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２４年　６月　５日現在

機関番号：13901
研究種目：基盤研究(B)
研究期間：2009～2011
課題番号：21300094
研究課題名（和文）
辞書自動編纂のためのテクノロジー
研究課題名（英文）
Technology for Automatic Dictionary Compilation
研究代表者
佐藤　理史（SATO SATOSHI）
名古屋大学・工学研究科・教授
研究者番号：30205918

研究成果の概要（和文）：本研究では、外国人名のアルファベット表記（原綴）とカタカナ表記の対応を示した、外国人名対訳辞典を自動編纂することを実現した。この辞典の編纂では、(1)カタカナ表記の人名を自動収集すること、(2)カタカナ表記に対応するアルファベット表記を自動推定すること、(3)見出し語を自動選定すること、(4)それぞれの見出し語に収録する実例（フルネームの人名対訳）を自動選定すること、(5)自動タイプセッティングにより書籍の形式にまとめること、のすべてをコンピュータプログラムによって実現した。この辞典は、15 万件の人名（フルネーム）対訳を収録している。

研究成果の概要（英文）：This study achieved automatic compilation of bilingual person-name dictionary, which contains person names in Latin alphabet and their Katakana transliterations. The compilation process is fully automated, which consists of the followings: (1) automatic collection of Katakana person-names from Japanese texts, (2) automatic back-transliteration of Katakana person-names, (3) automatic selection of headwords, (4) automatic selection of examples (translations of full-names) for every headword, and (5) automatic typesetting of the dictionary in book-style. This dictionary contains 150,000 full-name translations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	3,300,000	990,000	4,290,000
2010 年度	4,800,000	1,440,000	6,240,000
2011 年度	4,900,000	1,470,000	6,370,000
総　計	13,000,000	3,900,000	16,900,000

研究分野：　情報学
科研費の分科・細目：　情報学・図書館情報学・人文社会情報学
キーワード：　情報組織化、辞書編纂、人名抽出、トランスリタレーション

１．研究開始当初の背景

辞書は人間の知的活動を支える重要なツールである。国語辞書や外国語対訳辞書などの言語辞書以外にも、百科事典、専門分野の辞典や用語集、人名辞典、地名辞典など、多くの形態がある。これまで、辞書は人間の手によって編纂されてきた。しかしながら、編纂には膨大な労力がかかるため、多くの部数の販売が見込める少数の例外を除き、辞書の採算性は悪い。そのため、商業出版社は、社会が求める多様な辞書を提供したり、既存の辞書をタイムリーに改訂したりすることができない。

一方、辞書ユーザーにとっては、最新の情報が定期的に反映される辞書が望ましい。不特定多数のボランティアが執筆するウィキペディアは、その一つの解である。もう一つの解は、機械処理による更新の自動化であり、

その先に、辞書の自動編纂がある。現在、ウェブを通じて世の中の知識のかなりの部分にアクセスすることが可能であり、そこから、辞書の編纂や更新に必要な情報を収集することが、原理的には可能である。

辞書を辞書たらしめている本質は、網羅性・一貫性・信頼性にある。たとえば、『新明解国語辞典』などの小型国語辞典は、日常的に使う語彙を収録するという点で一貫しており、そのほとんどを網羅している。同時に、その記述にはほとんど誤りがなく信頼できる。これまでにも、我々の試みも含め、ウェブから情報を抽出してそれを辞書化するという研究は、多数存在する。しかし、そこでは、抽出した情報の信頼性が十分ではないと同時に、網羅性と一貫性という辞書に不可欠な要素が完全に欠落している。単に抽出した情報を束ねただけでは辞書にはならない。抽出した情報に対して、信頼性・網羅性・一貫性を生み出すための選別と組織化を行なうことが必要不可欠である。

最大の問題は、辞書編纂のための方法論が存在しないということである。辞書には多くの形態があるが、その設計・編纂においては、(a)辞書の収録対象範囲をどう設定するか、(b)見出し語集合をどう決めるか、(c)各項目に記述する情報をどう設定するか、(d)情報の信頼性をどう担保するか、(e)実際の利用を想定して全体の構成と検索機能をどう設計するか、など多くの共通する問題がある。辞書編纂の自動化を実現するためには、これらの問題に対する統一的な設計・編纂法が不可欠である。

２．研究の目的

本研究の目的は、以下の３点に集約される。

(1) 真に実用的な対訳辞書の実現

外国人名対訳辞書の自動編纂を本研究の中核と位置付け、外国人名の日本語訳を見つけたい時は誰もが最初に参照するデファクト・スタンダードとなる辞書の実現を目指す。我々がこれまでに、ウェブから自動収集した人名対訳数は2008年3月の時点で約19万件(精度 90%)であり、規模的にはすでに既存の辞書を凌駕している。収集した人名対訳に対してクリーニングを実施して 95%～99%の信頼性を確保するとともに、網羅性・一貫性という性質を付与するための選別と組織化を行ない、真に実用的な対訳辞書の自動編纂を世界で初めて実現する。

(2) 辞書編纂のための要素技術の高度化

対訳辞書編纂のための要素技術を、大きく、(a)見出し語（語・ターム・フレーズ）抽出、(b)関係（対訳関係、同義関係）抽出、(c)品詞・活用形推定、(d)信頼性・一貫性確保のためのチェック、(e)網羅度（母集団）推定、の５つに分け、既存技術の高度化と新技術の開発を行なう。多くの既存技術の精度は 80% 以下であり、そのままでは辞書の自動編纂には使えない。要素技術の到達目標は、最低で精度90%以上、できれば 95%以上である。まず、作るべき辞書を設定し、その実現に必要な要素技術の最低精度を決め、それを実現するための方法を考えるというトップダウン的アプローチで、精度向上を目指す。

(3) 辞書設計・編纂法の確立

複数の辞書の設計および自動編纂を通して、辞書設計・編纂法を確立する。現時点では、次の５ステップから構成することを想定している。(a)辞書使用者層の決定、(b)辞書サイズの決定、(c)見出し語集合の設計および選定、(d)各項目の記述内容の設計と項目の作成、(e)辞書の全体構造の設計とパッケージング（構造化・インデキシング）。

３．研究の方法

(1) 本研究の研究内容は、(a)実際に作成する辞書、(b)それを実現するための辞書設計・編纂法、および、(c)自動編纂技術、の３つに分かれる。外国人名対訳辞書の自動編纂を実現することを中心に研究を進め、その成果を生かして、カタカナ語辞書や日本語言い換え辞書の自動編纂にも取り組む。

(2) 外国人名対訳辞書を実現するためには、まず、外国人名対訳を大量に収集する必要がある。これには、日本語テキスト・英語テキストからの外国人名抽出、および、トランスリタレーションに基づく人名対訳推定が必要となる。これらを高精度(95%以上)で実現する方法を開発する。さらに、これら２つの要素技術を組み込んだ、対訳クローラーを実現する。

(3) 収集した対訳データに誤りが混入することは避けられない。そのため、収集したデータをクリーニングする技術が必要となる。特に、人名かどうかの判定を高精度で行うことは難しいと考えられるので、既存のデータベース（米国の国勢調査等）を利用した方法等も検討する。

(4) 実用的な辞書とするためには、見出し語集合の設計と選定が重要であるとが、その方法論は確立されていない。見出し語集合を自動的に選定する機構を実現することを通して、設計法・選定法について新たな知見を得る。

４．研究成果

本研究の主な成果は、以下のとおりである。

(1) 日本語テキストおよび英語テキストから、外国人名を自動的に収集する方法を実現

した。
実現した方法は、ブートストラップ的方法である。まず、簡単かつ信頼できるヒューリスティック（直後に「〜さん」を伴うカタカナ文字列）を用いて、日本語の新聞記事から外国人名を収集する。次に、収集した外国人名に基づき、カタカナ文字列が外国人名か否かを判定する人名フィルター（日本語テキスト用）を構成する。一方、次に述べる対訳推定を用いて、カタカナ表記から原綴を推定し、アルファベット表記の人名データを作成する。これに基づき、英単語列が人名か否かを判定する人名フィルター（英語テキスト用）を構成する。以上で、日本語テキストおよび英語テキストからの人名自動抽出が可能となる。人名フィルターの能力は、利用する人名データの質と量に依存するため、ブートストラップのループを何回か回し、能力向上を図る。最終的に得られた人名フィルターの精度は、0.96-0.97 であり、F 値は 0.93-0.94 である。

(2) ウェブに基づくトランスリタレーターを実現した。
外国人名の対訳を求めることは、トランスリタレーション（翻字または音訳）、または、逆トランスリタレーションを求めることに他ならない。トランスリタレーションは、大量の実例から作成した確率モデルに基づいて推定する方法が主流であるが、本研究では、ウェブを利用して高い精度で外国人名のトランスリタレーションを推定する方法を開発した。
ウェブ上には、外国人名の原綴とカタカナ訳が併記されたページが多数存在する。たとえば、カタカナ表記で検索エンジンを引き、その検索結果ページからアルファベット単語列を抜き出せば、その中に原綴が含まれている可能性が高い。ゆえに、その中から、カタカナ表記と対応する可能性がある単語列を選択することによって、原綴を求めることができる。さらに、出現数を加味することにより、高い精度での推定が可能となる。実現したトランスリタレーターの精度は、順方向(英日)で 98.5%、逆方向(日英)で 93.9%である。

(3) 人名抽出および対訳推定を組み合わせて、ウェブから人名対訳を自動収集する対訳クローラーを実現した。
ウェブ上には、複数の人名が列挙されているページが多数存在する。作成した人名対訳クローラーは、カタカナ表記の外国人名１件からスタートし、ウェブページを利用して、その対訳を推定すると同時に、そのページから新たな人名を収集する。これを繰り返すことにより、多数の人名対訳を自動的に収集することができる。実際に、このシステムを５か月間動かし、56 万件の人名対訳データを自動収集した。

(4) 非生産型トランスリタレーションの枠組みを提案し、その効率的なアルゴリズムを考案した。
ウェブに基づくトランスリタレーターは、入力からトランスリタレーションを生成するのではなく、入力とトランスリタレーションの関係となっている語を発見する方法である。このような考え方に基づき、入力のトランスリタレーションを、大規模な(数十万件)候補集合から選ぶという方式(非生産型トランスリタレーション)を考案した。同時に、動的計画法、プレフィックス・フィルタリング、反復型コスト束縛探索の３つの技法を用いて、これを効率的に実行するアルゴリズムを明らかにした。

(5) 見出し語設計・選定に関わる重要な概念として、メンバーシップ予測性という概念を提唱した。
辞書において、その価値を規定する最も重要なものの一つは、見出し語集合である。辞書編纂者は、設計の段階で「どのような語をどれだけ収録するか」を定めるが、利用者にとっては、実際に作成された見出し語集合が「集合としてどのような性質を持つか」がより重要である。
辞書の利用者は、ある語に対して、その語が載っていることを予想して辞書を引く。その予想が何度も外れると、それ以降、その辞書を使わなくなる。すなわち、辞書の見出し語集合は、利用者にとって、ある語がその集合に含まれるかどうかが予想可能であるような集合となっていなければならない。これをメンバーシップ予測性と呼ぶ。辞書の見出し語集合がメンバーシップ予想性を持つためには、見出し語集合に、ある種の潜在構造が仮定できることが必要である。このように、辞書の見出し語集合が持つべき性質を明らかにした。

(6) 『紬 2012：外国人名対訳辞典』の自動編纂を実現した。
自動収集した対訳データに対し、各種ヒューリスティックを適用するとともに、上記の見出し語集合に関する方法論に基づいて、見出し語集合の設計を行い、各種基礎資料等も併用して見出し語の自動選定を行い、最終的に『紬 2012: 外国人名対訳辞典』として結実させた。
この辞典は、外国人名を対象に、ラテンアルファベット表記（原綴）とカタカナ表記(カタカナ訳)の対応を示したものである。この辞典の見出し語は、人名を構成する要素

（姓と名）である。辞典は、(a)ラテンアルファベット表記の見出し語をＡからＺまで並べたもの(見出し語数 52,018 件)、(b)カタカナ表記の見出し語をアからンまで並べたもの(見出し語数 45,600 件)、の２つのパートからなり、150,744 件の実例数(人名対訳数)を収録している。辞典のサイズは、全 8 巻(3386 ページ)である。この他に、実例数を制限した縮約版（全２巻、1317 ページ）も作成した。

(7) カタカナ表記ゆれに対応した辞書引き方法を実現した。

日本語は表記ゆれが豊富な言語である。カタカナ外来語は、特にその傾向が顕著である。すべての表記を辞書に登録しておくのは、実際的ではない。そこで、規則的な表記ゆれは、辞書引きの段階で吸収し、辞書に登録されていない表記でも辞書を引けるようなシステムを実現した。

このシステムは、狭義の表記ゆれを生成する 102 個の規則と、広義の表記ゆれ(語形のゆれ)を生成する 114 個の規則を持ち、辞書に含まれていない文字列(表記)に対して、適切な辞書エントリーを推定する。これにより、あらかじめ辞書に登録しなければならない異表記の数を劇的に削減することができる。たとえば、UniDic-2.1.0 では、登録されている異表記のうちの 77%を削減することができる。この辞書引き法は、外国人名対訳辞典の検索においても有効である。

(8) 本研究の最大の貢献は、実用的な対訳辞書を自動的に編纂できることを実証した点にある。しかし、その一方で、自動編纂の実現のためには、要素技術を非常に高い精度で実現することが必要であると同時に、各種のヒューリスティックを適用して、信頼性を高めることが不可欠であることがわかった。このことは、人間による辞書の編纂において、常識的判断や経験等が多用されていることを示唆する。辞書自動編纂の方法論・技術を確立するためには、さらなる研究が不可欠である。

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 4 件）

① Satoshi Sato and Sayoko Kaide. A Person-Name Filter for Automatic Compilation of Bilingual Person-Name Lexicons. Proceedings of the Seventh Conference on International Language Resource and Evaluation, 査読有, 2010. http://www.lrec-conf.org/proceedings/lrec2010/pdf/343_Paper.pdf

② Satoshi Sato. Non-Productive Machine Transliteration. 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, 査読有, pp.16-19, 2010. http://dl.acm.org/citation.cfm?id=1937055.1937059.

③ Satoshi Sato. Web-Based Transliteration of Person Names. Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 査読有, pp.273-278, 2009, DOI: 10.1109/WI-IAT.2009.47.

④ Satoshi Sato. Crawling English-Japanese Person-Name Transliterations from the Web. Proceedings of the 18th International Conference on World Wide Web, 査読有, pp.1151-1152, 2009, DOI: 10.1145/1526709.1526902.

〔学会発表〕（計 6 件）

① Satoshi Sato. Dictionary Look-up with Katakana Variant Recognition. The 8th International Conference on Language and Evaluation. 2012.5.23, Istanbul, Turkey.

② 佐藤理史. 辞書の見出し語集合と代表性. 言語処理学会第 18 回年次大会, 2012.3.16, 広島市立大学(広島県).

③ 伊藤美咲姫, 佐藤理史, 駒谷和範. カタカナ表記ゆれに対応した辞書引きシステム. 言語処理学会第 18 回年次大会, 2012.3.15, 広島市立大学(広島県).

④ 松木久幸, 佐藤理史, 駒谷和範. 文末機能表現シソーラスの編纂に向けて--文末機能表現の網羅的生成--. 言語処理学会第 18 回年次大会, 2012.3.14, 広島市立大学(広島県).

⑤ 古武泰樹, 佐藤理史, 駒谷和範. オノマトペを言い換える表現の自動収集. 言語

処理学会第 17 回年次大会, 2011.3.10, 豊橋技術科学大学(愛知県).

⑥ 佐藤理史. 大規模候補リストを利用したトランスリタレーション. 言語処理学会第 16 回年次大会, 2010.3.11, 東京大学(東京都)

〔その他〕
ホームページ等
http://kotoba.nuee.nagoya-u.ac.jp

６．研究組織
(1)研究代表者
佐藤　理史（SATO SATOSHI）
名古屋大学・工学研究科・教授
研究者番号：30205918

(2)研究分担者なし

(3)連携研究者
藤田　篤（FUJITA ATSUSHI）
公立はこだて未来大学・システム情報科学部・准教授
研究者番号：10402801
(H21→H22)