チャットにおける対話支援用日英機械翻訳システムに関する研究

上野 哲志　　鈴木 理記也　　小川 均
立命館大学理工学部情報学科

1　はじめに

近年，インターネットや公衆回線網など国際的なネットワークが非常に身近になり，それらを媒体とした異言語間の逐次的コミュニケーションの支援は重要な研究課題となっている．

コンピュータ端末間における文字ベースのインタラクティブな通信システムにチャットと呼ばれるものがある．チャンネルの大半は英語を共通語としており，言語の壁が積極的なコミュニケーションの妨げとなっているのが現状である．

そこで，チャットにおいて，自動翻訳よる対話支援が行なえることは，積極的な国際コミュニケーションの促進に非常に有益である．特に和文英訳における労力の削減の効果は大きいと思われる．

しかしながら，これまでチャットのインターフェースという形での翻訳機構は研究成果として発表されていない．それは，チャットに用いられる文が話し言葉に近く，従来の論文翻訳などに利用されている書き言葉中心の文法では解析困難な文が多いためである．

しかし，チャットは音声対話システムと異なり，文字ベースの入力を得ることができるので，言いよどみ，音素的曖昧性が生じることはなく，異言語間のインタラクティブコミュニケーションの翻訳システムとしては，実現性が高いシステムであると考えられる．

本研究では，語彙主導型の文法解析を用いて構造解析を行ない，省略箇所の抽出や，補完に必要な情報の獲得を行なっている．その際，単語ごとに統語情報や意味情報を詳細に記述することによって，助詞の省略等話し言葉特有の文法形態に柔軟に対応できるようにした．

また，チャットの自然な入力感を損なわない為に，主語の補完や最低限の述語を自動的に補完することに注目した枠組みを取り入れた．省略の補完は照応先の決定を候補に対する重み付けによって行った．

2　チャット対話文の特徴

2.1　名詞・述語の省略

チャットは文字ベースの会議システムであるが，入力文は書き言葉というよりはむしろ話し言葉といってよい．その現象はチャットというシステムがリアルタイム性に優れていることに起因する．チャット上の会話文は入力された時間順にシーケンシャルに表示されるので，文章の入力までに要する時間は短いことが望まれる．

また，会話における文という性質上，相づちや慣用語が多く見受けられる．

その為，英語のチャンネル，日本語のチャンネルにかかわらず，入力される文は構造的に簡素な文が多い．

特に日本語の場合，文を短く簡素化する為に多くの省略が使われている．

下の例文の場合では，相づちをうつ例文2の目的格にあたる「チームを」という格が省略されている．これは文脈に基づき，読み手が自然に省略を補完することを期待しておこなわれるものである．

(例文)

1. 僕はチームがいまいち信じられないんです．
2. 私も信じられないです．

また，日本語は習慣的に主語を明示しないで省略する場合が多く，これはチャットにおいても同様である．本研究でサンプルとして用いたチャットにおける対話データ中の必須格要素の省略の7割は主語の省略であった．

(例文)

3. 今，(〜は) 何してるの？
4. (〜は) 学校で修論書いてる．

チャットのように口頭での会話に近い形式で文を生成していく場合，思いついた情報をシーケンシャルに出力していくので，前の文の情報を後から追加していく倒置的表現が多く用いられる．

(例文)

5. 何ができるの，パソコンで？

6. バージョンアップされたよ．NT 対応に．

例文5は述語「できる」と係り受けの関係にある格要素「パソコンで」が述語より後に来ている倒置的表現である．一般的な書き言葉では「パソコンで何ができるのですか？」となる．

例文6は句点で2つの文に区切られているが，後側は文というより，前文の格の構成要素が後から付け加えられる形となっている．

2.2 チャット独自の文法

チャットにおける入力文には幾つかの特殊なルールが見受けられる．リダイレクション記号や，URL の出現頻度が高く，以下のような特殊な用法で用いられている．

(例文)

9. A: 明日は何をしてるの？ ＞ B さん．

10. B: 寝てる．

11. A: http://www.xxx.ac.jp/peephole.html で私の今の状態がわかりますよ．

例文9は,「明日は何をしてるの？」という文をAさんに向けて発話していることを表わしている．例文9は主語が省略されており，受け答えを見ると，「Aさんは明日は何をしているの？」という文を表わしていると解釈されている．つまりこの場合，省略の補完要素が「Aさん」であることをリダイレクションは表しているという解釈ができる．

例文11では，URL が場所を表わす名詞として用いられている．

これらリダイレクション記号や URL は日本語，英語のチャンネルどちらにおいても，数多く用いられている．チャット上の文を解析する為にはこれらの記号の独特の使用法を文法として定める必要がある．

3 システムの構成

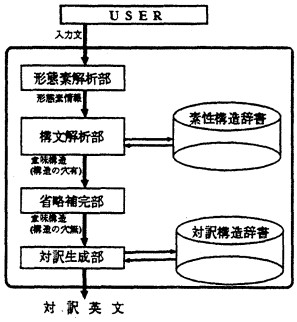

図 1: システム構成

形態素解析部は入力文から形態素ごとの情報を出力する．形態素解析には JUMAN を用いている．

構文解析部は形態素解析部から渡された情報を基に，各単語の素性構造を辞書から引き出し，各素性構造の単一化により文の意味構造を決定し，その結果を出力する．意味構造とは中間言語的な役割を果す．

省略補完部は得られた意味構造上の省略箇所の補完を行う．

対訳生成部は完全な形で与えられる意味構造を英語における構造に変換し，さらに文として表層化する．

3.1 意味構造の解析

本研究で作成するシステムで使用する日本語解析文法は日本語句構造文法 JPSG[Gunji, 87] に基づいて作成されている．

JPSG における日本語の句構造規則は次に示すただ一つである．

$$M \rightarrow C\ H$$

しかし，より広範囲の言語現象を効率良く解析するために，句構造規則を詳細化した．現在のところ，句構造規則として 80 の規則を設定している．

JPSG をベースとした本研究の日本語解析文法では，述語の下位範疇化素性と単一化する必須格補語が文中に存在しない場合は，統語的な穴として扱い単一化を進め，その結果最終的に伝播された穴を省略箇所として抽出する．

```
[sem, [[entity, 'らしい-auxv'], [tense, present],
       [core-sem, [[entity れる-passive],[tense, past],
                   [core-sem, [[entity, '向ける-1'],
                               [requisite, [[agen, [entity, '英タイムズ紙']],
                                            [obje, [[entity, '目'],
                                                    [[entity, 'の-adn'],
                                                     [fn-iden, '疑惑']]]],
                                            [rentai, [goal, 'GOAL']]]],
                               [option, [conn, '案の定']]]]]]]
```

図 2: 出力意味構造の例

3.2 省略の補完

チャット対話文における省略のうち，対訳生成に必要最低限な要素は自動的に補完する．本研究では，倒置表現も述語の省略として捉えている．

以下に本研究における省略の種類ごとの扱いを示し，補完すべき要素については，その具体的補完方法を示す．

チャットの対話における省略はゼロ代名詞が殆どである．本研究において目標としている翻訳とは，翻訳単語と構造の等価性を重視する翻訳である．したがって，目的言語において文法的に省略がゆるされており，読み手自身が自己の知識によって補完できる要素については省略のまま扱う．

しかし英文の場合，ゼロ代名詞のうち主語となるものは統語的に欠落してはならない．

ゼロ代名詞の照応先候補としては，

- 前出の名詞
- 一・二人称代名詞

をあげる．これら候補に以下に示すパラメータによって重みを与える．

1. 後置詞によって与えられる主題の重み．
2. 係り受けの関係にある述語の選択制限．
3. ゼロ代名詞 (省略箇所) との距離．
4. 固有名詞であるかどうか．
5. ゼロ代名詞であるかどうか．
6. 当該文の形態と先行詞候補との相性．

モデル文の省略関係を解析した経験に基づき，具体的なパラメータ値を以下のように定めた．対話においては，主題の提示が最も重要なので，特に主題の強さに着目して値を設定している．

表 1: 先行詞決定パラメータの値

		例	重み
1	表層表現		
	名詞 は/には	太郎はした.	30
	名詞 こそ/も/なら	太郎もした.	22
	名詞 が	太郎がした.	19
	名詞 を/に/から/と/で/へ/まで/までに/より	太郎にした.	15
	名詞 だ/です	太郎だ.	12
2	互いの意味素性が完全に不一致		-30
	互いの意味素性が親子関係		-10
3	係り受けにある述語が1つ離れるごとに		-2
4	固有名詞であれば		+5
5	ゼロ代名詞であれば		+5
6	(a). 一人称代名詞に		-30
	(b). 二人称代名詞に		+36
	(c). 一人称代名詞に		-20
	(d). 一人称代名詞に		+36

総合的な重みが最大のものを先行詞として決定する．

以下の例文 6 で省略されている主語の先行詞候補及びその重み付けを表 2に示す．

(例文)

1. A: ご存知の通り，ハンガリー辺りからヒルには不合理な戦略が続いたでしょ。
2. B: おお、その通り！そうなんですよ、何か不自然ですよね。
3. A: こりゃ、ひいきだぜって、すぐ思いこんだのです。
4. A: こっちのひいきめもあるんだけどね。
5. B: でも、私はそんなにヒルのファンではないのですが、ちょっとね。
6. A: 案の定、英タイムズ紙にも疑惑の目を向けられたらしい。

表 2: 先行詞候補スタック

総合	先行詞候補名	意味素性	ゼ	固	後置	得点	距離	文章
-	会話	(ACT)-			が	19	-20	
-	会話*	(ACT)-	5		が	19	-18	
-	放送中	(TIM)-			は	30	-16	
-	あと2人	(QUA)-			と	15	-16	
6	鈴鹿	(LOC)		5	で	15	-14	
-	6位	(QUA)-			に	15	-14	
15	ヒル*	(HUM)	5	5	が	19	-14	
15	ヒル*	(HUM)	5	5	が	19	-12	
25	ウィリアムズ	(ORG)		5	は	30	-10	
2	チーム	(ORG)			だ	12	-10	
11	チーム	(ORG)			が	19	-8	
-	わたし	(HUM)			も	22	-8	-20
14	ハンガリー辺り	(LOC)		5	から	15	-6	
29	ヒル	(HUM)		5	には	30	-6	
-	不合理な戦略	(DIV)-			が	19	-6	
0	わたし*	(HUM)	5		が	19	-4	-20
-	ひいきめ	(MEN)-			も	22	-2	
-	わたし*	(HUM)	5		に	15	-2	-20
10	わたし	(HUM)			は	30	0	-20
17	ヒルのファン	(HUM)		5	だ	12	0	

表2の 14 行目の「ヒル」は，提題助詞「には」によって主題として大きな重みを与えられ，固有名詞としての重みも与えられている．総合的に最高の重みを与えられ，先行詞として選択される．

次に述語の省略について説明する．

チャットの逐次性を考慮し，入力が複数文の場合，1 文ごとに訳出をおこなう．句点で区切られた倒置表現や付け足し文は省略を含む文の一種として扱い，その省略要素を補完する．

ゼロ代名詞同様，省略された述語と照応関係にある述語のうち，直接的照応関係にあるものを対話文中から特定することができる．特に倒置・付け足し表現は直接的な照応先が存在すると考えられる．

以下に本研究で取り扱う述語の省略の補完について示す．

1. 倒置・付け足し表現に伴なう述語の省略．

2. 繰り返しに伴なう述語の省略．

3. 「名詞＋格助詞＋名詞」による述語の省略.

述語が省略された文が後置詞句の場合 (1,2), 前文の述語を照応先の候補とし, 単一化によって照応関係を確かめる. 3 のような場合は特定のパターンとしてテンプレートを用意し, それにあてはめることによって述語の補完を行う.

3.3 対訳生成

省略の補完が行われた日本語意味構造を英語意味構造に変換し, 語順の決定を行い, 表層上の文として出力する.

構造変換においては従来の構文トランスファー方式を基盤としている.

日本語意味構造の各格要素を英語に変換し, 更に日英で格構造が異なる場合は格構造の変換を行う.

4 評価

今回, モデル対話文 182 文を構文解析し, 提案したアルゴリズム, 重み付けのパラメータによって省略の補完を行った.

モデル対話文は 182 文中に 143 個の省略を含んでおり, そのうち, ゼロ代名詞の省略 128 個, 述語の省略が 15 個であった. そのうち, 補完対象となるゼロ代名詞は 101 個であった.

補完結果を以下に示す.

- 正確に決定できた先行詞は 101 個中, 74 個であった.
- 述語の補完は倒置・付け足し表現及びパターンにあてはまるもの 12 個はすべて補完できた.

誤った補完を行ってしまったものは以下のような要因があげられる.

- 対話の性質上, 相づちによって過度の一人称ゼロ代名詞が候補として出現し, 誤って先行詞として選択された.
- 話題が定まらない会話初期に誤った先行詞選択が多かった.

これらの問題には, 以下のような解決法が考えられる.

- 候補とゼロ代名詞が係るそれぞれの述語の類似度による重みの加算

1.A: 落合が決勝打を打ったらしいよ。
2.B: 野茂が投げてたんだろ？
3.A: うん。でも、ホームランも打ったんだって。

例文 3 の「打った」は例文 1 の「打った」と類似度最大なので候補「落合」に対する重みを加算する.

述語の省略補完について, モデル対話文中の 3 文は, 3 つの省略形式のどれにもあてはまらなかった. これらは間接的照応であり, 省略の補完には読み手の知識が必要である. このような文は述語の補完をおこなわず, 最低限名詞句の訳出をおこなうことによりどれだけの意味が伝達可能か考察を深める必要がある.

A:「我々は車中泊です。」
B:「さすがに車では、、、」

上記の例文 B は「(車では) 泊れない。」もしくは,「泊まらない。」と補完できるが, これらの述語は対話文中には存在しない. とくに接尾辞「ない」の補完は直接的照応対象の探索による解決は不可能である.

5 おわりに

本研究では, チャットにおける日英対話を支援するための翻訳システムを考案・構築した.

素性の単一化による構文解析により抽出した省略箇所を, 照応関係にある先行詞を探すことによって補完することを試みた. 提案した手法によって約 7 割の補完を実現することができた.

今後は, 4 章に述べた諸問題の解決と共に, モデル文意外の翻訳を可能とするべく発展をおこなっていきたい.

参考文献

[1] 田代 敏久, 森元 逞：日本語会話文の言語解析実験, 情処学会自然言語処理研究会, 95-NL (1995)

[2] 加藤 恒昭：“自然言語インターフェースにおける省略の扱い”, 情報処理学会論文誌, Vol.34, No.9, pp1899-1908, (1993).

[3] 藤澤 伸二, 増山 繁, 内藤 昭三：“日本語文章における照応・省略現象の基本的検討”, 情報処理学会論文誌, Vol.34, No.9, pp1909-1918, (1993).

[4] “計算機用日本語基本動詞辞書 IPAL”, 情報処理振興事業協, (1987).

[5] “計算機用日本語基本名詞辞書 IPAL”, 情報処理振興事業協, (1996).