手話単語辞書の補完と日本語から手話への機械翻訳について

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

[概要] 変換方式の機械翻訳システムでは変換元と変換先の言語の機械可読辞書が必要である。日本語から手話への機械翻訳の場合、変換元の日本語の機械可読辞書は現実的な使用に耐えうるものが存在するが、変換先の手話は 600 語 ~ 数千語程度の規模の辞書しかない。本研究では日本語辞書を使って手話単語辞書を補完し機械翻訳に使用する手法を提案する。この手法で、単純に変換できなかった単語のうち、約 70% を補い、入力文に対して 95% の形態素が変換できる見込みがついた。

1 はじめに

昨年は手話ブームであったと言われている。聴覚障害者を主人公にしたドラマも放映され、手話が社会的に広まったような感がある。同時に聴覚障害への理解もある程度進み、職場や日常生活で手話が使えるようになりたいと考える健聴者が増えていることは確かである。それは手話サークルへ参加する人が増加していることからも明らかであろう。しかし、これらの人が必ずしも手話を習得できるわけではない。手話が独自の体系を持つ言語であることを考えれば、その習得に努力を要するのは明らかである。

手話は、日本語の構文と類似している日本語対応手話と、かなり異なっている日本手話に大別できる[1][現代 96]。日本語を習得した者にとって、後者の習得はかなり困難である。つまり手話の学習が英語のような外国語の学習と同じであるためである。本稿では、このような問題を工学的に解決するひとつの手法として機械翻訳 [安達 92] [西川 95] を考える。

機械翻訳では入力言語を解析し、中間言語へ変換し、対象言語を生成する。しかし、いまだに、あらゆる言語に適用して良い結果を得られるような汎用的な中間言語は開発されていない。そのため一般には入力文を入力言語の構造を反映した中間言語に変換し、それを対象言語の中間言語に変換する手法がとられている。この方法は様々な言語に柔軟に、現実的に対応できるが、非常に手間がかかる。手話は日本語とは違う言語であるが、完全に独立した存在であるとも考えにくい。本研究では最終的には、日本語と日本手話の相互機械翻訳の実現を目標としているが、その第一段階として、手話の表記に日本語を借用し [徳田 95]、日本語の構造にかなり近い手話へ変換を目指す。このため中間言語への変換に関わる労力を大幅に削減している。

それでも、日本語から手話への機械翻訳には様々な問題がある。特に現在の手話の状況では辞書の貧弱さが問題になる。現存する手話単語辞書では、最大でも約 6000 語 [ろう連 87] の見出し語しか含まれていない。また一般に入手できる機械可読辞書[2](MRD: Machine Readable Dictionary) になると約 600 語程度 [神田 94a] のものしか存在しない。これは日本語の辞書に比べると非常に少なく、機械翻訳を行う際の情報源としては、かなり心許ない。

そこで本稿では変換する単語が手話単語辞書にない場合には、日本語辞書を使って類似語を獲得し、手話単語へ変換する手法を検討した。これにより手話単語辞書の語彙の少なさを補完する。2 章で使用した辞書について述べ、それらを使って試作した日手機械翻訳システムについて 3 章で述べる。4 章でシステムの実験、評価を行う。

2 辞書について

一般に翻訳では変換元と変換先のそれぞれの言語の辞書が重要な役割を果たす。機械翻訳においても入力言語と対象言語の辞書が、変換の質や精度に関わってくる。本章では、本研究で使用した日本語と手話の辞書について述べる。

2.1 EDR 電子化辞書

本研究では日本語辞書として EDR 電子化辞書 [EDR95] を使用している。EDR 電子化辞書は単語辞書、コーパス、共起辞書、概念辞書など 8 つの辞書が含まれており、本研究では日本語単語辞書と概念辞書を使用している。

[1] 本稿では日本語との構文面からの相違という点から考えて、日本語対応手話と日本手話を両極端なものと考える。それらをすべてを包含するか、もしくは判別する必要がない場合に「手話」と呼ぶ。

[2] 計算機でテキストとして扱える形式で提供される辞書のこと。

日本語単語辞書は、日本語単語を見出し語とした辞書で、約 40 万レコードが格納されている。ひとつのレコードには、日本語見出し、かな表記、品詞、概念識別子[3]、日本語概念見出し[4]、日本語概念説明[5]、頻度[6] などの情報が含まれている。日本語単語辞書では語義を概念識別子として扱っており、語義の違いは概念識別子の違いとして表現されている。

概念辞書は概念識別子を階層的に体系化した辞書で、約 40 万レコードが格納されている。抽象的な概念を上層に、具体的な概念を下層に配置して、木構造に近い状態で整理されている。

2.2 手話単語辞書

現存する手話単語辞書には、地域によっては通用しなかったり、編集者が普及を目的に創作した語彙が含まれている。そこで市販の辞書 [ろう連 87] [ろう協 94] [神田 94a] より日本語見出しを抽出して選別を行い、機械可読な手話単語辞書としてまとめた。

選別の基準は、複数の辞書に掲載されていたり、初心者向けのテキストに載っている単語と、手話が容易に表現できる単語を採用した。この判定には石川県の手話サークルや聾唖者の協力を得ている。

3 日手機械翻訳システムの試作

辞書の有用性を調査する目的を兼ねて、日本語から手話への機械翻訳システムを試作した。システムの具体的な動作を次に示す。

step 1. 単語へ区切る 入力された日本語文を日本語形態素解析システム JUMAN[松本 94] で処理し、単語ごとに区切る。

step 2. 付属語を取り除く 次に助詞や助動詞、接尾辞などの付属語を取り除く。翻訳においても付属語の持つ情報は大切だと考えるが、今回のシステムでは単語の変換を中心に実験を行うこともあり、付属語を取り除き、処理の簡略化をはかっている。

step 3. 変換規則を適用する 月日に関する表現[7] と語尾の処理に変換規則を使っている。具体的には、入力単語が月日の表現や語尾と認識されたら、変換規則を適用して、変換が成功したとみなす。

step 4. 手話単語へ変換する 手話単語辞書を検索し、入力単語が辞書内に存在した場合は変換したとみなす。

step 5. 類似語を獲得する 残りの単語について、類似語による変換を試みる。具体的な方法は次の節で述べる。

step 6. 指文字表記に変換する 残りの単語について、品詞が名詞の場合には指文字表記に変換する。

名詞以外で類似語の獲得にも失敗して変換できなかった単語は、変換に失敗したとみなした。

3.1 日本語辞書による手話単語辞書の補完

システムに組み込まれた手話単語辞書の見出し語は約 3000 語程度である。

前述したように、現在のシステムは入力された単語が手話単語辞書内に存在しない場合に、日本語辞書を使って以下の 3 種類の方法によって類似語を獲得し、それを手話単語として使用する。

3.1.1 概念識別子による類似語の獲得

EDR 日本語単語辞書の見出し語には概念識別子が付加されている。概念識別子は語義を表しているため、同じ概念識別子を持つ見出し語は意味が似ていると考えられる。そこで、元の日本語単語の概念識別子と同じ概念識別子を持つ見出し語を獲得し、手話単語へ変換する方法を実装した。

図 1に 「学食」という日本語単語を 「食堂」という手話単語に変換する例を示す。 「学食」は手話単語辞書には存在しないが、同じ概念識別子 (3bcb25) を持つ 「食堂」は手話単語辞書にあるので、システムは 「学食」を 「食堂」に変換する。

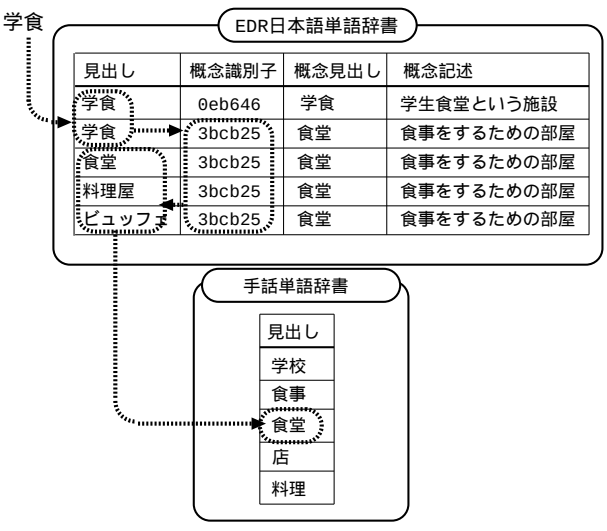

図 1: 例: 「学食」を 「食堂」に変換する

[3] EDR 電子化辞書内で定義されている 「概念」に割り振られた数文字列。

[4] 「概念」を把握しやすくするためにつけられた日本語のラベル

[5] 「概念」を簡潔な日本語文で説明したもの

[6] EDR コーパスでの見出し語の出現回数

[7] 「9 月 18 日」というように左手で月、右手で日を表現することを想定している

3.1.2 概念記述による類似語の獲得

概念識別子には、日本語概念記述という簡潔な日本語文で意味が説明されている。ここから自立語を抽出して手話単語へ変換する方法を実装した。具体的には日本語概念記述を JUMAN によって形態素解析し、主要な自立語を抽出規則により抜き出した。抽出規則は EDR 日本語辞書よりランダムに選択した 200 文の日本語概念記述に対して作成しており、名称や物の種類に関する名詞を抽出する規則である[8]。

図 2に 我慢する」を手話単語の 感情/抑える/耐える」に変換する例を示す[9]。この手法は、本システムの日本語から手話へ変換する手順をほとんどそのままに繰り返したものである。但し、日本語概念記述に自立語の抽出規則を適用する部分などが異なっている[10]。

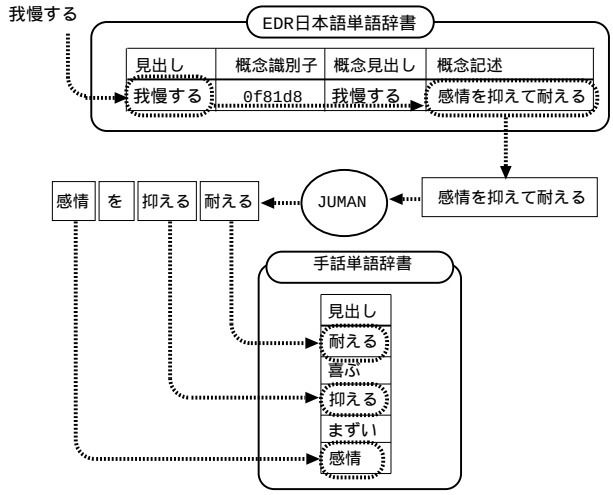

図 2: 例: 我慢する」を変換する

3.1.3 概念階層による類似語の獲得

EDR 電子化辞書では概念 (識別子) は抽象的な概念を上位に、具体的な概念を下位に持つような階層構造を構成していると仮定されている[11]。そこで元の日本語単語より抽象化された概念を持つ単語を使って手話単語への変換する方法を実装した。

鯖」を 魚」に変換する例を図 3に示す[12]。 鯖」は上位概念として 魚」を持つ。つまり 鯖」の抽象化した単語として 魚」が考えられ、システムは 鯖」を 魚」に変換する。

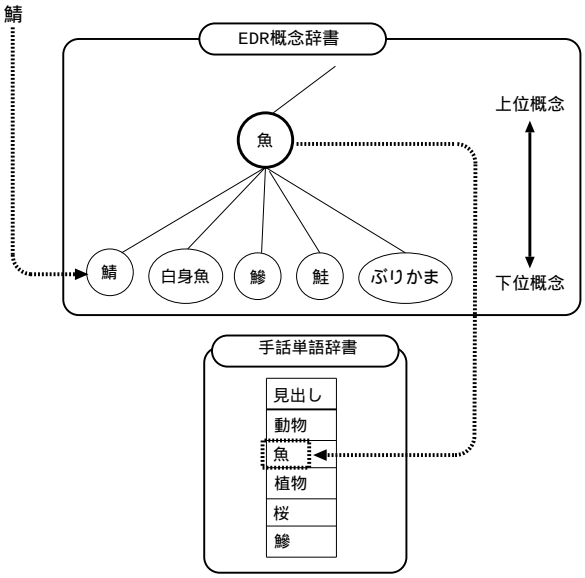

図 3: 例: 鯖」を 魚」に変換する

4 実験と評価

システムの実験の際に使用した入力文は、NHK の手話ニュースの 3 日分の日本語文章 [NHK94] であり、全部で 91 文を用意した。これらは JUMAN の処理により 1303 個の形態素に分割された。JUMAN 自身の形態素処理の失敗は手作業で修正し、本システムの変換部分に影響しないようにした。

現在のシステムは類似語の獲得の際に複数の単語が獲得できた場合の絞り込みが自動化されていないので、その部分は手作業で補っている。実験の結果を表 1に示す。これらの結果も手作業で変換の正しさを判定している。表中の 失敗」は、システムが変換できなかった形態素と、間違って変換した形態素の総数である。

変換が不要と判断された形態素を除いた総形態素数 880 個に対しては、約 94% が手話単語に変換されている。類似語を使って手話単語に変換を試みた形態素は、417 個[13]であり、そのうち 78%にあたる 325 個の形態素が本手法で手話単語に変換する見込みがついた。

類似語を使っても変換できなかった理由は、日本語辞書として使用した EDR 電子化辞書の日本語概念記述が不適切だった場合が多かった。例えば、 再現」に対する 再現する」のように見出し語の単なる繰り返しや、非常に長い文であるために形態素解析に失敗するような日本語概念記述からは、有用な類似語の獲得はできなかった。

[8] 例えば、 アメリカという国」や とんぼという昆虫」に対して、 という」の前後にある名詞を抽出する規則を作っている。

[9] この文には抽出規則は適用されていない。

[10] もう一つ、日本語概念記述内に見出し語と同じ単語があると無限ループに陥るため、その検査をしている点も異なる。

[11] 例えば、 熊」や 馬」は上位概念として 動物」を持ち、逆の関係は下位概念となる。

[12] EDR 概念辞書の内容は説明のために作ったもので、実際の辞書とは異なる点がある。

[13] 失敗」の 55 個の形態素のうち、手話単語の直接変換で失敗しているものが 1 つだけあった。

形態素数 1303 個					
手話単語に直接変換	変換規則で変換	類似語を使って変換	指文字表記	不必要形態素	変換失敗
297 個 (23%)	165 個 (13%)	325 個 (25%)	38 個 (3%)	423 個 (32%)	55 個 (4%)

表 1: 実験結果

指文字表記に変換された形態素は、ほとんどが外来語などカタカナで表記された単語が多く、品詞が名詞に限って指文字表記に変換する本システムの手法は実用的であると考えている。

5 今後の課題

現在のシステムでは語義の決定を行っていないため、変換精度が落ちていると考えられる。基本的に日本語と手話の語義は必ずしも 1 対 1 で対応するものではない。例えば日本語では 「楽しい」 「嬉しい」は違う単語であるが、手話では一つの単語とみなしても構わない。逆に日本語の 「開く」は、手話では対象により[14]により様々な単語に変化する。このように手話への翻訳を前提にした日本語の語義の決定と、それに対する手話側の語義の決定が必要である。そのためには手話単語辞書に日本語の語義との対応情報の付加と、入力文の意味解析が必要と考えている。

また、現在のシステムでは助詞などの付属語は取り除いているため、生成される手話が日本語対応手話になっている。さらに 「京都の手話を覚える」と 「京都で手話を覚える」は全く同じ結果を生成してしまう不都合も起きる。これを解決するためには構文上の変換規則を導入する必要があるだろう。今回の実験でもいくつかの変換規則を使っているが、これらは NHK の対訳コーパスから簡単なヒューリスティックスにより抽出したもので、語尾など単純な処理しか行っていない。構文規則を抽出するためには大量の対訳コーパスが必要であり、それらの言語資料から自動的に構文変換規則を抽出する方法を検討しなければならない。

本システムの出力には非手指動作が考慮されていない。手や腕の組合せで表現できることには限界があり、手話は日本語と同じ表現能力を持っているのだろうから、非手指動作を含めた単語の認識が必要と考えている。これらは辞書の構成から考え直さなくてはならない。

今後はさらに変換精度を高めるとともに、手話から日本語への変換方法も検討していきたい。

[14] 扉、箱、窓、ガレージを 「開ける」手話はそれぞれ異なる。

謝辞

手話単語辞書の作成や実験の評価に野々市町手話サークル 『てのひら』の皆さん、橋本諭氏、竹谷吉弘氏に多大な御協力を頂きました。深く感謝します。

参考文献

[徳田 95] 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 96] 徳田. 日本語から手話への機械翻訳に関する研究. 北陸先端科学技術大学院大学 修士論文. 1996.

[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.

[西川 95] 西川、寺内. 手話の空間配置を考慮した日本語 ・手話翻訳に関する検討. 信学技報,HC94-94, pp.55–62. 1995.

[EDR95] EDR 電子化辞書 version 1.0. 1995.

[神田 94a] 神田. 日本手話電子辞書. アルファメディア. 1994.

[神田 94b] 神田. 手話学講義. 福村出版. 1994.

[現代 96] 現代思想 4 月号増刊 ろう文化. Vol.24-05. 青土社. 1996.

[ろう協 94] 石川ろう協. 手話テキスト (入門編). 1995.

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

[NHK94] NHK ニュース 聴力障害者のみなさんへ. 1994 年 9 月.

[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.