外国人名のカタカナ表記推定

Katakana Transliteration of Person Names

名古屋大学大学院工学研究科教授
佐藤　理史

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 年より現職。現在、言語処理学会会長。

1　はじめに

2016 年のリオデジャネイロ・オリンピック開催時より、オリンピック参加者名簿の翻訳を支援するシステムを運用している。ここでの翻訳とは、ほとんどの国や地域（人名を漢字で表記する国を除く）からの参加者の英語表記に対し、カタカナ表記を定めることを意味する。カタカナ表記が付加された参加者名簿は、競技を実況中継する放送局や、競技結果を報道する通信社・新聞社にとって必須のものである。

そもそもの発端は、2012 年のロンドン・オリンピックにさかのぼる。我々が当時運用していた外国人名のカタカナ表記推定システムを利用できないかという問い合わせがあり、協力したことが始まりであった。それ以降、色々な経緯を経て新たなシステムを作成し、一部の関係者に対して公開してきたが、株式会社時事通信社のご協力とご理解を得て、平昌オリンピック開催に合わせて、システムを一般公開した。

2　問題の難しさ

オリンピックの参加者名簿の翻訳には、次のような難しさがある。

（1）人数が多い。1 万人以上である。

（2）国際オリンピック委員会（IOC）から英語版の暫定名簿が届くのは、開会式のおよそ 1 ヶ月前である。しかしながら、それ以降、随時アップデートが送られてくるので、開会式直前まで、参加者名簿は確定しない。

（3）国や地域の数が多い。200 以上の国や地域から選手・役員が参加する。

（4）人名に対して「言語」を特定するのはほとんど不可能である。たとえば、Michael は、欧米圏でよくある名前だが、それが何語であるかを判定することは不可能である。さらに悪いことに、英語・ドイツ語・フランス語読みで、それぞれ読み方が異なる。

（5）どんなカタカナ表記が「正解」であるかは、規定できない。一般に、許容できるカタカナ表記は複数考えられる。すでに、標準的なカタカナ表記が定着している人物に対しては、その表記が現実的な「正解」であるが、はじめてカタカナ表記される人物に対しては、許容できるカタカナ表記の中でどれを選ぶかには、自由度がある。

（6）名簿のうち、すくなくとも半数以上の人物は、これまで一度もカタカナ表記されたことがなかった人物である。

3　現在のシステム

上記のような状況と問題点を踏まえ、名前（姓名のいずれか一方）と国名を入力として、カタカナ訳候補の上位 5 件を提示するシステムを作成した。システムの入出力の様子を図 1 から図 3 に示す。システムの実現には、ウェブから自動収集した外国人名対訳データ（約 20 万件）と株式会社時事通信社提供の外国人名対訳データを利用している。前者のデータには国名は付与されていな

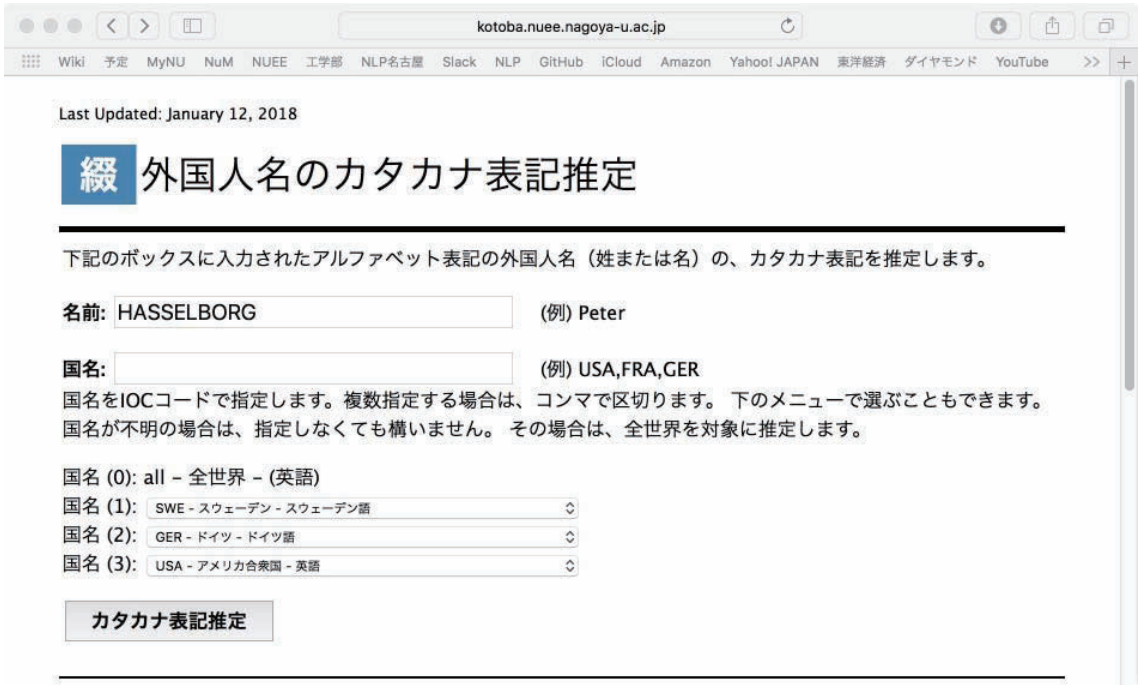
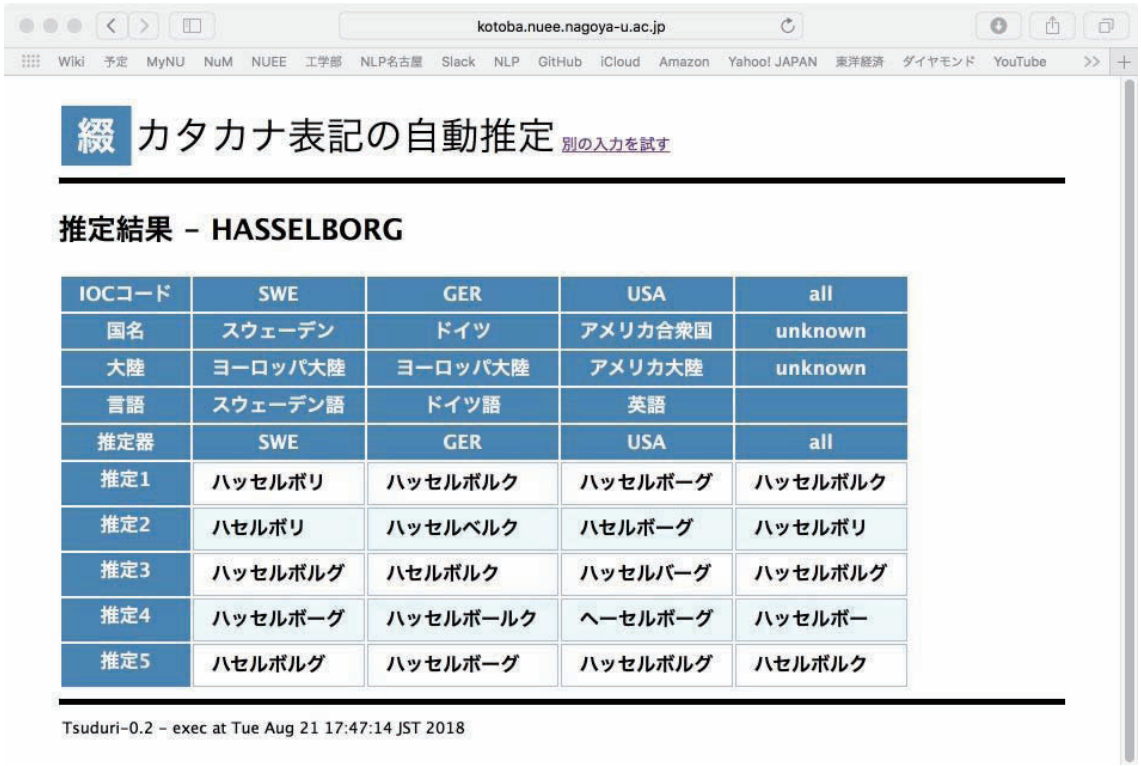


図１　システムの入力画面
名前に対して、国名を複数指定することができる



IOCコード	SWE	GER	USA	all
国名	スウェーデン	ドイツ	アメリカ合衆国	unknown
大陸	ヨーロッパ大陸	ヨーロッパ大陸	アメリカ大陸	unknown
言語	スウェーデン語	ドイツ語	英語	
推定器	SWE	GER	USA	all
推定1	ハッセルボリ	ハッセルボルク	ハッセルボーグ	ハッセルボルク
推定2	ハセルボリ	ハッセルベルク	ハセルボーグ	ハッセルボリ
推定3	ハッセルボルグ	ハセルボルク	ハッセルバーグ	ハッセルボルグ
推定4	ハッセルボーグ	ハッセルボールク	ヘーセルボーグ	ハッセルボー
推定5	ハセルボルグ	ハッセルボーグ	ハッセルボルグ	ハセルボルク

図２　システムの出力画面（未知の名前の場合）
それぞれの国ごとに、推定結果が表示される。平昌オリンピックのカーリングのスウェーデン代表の Anna HASSELBOG さんの読売新聞の訳は「アンナ・ハッセルボリ」である。（https://www.yomiuri.co.jp/olympic/2018/results/participant_3021688.html）

いが、後者のデータには国名（国籍）が付与されている。

システム作成上の最大の問題は、200 を超える国や地域に対して、未知の名前のカタカナ訳推定をどうやって実現するかである。当然のことながら、英語表記が同一であっても、それぞれの言語によって読み方が異なるものが多数ある。ただし、先に述べたように、国名と綴りから言語を推定することはほとんど不可能なので、国名に対して読みを推定することにするが、オリンピック

図３　システムの出力（既訳が存在する場合）
辞書に訳語が存在する場合は、その情報も表示する。ロシアなどでは「イリーナ」と訳すが、「イリナ」と訳す国もあることがわかる。

参加者の少ない国や地域に対しては、十分な数の国名付きデータが存在しない。そのため、まず、国名不明データを使って暫定的な推定器を作成し、次に、国名付きデータを使って、その推定器をその国用に修正するという方法を採用する。現在運用中のシステムは、MeCab を使って実装したシステム[1] だが、ニューラルネットによる実装も試みている[2]。

それぞれの機関（通信社等）は、それまでに翻訳した独自の対訳名簿を持っており、実際に参加者名簿を翻訳する際には、まずはそれと突き合わせて、人物単位で翻訳できるものを翻訳する。翻訳できなかったものは、（予算が潤沢にある場合は）外部の翻訳者に翻訳を依頼するようである。しかしながら、最終段階で参加者名簿に修正があった場合などは、その場で翻訳しなければならず、我々のシステムの出番となる。つまり、システムの利用は、翻訳のメインストリームではなく、困った時にお助けという位置付けである。候補を複数出力すること、名前に既訳があった場合はその情報を出力すること、複数の国に対する結果を同時に表示できることなどは、そのような利用において、人間の最終判断を手助けすることを意図している。

4 運用上の問題点

このようなシステムは、定期的なアップデートが不可欠である。新たに提供されるデータがクリーンであれば、ほぼ rake コマンドを実行するだけで、200 を超える推定器を生成できるようにしているが（生成には数十時間を要する）、人間が作成したデータがクリーンであることは「ありえない」ので、これまでの研究のノウハウが詰まったチェックプログラムを走らせて、引っかかったものを目視で調べるという作業が必ず必要となる。そ

して、誤りと思われるデータのリストを作り、データ提供元に確認を求めるという工程がどうしても発生する。

データのクリーニングにおいては、人名のカタカナ表記の規範がほとんど存在しないことが問題となる。外来語の表記に使うカタカナに関しては、「外来語の表記（平成３年６月28日）内閣告示第二号」[3] という公式のガイドラインがあるが、残念ながら、実際に使われているカタカナ表記を十分に反映しているとは言えない。実際にデータを調べると、複数の人間によって翻訳されたカタカナ訳（表記）では、ウ濁点（ヴ）の使用、小書文字（小さなアイウエオヤユヨ）の使用などが統一されておらず、ばらつきが存在する。この問題を回避するために、我々のシステムでは、使用できるカタカナ表記の範囲を厳格に定め、それから逸脱したカタカナ表記は訓練データとして採用しない（あるいは、正規化したものを採用する）こととしている。これにより、推定するカタカナ訳もその範囲を逸脱しないことを保証している。

いまのところ、本システムを2020年の東京オリンピック開催時までは運用する予定である。人名のカタカナ訳が必要になった場合は、試していただければ幸いである。URLは、http://kotoba.nuee.nagoya-u.ac.jpである。

参考文献

[1] 佐藤理史.『言語処理システムを作る』. 近代科学社, 2017.

[2] Dawoon JUNG and Satoshi SATO. Country Adaptation in Neural Machine Transliteration of Person Names. The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2L4-04, 2018.

[3] 文化庁編.『新訂　公用文の書き表し方の基準（資料集）』. 第一法規株式会社, 2011.