対訳テキストからの翻訳知識の獲得と機械翻訳システムへの適用

北村 美穂子　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

情報流通の国際化が進むとともに，翻訳の必要性は高まる一方である．これに答えるために，機械翻訳システムの利用が注目されている．しかし，翻訳精度を支えるために必要な翻訳知識は，人手によって大量に作成しなければならず，翻訳精度の向上には限界がある．このため，翻訳知識をいかに効率良く，高精度で獲得するかという点は，機械翻訳システムの実用化において最も重要な課題となっている．

この問題を解決するために，「用例に基づく機械翻訳」[5][9]，「統計に基づく機械翻訳」[1] などの大量の対訳テキストの利用に注目した機械翻訳システムの研究が活発に進められてきた．しかし，前者は用例の一般化や用例の組合せ問題のため，一方，後者では，言語体系が大きく異なる言語間の統計的翻訳モデルの構築の難しさのため，これらの方式だけで翻訳することは難しく，これらの方式や従来の変換方式を統合した新しい枠組の機械翻訳も提案されている[2][6]．

文献[12][13]では，統計的手法と従来の構文解析技術を融合した翻訳知識の獲得方法，つまり，対訳テキスト全体の統計的分析結果と対訳テキスト内の各対訳文の構造照合結果に基づいて翻訳辞書を自動作成する方法を提案した．

本稿は，この方法を発展させ，対訳テキストから自動獲得した知識を利用した機械翻訳システムの構築について述べる．

2 翻訳知識の自動獲得

翻訳知識の自動獲得方法は文献[12][13]に基づくが，その概要を以下に述べる．

翻訳知識の獲得に必要な材料は，

- 獲得対象とする対訳テキスト
- 各文を構文解析する際に必要となる各言語の基本的な文法規則および辞書
- 文構造の対応付けを行なう際に用いる対訳辞書

である．なお，獲得対象とする対訳テキストは，文対応済みであることが望ましいが，そうでなくても既に提案されている文の対応付けの方法[7] を利用することにより，文対応済みの対訳テキストを得ることができる．

翻訳知識の自動獲得手順は以下の通りである．最初に，統計的手法に基づいて，文対応済みの対訳テキストと対訳辞書からそのテキストに依存した対訳単語間の類似度を計算する．これは，文献[3]の単語間の類似度計算方法を応用して求めた0から1までの類似度を示す値である．

次に，対訳テキスト中の各対訳文を各言語の文法規則と辞書にしたがって構文解析する．さらに，構文解析結果に対し，上記で求めた単語間の類似度に基づいて部分構造間での対応付けを行ない，その中で最も類似性の高い対応付けの結果を求める．

最後に，上記の結果から，作成したい単語を含む結果を取り出し，それをもとに各単語の訳語選択辞書，翻訳定型パターン辞書を作成する．ただし結果の信頼性を保持するため，辞書作成に用いる結果は出現回数2回以上の結果に限定する．

3 対訳テキスト知識に基づく機械翻訳システム

対訳テキストから獲得した翻訳知識は，商用化されている機械翻訳システムのユーザ辞書などに使用することができるが，より効率良く利用するために，この翻訳知識を変換辞書とする対訳テキスト知識に基づく変換方式の機械翻訳システムを提案する．このシステムを以後「対訳テキスト知識に基づいた機械翻訳」と呼ぶ．

3.1 翻訳処理方式

この機械翻訳システムの翻訳方式は変換方式を基にするが，変換処理に用いられる変換辞書は，対訳テキスト，つまり用例から獲得したものである．また，その獲得時には，文法規則や辞書の言語知識だけでなく，統計的手法から得られる言語知識も利用している．ゆえに，変換方式，用例に基づく方式，統計に基づく方式の3つの利点を生かした統合方式の機械翻訳ということもできる．

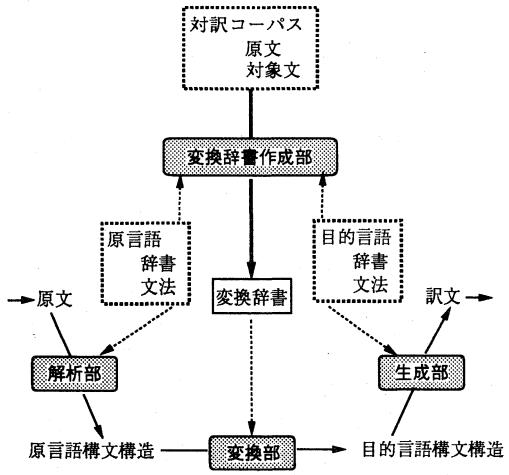

図 1: 対訳テキスト知識に基づく機械翻訳システムの概観

このシステムの概観を図1に示す.

翻訳処理に必要な材料 (図1の点線四角内) は，2. で説明した翻訳知識の獲得方法と同様，翻訳の対象と同種または同分野の対訳テキストならびに解析，生成用の文法辞書および辞書である.

翻訳処理は以下の処理部 (図1の網掛けした四角内) から構成される.

変換辞書作成部 翻訳の対象と同種または同分野の対訳テキストから，文献(12)の獲得方法に基づいて，変換辞書を作成する.

解析部 翻訳知識の獲得時に用いた文法，辞書と同じものを用いて入力文を構文解析し，原言語の素性構造を得る．構文解析結果に曖昧性がある場合は，すべての素性構造を得る.

変換部 原言語の素性構造を変換辞書作成部で作成した変換辞書を用いて目的言語の素性構造に変換する．構文解析結果に曖昧性がある場合や適合する変換辞書が複数存在する場合は，より詳細な情報を持つ変換辞書によって変換された結果を優先する.

生成部 翻訳知識の獲得時に用いた文法，辞書と同じものを用いて変換部から渡された目的言語の素性構造を生成し，訳文を得る.

英語訳語	が	を	に
give(58)	[こそあど],[本体],[社寺 学校],[店],[異同]	[異同],[開始 終了][時機],[範囲 席 跡][数],[単位][感情 気分][承認 肯定],[言語][伝達 報知],...	[こそあど][本体][範囲 席 跡][相手 仲間][社寺 学校][店 旅館]
affect(8)	[変化]	[因果]	[売買]
confer(6)	[学校]	[是非]	[店 旅館]
furnish(3)	[異同],[学校],[店]	[学校],[感情 気分]	[範囲 席 跡]
render(1)	[異同]	[救護 世話]	[範囲 席 跡]
afford(1)	[調和]		
provide(1)	[異同]		

() 内は各単語の出現回数，[] 内は各分類番号の意味[4]

日本語文パターン	英文パターン
[1] が [2] に影響を与える (17)	[1] affect [2]
[1] が [2] に同意を与える (2)	[1] assent to [2]
[1] が [2] に承認を与える (1)	[1] authorize [2]
[1] が [2] に [3] の必要量を与える (1)	[1] furnish [2] with [3]
[1] が [2] に承認を与える (1)	[1] authorize [2]
[1] が [2] に [3] の必要量を与える (1)	[1] furnish [2] with [3]

表 1: 取引条件文「与える」の翻訳規則の例

4 機械翻訳システムの試作

前項の特徴を持った機械翻訳システムの試作実験を Prolog [1] 上で構文解析システム (SAX) [2] を用いて行なった．処理の中心となる変換辞書の作成実験，変換方式の実現方法について報告する.

4.1 翻訳規則の抽出

文献(12)に基づいて日英翻訳のための変換辞書を作成した．なお，本稿では分野依存性を持った対訳テキストにおける本方法の有効性を確認するため，取引条件表現法辞典(11)の対訳文 9,804 文を対象にした．ただし，対訳文に並列文や複文が含まれると構文構造の曖昧性が極端に増え，構造照合できないという問題が生じたため，単文に分割するという前編集を施した.

対訳辞書には，パブリックドメインとして提供されている Jim Breen による EDICT1994 [3] と和英辞典(10)を合わせた 93,106 の対訳単語を用いた．また，変換辞書作成に用いるシソーラスには，分類語彙表 (8)の最下位の分類番号のクラスを用いた.

ここで獲得した対象テキスト中の動詞「与える」における訳語選択規則と翻訳定型パターンの例を表1に示す.

[1] SICStus Prolog 2.1
[2] 現在の文法規則数は，英語 121 規則，日本語 49 規則である.
[3] monu6.cc.monash.edu.au. の pub/nihongo から入手した.

4.2 変換辞書・変換処理

上記の結果を以下に定義する変換辞書の形式に変換したものを変換辞書として利用する．

trans_dict (見出し, 原言語素性構造,
目的言語素性構造, 条件).

見出し 辞書引きのキーとなる見出しを記述する．ただし，これは第 2 項の原言語側素性構造の根節点の *pred* [5] の値に含まれなければならない．

原言語素性構造 見出しを根節点の *pred* に持つ翻訳辞書の原言語側の素性構造を，文献(4)にしたがって記述する．

目的言語素性構造 翻訳辞書の目的言語側の素性構造を文献(4)にしたがって記述する．

条件 本変換辞書が適用される場合の条件を記述する．(本実験では，根節点の単語に依存する単語のシソーラスのクラスの条件が相当する．)

上記の定義にしたがった変換辞書の例を図 2に記す．このように，本辞書では，“confer:与える” のような訳語選択規則を伴った単語間の変換の場合も，“[1] が [2] に報酬を与える:[1] compensate [2]” のような複数単語に関する変換も，素性構造のパターン変換を用いて統一的に処理することができる．

一方，変換処理は，入力文を構文解析した結果得られた素性構造に対して，先に述べた変換辞書を根節点から素性構造の枝に沿って順に適用していき，再帰的に部分木の構造の変換を行なっていく．

4.3 翻訳例

本方法にしたがった特徴的な翻訳例を図 2に示す．

日本語文: 会社が販売店にすべての権利を 与える.

>適合する「与える」の変換辞書

```
tr_dict(与える, [ pred:与える (動詞),
                  が:X,
                  を:Y,
                  に:Z |Remain ],
                [ pred:confer(e_v),
                  subj:X,
                  obj1:Y,
                  on:Z |Remain ],
                ( checksem(X,[12630]),
                  checksem(X,[12650]),
                  checksem(Y,[11343]) ) ).
```

翻訳結果: Company confers all rights on Distributor.

日本語文: 会社は代理店に報酬を与える.

>適合する「与える」の変換辞書

```
tr_dict(与える, [ pred:与える (動詞),
                  が:X,
                  を:[pred:報酬 (普通名詞)],
                  に:Z |Remain ],
                [ pred:compensate(e_v),


                  subj:X,
                  obj1:Y |Remain ],
                true ).
```

翻訳結果: Company compensates agent.

図 2: 変換辞書および翻訳例

5 考察

この機械翻訳システムの特徴は，利用する変換辞書はすべて対訳テキストから獲得するという点である．したがって，対訳テキストに出現した単語や言語表現は対訳テキストの翻訳結果に忠実に翻訳することができる．さらに，対訳テキストが増えるほど，獲得できる変換辞書は増え，質の高い翻訳が期待できる．もちろん，変換辞書の追加は，従来のような手作業とは異なるため，辞書間に矛盾が生じるような問題はない．

適合する変換辞書が複数存在するという変換結果の曖昧性や構文解析結果の曖昧性の問題は，より詳細な情報を持つ変換辞書による変換結果を優先することによって解消できる．また，解析部，生成部は，変換辞書の抽出の際に用いる文法規則，辞書と同じものを用いるため，解析，変換，生成規則の間には矛盾は生じない．

変換処理に用いられる変換辞書は，単語間の変換も，複数単語に関わる変換も，素性構造のパターン変換を用いて統一的に処理される．したがって，単語間の変換も，訳語選択の条件を必要とする変換も，単語間の依存関係やカテゴリが変化する構造変換を伴うような特殊な変換も同じ方法で処理することができる．

また，変換辞書を抽出する際に用いられる構造照合結果は，対訳間の方向性はない．したがって，構造照合結果から変換辞書を作る際，両方向の変換辞書を作成すれば，双方向の翻訳も可能である．

さらに，表 1から，分野依存性を持った対訳テキスト

[5] *pred* の値はその節点に依存する単語の標準形見出しである．

を獲得対象にすることによって，その分野に依存した翻訳規則が獲得できることがわかる．翻訳は一般に分野やジャンルによって訳語や形態を変化させるが，本システムでは，対象とする対訳テキストを種類別，目的別，分野別，作者別などそのテキストの特徴にしたがって分類するだけで，対訳テキスト固有の知識を基にして翻訳することが可能となり，その結果自然な翻訳結果を生成することができる．

しかし，本システムでは次の課題が残されている．第一に，対訳テキストの疎性の問題が挙げられる．語彙，表現形式がかなり限定された分野依存性を持った取引条件文の対訳テキストを獲得対象にした場合でも表1の“afford”, “render” のような出現回数一回の重要語が数多く存在した．今後，対象とする対訳テキストを増やすことによって解決されると思われる．

次に，翻訳定型パターンに依存する単語の変換後の依存関係が特定できないという問題がある．これは，翻訳例図2において「会社は代理店に過剰な報酬を与える.」が入力された場合，現時点の変換辞書では,「過剰な」の変換後の依存関係を特定できないため翻訳することができない．この問題を解消するためには，翻訳知識の獲得の際に，翻訳定型パターン内の単語間の対応関係も獲得する必要がある．

最後に，主辞の情報の訳語選択への利用の問題がある．“paper” には,「新聞」「紙」「論文」など種々の訳語が存在する．しかし，以下の例では，主辞となる動詞“submit” によって，“paper” の訳語を「論文」と決定することができる．

The committee submitted a paper.
委員会は論文を提示した.

この例のように，名詞の訳語はその主辞となる動詞によって決定されることが多い．しかし，現時点の変換辞書では，翻訳対象となる単語の親の情報を条件とすることはできないため，今後の拡張が必要となる．

6 結論

本稿は，対訳テキストの統計的分析結果とテキスト内の対訳文の構造照合結果に基づいて自動獲得した翻訳知識を適用した「対訳テキスト知識を基づく機械翻訳システム」を提案した．

また本システムの有効性を確かめるため，分野依存性を持った実用的な対訳テキストである取引条件文における変換辞書を自動作成し，その辞書を利用した機械翻訳システムの試作実験を行なった．その結果，翻訳したい分野の対訳テキストを準備するだけで，その対訳テキストの翻訳知識を獲得しそれを利用して翻訳する機械翻訳システムの実現の見込みが得られた．一方で，そのシステムの特徴，問題点を明確することができた．

自動獲得する翻訳知識を質的にも量的にも向上させるためには，今後さらに大規模な対訳テキストが必要であるが，本システムの翻訳結果を後編集した対訳文を利用することにより，本システムの精度を向上させることができると考える．

参考文献

(1) P. Brown, J Cocke, S.D. Pietra, V.J.D Pietra, F Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, 1990.

(2) S.A.D. Brown, P.F.and Pietra and V.J.D. Pietra. Analysis, Statical Transfer, and Synthesis in Machine Translation. In *Theoretical and Methodological Issues in Machine Translation (TMI92)*, pages 83–100, 1992.

(3) M. Kay and M. Röscheisen. Text-Translation Alignment. *Computational Linguistics*, 19(1):121–142, 1993.

(4) Y. Matsumoto, H. Ishimoto, and T. Utsuro. Structural Matching of Parallel Texts. In *31st Annual Meeting of the Association for Computational Linguistics * Proceedings of the Conference (ACL93)*, pages 23–30, 1993.

(5) E. Sumita, H. Iida, and H. Kohiyama. Translating with Examples: A New Approach to Machine Translation. In *3rd Int. Conf. on Theoretical and Methodological Issues in Machine Translation*, 1990.

(6) J. Tsujii and S Ananiadou. Lbmt, kbmt, sbmt and ebmt –an attempt towards integration–. In *ATR International Workshop on Speech Translation*, 1993.

(7) T. Utsuro, H. Ikeda, M. Yamane, Y. Matsumoto, and M. Nagao. Bilingual Text Marthing Using Bilingual Dictionary and Statistics. *Coling-94*, 2:1076–1082, 1994.

(8) 国立国語研究所. 分類語彙表. 秀英出版, 1994.

(9) 佐藤理史. MBT1: 実例に基づく訳語翻訳. 人工知能学会誌, 6(4):592–600, July 1991.

(10) 清水護, 成田成寿. 和英辞典. 講談社学術文庫, 1979.

(11) 石上進. 取引条件表現法辞典 電子ブック版 第*1*巻 物品取引. 国際事業開発株式会社, 1992.

(12) 北村 美穂子, 松本裕治. 対訳テキストを利用した翻訳辞書の自動作成. In「自然言語処理における学習」シンポジウム論文集, pages 150–157, 1994.

(13) 北村美穂子, 松本裕治. 二言語対訳コーパスからの翻訳知識の自動獲得. 人工知能学会全国大会 *(第8回)* 論文集 , pages 645–648, 1994.