うと，「中国人」「アメリカ人」などあらゆる国名に対して同様の表現を登録する必要があるが，それは現実的に不可能である．このような複合語を一切登録しないという立場もある．しかし，それを徹底して行うと，上記のような連語をほとんど登録することができなくなり，様々な自然言語処理応用にとって使いにくい辞書となってしまう．(3)で行うのは，(2)の辞書に登録されている複合表現（あるいは，利用者が特に指定した種類の複合表現）と同様の構造をもった語が新しい文書中に現れた際に，それらを動的にまとめ上げる汎用的な用語解析システムを構築することである．用語解析システムは，機械学習に基づいた手法により実装する予定であり，並行して，複合表現や用語の事例集を作成する．

このような言語資源と言語処理システムを構築することにより，語構成を考慮した系統的な語彙・用語解析か可能となり，種々の応用目的に合致した汎用性の高い単語分かち書きと品詞・意味分類の処理を可能にすることが本研究の目的である．

３．研究の方法

本研究は，主に次の３つの項目に分けて研究を行った．

(1)複合語の語構成の整理と分類と分類基準の策定
(2)言語解析用辞書の語彙項目の拡充および各登録語の内部構造記述を行うための辞書管理システムの開発
(3)複合表現・専門用語の内部構造自動解析手法の開発と解析ツールの設計

複合語の語構成の整理については，統語的な構造と意味的関係の２つの視点から分類を行った．特に中国語の複合語について，こ

れらの両視点から整理した．日本語については，統語的視点からのみ分類し，４つの統語的係り受け関係を規定した．(2)に関しては，語の品詞や読みなどの基本情報以外に，意味ラベルなどの付加情報，および，内部構造を木構造によって表現できる辞書管理システムを開発する．このシステムには，語の内部構造をアノテーションするためのインタフェースを構築し，これを用いて内部構造記述の作業を行う環境を整備し，具体的なアノテーション作業を通じてシステムの仕様改善を行う．(3)に関しては，人手でアノテーションを施した用語集を学習データとして，機械学習に基づいた用語の内部構造自動解析手法を設計する．特に，複雑な専門用語等で生じる文字単位の縮約現象に対応するため，文字レベルの係り受け解析によって，用語の内部構造解析を実行する手法を開発する．

４．研究成果

日本語および中国語の複合語の語構成分類を統語および意味関係に基づいて行い，内部構造情報の記述法を考案した．また，複合語の内部構造記述を行うためのインタフェースと検索機能を備えた辞書管理システム「Cradle」を開発した．本研究で行った具体的な研究成果は以下の通りである．

(1)統語および意味分類にもとづく複合語の構造の分析と複合語内部構造記述法の設定：中国語の複合語を構成する構成語間の統語および意味関係に基づく分類を行った．また，日本語の複合語については，主に統語的構造の視点から４種類の係り受け関係を用いた内部構造記述を行った．日本語専門用語に対して，文字単位による係り受けによって，文字レベルの縮約を伴う複合現象の解析を可能にするタグ付け手法を提案した．

(2)言語解析用辞書の語彙項目の拡充と語構成記述の表示機能をもったユーザインタフェースの開発：綴り，読み，品詞，構成語など複合語のもつ統語的情報だけでなく，意味クラス情報や内部構造を記述することのできる辞書管理システムを開発した．また，各語がもつ様々な情報を指定して，任意の語を検索し，内部構造等の情報表示を行う機能，および，複合語の内部構造タグ付け支援機能を実装した．

(3)専門用語および固有表現の自動抽出手法の開発：Wikipedia からの固有表現の自動抽出，および，専門分野の文書からの用語抽出と自動分類に関する研究を行った．

(4) 複合語・専門用語の自動解析手法：複合語の内部構造の自動解析を行うために，機械学習に基づく解析手法を開発した．また，機械学習の訓練およびテストデータとして用いるため，日本語および中国語それぞれ約 800 語の複合語・専門用語の内部構造タグ付きデータを構築した．日本語の複合語解析のために，４種類の係り受け関係以外に，文字レベルの係り受け関係２種類を定義し，合計６種類の文字単位の係り受け解析として，複合語の内部構造解析を行う手法を提案した．

また，上記の研究を通じて，これまで開発してきた日本語，中国語，英語の辞書の拡張を行い，それぞれ約 30 万語，13 万語，10 万語を含む辞書を構築し，研究利用目的での公開を行った．

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７件）

① 渡邉陽太郎，浅原正幸，松本裕治，グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類，人工知能学会論文誌, 23 巻, 245-254, 2008, 査読有
② Jia Lu, Masayuki Asahara and Yuji Matsumoto, Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Derived Words, International Journal of Computer Processing of Language, Vol.21, 101-122, 2008, 査読有
③ 中川哲治，松本裕治，大域的な情報を用いた未知語の品詞推定，情報処理学会論文誌，49 巻，1437-1450，2008，査読有
④ 原 一夫, 新保 仁, 松本 裕治，アラインメントと機械学習を応用した並列句解析，人工知能学会論文誌，22 巻，248-255，2007，査読有
⑤ 平野徹, 飯田龍, 藤田篤, 乾健太郎, 松本裕治，動詞項構造辞書への大規模用例付与, 自然言語処理, 13 巻, 113-132, 2006, 査読有
⑥ 北村美穂子, 松本裕治，言語資源を活用した実用的な対訳表現抽出，自然言語処理，13 巻，3-25，2006，査読有
⑦ Chooi-Ling Goh, Masayuki Asahara and Yuji Matsumoto, Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing , Journal of Chinese Language and Computing, Vol.16, 85-106, 2006, 査読有

〔学会発表〕（計４件）

① 松本裕治，コーパス自動解析ツールと利用環境について，関西言語学会第33回大会，2008，査読無
② 松本裕治，語構成を考慮した多言語の辞書および解析システム,『言語処理技術の

深化と理論・応用の新展開』科研・合同シンポジウム，2008，査読無

③ Jia Lu, Masayuki Asahara and Yuji Matsumoto, Analyzing Chinese Synthetic Words with Tree-based Information and a Survey on Chinese Morphologically Derived Words, SIGHAN Workshop on Chinese Language Processing, 2008, 査読有

④ Chooi-Ling Goh, Jia Lu, Yuchang Cheng, Masayuki Asahara and Yuji Matsumoto, The Construction of a Dictionary for a Two-layer Chinese Morphological Analyzer, Proceedings of the 20th Pacific Asia Conference on Language Information and Computation, 332-340, 2006, 査読有

〔図書〕（計０件）
〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕

６．研究組織
(1)研究代表者
松本　裕治（MATSUMOTO YUJI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：10211575

(2)研究分担者
乾　健太郎（INUI KENTARO）
奈良先端科学技術大学院大学・情報科学研究科・准教授
研究者番号：60272689

浅原　正幸（ASAHARA MASAYUKI）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：80379528

橋本　喜代太（HASHIMOTO KIYOTA）
大阪府立大学・人間社会学部・准教授
研究者番号：50278818

(3)連携研究者
（　　）

研究者番号：