

「協調融合機械翻訳」における訳語選択¹

山田節夫 山本和英 飯田仁

{syamada, yamamoto, iida}@itl.atr.co.jp

ATR 音声翻訳通信研究所

1. はじめに

機械翻訳において、意味を正しく伝えるように翻訳するためには、全く同じ原言語表現であってもその文の意味に適した訳語の選択が必要である。ルールベース機械翻訳では、内容語に関しては、あらかじめ辞書項目に訳語選択条件を登録しておいて、翻訳する際にその条件によって訳語選択をしている。機能語に関しては、内容語に関係がある場合は辞書項目に登録し、そうでない場合はルールやパターンに記述して、原言語の解析、変換時に訳語を決定している。用例ベース機械翻訳では、内容語、機能語に関わらず、登録してある用例やパターンが原言語における語や解析結果と最も類似した訳語を選択している。しかし、用例とルールを相互利用する機械翻訳では、これら訳語選択方法を直接取り入れることは難しい。このような翻訳方式である[1]では、原言語と目的言語の解析木レベルでの対応関係を様々な抽象化で保存しておいて、原言語が入力されると原言語側の構造と語彙の情報から保存してある対応関係の中で最も適したものを選ぶ方法で訳語選択を実現している。

協調融合機械翻訳[2]も用例ベースとルールベースの融合した翻訳システムであるが、[1]と違って様々な抽象レベルのパターンを持たないので、直接この方法を適用するのは難しい。

本稿では、協調融合翻訳に適した訳語選択方法を述べる。この手法は単に用例を検索するのではなく、ルールの中に用例を埋め込み、入力された語と用例との意味距離計算をすることによって訳語選択を実現している。

2. 協調融合翻訳システム

協調融合翻訳では、変換処理を中心に形態

```
(原言語パターン)
⇒
  ((目的言語パターン1)
   ((原言語用例1)
    (原言語用例2)
    ...))
  (目的言語パターン2)
  ...)
```

図1 ルール

素処理、解析処理、生成処理など各処理が翻訳のために協調した処理を行い翻訳結果を作り出す。形態素処理は、コーパスからの学習、及び単語と品詞の混合 n-gram 出現頻度を学習し、その確率を用いて形態素単位や品詞を認定している[3]。内容語が接続する場合は、機能語に相当するマーカーを挿入することもある。解析処理は、チャート解析法を基にボトムアップで左から右への解析をしている[4]。解析木の曖昧性は、ルールに記述してある用例との意味距離計算によって適用ルールを決定し、解析木を構築している。目的言語の解析結果から形態素合成や語順の適正化など目的言語に依存したルールを用いて生成を行っている[5]。変換処理については、次節で詳しく述べる。

2.1. 変換処理

変換処理は原言語の解析処理と同時に、原言語の解析結果から目的言語の解析木へと漸次的に変換している。図1に示すように、ルールは1つの原言語パターンに対し目的言語パターンと原言語の用例から構成されている。原言語パターンは、機能語またはマーカーの前か後ろを変項とした形になっている。例えば、機能語“を”のパターンは、前と後ろに変項

¹ Word Selection on Cooperative Integrated Machine Translation; S.Yamada, K.Yamamoto, H.Iida

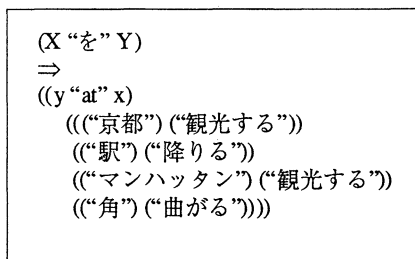


図2 “を” のルール例1

がある場合で(X “を” Y)と記述している。目的言語パターンは、原言語パターンの変項と対応するように変換後のパターンが記述されている。例えば、(X “を” Y)のパターンに対して(y “at” x)と記述すると、“を”は“at”と訳される。原言語の変項 X は“を”の前に出現しているが、それに対する目的言語の変項 x は“at”の後ろで生成される。原言語用例は、原言語パターンと目的言語パターンに合った用例を訓練文から取ってきて記述する。例えば、訓練文が「京都を観光したいのですが」で、その対訳が「I would like to go sightseeing at Kyoto」の場合、変項 X に対しては“京都”、変項 Y に対しては“観光する”が用例となり、ルールの中に記述される。図2に“を”のパターンの一部を示す。

機械翻訳システムへの入力文の単語が既に用例としてルールの中にある場合、解析処理時にそのルールが選ばれ、目的言語のパターンが決まる。例えば、上記例で「京都を観光する」は、図2のルールの(X “を” Y)のパターンに“京都”、“観光する”が記述されているので、このルールが選ばれ、目的言語のパターンは(y “at” x)となる。目的言語のパターンが決まると、y と x に当てはまる訳は、変換辞書を検索して目的言語へと訳される。この例の場合、変換辞書には“京都”の訳は“Kyoto”、“観光する”の訳は“go sightseeing”と登録されているので、「京都を観光する」に対する目的言語列は「go sightseeing at Kyoto」となる。

一方、機械翻訳システムへの入力文の単語が用例としてルールの中にない場合、入力語の単語と用例の意味距離計算によって、最も近い用例が含まれるルールが解析処理時に選

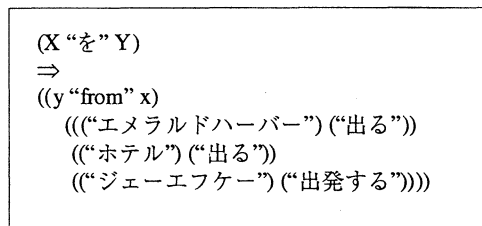


図3 “を” のルール例2

ばれる。各内容語には角川類語新辞典における分類体系に基づいた意味コードが付与されており、内容語間の意味距離計算は、付与されている意味コードによって計算される[6]。例えば、「長野を周遊したいのですが」を入力とした場合、図2に示すルールの用例には、“長野”と“周遊する”は含まれていないが、意味距離計算の結果((“京都”) (“観光する”))の対と((“長野”) (“周遊する”))の対が最も近いと判断され、図2のルールが選ばれる。入力文全てに関しては、個々のルールを選ぶ時の意味距離計算の総和が最小となる組み合わせ結果がその解析木となる。

3. 訳語選択

協調融合機械翻訳では、機能語と内容語によって訳語選択の方法が違う。本章では、機能語と内容語に分けて訳語選択方法を述べる。

3.1. 機能語の訳語選択

2章で述べたように、機能語はルールの原言語パターンに含まれる。したがって、ルールに記述する原言語の用例によって、目的言語のパターンを変えれば、2章で述べた枠組みそのまま機能語の訳語選択が可能となる。例えば、(X “を” Y)のパターンに対して図2では(y “at” x)のパターンしかないが、これに図3に示すパターンを追加すれば、用例との類似性によって目的言語のパターンは(y “at” x)か(y “from” x)のどちらかが選ばれる。「ホテルを昼に出たいのですが」を入力文した時、「ホテルを出る」に対して、目的言語のパターンは(y “from” x)が選ばれ、全体は「I would like to leave at noon from the hotel」と訳出される。

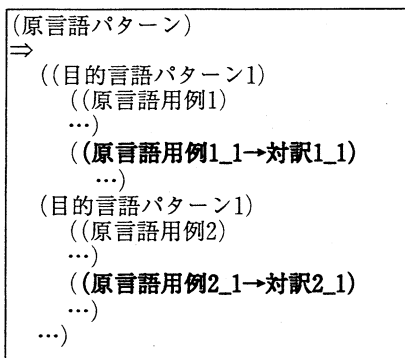


図4 内容語の訳語選択を考慮したルール

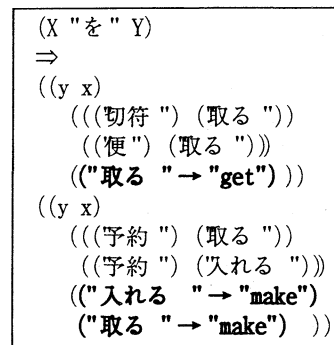


図5 “を” のルール例3

3.2.内容語の訳語選択

2章で述べた枠組みそのままでは内容語は一意に決まってしまう。しかし、用例を記述する部分を拡張すると、この枠組みでも内容語の訳語選択が可能となる。用例自身を条件にしてその用例の訳語をルールに記述することにより内容語の訳語選択を可能にしている。意味距離計算の結果、訳語付き用例が選ばれた時に同時にその訳語も選ばれるのである。これを図4に示す。目的言語のパターンが同じでも、内容語の意味によって訳を変えるためにルールの記述を別扱いにしている。

• 用言の場合

例えば、図5に示すように、(X “を” Y)のパターンに対して、目的言語のパターンが(y x)と同じにもかかわらずルールの記述を別扱いすることによって、“取る”の訳し分けを可能にしている。この例の場合では、Xの用例でYの訳語選択をしている。入力文が「それでは早速チケットをお取りいたします」の時は、図5の上のルールが選ばれ、“取る”の訳は“get”となり、その訳は「Well, I will get a ticket at once」となる。一方、「予約をお取りいたします」の場合は、図5の下のルールが選ばれ、“取る”の訳は“make”となり、その訳は「I will make a reservation」となる。

• 体言の場合

体言に対してもこの方法は有効で、例えば、変換辞書に“場所”の訳が“place”と登録してあっても、図6に示すルールを記述してお

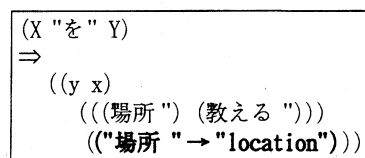


図6 “を” のルール例4

けば、「ホテルの場所を教えてください」が入力された場合、“場所”の訳は“location”となり、その訳は「Please tell me the location of the hotel」となる。

また、意味距離計算をしてルールの選択をするのと同様に用例にない入力語でも、入力された語と訳語選択条件となっている用例の意味が近ければ、その用例が含まれるルールが選ばれるので、ルールが選ばれるのと同時に訳語選択ができる。

4.考察

日英旅行会話集からオープン文延べ 629 文に対して内容語の訳語選択機能を入れたものと入れないもので、訳の質に関して比較したところ、表1に示す結果となった。訳語が良くなったのは30.5%、悪くなったのは1.4%、訳の質が変わらなかったのは68.1%となった。協調融合翻訳の枠組みのままルールの中の用例に対訳を追加するだけでも効果があると言える。なお、この比較実験に用いたルール数は777、用例数は10000、訓練文数は異なりで3000文である。

訳語選択は、3章で述べた名詞や動詞の他に、旅行会話では頻出の副詞についても効果がある。

● “ちょっと”の訳語選択例1

例えば、「ちょっとお尋ねしたいんです」の場合、訳語選択機能を入れないと、変換辞書には“ちょっと”の訳が“a little”と登録してあるので、訳語選択機能がないと“ちょっと”の訳は“a little”となるが、用例の((“ちょっと”) (“伺う”))に(“ちょっと”→“something”)が付与されていると、意味距離計算の結果、((“ちょっと”) (“伺う”))と((“ちょっと”) (“尋ねる”))がマッチするので、例文の訳は「I would like to ask something」となる。一方、「ちょっと遠いです」の場合は、訳語選択機能を入れても異なるルールが選択されるため、全体の訳は「It's a little far」となり、どちらも正しく訳し分けができる。

● “ちょっと”の訳語選択例2

また、この機能を応用して対訳を意図的に出さない訳し分けもできる。例えば、「ちょっと都合でいけません」の場合、“ちょっと”は話し言葉特有の間投詞的な意味合いがあるので、特に訳出する必要はない。用例の((“ちょっと”) (“都合”))に(“ちょっと”→“”)が付与されていると、全体の訳は「I can't go for some reason」となり、“ちょっと”は訳されない。もちろん、細かい文意を伝えるには不十分であるが、原言語の内容を伝えるには十分である。

● 訳の質が悪くなった例

訳語選択機能で訳文が悪くなったものでは、用例が少ないため意味距離計算で間違ったルールが当たっていた。例えば、用例の((“サービス”) (“する”))に(“する”→“provide”)が付与されている場合、「そのようなサービスはしておりません」の訳出は「We don't provide such service」になるが、「お買い物は後でしますか」の訳出は((“買い物”) (“する”))が((“サービス”) (“する”))の用例にマッチするので、“する”の標準訳“do”とはならず「Do you provide shopping after」になってしまう。

5.おわりに

本稿では、協調融合翻訳に適した訳語選択方法を述べた。この手法を用いれば、機能語に関してはこの枠組みそのまま訳語選択が実現でき、内容語に関しては、ルールの中の

表1 内容語の訳語選択機能の有無による差分評価結果

	文数	%
良くなった文	192	30.5
悪くなった文	9	1.4
訳の質が変わらなかった文	428	68.1

用例に対訳を追加するだけで訳語選択が実現できる。

日英旅行会話集から延べ 629 文に対して内容語の訳語選択機能を入れたものと入れないものの比較実験を行った。1.4%の訳が悪化したものの、30.5%の訳が改善し、この訳語選択方法が有効であることが確認された。

今後は、協調融合翻訳方式に基づく枠組みの利点を活かしつつ選択条件の拡張などを検討する予定である。

謝辞

本稿の作成にあたり有益なご指導をいただいた NTT コミュニケーション科学研究所の古瀬蔵氏に感謝いたします。

参考文献

- [1] H.Watanabe: An Example-Based Transfer System, Ph.D thesis, Kyoto University, 1996
- [2] 古瀬, 隅田, 飯田: 経験的知識を活用する変換主導型機械翻訳, 情報処理学会論文誌, Vol.35, No.3, pp.414-425, 1994
- [3] 山本, 河井, 隅田, 古瀬: 単語と品詞の混合 n-gram を用いた形態素解析, 情報処理学会第 54 回全国大会, 1C-02, pp.51-52, 1997
- [4] O.Furuse, H.Iida: Incremental Translation Utilizing Constituent Boundary Patterns, Proc of COLING-96, pp.412-417, 1996
- [5] S.Akamine, O.Furuse, H.Iida: Integration of Example-Based Transfer and Rule-Based Generation, Proc of ANLP'94, pp.196-197, 1994
- [6] E.Sumita, H.Iida: Example-Based Transfer of Japanese Adonominal Particles into English, IEICE Transactions on Information and Systems, E75-D, No.4, pp.585-594, 1992