2-3-2　アジア言語処理
2-3-2　Asian Language Processing

丁 塵辰
DING Chenchen

現在、自然言語処理技術における研究・応用はヨーロッパ言語及び一部のアジア言語に集中しており、広く世界の言語を翻訳するという究極の目的を実現するためには、研究が未熟である言語の抱える中核的な課題を解決し、実用レベルにしなければならない。本稿では、アジア地域、特に東南アジア諸国連合（ASEAN）地域の諸言語を対象とする言語処理技術の近年の発展について述べる。

The research and applications of natural language processing technology are concentrated on European languages and some major Asian languages. To realize an ultimate goal of universal translation of the world's languages, it is important to solve essential problems and develop practical applications for those understudied languages. This paper describes recent developments in language processing technology in the Asian region, especially in the region of the Association of Southeast Asian Nations (ASEAN).

1　背景

近年、深層学習などの機械学習理論の実用化大規模データの蓄積及び計算機の演算処理能力向上に従い、自然言語処理という分野が一層脚光を浴びている。学界における研究及び業界における応用技術開発が共に劇的に進展している。しかしながら、自然言語処理技術の研究・応用はヨーロッパ言語及びアジアの日中韓に偏しており、他の言語に関する研究は、未熟または未開拓と言える。日進月歩の言語処理技術にもかかわらず、世界中各言語における資源整備・処理技術の格差が広がる。

これら言語間研究上の巨大な格差は、処理技術より、下敷きとなる言語資源整備の段階から生じる。英語・日本語は1990年代から大規模言語リソースが整備されており、多数のヨーロッパ言語について2000年代以来整備されつつある。先端の言語処理技術は極端に言語資源に依存する一方、多数のアジア言語は資源が極めて不足している状態であり、研究は展開できない状態となっている。

アジア地域の言語処理技術及び実用化を向上するために、2016年以来、NICT・先進的音声翻訳研究開発推進センター（ASTREC）にてASEAN地域の東南アジア諸言語を中心とするデータ整備・研究開発を推進している。本稿ではこの期間の研究活動を紹介する。以下は「言語資源の整備」、「解析技術の発展」、「翻訳性能の評価」に分けて詳細を述べる。

2　言語資源の整備

「アジア言語ツリーバンク」[1]は東南アジア諸国連合（ASEAN）の公式言語を中心とする、言語学的情報付きの大規模データセットである。具体的に、東南アジア大陸部のビルマ（ミャンマー）語、タイ語、クメール（カンボジア）語、ラーオ語、ベトナム語及び島嶼（とうしょ）部のマレー・インドネシア語、タガログ（フィリピン）語である。翻訳は英語・日本語・中国語・ヒンディー語・ベンガル語にも拡張した（表1）。本プロジェクトは2016年に発足し、2016年度から2019年度の3年間、ICT Virtual Organization of ASEAN Institutes and NICT（ASEAN IVO）プロジェクトに採択され、ASEAN地域各地の大学・研究機構との広範囲の連携により推進した。ASEAN IVO Forum 2019にて、最優秀貢献賞（Excellent Contribution Award）を受賞した。

具体的に、20,000文の新聞記事を上述諸言語に翻訳した多言語対訳データを下敷きとしている。それぞれの言語の言語学的な分析に基づき、工学的自動処理に有益な情報を付け加える。具体的には、表層的な分かち書き・品詞情報及び深層的な構文情報が挙げられる（図1）。このデータセットの整備により、ASEAN地域の多数の言語における自動処理が初めて可能になり、

表 1　アジア言語ツリーバンクにおける 13 言語対訳データの一例。言語コード順で並べる。

コード	言語名	例文
bn	ベンガル語	ফ্রান্সের প্যারিসের পার্ক দি প্রিন্সেস-এ হওয়া ২০০৭-এর রাগবি বিশ্বকাপের পুল সি-তে ইটালি পর্তুগালকে ৩১-৫ গোলে হারিয়েছে।
en	英語	Italy have defeated Portugal 31-5 in Pool C of the 2007 Rugby World Cup at Parc des Princes, Paris, France.
hi	ヒンディー語	2007 में फ़्रांस, पेरिस के पार्क डेस प्रिंसेस में हुए रग्बी विश्व कप के पूल C में इटली ने पुर्तगाल को 31-5 से हराया।
id	インドネシア語	Italia berhasil mengalahkan Portugal 31-5 di grup C dalam Piala Dunia Rugby 2007 di Parc des Princes, Paris, Perancis.
ja	日本語	フランスのパリ、パルク・デ・プランスで行われた 2007 年ラグビーワールドカップのプール C で、イタリアは 31 対 5 でポルトガルを下した。
km	クメール（カンボジア）語	អ៊ីតាលីបានឈ្នះលើព័រទុយហ្គាល់ 31-5 ក្នុងប៉ូលCនៃពិធីប្រកួតពានរង្វាន់ពិភពលោកនៃកីឡាបាល់ឱបឆ្នាំ2007ដែលប្រព្រឹត្តនៅប៉ារីសរដូវ ក្រុងប៉ារីស បារាំង។
lo	ラーオ語	ອິຕາລີໄດ້ເສຍໃຫ້ປຶ້ອກຕຸຍການ 31 ຕໍ່ 5 ໃນພູລ C ຂອງ ການແຂ່ງຂັນຣັກບີ້ລະດັບໂລກປີ 2007 ທີ່ ປາກເດແພຣ້ງ ປາຣີ ປະເທດຝຣັ່ງ.
ms	マレー語	Itali telah mengalahkan Portugal 31-5 dalam Pool C pada Piala Dunia Ragbi 2007 di Parc des Princes, Paris, Perancis.
my	ビルマ（ミャンマー）語	ပြင်သစ်နိုင်ငံ ပါရီမြို့ ပါ့ဒက်စ် ပရင့်စက် ၌ ၂၀၀၇ခုနှစ် ရပ်ဘီ ကမ္ဘာ ဖလား တွင် အီတလီ သည် ပေါ်တူဂီ ကို ၃၁-၅ ဂိုး ဖြင့် ရေကူးကန် စီ တွင် ရှုံးနိမ့်သွားပါသည် ။
th	タイ語	อิตาลีได้เอาชนะโปรตุเกสด้วยคะแนน31ต่อ5 ในกลุ่มc ของการแข่งขันรักบี้เวิลด์คัพปี2007 ที่สนามปาร์กเดแพร็งส์ ที่กรุงปารีส ประเทศฝรั่งเศส
tl	タガログ（フィリピン）語	Natalo ng Italya ang Portugal sa puntos na 31-5 sa Grupong C noong 2007 sa Pandaigdigang laro ng Ragbi sa Parc des Princes, Paris, France.
vi	ベトナム語	Ý đã đánh bại Bồ Đào Nha với tỉ số 31-5 ở Bảng C Giải vô địch Rugby thế giới 2007 tại Parc des Princes, Pari, Pháp.
zh	中国語	意大利在法国巴黎王子公园体育场举办的 2007 年橄榄球世界杯 C 组以 31-5 击败葡萄牙。

更に高度な自動翻訳の実現につながる。

3　解析技術の発展

日本語のテキストを処理する際、分かち書き、専門用語・固有表現の同定、構文解析などの自動解析が求められる。上述したアジア言語ツリーバンクの整備により、このような基本解析技術が ASEAN 諸言語にも発展できるようになっている。特にビルマ語、クメール語のようなデータ・従来研究が極めて少ない言語に、ゼロから研究の基盤を構築した。代表的な研究活動を以下のいくつかを挙げる。

3.1　解析向けの表層的表記手法の設計

言語は形態的な特徴によりおおむね、語形変化を多く持つ「屈折語」、接辞などの語尾を重ねる「膠着語（こうちゃく）」及び語形変化一切無しの「孤立語」に分類できる。日本語は典型的な膠着語であることに対して、東南アジア大陸部にある多くの言語は「孤立語」の特徴が強い。こういう言語に「語」と「品詞」の定義自体が曖昧であり [2]、これらの概念をはっきり整理するのは、高度な言語学的分析が必要とされる一方、工学処理に効率的に有用な情報を導入するのは課題とされる。これに対して、「NOVA」[3] といった表記体系を提案した。基本タグを設け、さらに括弧表記による単語境目の曖昧性を一定程度に許容する（図 1 には単語番号の上 1 段の記号列）。言語学に特に詳しくない母語話者によりも、語感だけでスクラッチからデータの初歩的な整備を実現できている。これをベースに、表記の更なる精密化を段階的に図る（図 1 には上部の木構造）。この手法により、ビルマ語、クメール語のデータ整理はゼロから素早くできるようになっていた [4]–[6]。

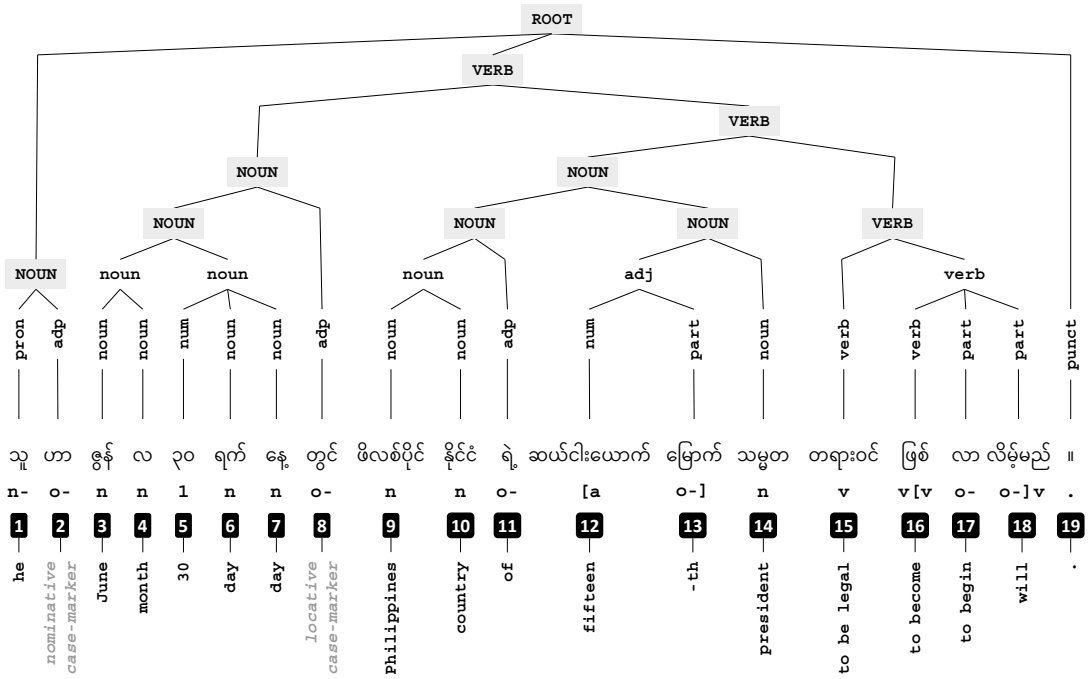

図 1　言語学的情報付きの一例 (ビルマ語) [5]。数字は分かち書き後の形態素の順番を示している。数字の下はそれぞれの形態素の英訳・解釈をつけている。数字以上の部分は各段により、NOVA の表記、形態素そのもの、更に詳し品詞情報、各レイヤの意味単位 (木構造)。

3.2　単語分かち書きと構文解析の統合

上述したように「語」と「品詞」の曖昧さについて、更に展開するとすなわち「意味単位」の同定につながる。日本語、中国語のような処理技術が進んでいる言語において、複数の分かち書き・品詞体系が確立され、運用上のネックとなる場合もある。近年技術の発展により、文字単位のような語より小さい粒度から直接モデルを学習することができるようになり、ある程度人間による定義の曖昧さを回避できていた。

「NOVA」は設計する時点で単語境目にある曖昧性を許容する。すなわち「意味単位」の粒度を従来の「語」「節」「句」「文」など定義を回避し、「意味単位」の入れ子構造のみに注目している。これは工学処理に向けて言語における根本的な構造情報を導入するためである。実際の工学処理上、分かち書き、品詞付与、構文解析などの複数のツールの開発により、一つの言語にモデルを一つに統合し訓練すればよいことになる。更に高度なタスクに解析情報を提供でき、ツールの開発・保守の労力を低減できている。

3.3　人名地名の転写・翻字の自動処理

実世界のテキストデータに、大量の人名・地名・専門用語が含まれている。特に多くのアジア言語はそれぞれ固有な文字体系を持っており、外来語の表記に揺れがある。日本語のカタカナ表記にも同じような問題があることに対して、東南アジア地域に使われるより

図 2　クメール語の文字・音韻構造による分析例 [7]。クメール語人名とローマ字表記を構成要素・文字単位に分解して、更に補助記号を挿入した。数字は両方対応している記号を示している。分かち書きのスペース (4 番) も普通の一文字として扱う。クメール語側に楔形が示した記号二つとローマ字側の 7 番、9 番の記号は対応付けのために挿入した補助記号である。このような分解・挿入・対応付けは、文字・音韻構造の事前知識を自動処理のために導入した。統計的なモデルはより簡単・正確に学習できるようになった。

複雑な文字体系において、この問題が更に深刻になる。現代社会で流行文化により外来語の移り変わりも加速しているので、単なる辞書の編集で問題の解決にたどり着けるのが困難である。

この問題の解決には、データに基づくアプローチが求められる。前述したツリーバンクは言語の形態的・構文的な情報に注目することに対して、ここで言語の音韻的・正書法的な特徴に注目する。ASEAN 地域の連携により、大規模な人名転写・翻字データセットを整備しつつある [7]–[9]。典型的な成果として、クメール語の文字・音韻構造を分析し、そのローマ字表記に関する研究 [7] (図 2) は、「International Conference of the Pacific Association for Computational Linguistics」の「Best Paper Award」を受賞した。

4　翻訳性能の評価

アジア翻訳ワークショップは2014年発足した、アジア言語を中心とする最先端翻訳技術の評価キャンペーンである。発足当時、日本語と中国語の技術・特許文書の翻訳タスクを中心としていたが、徐々にアジアの多言語・多分野の翻訳に展開していく。アジア諸言語のデータ整備に従い、ビルマ語 [10]（2018 年）、クメール語 [11]（2019 年）及びインド語群の多くの言語の翻訳タスクを導入した。

これらのアジア言語の翻訳タスクに対して、Facebook AI（現在 Meta AI）を含む世界中の先端研究チームが興味を示し、参加することになった。アジアの未開拓の言語と最先端の翻訳技術と結び付けている。翻訳性能上、ビルマ語・クメール語の新聞・法律文書において、おおよそ2010年代の日本語における自動翻訳性能に達していることが分かった。

5　終わりに

本稿では、近年 NICT・ASTREC にてアジア言語処理に関する研究活動を紹介した。アジア言語ツリーバンクプロジェクトを主軸としてデータ整備より着手し、複数のアジア言語におけるデータ不足の現状を大幅に改善した。これに基づき、基盤解析技術及び自動翻訳等の応用技術を発展させた。今までの研究により、諸アジア言語の処理技術を短時間で引き上げた。

【参考文献】

1 Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding, “Introduction of the Asian Language Treebank,” Proc. of O-COCOSDA, pp.1–6, 2016.

2 Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita, “Word Segmentation for Burmese (Myanmar),” ACM Transactions on Asian and Low-Resource Language Information Processing, vol.15, Issue 4, Article no.22, 2016.

3 Chenchen Ding, Masao Utiyama, and Eiichiro Sumita, “NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol.18, Issue 2, Article no.17, 2018.

4 Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita, “Towards Burmese (Myanmar) Morphological Analysis: Syllable-based Tokenization and Part-of-Speech Tagging,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol.19, Issue 1, Article no.5, 2019.

5 Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita,“A Burmese (Myanmar) Treebank: Guideline and Analysis,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol.19, Issue 3, Article no.40, 2020.

6 Hour Kaing, Chenchen Ding, Masao Utiyama, Eiichiro Sumita, Sethserey Sam, Sopheap Seng, Katsuhito Sudoh, and Satoshi Nakamura, “Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion,” ACM Transactions on Asian and Low-Resource Language Information Processing, vol.20, Issue 6, Article no.104, 2021.

7 Chenchen Ding, Vichet Chea, Masao Utiyama, Eiichiro Sumita, Sethserey Sam, and Sopheap Seng.,“Statistical Khmer Name Romanization,” Proc. of PACLING, CCIS 781, pp.179–-190, 2018.

8 Chenchen Ding, Win Pa Pa, Masao Utiyama, and Eiichiro Sumita,“Burmese (Myanmar) Name Romanization: A Sub-syllabic Segmentation Scheme for Statistical Solutions,” Proc. of PACLING, CCIS 781, pp.191–202, 2018.

9 Aye Myat Mon, Chenchen Ding, Hour Kaing, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita,“A Myanmar (Burmese)-English Named Entity Transliteration Dictionary,” Proc. of LREC, pp.2973–2976, 2020.

10 Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi,“Overview of the 5th Workshop on Asian Translation,” Proc. of PACLIC, pp.904--944, 2018.

11 Toshiaki Nakazawa, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi, “Overview of the 6th Workshop on Asian Translation,” Proc. of WAT, pp.1–35, 2019.

丁 塵辰（てい じんしん）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的翻訳技術研究室
主任研究員
博士（工学）
計算言語学、自然言語処理

【受賞歴】
2017 年　Pacific Association for Computational Linguistics, Best Paper Award