2020東京オリンピック参加者名簿の翻訳

Translation of the List of Participants in the 2020 Tokyo Olympic Games

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。

1 はじめに

外国人名翻訳の最前線、それは、夏季オリンピックの参加者名簿の翻訳である。そこでは、200 カ国以上、1 万人以上の人名の翻訳を、数週間という短期間で行うことが要求される。このようなタスクにこそ、機械による支援が適用されるべきであろう。筆者は、過去に外国人名対訳辞書の自動編纂 [1] やトランスリタレーションの研究 [2] を行った経験があり、縁あって、2020 東京オリンピックの参加者名簿の翻訳支援を実施した。

外国人名にカタカナ訳を付与するというタスクは、一般に、トランスリタレーション、すなわち、「アルファベッ表記に対して、正解カタカナ訳を推定する」問題だと認識されている。しかし、オリンピック参加者名簿の翻訳という現実のタスクに取り組む過程で、その認識は、現実のタスクの本質を捉えていないことに気がついた。このタスクの本質は、「既訳（つまり、正解）が存在しない外国人名に対して、社会的に受け入れられるカタカナ訳を新たに作る」ことにある。本稿では、2020 東京オリンピック参加者名簿の翻訳支援の内容とその経験から学んだことを述べる。

2 オリンピック参加者名簿の翻訳の過酷さ

オリンピック参加者名簿の翻訳には、外国人名翻訳自体の難しさと、オリンピック参加者名簿固有の過酷さが存在する。

外国人名の翻訳（カタカナ訳付与）は、もっぱら発音に基づいて行われるが、日本語にない音が含まれる場合は、日本語の音で近似する必要がある。そこには自由度があり、訳者あるいは翻訳する組織の意向が反映される。たとえば、テニス選手の Andy Murray の姓の訳は統一されておらず、「マリー」と「マレー」が存在する。もうひとつの問題は、人名の大半は 2 語であるため、アルファベット表記だけからは原言語を特定できない（そのため、発音を同定できない）ことにある。スポーツ選手の場合は、所属する国はほぼ特定できるが、国が特定できたとしても氏名の原言語は一意には決定できない。

一方、オリンピック参加者名簿固有の過酷さは、その分量と時間的制約にある。2020 東京オリンピックの参加者は、ウィキペディアによれば 205 カ国 11,092 人である。参加者名簿は、その候補リストが国際オリンピック委員会（IOC）から開催の約 1 カ月前から数回にわたって送られてくる（回を重ねる毎に更新される）。参加者がほぼ確定するのは開幕直前であるため、名簿が確定してから翻訳するのでは遅すぎる。それゆえ、候補リストの段階から翻訳作業を進めることになるが、候補リストのサイズは実際の参加者数の数倍で、東京オリンピックの場合は約 3 倍であった。

日本において、オリンピック参加者名簿を翻訳する主な組織は、（1）放送局から委託を受けた NHK の関連会社と、（2）2 つの通信社（時事通信社と共同通信社）である。以前は、この 3 組織は独立に翻訳を行っていたそうで、同一人物のカタカナ訳が 3 組織で一致しないものが少なからず存在したようである。東京オリンピック開催に合わせて、3 者間でカタカナ訳のすりあわ

せが行われたと聞いているが、カタカナ訳の不一致問題は、完全には解消されていないものと推察する。

ここで、外国人名翻訳が、通常のテキストの翻訳と大きく異なる点を指摘しておきたい。テキスト翻訳のゴールは「意味が通じること」であるが、外国人名翻訳は「人物が特定できること」である。つまり、カタカナ訳は、日本語におけるその人物の識別子として機能する。そのため、同一人物に対して複数の訳が存在することは好ましくない。そのゆえ、まず、既訳の有無を確かめることが重要となる。

既訳が存在する場合は、それに従うのが基本である。この原則に従えば、人名に対するカタカナ訳付与は 1 回限りとなる。一方、既訳が存在しない場合は、アルファベット表記から発音を推測し、それを近似的に表すカタカナ表記を定める。そこには恣意性が存在し、客観的な正解というものは存在しない。

だからといってどんな表記を採用してよいわけではなく、発音に反映した妥当な表記を採用する必要がある。この表記が社会的に受け入れられれば、その人物に対する日本語の識別子（既訳）として定着する。つまり、人名翻訳の本質は、「定着することが期待される識別子（カタカナ訳）を新たに設定すること」なのである。

3 支援システムの基本方針

時事通信社から話があって、東京オリンピックの参加者名簿の翻訳支援プロジェクトを開始したのは、2015 年に遡る。この年よりシステム設計・開発を進め、リオデジャネイロ・オリンピック（2016 年）、平昌オリンピック（2018 年）等で経験を積み、東京オリンピックに備えた。東京オリンピックは 1 年延期されたため、実際の作業は 2021 年にずれ込んだ。その後、北京オリンピック（2022 年）の参加者名簿の翻訳支援も行った。

支援システムの設計にあたっては、次の 3 つの基本方針を定めた。

（1）カタカナ表記を統制する

日本語は表記に対して寛容な言語であり、いわゆる「表記ゆれ」が多数存在する。外国語の表記のガイドラインとして、平成 3 年 6 月 28 日の内閣告示二号『外来語の表記』[3] が存在するが、現実にはこのガイドラインを逸脱した表記が多数存在する。支援システムでは、使用するカタカナ表記の範囲を厳密に定め、出力するカタカナ訳を統制する。

（2）既訳辞書による翻訳を優先する

既訳辞書を整備し、既訳が存在する場合は、既訳を採用する。

（3）国毎のトランスリタレータを用意する

同一のアルファベット表記であっても、原言語によってカタカナ訳が異なる場合がある（たとえば、Peter の訳としては、「ピーター」、「ペーター」、「ペテル」の 3 種類が存在する）。しかし、原言語はわからないため、国毎にトランスリタレータを用意し、訳し分けを実現する。

既訳辞書の整備は、時事通信社が担当した、時事通信社の人名翻訳の基本方針は「1 国 1 名前（姓または名）に対して、カタカナ訳はひとつ」である。つまり、国と名前が定まれば、カタカナ訳が一意に定まるように統制する。

4 綴 2021 と裕 2019

外国人名翻訳支援を実現するために、「綴 2021」と「裕 2019」の 2 つのシステムを作成した。

「綴 2021」は翻訳を担当するシステムで、既訳辞書と 208 カ国のそれぞれに対するトランスリタレータから構成されており、人名リストを一括で翻訳する「一括翻訳サービス」と、ユーザーの入力に対して翻訳結果を提示する「ウェブサービス」の 2 つのサービスを提供する。既訳辞書はリレーショナルデータベースに格納され、それぞれにエントリは、国情報（IOC コード）を持つ。トランスリタレータのソフトウェアには MeCab [4] を用い、それぞれの国に対して国別モデル（MeCab 辞書）を用意した。

この国別モデルの作成を担当するのが「裕 2019」で、その中核機能は、対訳例（アルファベット表記とカタカナ表記のペア）の部分対応を推定して、MeCab の学習用コーパスを作成する機能である。国別モデルの作成では、まず、国情報が不明の約 13 万件の名前対訳データを用いてベースモデルを作成する。そののち、国情報付きの名前対訳データ（既訳辞書）を用いて追加学習を行い、それぞれの国別モデルを作成する。これらのシステ

ムの技術的詳細は、論文 [5] を参照されたい。

5 暫定訳はどのぐらい採用されたか

2022 年の北京オリンピック終了後、時事通信社が最終的に作成した翻訳名簿を入手して、支援システムで翻訳した結果がどのぐらい採用されたかを調べた。

時事通信社に提供した一括翻訳サービスの翻訳結果には、その後の修正作業を支援するために、既訳辞書による翻訳結果やトランスリタレータによる翻訳結果(複数)が含まれる。そのひとつに、ある基準に基づいて決定された「暫定訳」がある。暫定訳がそのまま採用された場合は、確認作業のみで、修正作業は生じなかったことを意味する。そのため、暫定訳の採用率は支援の効果を評価する適切な指標となりうる。

表 1 に東京オリンピックの暫定訳の採用率を示す。この表では、氏名単位で評価した場合と、名前(姓または名)単位で評価した場合の両方を示している。氏名単位の場合は、以下の 4 種類に分類されている。

(1) 完全一致:氏名単位の既訳が辞書に存在し、それを暫定訳とした(既訳辞書による翻訳)

(2) 一致:氏名単位の既訳は存在しなかったが、姓および名の両方の名前単位の既訳が辞書にそれぞれ 1 つだけ存在した。それを組み合わせたものを暫定訳とした(既訳辞書による合成翻訳)

(3) 重複:姓または名のどちらかの既訳が辞書に複数存在した。頻度が高い方を暫定訳として採用した(既訳辞書による合成翻訳)

(4) 推定:姓または名のどちらかの既訳が辞書に存在しなかったため、トランスリタレータの出力(1 位)を暫定訳として採用した(カタカナ訳を新たに作成)

この表が示すように、氏名単位の採用率は全体では 90.4% であり、トランスリタレータを利用した場合でも 88.7% であった。これらの値より、翻訳支援は十分に機能したと考えてよいだろう。なお、ここでは、「完全一致」や「一致」の採用率が 100% ではないことに注意したい。これは、既訳がさまざまな事情(たとえば、他のメディアとの整合性)で修正されることを意味している。

名前単位の採用率は全体では 94.0% で、トランスリタレータの翻訳結果に限った場合は 92.3% であった。既訳辞書は、それまでのオリンピックや主要な国際大会等の翻訳名簿などを元に、参加者予想を加味して作成されたが、そのような既訳辞書を用いても、名前単位のカバー率が 40.2% にとどまることに注目したい。人名翻訳は、既訳辞書を整備すれば十分なカバー率が得られると思われるかもしれないが、オリンピック参加者名簿の翻訳に対しては、それは的外れな予想である。

6 経験から学んだこと

このプロジェクトを実施した経験から多くのことを学んだ。

(1) カタカナ表記の統制が重要である

暫定訳の採用率がこんなに高かったのは驚きであった。その一番の理由は、カタカナ表記の統制を厳格に行い、それを既訳辞書および対訳データに適用したことにあると思われる。新たなカタカナ訳を付与する場合は、標準的な規則に基づいてカタカナ訳を定めると考えられるが、カタカナ表記の統制は規則性の発見を容易にする。

表 1　東京オリンピックの暫定訳採用率

		採用数	採用率	修正数	計	割合
氏名単位	完全一致	2,552	94.0%	164	2,716	23.7%
	一致	726	95.2%	37	763	6.7%
	重複	5	100.0%	0	5	
	推定	7,091	88.7%	900	7,991	69.6%
	計	10,374	90.4%	1,101	11,475	100.0%
名前単位	辞書	7,528	96.6%	268	7,796	40.2%
	推定	10.693	92.3%	897	11,590	59.8%
	計	18,221	94.0%	1,165	19,386	100.0%

（2）安定したソフトウェアを用いる

長期に渡るプロジェクトでは、安定して動作し、よく保守されているソフトウェアを用いることが非常に重要である。トランスリタレータとして MeCab を用いたのは、ベンチマーク性能ではなく、安定性・移植性・実行速度を重視したためである。

（3）十分なリハーサルを行う

タスクの実行に時間的制約が課せられる場合は、十分にリハーサルを行うことが重要になる。2016 年のリオ・オリンピック、2018 年の平昌オリンピックの経験は、本番で非常に役に立った。

（4）現実のタスクは、ベンチマークとは異なる

研究では、公開されているデータセットを用い、正解が明確に与えられている環境でベンチマーク性能を競うことが多い。しかしながら、現実のタスクでは、しばしば正解は存在しない。自然言語処理の実社会応用においては、「ことば」が社会的存在であることを考慮したシステム設計・運用が重要となる。

参考文献

[1] 佐藤理史 . 外国人名対訳辞書の自動編纂を目指して . Japio YEAR BOOK 2009, pp250-253, 2009.

[2] 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 . Japio YEAR BOOK 2010, pp258-261, 2010.

[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.

[4] MeCab：Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/

[5] 佐藤理史 . 2020 東京オリンピック参加者名簿の翻訳、自然言語処理、Vol.30, No.2, pp748-772, 2023.