日本語からハンガリー語への機械翻訳システムの研究開発

山形大学　　VARGA István、山本 広志

本研究は初の日本語からハンガリー語への機械翻訳の研究開発で、実験的翻訳システムを作成した。日本語とハンガリー語の構造や文法の比較分析を行ない、日本語とハンガリー語に最適な翻訳システムを考案し、特定の分野の翻訳システムとデータベースを開発した。さらに、本翻訳システムの評価を行った。

I　序

機械翻訳

翻訳を機械化する考えは17世紀まで遡ることができるが、実際の方法は20世紀までなかった。[1] 現在まで様々な方法[2~6]が提案されたが、全言語を翻訳できる統一された方法はない。対象となる言語の構造や単語の意味の範囲、語順などの特徴がそれぞれ異なるためである。例えば、同じ語族や言語構造が近い言語の場合は文法規則に基づく機械翻訳方法が効果的であるが、構造が異なる言語の場合はコーパスに基づく方法の方が将来性があるとされている。[2] それゆえに、対象となる言語によって最も将来性がある翻訳方法を研究する必要がある。

ハンガリー語と日本語

ハンガリー語は、主に中部ヨーロッパで話されている言語である。ウラル語族のフィン・ウゴル語派に分類され、ヨーロッパで話される言語の多くが属するインド・ヨーロッパ語族と系統が異なるため、アジア系の言語と呼ばれる。[7,8] しかし、ウラル語族は、日本語の属するアルタイ語族との系統関係が実証されてはいない。[8]

ハンガリー語は語幹に機能語や接続語が付き、形態的に膠着語に分類される。日本語も膠着語であるため、ハンガリー語と統語論的や形態論的な類似点があると思われる。統語論的比較では、日本語の語順はSOV(主語+目的語+動詞)であり、ハンガリー語も基本的な語順はSOVで特定の節を強調する際には語順が変わる。両言語とも比較的自由に語順を変えることができ、文中の依存関係は変化しない。日本語で文中の語順を変えた場合、ハンガリー語訳の語順も意味を変えず同じように変えることができる。

形態論的比較では、日本語とハンガリー語の動詞はともに現在時制と過去時制しかない。また、両言語とも膠着語であり、平叙形、仮定形、命令形[9]といったハンガリー語の活用にも日本語と類似性があるが、ハンガリー語の場合は動詞が人称と数によっても活用する。日本語の形容詞、形容動詞と動詞の活用には類似性があるが、ハンガリー語の形容詞と動詞の活用は全く異なる。また日本語の形容詞、形容動詞とハンガリー語の形容詞にはともに副詞的に使える連用形が

あるという点は似ているが、ハンガリー語では形容詞の比較級と最上級が活用によって形成されるという点が異なる。日本語とハンガリー語ともに副詞は活用しない。

日本語の名詞に格助詞を付けると節に意味づけがされる。ハンガリー語の名詞の場合は人称と数と格[10]によって異なる接尾辞が付く。

II　研究方法

日本語とハンガリー語の実際の構造的、文法的などの相違の研究は、効果的な日本語-ハンガリー語の機械翻訳システムの基礎となる。そのため日本語とハンガリー語の特徴をふまえて実験的機械翻訳システムと翻訳データベースを作成した。対象となる言語によって機械翻訳の最適な方法が異なるため、日本語からハンガリー語への翻訳に最適な方法を考案した。

III　研究結果

本研究の翻訳方法

日本語とハンガリー語の文法には共通点があるが、日本語とハンガリー語の形態論的違いは大きい。そのため、文法規則に基づく機械翻訳方法では成功率が低いと考えられる。その上、ハンガリー語には複雑な活用や語形変化があり、また文法的例外が非常に多いため、日本語からハンガリー語への機械翻訳システムに文法規則に基づく方法は不適切であると考えられる。そこで日本語－ハンガリー語のコーパスに基づく方法で機械翻訳システムを作成した。

今のところコンピュータ用の日本語とハンガリー語の二ヶ国語コーパスは存在しない。一人の作成者が二ヶ国語のコーパスを作成するには通常数年を要するため、本システムでは分野を限定してコーパスを作成した。天気予報で利用されている文では単語や表現が少ない一方、文の構造は一般的な文と大きく変わらない。そこでこの分野のコーパスを作成することにし、機械翻訳の実験を行った。

また、日本語とハンガリー語の単語の意味する範囲は異なり、また様々な概念がしばしば一語では表現できないため、節の辞書を作成した。

本研究の実験的翻訳システムでは、元の日本語の文にある依存関係を判断せずに翻訳を行なう。日本語とハンガリー語の語順は近く、またハンガリー語は一つの文にいくつかの正しい語順がある。そのため、このことが翻訳結果にどう影響するかということを検討する必要がある。また、単語の辞書ではなく節の辞書を利用することによって言語特有の表現や連語が正しく翻訳されると考えた。

本研究のシステムは機械翻訳を境界判断・言語置換の二段階で行う。

1 境界判断

文分割アルゴリズムは一つの文にいく通りかの形態論的に正しい分割を行う。第一段階として、日本語の文の中で日本語シソーラスに登録されている全ての形態素を抽出する。第二段階と

して不適切な形態素を削除する。第三段階では削除されなかった形態素から元の文を再現できるかどうかを確認する。第三段階でも複数の候補が残った場合は、全ての候補を次の言語置換に渡す。

2 言語置換

言語の置換は統計的翻訳方法にある翻訳モデルと言語モデル[3]を元にし、二段階で行う。翻訳モデルでは「類語大辞典」[11]にある三階層の日本語シソーラスと節の辞書を用い例文に基づく翻訳方法によってハンガリー語の訳文の候補を複数組み立てる。生成したハンガリー語文と元の日本語の文の忠実性を抽出された節の数によって数値で評価する。次に言語モデルに基づいて、ハンガリー語の訳文にある単語の共起からn-gram(n≦3)によって理解容易性を計算する。[3] 忠実性の値と理解容易性の値によって最も可能性が高い翻訳文が選択される。

翻訳評価

本システムを評価するため、二ヶ国語のコーパスに用いたのと同じ天気予報でコーパスに保存されていない、文字数が平均18.3字の新たな100文を利用した。82%の文は分割に成功し、平均で2.1種類の分割が出力された。18%の文が正しく分割されなかった原因は、シソーラスに保存されていない地名、形態素や活用形があったためだった。

次に翻訳を評価するため訳文を五段階に分類した。本研究のシステムは実験的システムであるため翻訳方法が単純だが、正しく分割された文の中で「正しく翻訳した文」と「訳文が間違っているが、全体の意味は理解できる文」は合わせて49.8%で、「翻訳文が間違っているが、部分的には理解できる」は24.3%で、「少数の単語が翻訳されたが文を理解できない」と「完全に失 敗した翻訳文」は25.6%であった。

依存関係の判断を行わなかったため、複文や重文の多くは部分的に翻訳されたが、訳文にある節の関係が不明、あるいは語順や節の順が異なったため文を全体的に理解できない場合が少なくなかった。しかし、ほとんどの短い文の場合は日本語の文とハンガリー語の文の語順が近かったため、生成されたハンガリー語訳にある依存関係は正しかった。複文や重文を正しく翻訳するためには翻訳モデルと言語モデルの改良を行う必要がある。

23%の単語や形態素が辞書から抽出されなかった。また、単語の訳が抽出されても不適切な訳が選ばれる場合が少数あった。コーパスや節の辞書が小さいことが原因と考えられる。コーパスと節の辞書の改良が必要である。

意味カテゴリーを利用している検索モデルは本システムの重要な部分で、それ無しでは多くの適切な節が抽出されない。したがって、検索モデルを利用しない場合は結果が大きく変化すると予測したが、実際にはさほど大きな変化は見られなかった。これは、コーパスが小さいために検索モデルにそれほど影響が現れなかったことが原因だと考えられる。

IV　まとめ

本研究では初の日本語からハンガリー語への機械翻訳システムの研究開発を行なった。現在までに開発された翻訳方法、また日本語とハンガリー語の比較研究を行ない、天気予報の分野のコーパスに基づく実験的翻訳システムを作成した。日本語とハンガリー語の特徴から、依存関係を判断せずに正しい訳文が生成されるということが分かった。しかし、複文や重文の翻訳成功率が低い、また抽出できなかった単語があり節の辞書の大きさが不十分であるということが課題として明らかになった。

参考文献

1. Hutchins, W. John: 「Two precursors of machine translation: Artsrouni and Trojanskij」, International Journal of Translation **16**(1), 11-31, (2004).
2. Nagao M.: 「A Framework of a Mechanical Translation between Japanese and English by Analogy Principle」, Artificial and Human Intelligence, Elsevier, (1984).
3. Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P:「A Statistical Approach to Language Translation」, In COLING-88, **1**, 71-76, (1988).
4. Turcato, D., Popowich, F., McFetridge, P., Toole, J.:「A Unified Example-Based and Lexicalist Approach to Machine Translation」, In Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation, 33-43, (1999).
5. Shirai S., Bond, F., Takahashi Y.:「A Hybrid Rule and Example-based Method for Machine Translation」, NLPRS-97, 49-54, (1997).
6. Ikehara S., Shirai S., Yokoo A., Nakaiwa H.:「Toward an MT system without pre-editing – effects of new methods in ALT J/E」, MT Summit III, 101-106, (1991).
7. Stemler Ágnes: 「Tudománytörténet: Mátyás Flórián munkásságáról」, Magyar Nyelv **5**(1), 84-91, (2004).
8. Érdi Miklós: 「A sumír, ural-altaji, magyar rokonság kutatásának története」, Gilgamesh, New York, (1974).
9. Tompa J.:「A mai magyar nyelv rendszere」, Akadémiai Kiadó, Budapest, (1970).
10. Abondolo, D.: 「The Major Languages of Eastern Europe」, Routledge, London, (1987).
11. 柴田武、山田進：「類語大辞典」, 講談社, (2002).