Japio

外国人名対訳辞書の自動編纂を目指して

名古屋大学大学院工学研究科教授
佐藤　理史

PROFILE
京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005年6月より名古屋大学大学院工学研究科電子情報システム専攻教授。

1 はじめに

「ことば」に関する知識を集積したもの、それが辞書である。知の世界への扉を開く「ことば」がわからなければ、最初の一歩を踏み出すこともかなわない。我々にとって、辞書は、頼りがいのある最強のリファレンス・ツールである。

この辞書を、計算機を用いて自動的に編纂することはできるだろうか。

ウェブの普及と巨大化により、品質を無視すれば、とにかく膨大な量のテキストにアクセスすることが可能となった。それらをうまく利用すれば、辞書を自動的に編纂できるかもしれないと考えるのは、それほど突飛なことではない。しかしながら、辞書を辞書たらしめている3つの基本的な性質、すなわち、**一貫性、網羅性、信頼性**を持つ、真に実用的な辞書を自動的に作ることは、まったくもって容易ではない。辞書の自動編纂は、非常にチャレンジングな研究テーマである。

我々が現在、取り組んでいるのは、外国語固有名詞の対訳辞書の自動編纂である。固有名詞は、数が多く、かつ、新たなものが出現するため、自動編纂のニーズがある。その一方で、固有名詞は「もの」を指し示すための単なるラベルであるため、「語を定義すること」に付随する難しさを回避することができる。さらに、外国語の固有名詞の和訳のほとんどは、その発音を日本語的な音に近似してカタカナで表したもの（翻字）であり、正しい訳語であるか否かを、比較的容易に判定できる。このような理由により、各種の辞書の中で、最も自動編纂が容易と考えられる。

我々は、ターゲットを外国人名対訳辞書に定め、2006年の秋から本格的に研究を開始した。最初の目標は、大量の人名対訳対を自動収集することである。2006年度に4.2万件[1]、2007年度に19万件[2]、2008年度に50万件の人名対訳対を収集することを実現した[3]。一方、収集した大量の人名対訳対を取捨選択・整理して対訳辞書という形にまとめあげる編集処理の自動化は、ようやくそれを研究できるようになってきたという状況である。本稿では、『紬』と名付けた、我々の対訳辞書自動編纂システムの現状について述べる。

2 外国人名翻訳の特徴

システムを説明する前に、外国人名翻訳の特徴を簡単にまとめておこう。

(1) 同一人物は、完全に同じカタカナ綴で表現されるべきである。

外国人名の和訳であるカタカナ綴は、日本語において、その人物を指し示すIDの役割を果たすことになる。同一人物が複数の異なるカタカナ綴で表現された場合、読み手は、それらが同一人物を指し示していることを正しく認識できない可能性がある。このため、もし、ある人物に、既に定着した標準的なカタカナ綴が存在する場合は、（それが原語の発音とは異なるとしても）そのカ

タカナ綴に従うことが望ましい。しかし、現実には、この原則は守られておらず、複数のカタカナ綴を持つ外国人名が多数存在する。

(2) 色々な言語由来の人名が英語テキストに現れる。

英語のテキストには、英語以外の言語由来の人名が多数出現し、英日翻訳では、これらの人名も翻訳しなければならない。このような人名の和訳は、英語の発音ではなく、元々の言語の発音に基づくカタカナ綴となることが多い。たとえば、Michael Schumacher は、マイケルではなく、ミハエル・シューマッハと綴るのが普通である。このため、人名の翻訳は、まずはフルネームの単位で考える必要がある。

以上の説明から明らかなように、外国人名翻訳のポイントは、「フルネームという単位で、既に定着したカタカナ訳が存在するかどうか」を知ることである。存在する場合は、その訳に従えばよい。存在しない場合に、はじめて新たなカタカナ訳を作る必要が生じる。我々の目標は、この「定着した既訳の有無」に対して、必要十分な情報を提供する辞書を編纂することにある。

3 紬システム

紬システムの構成を図 1 に示す。この図に示すように、システムは、対訳収集、辞書編纂、辞書本体から成る。

自動編纂した外国人名辞書は、現在、kotoba.nuee.nagoya-u.ac.jp で公開している。辞書の検索例を図 2 に示す。この図に示すように、ひとつの外国人名に対して、複数のカタカナ綴が存在する場合、どんなカタカナ綴があるのか、その中でどのカタカナ綴が最もよく使われているか、という情報を提示する。また、実際に、どのようなウェブページから訳語対を抽出したのかも、併せて表示する。

現在のシステムの主要部は、紬クローラーと呼ばれる対訳収集モジュールである。このモジュールでは、単言語（アルファベット綴またはカタカナ綴）の人名に対

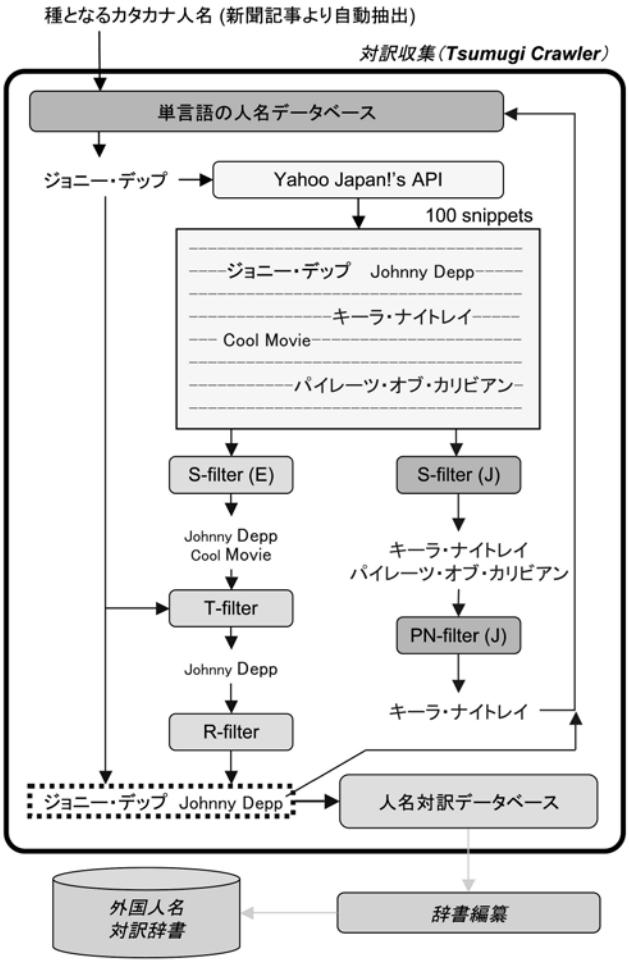

図 1　紬システムの構成

図 2　辞書インタフェース

して、検索エンジンを利用して、その対訳を推定することを行なう。この対訳推定の中核は、3つのフィルタ（S-filter, T-filter, R-filter）である。最初のS-filterは、検索エンジンの出力テキストから、人名となりうる可能性がある候補文字列を抽出する。T-filterは、それぞれの候補が、元の人名の訳となる可能性があるかどうか（翻字関係にあるかどうか）を調べ、可能性が高いもののみを残す。最後のR-filterは、検索エンジンの出力テキスト中の頻度を利用し、頻度最上位の候補と2位以下の有望な候補を残す。こうして対訳が得られると、それを入力として、逆方向に対訳推定を行なう。

対訳推定処理と同時に、検索エンジンの出力テキストから、新たなカタカナ人名の収集を行なう。この処理では、前述のS-filterのほかに、PN-filterを使用する。PN-filterは、与えられた文字列が人名であるかどうかを、大量の訓練例から学習した「人名らしさ」の指標に基づいて判定し[5]、可能性が高いもののみを残すフィルタである。

辞書編纂モジュールは、収集した対訳集合の中から、辞書のエントリーを決定することを行なう。現在は、それぞれの対訳に対して、ウェブ上の実例数を見積もり、競合する綴のうち、相対的に実例数が低い綴を削除することを行なっている（図3）。

なお、辞書検索においてユーザーが入力した人名が辞書に存在しなかった場合は、その場でウェブを調査して対訳を推定し、その結果をユーザーに提示する。

4 おわりに

現在公開している辞書のサイズは、標準表記と考えられる対訳数28万件、異表記を含めた対訳総数40万件（アルファベット表記31万件、カタカナ表記37万件）である。書籍体の外国人名辞典のエントリ数約2万件、ウィキペディアで調べることができる固有名詞の英日対訳数12.5万件と比較して、サイズ的には、かなり大きなものとなっている。ウェブで調査可能な人名対訳のどのくらいの割合をカバーしているか（いわゆる再現率）を調べるのは非常に難しいが、書籍体の洋書とその訳書から抽出したテストデータを用いた評価では、66%という結果が得られている。

研究開始から3年が経過し、対訳収集は、ほぼ目処が立ってきた。今後は、数から質へ、研究をシフトしていく必要がある。最初に述べたように、辞書を辞書たらしめているのは、一貫性、網羅性、信頼性という3つの基本的な性質であり、収集した対訳集合を束ねただけ

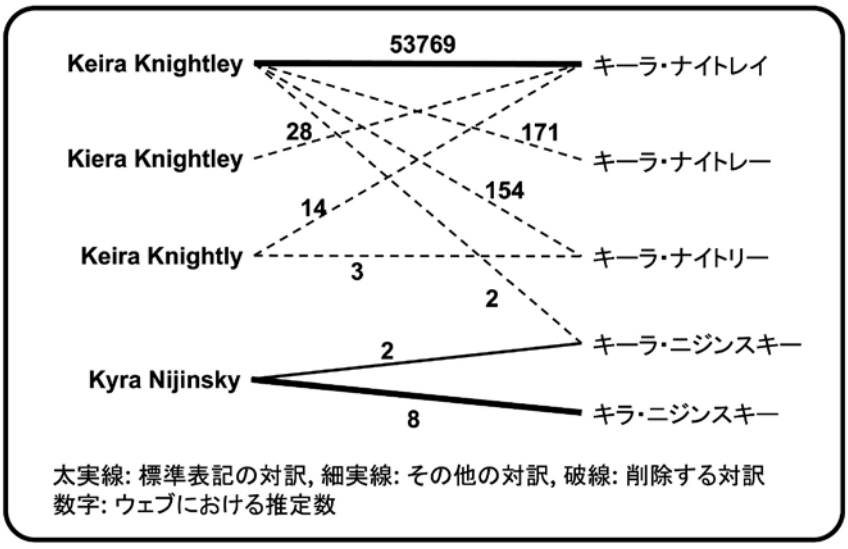

図3　辞書編集による対訳エントリーの絞り込み

では、実用的な辞書とはならない。辞書にこれらの性質を付与するためには、辞書編纂において何を行なわなければならないか。それは、どうすれば自動化できるのか。実用的な対訳辞書の自動編纂の実現は、まだ、道半ばである。

参考文献

[1] 榊原洋平，佐藤理史．ウェブを用いた外国人名事典の自動編纂．言語処理学会第 13 回年次大会発表論文集，pp. 879-882, 2007.

[2] 榊原洋平，佐藤理史．外国人名対訳辞典の大規模化 ― 15 万件の自動編纂 ―．言語処理学会第 14 回年次大会発表論文集，pp.833-836, 2008.

[3] 佐藤理史．外国人名対訳辞書の自動編纂：現状と展望．言語処理学会第 15 回年次大会論文集，pp.304-307, 2009.

[4] Satoshi Sato. Crawling English-Japanese Person-Name transliterations from the Web. Proc. of WWW-2009, p1151-1152, 2009.

[5] 開出紗代子，佐藤理史．生起確率の差を用いた人名判定．言語処理学会第 15 回年次大会論文集，pp.12-15, 2009.