日本語から手話表記文への変換に関する検討

徳田 昌晃 , 奥村 学
tokuda@jaist.ac.jp , oku@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

聴覚障害者の中でも幼児期に音声言語が習得できなかった場合、手話を母国語として習得する者が多い。手話は手、腕、顔の表情などを駆使する言語であり、単に日本語や英語などの音声言語と表現方法が違うだけではなく、独自の体系を持つ言語であり、近年ではこのような手話を日本手話と呼んでいる。

手話を扱った研究では、画像処理やデータグローブによる手話の認識やCGによる手話画像の合成が盛んに行われているが、これらは手話と日本語の単語が1対1で対応し、構文も日本語と同じである事を仮定した日本語対応手話を扱ったものが多い。聾者が使う手話はほとんどが日本手話であり、日本語と手話間の機械翻訳システム (以下、日手機械翻訳システム) を実現するためには、構文変換を含めた言語処理が必要になる。

本稿では日本語から手話単語列への変換方法を中心に、日手機械翻訳システム「手話ん」を試作し評価を行う。

2 日手機械翻訳システムの概要

本研究の最終目的は日本語から手話への機械翻訳システムを実現することである。現在考えている手話通訳システムの概観を図1に示す。

2.1 手話表記法 (文)

手話は視覚的な言語であるため、計算機で扱うには入出力が壁となる。手話の認識や画像合成の研究は数多く行われているが、未だに実用的な入出力装置は実現していない。また、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。しかし、[長嶋 95] や [本名 90] で使われている表記法は手話を正確に再現する事を目的としており、記号処理や人間が直接読み書きすることには向いていなかった。そこで手話の記述法として「手話表記法 (文)」を提案した [徳田 95a] [徳田 95b]。これは、日本語単語をラベルとして使用し、手話を日本語文字列として記述する手法である。具体例を図2に示す。日本語ラベルには人間が直接読み書きに使用するため、直観的にわかりやすいものを選び、使用する特殊記号も極力少なくした。

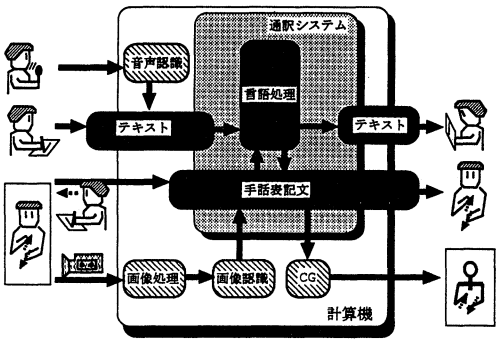
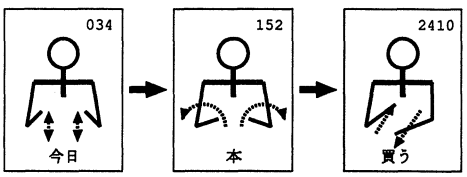


図 1: 日本語から手話への機械翻訳システムの概観



手話表記文：今日/本/買う
(日本語：今日、本を買った)

図 2: 手話表記文の例

2.2 手話単語辞書

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、変

換方式の機械翻訳が適切と考えた。この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。一方の手話単語辞書は既存のものは語彙数が少なく、収録されている単語にも方言や死語などが多く含まれているため、直接利用するには不十分である。

そこで本研究では手話の単語集として最もよく使われている「わたしたちの手話」(収録語数約 6000 語[1])[全ろう 87] を基本に電子手話辞書 (収録語数約 700 語)[神田 94] やテキスト (収録語数 342 語)[石 94] の収録単語から日本語見出しを抽出した。その中から、ろうあ者や手話通訳の経験者の意見を参考にしながら、日本語見出しを見て躊躇する事なく手話が表現出来る単語を 3162 語選定し、読み、分類などの情報を付加し、機械可読な単語辞書としてまとめた。

3 日本語から手話への変換手法

本稿では日本語から手話への変換のみを考える。入力は日本語文で手話表記文を出力することを目標にする。以下、変換手法を手順に沿って述べる。

3.1 形態素解析と不必要形態素の除去

最初に入力文の形態素解析を行う。解析には日本語形態素解析ツール JUMAN[松本 94] を使用した。

次に、形態素解析を行った入力文から、明らかに必要のない形態素 (助詞、句読点など) を除去する。

3.2 変換規則での処理

本稿で使う変換規則は適用する範囲が句単位のものと形態素単位のものに分けられる。前者の句全体に適用する規則は、月日の表現に関するもので、手話では特殊な表現としておおむね認められているもので、変換規則で処理することにした。後者の形態素単位で適用する変換規則は、動詞の語尾や、接尾辞に関するものである。これらの形態素は方向性を表現する記号への変換が適当と思われるため、変換規則で処理することにした。

3.3 手話単語辞書での処理

変換規則で処理しなかった形態素について、手話単語辞書の日本語見出しを検索し、一致するものがあればその見出し語を結果として出力する。また数字は手話単語として認識し、そのまま出力する。

[1]収録語数は日本語の見出し語の数

3.4 EDR 電子化辞書を使った類似単語の獲得

残りの形態素は EDR 電子化辞書 [EDR95] を使って類似した単語を獲得し、手話単語への変換を試みる。類似単語は次に述べる 3 つの方法の順で獲得し、変換に使用することにする。

方法 A：同じ概念識別子をもつ見出し語　EDR 日本語単語辞書には見出し語ごとに概念識別子が付けられている。概念識別子とは、EDR 電子化辞書内で一意に割り振られた数であり、同じ概念を含む見出し語には同じ概念識別子が割り振られている。そこで、元の見出し語の概念識別子と同じ概念識別子は意味が似ている事を利用し手話単語へ変換する。

「学食」を「食堂」に変換する具体例を図 3に示す[2]。学食の概念識別子を調べると概念識別子 3bcb25 を持つ見出し語は「学食」の他に「食堂」「料理屋」「ビュッフェ」の 3 つがある。手話単語辞書には「食堂」のみが含まれているのでシステムは「学食」を「食堂」に変換する。

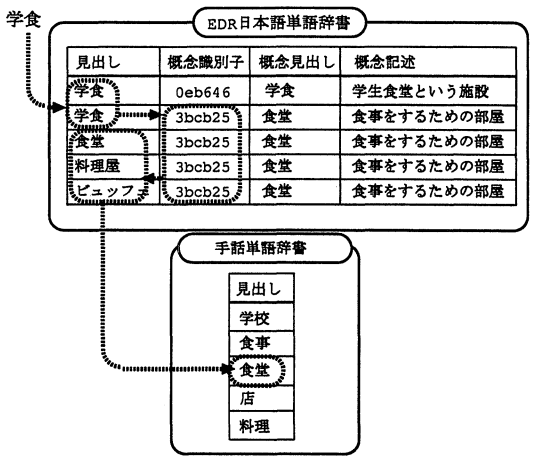


図 3: 概念識別子による類似語の獲得

方法 B：概念記述から抽出した語　概念識別子には概念記述が付けられている。これは人間が概念識別子を理解しやすくするために付加されているもので、概念を日本語文で解説したようなものになっている。これ

[2]図の EDR 辞書のイメージは説明用に作ったもので、一部実際の辞書の構成と異なる部分がある。

を見出し語に関する説明文であるとみなして、これより単語を抽出し手話単語に変換する。

概念記述は類似した文型が多く見られるので、効率良く単語を抽出するために抽出規則を作り、手話単語へ変換できるような単語を取り出した。

方法C：上位概念の見出し語　EDR 概念辞書では概念を階層化し、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、見出し語の概念識別子より上位の概念識別子を獲得し、その見出し語および概念記述を使って、類似単語を得ることにする。

3.5　指文字表記への変換

手話では外来語などの明らかに手話単語が存在しない語を指文字を使い表現することが多い。そこで類似語も獲得できずに残った形態素について、品詞が名詞ならば指文字表記に変換する。品詞が名詞以外なら失敗とする。

変換処理までが終了すると、整形して区切り文字を挿入し、手話表記文を出力する。

4　実験

日本語から手話表記文への翻訳システムとして「手話ん」を試作し、評価した。

入力データとして「NHK ニュース聴力障害者のみなさんへ」の 3 日間の放送分を元にして対訳コーパスを 91 文作成した。今回の変換手法は表層情報しか用いないので、ニュース文を採用する事で日本語文が既存の日本語処理技術で扱いやすい事と、手話表現が日本語の構文に近い事を期待した。日本語文は JUMAN で処理した後、形態素区切りなどの誤りを修正した後、「手話ん」に入力した。

現在の「手話ん」は変換時に語義を考慮していない。また EDR 電子化辞書によって複数の類似語が獲得できた場合の絞り込みが自動化されていない。そこで、それらの部分を次の基準で手作業で変換に成功したかを判定した。

1. 入力された単語の語義と出力された手話単語の語義がほぼ同じとみなせること。
2. 出力された手話表記文が手話として理解できるものであること。

実験結果を表1に示す。表中の項目の内容を次に示す。

手話単語　手話単語に変換できた形態素数
数字　数字 (手話単語) に変換できた形態素数
変換規則 A　形態素ごとに適用する変換規則で変換した形態素数
変換規則 B　句全体 (月日に関する) の変換規則で変換した形態素数
候補語を獲得　EDR 電子化辞書より類似語を獲得できた形態素数
指文字　指文字表記に変換された形態素数
不要形態素　形態素解析終了語、変換が必要ないと判断された形態素数
失敗　システムが変換できなかった形態素数

5　考察

5.1　類似語の獲得について

EDR 電子化辞書による類似語は、ほとんどが概念記述で獲得することにり、変換に失敗したものは概念記述に原因があるものが大部分を占めた。その主要な原因は次の 2 つである。

EDR 電子化辞書の概念記述自体がよくなかったもの　「再現」に対する「再現する」のように概念記述自体が見出し語で構成されているものがあった。これらは EDR 日本語単語辞書の概念記述自体が原因であるから、辞書を変更するか、改良することで解決する事が考えられる。

概念説明が適切でないもの　概念記述が日本語文が冗長で手話表現への変換に不適切なものがあった。具体的な例では、「外国人」の手話表現は「世界[3]/人 (外国の人)」もしくは「日本人/ 他 (日本人以外の人」) と表現したい。しかし、概念記述は「ある国の国籍を持っていない人」なので、このような表現は容易には出力できないと思われる。これらの記述が適切でないという判断は主観的判断による。

これらは概念記述を再び「手話ん」に数回処理させれば理想的な表現が生成できるかもしれないが、現在の「手話ん」の精度では結果に期待できない。今後「手話ん」の変換制度を向上させるか、他の日本語辞書を使い別の語釈文を使って解決する事が考えられる。

5.2　指文字表記への変換について

指文字表記への変換に失敗と判断した形態素は形式名詞であり、これらは EDR 電子化辞書で概念記述などが省略されており類似語の獲得に失敗していた。これらは変換規則による処理が妥当と考えている。

[3] 手話単語の「世界」は「外国」という意味を持っている。

形態素数 1303 個							
手話単語	数字	変換規則 A	変換規則 B	候補語を獲得	指文字	不要	失敗
267 個	30 個	149 個	16 個	325 個	38 個	423 個	55 個

表 1: 結果

5.3 システムが「変換失敗」とする形態素について

現在のシステムで失敗となる形態素は次の 3 つに分類される。

ひらがな表記による失敗 漢字で記述してあれば変換できたが、入力がひらがなであったため変換に失敗した形態素があった。漢字で入力するように心がけるか、読みの情報使うことで解決できるだろう。

語尾への対応が不完全 語尾が受身形のときに起こる失敗で、「売られる」「言われる」が該当する。形態素解析で語尾として分離していれば、現在の「手話ん」は語尾の変換規則で対応できるので、形態素解析の精度が上がれば解決するものと考える。

辞書に見出し語がない 手話単語辞書にも EDR 日本語単語辞書にも見出し語として存在しないもので、「ごきげんよう」「では」が該当する。辞書を補強することで解決できると考えている。

6 おわりに

現在の「手話ん」は類似語の獲得処理を自動化していない。間違った語義で類似語を獲得し、変換してしまう誤りが多いためである。正確な語義が決定できれば、類似語が複数獲得できた場合でも絞りこみができ、信頼性もあがるため再度類似語を獲得する手法も適用できるだろう。しかし、ある単語の語義は手話と日本語で同一ではない。そこで手話への変換を見越して、日本語の語義のグループ化する必要がある。この作業は手作業で行うには限界があるので自動化する必要がある。また、入力された形態素の語義がどのグループの語義に当たるかを決定する手法を検討する必要がある。さらに現在の手話単語辞書には複数の語義がある単語は印があるので、手話での語義の曖昧性がある場合の決定方法を検討しなければならない。

「手話ん」では手話単語や変換規則に適合しなかった単語は EDR 電子化辞書を使って類似語の獲得を試みる。この際に複数の類似語が獲得できる。最終的な出力では、これらを一つに絞り込む必要がある。EDR 電子化辞書にはコーパスに単語が出現した回数の情報が付与されているので、この情報を使うか、何らかのヒューリスティックスを使用することを検討している。

謝辞

手話表現などで貴重な御意見を頂いた、野々市町手話サークル「てのひら」、野々市町聴覚障害者福祉協会、石川県聴覚言語障害者福祉協会青年部の皆さんに感謝します。

参考文献

[徳田 95a] 徳田、奥村. 日本語から日本手話への機械翻訳に関する検討. 日本手話学会 21 回全国大会予稿集, pp.36–39. 1995.

[徳田 95b] 徳田、奥村. 手話表記法の提案と日本語から手話への変換方法の検討. 情報処理学会 第 51 回 全国大会, 3-123. 1995.

[長嶋 95] 藤井、岩波、亀井、長嶋. 2 次元の時空間画像による手話の大局的な調動認識. 第 11 回ヒューマン・インターフェース・シンポジウム, pp.197–202. 1995.

[本名 90] 本名、加藤. 手話の表記法について. 日本手話研究所所報, No.4. 1990.

[安達 92] 安達. 手話通訳のためのニュース文の話しコトバへの変換処理. 信学技報, NLC92-47, PP.17–24. 1992.

[EDR95] EDR 電子化辞書 第 2 版. 1995.

[神田 94] 神田. 日本手話電子辞書. アルファメディア. 1994.

[石 94] 石川ろう協. 手話テキスト (入門編). 1995.

[全ろう 87] 全日本ろうあ連盟. わたしたちの手話 1-10. 1987-90.

[松本 94] 松本、黒橋、宇津呂、妙木、長尾. 日本語形態素解析システム JUMAN 使用説明書 version 2.0. 1994.