日本語から手話への機械翻訳に関する研究

徳田 昌晃
北陸先端科学技術大学院大学 情報科学研究科

1996 年 2 月 15 日

要約

子供の頃に音声言語が習得できなかった聾唖者が使う言語に手話がある。手話は手、腕、表情などを駆使する言語であり、日本語や英語などの音声言語とは表現が大幅に違うが、近年の言語学の研究成果によれば、手話もまた独自の体系を持つ言語であることが明らかになっている。

手話が他の言語と違うことを特徴づけているのは視覚的な言語であることである。そのため、計算機で扱うためには入出力が壁となる。画像処理技術を用いた動作認識や手話の画像合成、特殊な装置を使った手話の単語認識の研究は数多く行われているが、未だに実用的な入出力が出来るようにはなっていない。また、これらの手話の入出力処理だけでも現在の計算機にとってはかなりの負荷となる。この問題とは別に手話を自然言語として処理するためには、計算機の内部表現として手話を記号列で表現する必要がある。手話を記録する際に使用する記述法や手話と日本語との間の翻訳システムは社会的な要請もあり、それらを実現する意義は大きい。そこで本研究では、手話の記述法として 「手話表記法 (文)」を提案する。手話表記法は、日本語単語をラベルとして使用して手話を文字列として記述する手法であり、手話単語の語順、手の方向、表現回数を明示的に記述することができる。手話表記文は手話単語と手話の実際の表現に関する情報のつながりを記録しておく必要があるが、計算機で容易に手話を扱えるように、また人間が直接読み書きできるように考案されている。システムは日本語から手話表記文への翻訳のみを扱い、既存の研究によって手話の入出力データを手話表記文に簡単に変換できることを仮定することにより、副次的に本研究では画像処理のような手話の入出力処理を扱わないことにする。

本研究で扱う機械翻訳は変換方式によるものである。手話はかなりの影響を日本語に受けていることと、手話表記文と日本語には表記上類似性があるので、入力された原言語の統語構造を解析して目標言語の訳文を生成する変換方式が適切と考えたためである。変換方式では、両言語の構文規則、単語辞書が必要となる。試作した日手機械翻訳システム 「手話ん」では原言語側の日本語の処理や辞書と構文規則は既存のものを使う。目標言語側の手話の辞書は、市販されている手話単語辞書より見出し語を収集し、聾唖者や手話通訳者の協力を得て、約 3000 語の手話単語を選定し、機械可読な手話単語辞書として作成し使用する。今回のシステムの実験には入力文としてニュース文を使うことで、手話表記文と日

本語文の構文がほぼ同じであることを期待し、単語の変換を中心にシステムの実装と検討を行った。

試作した日手機械翻訳システム「手話ん」は次の 4 段階の処理を行い、日本語から手話への変換を実現する。

1. 形態素解析と不要形態素の除去
2. 変換規則の適用と手話単語への直接変換
3. 類似語を使った手話単語への変換
4. 指文字表記への変換

形態素解析は日本語形態素解析システム JUMAN を使用する。「手話ん」は解析後の結果より、変換不要の形態素を取り除く。

次にそれぞれの形態素が手話単語辞書に存在するかを調べ、存在したら手話単語へ変換する。同時に月日の表現は句の変換規則を、活用語尾には形態素の変換規則を適用し手話表記文に変換する。

日本語から手話への翻訳の際に問題になるのが、手話単語辞書の語彙の少なさである。「手話ん」では入力された日本語単語に対して該当する手話単語が辞書に存在しない場合に、計算機可読な日本語辞書である EDR 電子化辞書を使用し、次の 3 つの方法で類似した日本語単語を獲得し手話単語に変換する方法を検討した。

1. EDR 日本語単語辞書には見出し語ごとに概念識別子が付加されている。同じ概念識別子を持つ単語は同じ意味を表現する単語とみなすことができる。そこで、元の単語と同じ概念識別子を持つ単語を類似語として獲得する。
2. 概念識別子には概念記述と呼ばれる単語の説明文が付加されている。概念記述は比較的定型文であることが多かったので、元の単語の概念記述より主要な単語を抽出規則により取り出し、それらを手話単語に変換する。
3. EDR 概念辞書では概念識別子を階層化し、抽象的な概念を上位に、具体的な概念を下位に分類している。そこで元の単語の上位概念にあたる単語と概念記述より類似語を獲得する。

今回のシステム「手話ん」は複数の類似語が獲得できた場合の絞り込みが自動化できなかったので、実験の際には手作業でそれらの中に適切な手話単語が含まれているかを確認した。

最後にこれらの方法でも獲得できなかった単語は品詞が名詞ならば指文字表現に変換される。指文字表現は外来語など明らかに手話単語が存在しない場合によく使われる方法である。

実験には入力文として NHK の手話ニュースを使用した。手話ニュースは日本語と手話の両方のナレーションがつくので対訳文が得やすく、ニュースのために様々な単語の手話表現を得られることを期待した。なお、形態素解析の結果が「手話ん」の実験に影響しな

いように、形態素の区切りなどを修正している。実験の結果、入力された形態素のうち、32%が変換不要、26%が手話単語へ直接変換でき、12%が変換規則により変換できた。さらに 26%は類似語の獲得により変換できると思われる。残りの形態素の 3%は指文字表記に変換され、最終的に変換に失敗したのは全体の 4%であった。文章全体で約 95%が手話表記文に変換できたことになる。変換に失敗したものは EDR 電子化辞書の概念記述自体が適切でなかったものと入力される形態 (ひらがな表記や形態素区切りの誤り) に起因するものが多かった。

今後は複数の類似語候補の絞り込みの自動化の手法などを検討していきたい。

キーワード: 手話、機械翻訳、計算機可読辞書、類似語