機械翻訳用超大規模辞書データ資源

著者	春遍 雀來
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	148-153
発行年	2017
URL	http://doi.org/10.15084/00001468

機械翻訳用超大規模辞書データ資源

春遍雀來（日中韓辭典研究所）

"Very Large Scale Lexical Resources for Machine Translation"

Jack HALPERN (The CJK Dictionary Institute, Inc. (CJKI))

要旨

情報交流の国際化に伴い多言語情報の充実は今や喫緊の課題である。特に固有名詞や **POI** (points of interest)は膨大な数量に加え頻繁な名称変更にも対応する必要があるため，正確で充実した多言語辞書データ資源が必須だ。そこで，機械翻訳の作業効率と精度を格段に向上させる，**超大規模辞書データ資源(Very Large Scale Lexica: VLSL)**の構築例として，固有名詞・専門用語等を含む日中韓英辞書データベースや多言語固有名詞辞書データベースを紹介する。VLSL は情報検索・形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野に応用が可能で更なる展開が期待される。

１．はじめに

近年，科学技術・学術・文化等の多方面で諸外国との相互理解・交流の重要性が再認識されている。2020 年の東京オリンピック開催に向け，多言語情報の充実は今や喫緊の課題となっている。IT 技術の発達に伴い，多言語情報は企業から一般ユーザーまで広く活用されるようになったが，そのような技術に不可欠なのが豊富な情報を包括した大規模な辞書データ資源である。

当研究所は，日中韓英を中心とする各種の辞書データベースの構築を行っており，固有名詞・専門用語の他，日本語の語彙・異表記等も含め約 2400 万項目を収録している。また，IT 関連の大手企業に広く採用されている中日・日中専門用語データベースは 20 分野に亘る専門用語を網羅した日中対訳辞書である。更に，動詞と形容詞・形容動詞を扱った日本語全活用辞典（J_FULEX）の開発もある。

これらの辞書資源は，人力による翻訳や機械翻訳の作業効率と精度を格段に向上させてきた一方，形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野で応用されている。

２．多言語固有名詞辞書データベース

辞書データ資源は翻訳のみならず，各種の言語データ処理の場面でも活用される。例えば自然言語処理に於いて特に扱いが難しい固有名詞では，多数の異表記（アラブ人名「アブドゥル・ラフマーン」には千通り以上のアルファベット表記がある）や平仮名表記の中国語訳（市町村名等）に対応しなければならない。また，一般語彙に於ける同義語（「ソフトウェア」は簡体字では「**软**件」，繁体字では「軟體」と表記）を処理する際にも辞書データベースは有用である。これらの点を踏まえ，当研究所では専門用語を含む膨大な辞書データベースの構築・拡張を続けている。

３．POI の辞書データベースと機械翻訳

地名や POI（points of interest = ホテル，公園，大学，施設等）は数が膨大である上，名称が変更される場合もある。各言語体系に基づく正しい表記が必要であるため，アルゴリズムによる全面的な自動処理での生成は不可能で，辞書データベースが必須となる。最先端のニューラル機械翻訳（NMT）ですら，POI の辞書データベースなしには学校名・道路情報での翻訳がほぼ不可能な事が Google の抜き取り調査から明らかになった。

POI の辞書データベースを含む超大規模辞書データ資源の構築は半自動的に行われ，結果に求められる精度と費用を勘案して自動翻訳と人間翻訳の割合を決定する事になる。特に固有名詞の翻訳作業では字訳・音訳・意訳・意音訳による自動変換と，人間翻訳という５通りの手法が数えられ，実際にはこれらの多様な組み合わせが可能である。つまり自動処理の割合が高く，安価で速いが精度が上がりにくいものから，人間翻訳で高価だが翻訳としての正確さを期す（定訳を選択する）ものまで様々である。

日本の地名・公共施設名

日本語	成田国際空港	京都府庁
中国語(簡体字)	成田国际机场	京都府厅
中国語(繁体字)	成田國際機場	京都府廳
韓国語	나리타국제공항	교토부청
英語	Narita International Airport	Kyoto Prefectural Office
アラビア語	مطار ناريتا الدولي	مكتب محافظة كيوطو
インドネシア語	Bandar Udara Internasional Narita	Kantor Pemerintahan Kyoto
ベトナム語	Sân bay quốc tế Narita	Tòa nhà chính quyền tỉnh Kyoto
タイ語	สนามบินนานาชาตินาริตะ	ที่ว่าการจังหวัดเกียวโต
ヒンディー語	नारिता अंतर्राष्ट्रीय हवाई अड्डा	क्योतो प्रीफ़ेक्चर मुख्यालय
ロシア語	Международный аэропорт Нарита	администрация префектуры Киото
ドイツ語	Internationaler Flughafen Narita	Präfekturverwaltung Kyoto
ポルトガル語	Aeroporto Internacional de Narita	Sede do Governo de Quioto
スペイン語	Aeropuerto Internacional de Narita	Oficina Prefectural de Kyoto
フランス語	Aéroport international de Narita	Préfecture de Kyoto
イタリア語	Aeroporto Internazionale di Narita	Sede del Governo prefettizio di Kyoto

４．日本語異表記データベース

日本語は表記の幅が広い言語であり，日本語異表記の種類には，漢字表記・平仮名表記・片仮名表記・交ぜ書き等がある。更に片仮名語の異表記（コンピュータとコンピューター，メイドとメード等）も多数出現する。また，同音異形異義語の具体例には，うまい＝美味い，上手い，巧い等，意味や表記の揺れが認められる。更に，日本語を扱う際には意味互換性の度合いや同訓異字への対応，異表記の類型（送り仮名や文字種等），詳細な属性等きめ細やかな配慮が常に求められる。

当研究所は自然言語処理で課題となるこれら異表記の問題を，データベースに全てを包括する事によって解消している。各種国語辞典・内閣告示・新聞や公用文に見られる表記・出現頻度等，様々な角度から総合的に判断した「代表表記」を定める作業が，現在も進行中である。

日本語異表記辞書データサンプル

ID	読　み	POS	SUB_ID	表　記	代表表記
F000043	あっせん	VN	a	斡旋	あっせん
			b	あっせん	
			c	あっ旋	
F000690	あかとんぼ	NC	a	赤とんぼ	赤とんぼ
			b	赤トンボ	
			c	赤蜻蛉	
			d	アカトンボ	
			e	あかとんぼ	
F000853	あきかん	NC	a	空き缶	空き缶
			b	空缶	
			c	明き罐	
			d	あき缶	
			e	あき罐	
			f	空きかん	
			g	空きカン	
			h	空き罐	
			i	空罐	
			j	空き鑵	
			k	空鑵	
F001543	あじつけ	VN	a	味つけ	味付け
			b	味付け	
			c	味付	

５．固有名詞情報と VLSL(超大規模辞書データ資源)

当研究所では，こうした異表記を網羅する日中韓英各語とアラビア語の大規模な辞書データ資源を提供しており，世界の大手企業もこれを採用している。中国語のデータベースには検証済みの正確なピンインも収録されている。先進的な計算辞書学の手法によって構築・維持された当研究所のデータ資源は，固有名詞・専門用語のほか，日本語の語彙・異表記・音韻等も含め，約2400 万項目に上る。

日中韓英固有名詞データベースの収録語数

	日英	日中	日韓
中国人名	1,000,000	1,000,000	1,000,000
中国地名	2,400	5,600	3,000
韓国人名	13,000	2,100	13,000
韓国地名	5,900	2,000	5,900
日本人名	390,000	281,000	390,000
日本人姓	150,000	91,000	150,000
日本地名	77,000	74,000	77,000
西洋人名	31,000	38,000	10,000
西洋地名	1,100	2,500	1,800
合　計	1,670,400	1,496,200	1,650,700

「日中韓英固有名詞データベース」は日中韓英語の各種固有名詞辞典を含み，総計1100 万項目に及ぶ大規模なデータベースである。その用途は機械翻訳，情報検索，形態素解析，電子辞書，入力システム，固有名認識等多岐に亘る。

日中専門用語データベース

分野	中国語	日本語
医学	肾上腺素能受体	アドレナリン受容体
生物	亲和性	親和性
生物	亲和层析法	アフィニティークロマトグラフィ
生物	琼脂扩散法	寒天拡散法
生物	琼脂糖	アガロース
生物	琼脂胶	アガロペクチン
生物	类蛋白	アルブミノイド
医学	类天花	アラストリム
医学	变应性试验	アレルギー試験
医学	变应性肉芽肿	アレルギー性肉芽腫

「中日日中専門用語データベース」は日中二ヶ国語の双方向対訳辞書である。コンピュータ科学からバイオテクノロジーに至る 20 分野に亘る幅広い専門用語を収録しており，収録語は中日・日中それぞれ約 80 万語，総計約 160 万語に及ぶ。その用途は特許翻訳を含む各種翻訳業務，用語の抽出やインデックス作成に役立つ情報検索アプリケーション，形態素解析や分節システム等，各種の自然言語処理アプリケーション，スマートフォンアプリケーションや電子辞書・CD-ROM 等多岐に亘る。

6．まとめ

POI の辞書データベースを含む VLSL（超大規模辞書データ資源）は各種の自然言語処理に向いており，とりわけ機械翻訳に有効である。コンピュータメモリーが無制限に拡大可能になった今日，自然言語処理に於いてはアルゴリズムやコーパスのみに過度に依存する必要はもはやない。VLSL や POI の辞書データベースの効果的な活用は固有名詞の翻訳精度を大幅に向上させるばかりではなく，情報検索や形態素解析・固有表現認識・用語抽出等，自然言語処理の幅広い分野に応用が可能であり，更なる展開が期待されるのである。