

気象情報を対象とした日本語一手話 CG 翻訳の 主観評価実験

加藤 直人[†] 宮崎 太郎[†] 金子 浩之[†] 井上 誠喜[†]
梅田 修一[†] 清水 俊宏[†] 比留間 伸行[†] 長嶋 祐二[‡]

[†] N H K 放送技術研究所

[‡] 工学院大学情報学部

E-mail: [†] {katou.n-ga, kaneko.h-dk, inoue.s-li, umeda.s-hg, shimizu.t-hy, hiruma.n-dy}@nhk.or.jp

[‡] nagasima@cc.kogakuin.ac.jp

1 はじめに

手話サービスの拡充をめざし、テキストから手話 CG に自動翻訳する研究を進めている [1]。手話は、先天的あるいは幼少時に失聴した聾者にとって母語であり、日本語より理解しやすい。しかしながら、字幕放送など日本語による情報提供は増えているものの、手話では少ない。これは、日本語から手話への“翻訳”という作業を伴うからである。日本語と手話は異なる言語であるので、英語への翻訳のように、手話への翻訳が必要となる。しかしながら、手話へ翻訳できる人材は、英語の場合に比べはるかに少ない*1。

我々はニュースの手話翻訳を目指している。当面の目標としては天気や気象災害などのような気象情報を手話翻訳の対象としている。これは、地震や台風のニュースなど気象情報は特に重要であるからである。さらに、気象災害が深夜や早朝に起きた場合に、手話通訳者を確保することは困難であると考えられるからである。

本報告では、気象情報を対象として、日本語一手話 CG 翻訳の主観評価実験について述べる。翻訳精度の改善にはその評価が欠かせないが、手話翻訳の研究は行われてきてはいるものの [2] [3] [4]、その評価を行った研究はない。これは、翻訳評価を行うためにはある程度の数の翻訳結果が必要となるが、従来の翻訳システムは語彙規模や扱える言語現象が少ないために、本格的な翻訳評価ができるまでの翻訳結果を得ることが難しかったからである。一方、我々の翻訳システムは気象情報という分野に限られているものの、語彙数や扱える表現が多いので、従来に比べて幅広い文を翻訳することができる。本報告での実験では評価文は 100 文と従来にはない規模の評価実験となっている。評価は、13 人の評価者に、入力 of 日本語テキストと、出力 of 手話 CG を比較してもらい、5 段階の主観評価とともに、問題点・改善点を記述してもらった。

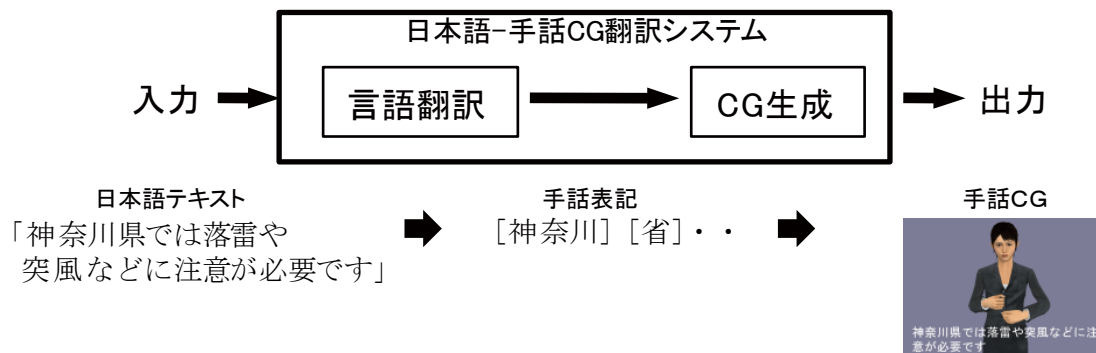


図1 日本語・手話 CG 翻訳システム

*1 例えば、手話通訳士は全国で 3,000 人程度である。

2 日本語-手話 CG 翻訳

我々の日本語-手話 CG 翻訳システムについて簡単に説明する。詳細は文献 [5] を参照されたい。

翻訳システムの概要を図 1 に示す。システムは大きく分けて、言語翻訳と CG 生成から構成される。言語翻訳では、システムに日本語テキストが入力されると、手話表記に翻訳する。ここで、手話表記とは手話動作を単語ごとに書き起こしたものである。日本語テキストを直接手話 CG に変換することは困難であるので、手話表記を用いている。ただし、手話にはその表記法が確立されておらず、我々の手話表記も手話動作を完全に記述したものではない。詳しくは文献 [6] を参照されたい。一方、CG 生成では、言語翻訳の結果から、各単語を日本語手話対訳辞書 [7] を使って変換し、各手話単語のモーションを線形補間で繋いでいくことにより文を生成する。我々の日本語-手話対訳辞書には日本語語彙約 86,600 語、手話語彙約 4,900 語が登録されている。手話語彙は、光学式モーションキャプチャによって取得した動作データと、骨格構造を持った人体 CG モデルで構成されている。これは映画やビデオゲームで用いられる手法であり、自然な動作を再現できる。また、CG の描画は映像コンテンツ記述言語 TVML [8] を使っているため、容易に動作編集が可能となる。

言語翻訳では用例翻訳と統計翻訳を用いている。用例翻訳は、用例があれば精度よく翻訳可能であるものの、用例がないと翻訳できないという問題がある。一方、統計翻訳は翻訳可能な文は多いが、翻訳精度は保証されない。気象情報は定型的な表現が多く、用例翻訳が有効である。我々の用例は節・句単位であり、NHK 手話ニュースの気象情報に出現した文から人手で節・句に分割した。現在までのところ、2009 年 4 月～2011 年 2 月までの約 5,300 文から、異なりで約 3 万個の用例を作成した。統計翻訳は用例翻訳で翻訳できなかった部分に対して行う。統計翻訳の学習データには、用例翻訳に使っている節・句単位の用例を用いている。

*2 CODA (Children of Deaf Adults) とは、聾者の両親を持つ聴者のことである。

*3 デフファミリーとは、家族全員が聾者のことである。

3 主観評価実験

3.1 評価概要

気象情報を対象に日本語-手話 CG 翻訳システムによる翻訳を行い、その翻訳結果に対して主観評価実験を行った。

評価者は日本語と手話がわかる 13 名である。評価者の情報を表 1 に示す。2 名のみが聴者であるが、CODA*² であるので手話の能力は高い。また、聾者にはデフファミリー*³ の者や兄弟に聾者がいる者が多く、日常的に手話を使用している。また、後述するように、評価には日本語によるコメントを記入すること行ってもらっているため、聾者の日本語能力も問題ないと考えられる。

評価データ 100 文は次のようにして作成した。まず、2009 年 4 月～2011 年 2 月において NHK 手話ニュースの気象情報に出現した約 5,300 文からランダムに 100 文を選んだ。その残りの文から、用例翻訳に用いる文を選択し、節・句単位の用例を作成した。評価文 100 文は日本語-手話 CG 翻訳システムで翻訳し、計算機で再生できる映像ファイルに変換した。評価文の一例を図 2 に示す。手話 CG は画面中心に配置し、日本語（入力）は画面左上に配置した。

評価は、5 段階評価とともに、問題点や改善などのコメントを記述してもらった。評価作業を始める前に、評価に関する注意事項を書いた A4 1 枚の紙を読んでもらい、分からないことがあれば、聾者には手話で説明することを手話通訳者にお願いした。また、映像は何回繰り返し見てもよいなど、映像を再生する際の制約は設けなかった。

5 段階評価の基準を図 3 に示す。聾者に対しては、図 3 の日本語を読んで分からない箇所があれば、手話での補足説明をお願いした。



図 2 評価文の例

評価5

原文(日本語)の意味を正しく理解できる手話である。手話として問題ない。

評価4

一部不適切な訳語がある、あるいは、一部におかしな手話や不要な手話はあるが、原文(日本語)の意味を十分理解できる手話である。

評価3

原文(日本語)の情報は大体含まれており、おかしな手話や不要な手話はあるが、原文(日本語)を概ね理解できる手話である。

評価2

原文(日本語)の情報があまり含まれていない、あるいは、おかしな手話や不要な手話が多いため、原文(日本語)を部分的にしか理解できない手話である。

評価1

原文(日本語)の情報が失われており、原文(日本語)をまったく理解できない手話である。

図3 5段階評価の基準

3.2 評価結果

各評価者による評価値の平均を表2に示す。表2を見ると、評価者全体の平均は3.30であるが、個々の平均は2.06から4.15までと評価者によってかなり差があることがわかる。そこで、評価者間の結果の一致度を調べるために、評価者2人ごとの κ 係数を求めた。表3に示す。表3を見ると、ほとんどの評価者間で κ 係数の値が低く、評価者による評価のゆれが大きいことがわかる。さらに、評価者間の傾向を調べるために相関係数を求めた。表4に示す。表4を見ると、相関係数が0.5前後である場合が多く、評価者間の評価結果には比較的相関はあることがわかる。評価値が一致しない原因には、手話翻訳の精度にはさまざまな要因が絡んでくるにもかかわらず、一つの基準で評価したことが考えられる。例えば、音声翻訳の評価では、言語翻訳と音声合成では別に行い、さらに言語翻訳の評価は正確さ(adequacy)と流暢さ(fluency)と2つに分けて行う。このような評価が可能なのは、音声をテキストに書き起こすことができるからである。一方、手話はまだその表記法が確立されていないために、音声翻訳のように3つに切り分けて評価することが困難である。

5段階評価が一致しなかった例には「台風の影響で各地では被害が相次いでいます」という短い文があった。その5段階評価は5が3人、4が3人、3が5人、2が2人であった。各評価者のコメン

トを見たところ、評価値が3や2を付けた評価者に、『各地』の前に『日本』や『九州』が必要」という指摘があった。これは『各地』がどの地域なのかが明確でないということである。本来は文脈があるので、そのような情報を付加して翻訳していない。しかし、今回の評価文はランダムに選んであるために文脈の情報がないという問題がある。今後は評価データに連続した文を選択することも必要であろう。

4 おわりに

現在開発を進めている、日本語一手話CG翻訳システムの主観評価実験について述べた。その結果、現在のシステムの翻訳精度を得ることができた。しかしながら、評価者によって値のバラツキが多いなど手話翻訳の主観評価にも課題があることが明らかになった。

今後はさらに評価結果の分析を進めて翻訳システムの改良を行い、翻訳精度の向上をはかりたい。特に、評価者のコメントの分析を詳細に行い、改善点を洗い出したい。また、主観評価方法についても検討していきたい。

参考文献

- [1] 加藤直人, 金子浩之, 井上誠喜, 梅田修一, 比留間伸行, 長嶋祐二, “用例利用による日本語一手話CG翻訳システム,” 電子情報通信学会 HCG シンポジウム 2011, 1-1, 2011.
- [2] 日立製作所, 手話アニメーションソフト MimehandII, <http://www.hitachi.co.jp/Prod/comp/app/shuwa/mimehand/index.html>
- [3] 田丸満枝, 猪木誠二, 山本泰則, 宇治谷恵, 杉田繁治, “モバイル環境での手話アニメーションによる情報提供,” 電子情報通信学会 HCG シンポジウム 2002, pp. 453-454, 2002.
- [4] 松本忠博, 池田尚志, “日本語テキストから手話テキストへの機械翻訳の試み,” 自然言語処理, Vol. 15, No. 1, pp. 23-51, 2008.
- [5] 加藤直人, 宮崎太郎, 金子浩之, 井上誠喜, 梅田修一, 比留間伸行, 長嶋祐二, “気象情報の日本語-手話CG翻訳,” 言語処理学会年次大会, PA1-21, pp. 275-278, 2012.
- [6] 加藤直人, “手話ニュースコーパスの構築,” 言語処理学会年次大会, PA2-5, pp. 494-497, 2010.
- [7] 加藤直人, 金子浩之, 井上誠喜, 清水俊宏, 長嶋祐二, “日本語一手話対訳辞書の構築 ~日本語語彙の拡張~, ” 電子情報通信学会 HCG シンポジウム 2009, 1-3, 2009.
- [8] <http://www.nhk.or.jp/str1/tvml/index.html>

表 1 評価者の情報

評価者	年齢	聴者/聾者	生育地	手話ニュース	その他
A	40代	聴者	神奈川	よく見る	手話通訳士、CODA
B	30代	聴者	東京	ときどき見る	手話通訳士、CODA
C	20代	聾者	東京	週1回程度見る	デフファミリー
D	40代	聾者	鹿児島	毎日見る	デフファミリー
E	30代	聾者	東京	よく見る	デフファミリー
F	30代	聾者	東京	ときどき見る	デフファミリー
G	20代	聾者	東京	よく見る	両親聴者
H	40代	聾者	神奈川	あまり見ない	デフファミリー
I	30代	聾者	東京	ときどき見る	両親聴者
J	30代	聾者	鹿児島	ときどき見る	両親聴者、兄弟にろう者
K	20代	聾者	鹿児島	ほとんど見ない	両親聴者
L	30代	聾者	鹿児島	ときどき見る	両親聴者、兄弟にろう者
M	40代	聾者	鹿児島	ときどき見る	両親聴者

表 2 各評価者の 5 段階評価の平均

A	B	C	D	E	F	G	H	I	J	K	L	M	全平均
3.42	2.5	3.46	3.01	3.66	4.15	3.37	3.26	3.39	4.02	2.06	2.65	3.89	3.3

表 3 各評価者間の κ 係数

												L	M	
													-0.03	
												K	0.1	-0.03
											J	-0.013	0.03	0.25
										I	0.05	0.02	0.11	0.12
									H	0.08	0.14	-0.02	0.08	0.12
								G	0.1	0.15	0.23	0.05	0.1	0.16
							F	0.18	0.14	0.12	0.37	-0.04	-0.05	0.33
						E	0.29	0.22	0.22	0.08	0.32	0.01	0.03	0.12
					D	0.11	-0.02	0.17	0.18	0.19	0.02	0.05	0.21	0.01
				C	0.07	0.19	0.21	0.23	0.15	0.26	0.28	0.06	0.17	0.17
			B	0.02	0.2	0.03	-0.05	0.09	0.06	0.02	0	0.24	0.29	-0.04
A	-0.06	0.14	0.26	0.07	0	0.05	0.1	0.29	0.01	-0.06	0.08	0.08	0.04	

表 4 各評価者間の相関係数

											L	M
											0.37	0.46
										K	0.37	0.37
									J	0.42	0.62	0.43
								I	0.69	0.3	0.54	0.52
						H	0.52	0.45	0.39	0.41	0.44	
					G	0.47	0.65	0.51	0.39	0.55	0.48	
				F	0.59	0.46	0.6	0.66	0.45	0.62	0.47	
			E	0.52	0.53	0.48	0.56	0.61	0.27	0.4	0.35	
		D	0.56	0.73	0.61	0.55	0.65	0.67	0.5	0.59	0.56	
	C	0.67	0.65	0.7	0.64	0.52	0.68	0.74	0.41	0.58	0.44	
	B	0.51	0.66	0.45	0.54	0.62	0.51	0.6	0.53	0.39	0.62	0.53
A	0.58	0.56	0.7	0.42	0.6	0.66	0.43	0.58	0.53	0.53	0.57	0.57