直訳調の訳を生成する機械翻訳

後藤 功雄　　隅田 英一郎
情報通信研究機構

1　はじめに

機械翻訳では，大きく語順を入れ替える必要がある文の翻訳は課題となっている．大きく語順を入れ替える必要性は日英のように語順が大きく異なる言語間で長い文を翻訳する場合にしばしば起こる．たとえば，次の日本語特許文

> 待機系の通信制御装置1は，ＴＥＩ／ＩＤチェック要求に対して，運用系が正しく応答するかどうかを運用系のＴＥＩ値をもとに監視する。

の英訳は以下のようになる．

> The stand-by communication control unit 1 monitors if the active unit responds correctly to a TEI/ID check request based on the TEI value of the active unit.

動詞や格要素の語順が日英で大きく入れ替わっている．

そこで，翻訳時に語順を適切に決定する方法が必要である．入力文の構文構造と同じ構文構造を持つ訳，つまり，直訳調の訳を生成することで適切な語順を決定できる可能性がある．特に，特許の明細書，法律文書，学術論文など論理性が重要視される文を機械翻訳する場合，元の文の論理を保持するためには直訳調に翻訳することが望ましいと考えられる．

本論文では，格構造に基づく構文構造を利用して直訳調の訳を生成する翻訳手法を提案する．提案手法は，入力文の格構造を保持することを優先したシステムデザインにより直訳調の訳を生成する．

2　提案手法

2.1　概要

提案手法は，直訳調の訳を生成するために，入力文の格構造を保持することを優先して訳を生成する．具体的には，入力文の格構造と同じ格構造を持つという制約のもとで訳文を生成する．

翻訳処理の概要は次の通りである．

1. 入力文の格構造を解析する．
2. 格構造の構成要素毎に翻訳候補を生成する．
3. 入力文の格構造と目的言語の格構造が同じという制約の中で，コーパスの統計量や文法に基づく知識を用いて訳語選択，語順の決定，動詞の表層形の生成を行なって訳文を生成する．

2.2　詳細

翻訳処理の詳細について述べる．日英翻訳の例を用いて手法を説明する．

はじめに，本手法で用いる格構造について説明する．格構造の格に相当するものをここでは「関係」と呼ぶ．「関係」には主に表層格を用いる．具体的には，「関係」として，機能語相当表現（助詞や前置詞など），Subject, Object，係り先の主辞の品詞を用いる．格構造の構成要素をノードと呼ぶ．ノードが動詞の場合は，能動態／受動態を区別して扱う．

以下，図1の翻訳処理の概要図を用いて各処理の内容を説明する．

2.2.1　入力文の解析

この処理は図1の (a) である．はじめに，入力文を係り受け解析する．

次に，解析結果に対して，人手で作成したパターン，およびパラレルコーパスを GIZA++ [6] でアラインメントして半数以上が英語の「関係」にアラインメントされた日本語表現の集合を用いて，一部の文節を機能語相当表現と識別して，前の文節の一部とする．例えば，「コネクタ（を）／介し（て）」という２つ文節を「コネクタ（を介して）」という1つの文節にまとめる．ここで，／は文節の区切り，丸括弧は機能語相当表現を示している．

さらに，ルールを用いて「関係」および述部の否定／肯定，時制，受動態／能動態，モダリティを識別する．なお，助詞が「は」の場合は，ガ格，ヲ格，提題のいずれかを係り先の自動詞／他動詞の区別，態，同じ係り先に係る格および日本語のモノリンガルコーパスを用いて推定する．

特許文の場合には名詞に後続する数値が頻出する．これらの数値は別扱いとしてこの段階で取り除き，後の生成時に名詞の後に挿入する．

2.2.2　構成要素の翻訳候補の獲得

この処理は図1の (b) である．ノード毎に内容語の翻訳候補を対訳辞書とパラレルコーパスから構築したフレーズテーブルを用いて獲得する．翻訳候補の獲得の手順を次に示す．対訳辞書に登録があれば，その訳を獲得する．なければフレーズテーブルからその訳を獲得する．アラビア数字とアルファベット類は ascii 文字へ変換して翻訳候補とする．名詞表現の場合は，基本形で索引を構築した英語モノリンガルコーパス中の名詞表現を検索して，冠詞を含めた表層形を獲得する．これは，複合名詞の場合には，要素合成法 [7] による翻訳候補の獲得として機能する．

また，次の例外ルールを適用して，日本語の格構造に英語で必要とされるノードが存在しない場合に，英語の格構造に合わせてノードを追加する．日本語の名詞文，形容詞文に訳候補 “be” のノードを追加する．存在を示す文には，訳候補 “be” のノードとその主語となる訳候補 “there” のノードを追加する．

2.2.3　目的言語の「関係」候補の獲得

この処理は図1の (c) である．入力文の親子関係のノードペアから，入力言語の親の主辞・親子間の「関係」・子の主辞と訳候補の親の主辞・子の主辞の５つ組みを取得する．アラインメントしたパラレルコーパス中で，親の主辞と子の主辞が親子関係を保持してアラインメントされた対訳の親子ノードペアから対訳の３つ組みをあらかじめ獲得しておく．この対訳の３つ組みを用いて，入力文中の５つ組みと一致する対訳の３つ組みの頻

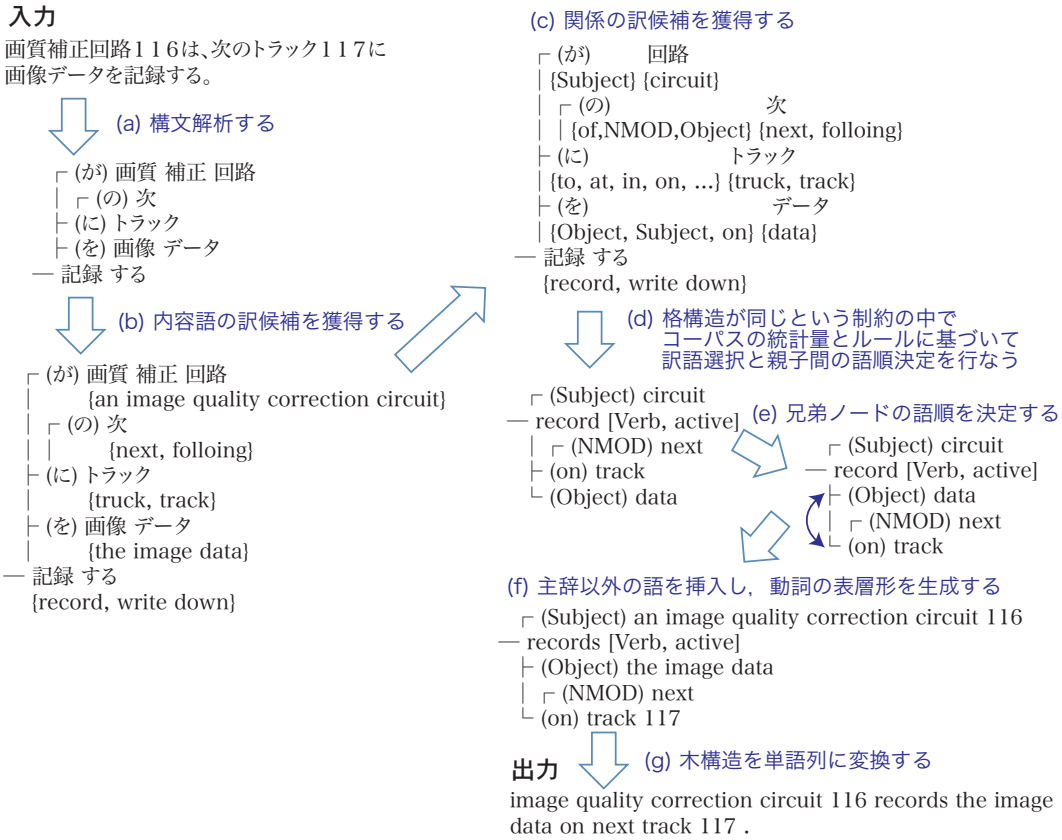

図 1　翻訳処理の概要図

度を獲得する．最も頻度が高い対訳の３つ組みの目的言語側の「関係」を候補として獲得する．さらに，主辞を品詞に汎化または主辞を省略した場合も同様にして目的言語側の「関係」候補を獲得する．また，人手で作成した対訳の「関係」ルールからも目的言語の「関係」を獲得する．

一部の「関係」(主に接続助詞などからなるもの) については，人手で作成した対訳の「関係」ルールから獲得した目的言語の「関係」候補のみを用いる．

2.2.4　訳語選択と親子間の語順決定

この処理は図 1 の (d) である．この処理で，主辞の訳語と親子間の語順が決まる．主辞の訳語，および動詞とその格要素との語順は，コーパスの統計量にもとづいて決める．

まず，注目しているノード（現ノード）への子ノードからの文脈の影響として，ノード中の全ての訳候補の主辞に対して式 (1) の b_{h_a} を計算する．b_{h_a} の計算は，木構造のリーフからルートの順に計算する．子ノードがないリーフの主辞の b_{h_a} は，そのノード中の訳語候補の主辞の種類数で 1 を割った値とする．

$$b_{h_a} = \frac{v_{h_a}}{\sum_{h_a} v_{h_a}} \tag{1}$$

$$v_{h_a} = \prod_{i \in C} (u_{i,h_a})^{1/|C|} \tag{2}$$

$$u_{i,h_a} = \max_{r,o,h_c} \left(\prod_j f_j(h_a, r, o, h_c, g_a, g_r, g_c)^{w_j} b_{h_c} \right) \tag{3}$$

ここで，h_a は現ノードの訳候補の主辞，C は現ノードの子ノードの集合，$|C|$ は C に含まれるノード数，i は現ノードの子ノード，h_c は子ノード i の訳候補の主辞，r はノード間の目的言語の「関係」，o は親子ノード間の目的言語側の語順，g_a は現ノードの入力言語の主辞，g_r はノード間の入力言語の「関係」，g_c は子ノード i の入力言語の主辞を表す．また，f_j は素性関数で，パラレルコーパスおよびモノリンガルコーパスから得られる統計量を返す．f_j には，r を含む統計量もしくは，h_c と g_c を含む統計量を用いる．r を含む統計量として，アラインメントしたパラレルコーパスから獲得した対訳の $\langle h_a, r, o, h_c, g_a, g_r, g_c \rangle$ の頻度，およびこれらの主辞を品詞に汎化した頻度や主辞を省略した頻度，またモノリンガルコーパス中の $\langle h_a, r, o, h_c \rangle$ の頻度を用いる．h_c と g_c を含む統計量として，アラインメントしたパラレルコーパスから獲得した対訳の $\langle h_c, g_c \rangle$ の頻度，および IBM model 1[1] の h_c の翻訳確率を用いる．

式 (3) の max は 1 つの子ノード中で尤もらしい主辞の訳語候補を選択している．式 (2) で平均をとっているのは全ての子ノードからの文脈の影響をまとめている．

次に，式 (4) および式 (5) を用いて木構造のルートからリーフの順に目的言語の「関係」，親子ノード間の語

順，主辞の訳語を決める．現ノードがルートの場合は，

$$\underset{h_a}{\operatorname{argmax}}\left(\prod_j f_j(h_a, g_a)^{w_j} b_{h_a}\right) \tag{4}$$

とし，それ以外の場合は，

$$\underset{r,o,h_a}{\operatorname{argmax}}\left(\prod_j f_j(h_p, r, o, h_a, g_p, g_r, g_a)^{w_j} b_{h_a}\right) \tag{5}$$

とする．ここで，h_p は親ノードの訳候補の主辞，g_p は親ノードの入力言語の主辞を表す．式 (5) は，注目しているノード（現ノード）への親ノードからの文脈の影響も考慮している．

ただし，式 (5) より次の規則を優先する．入力言語の主辞が動詞の場合は動詞の訳候補を優先する．入力言語の主辞が動詞で受動態の場合は受動態の訳候補を優先する．入力言語の動詞にガ格とヲ格が係っている場合，ガ格の目的言語の「関係」候補には Subject，ヲ格の目的言語の「関係」候補には Object，動詞は能動態を優先する．入力言語の関係が「名詞」(名詞を修飾し，機能語相当表現が存在しない）で，子ノードが木構造のリーフでない場合の語順は，親ノードが前で子ノードが後を優先する．

なお，式 (5) の括弧の中は，式 (3) の括弧の中と基本的に同じである．違いは，親子関係のノードのうち現ノードが親ノードであるか子ノードであるかである．つまり，式 (3) の h_a を h_p，g_a を g_p，h_c を h_a，g_c を g_a に読み替えれば括弧の中は同じになる．式 (4) の素性関数は，式 (5) の素性関数のうち h_a と g_a のみを利用するものである．

2.2.5 受動態への構造変換

目的言語の動詞に「関係」が Subject の子ノードがない場合は，受動態に構造を変換する．

2.2.6 兄弟間の語順決定

この処理は図 1 の (e) である．親ノードとの相対位置が同じ子ノードが複数存在する場合は，兄弟ノード間の語順を推定する．モノリンガルコーパス中で，親ノードおよび相対位置が同じ 2 つの子ノードという 3 つ組のノードの出現頻度を用いて，2 つの子ノードのうち，どちらが親ノードの近くに出現しやすいかを調べて，兄弟間の語順を決定する．この処理は内元ら [8] の手法と基本的な考え方は同じであるが，以下の違いがある．ここでは，確率モデルを構築せずに頻度をそのまま利用する．また，兄弟ノードが 3 つ以上ある場合は，全ての語順の組合せを評価するのではなく，全ての兄弟ノードのペアでどちらが親ノードに近いかを調べて，親ノードに近いペア数で兄弟ノードの語順を決定する．

2.2.7 訳文の生成

この処理は図 1 の (f) と (g) である．入力文の時制，否定／肯定，モダリティ，Subject の人称，目的言語の構造を反映させた述部の表層形を文法の知識を用いて生成する．「関係」が “of” または主辞が動詞の子ノードを持つ名詞のノードの冠詞には “the” を追加する．ただし，名詞に数値が後続する場合は冠詞を除く．木構造から単語列を生成する際に同じ単語が連続する場合，および親ノードの末尾と子ノードの主辞が同じ場合は，それらの重複を除く．最後に木構造から単語列を生成して，訳文を出力する．

表 1　コーパス

目的	言語	期間	文数
翻訳用	日英	1993-2000	1,798,571
	英語	1993-2000	147,063,894
	日本語	1993-2000	242,200,316
テスト	日本語	2001-2002	100

表 2　対訳辞書

日英対訳辞書	日本語見出し語数
EDR V3.0	364,430
Cross Language	1,539,048
JMDict	129,128

表 3　ツール

目的	名前
日本語形態素解析	Mecab[*1]
日本語依存構造解析	Cabocha[*2]
英語形態素解析	Tagger[*3]
英語依存構造解析	MSTParser[*4]

3 実験

3.1 実験設定

特許文の日英翻訳実験を行なった．

3.1.1 リソース

NTCIR-7 日英特許翻訳 [3] のデータを用いた．実験で用いたコーパスを表 1 に示す．

テストデータは以下のようにして選択した．語順を大きく入れ替える必要がありそうな文をテストデータとして選択することを考えた．そこで，比較的文が長い特許文で中程度の長さの文をテストデータとして選択した．具体的には，NTCIR-7 日英特許翻訳のフォーマルランのテストデータを文字数でソートして中央の順位の文を中央に含む 100 文をテストデータとして選択した．

実験で用いた対訳辞書を表 2 に示す．これに加えて，英辞郎日英対訳辞書と対訳コーパスから要素合成法 [7] により獲得した対訳の名詞表現も辞書として用いた．さらに，GIZA++ でパラレルコーパスをアラインメントした結果から次の 2 つの方法で獲得した対訳表現も辞書として用いた．1 つはアラインメントされた対訳のうち日本語側が主辞のもので，もう 1 つは Moses [4] のフレーズテーブルで日本語側が 1 形態素の対訳である．これらをマージした辞書を用いた．

IPAL 動詞辞書を日本語動詞の自動詞／他動詞の識別に用いた．英語形態素データベース [5] を基本形，動詞の変化形，名詞の人称の獲得に利用した．

実験で用いたツールを表 3 に示す．日本語数値表現は 1 つの形態素にまとめた．日本語の係り受け解析結果は，人手で作成したパターンとモノリンガルコーパス中の頻度を用いて解析結果の一部を自動修正した．

なお，式 (3),(4),(5) の w_j の値は，r を含む統計量を用いる f_j の種類数を c_1 とすると，それらの w_j には $1/c_1$，

[*1] http://mecab.sourceforge.net/
[*2] http://sourceforge.net/projects/cabocha/
[*3] http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/postagger/
[*4] http://sourceforge.net/projects/mstparser/

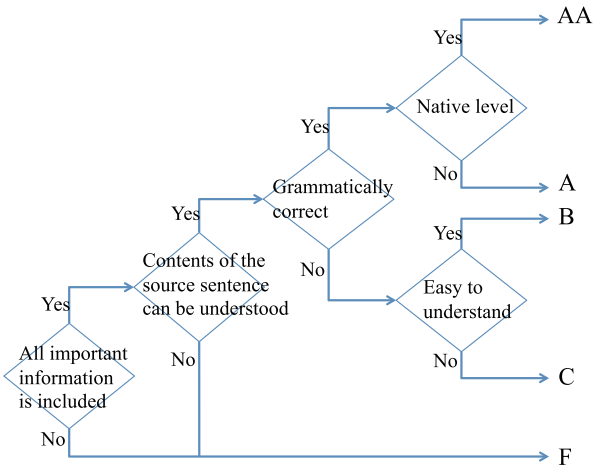

図 2　Acceptability

表 4　評価結果 (%)

	AA	A 以上	B 以上	C 以上
提案手法	1	5	20	44
階層フレーズ SMT	1	7	24	41

h_c と g_c の統計量を用いる f_j の種類数を c_2 とすると，それらの w_j には $1/c_2$ を用いた．

3.1.2　評価方法

5 段階の主観評価を行なった．評価基準には図 2 に示す Acceptability を用いた．日本語を理解できる英語ネイティブスピーカー 2 人が入力文とその翻訳結果を半分ずつ評価した．

3.1.3　比較手法

語順の違いをモデル化することができる最新の統計翻訳手法の 1 つである階層フレーズベース統計翻訳（階層フレーズ SMT）[2] との比較を実施した．この手法の実装として Moses[4] を用いた．翻訳モデルの訓練データには，表 1 のパラレルコーパスに，提案手法で利用した 3 つの既存辞書および英辞郎と要素合成法を用いて獲得した対訳の名詞表現をマージした辞書を追加して用いた．言語モデルには，表 1 の英語モノリンガルコーパスから学習した 5-gram を用いた．NTCIR-7 ドライランのデータを用いてパラメータを調整した．システムの設定は次の点を除いて標準設定を用いた．変更点はルール作成時に変数にする最小の単語数で，標準の 2 単語を 1 単語に設定した．これは，1 単語に設定することによって，数値部分だけを変数として扱えるようになるので翻訳精度が高くなると考えたためである．

3.2　評価結果と分析

評価結果を表 4 に示す．提案手法は，階層フレーズ SMT と比較して同程度の翻訳精度が得られた．

提案手法が階層フレーズ SMT より高い評価になった例を示す．

入力文：分割された４つのフォトダイオードＰＤａ～ＰＤｄは、先に説明した非点収差法などでフォーカス誤差信号を読み取っている。

階層フレーズ SMT：four photo diodes pda to pdd divided as described above , the focusing error signal by an astigmatism method or the like is read .

提案手法：the four divided photodiode PDa ~ PDd reads focusing error signal by the astigmatism method described previously .

提案手法で適切に訳せなかった例とその原因を示す．

入力：「入出力制御装置の構成」

出力：“the structure an input/output control apparatus”

名詞間の「関係」の訳の精度が高くないことが原因である．

入力：「ピン４５、タペット４４及びピストン３３を往復運動させる。」

「ピン４５」の係り先が「往復運動させる」と誤って解析され，「ピン４５」が主語として扱われた．

入力：「A は…B し、…C する。」この場合，A は B と C の両方に係るが，構文解析結果では C に係っているという情報しか得られないため，B の主語がない翻訳結果になった．

入力：「…消費されてしまうことになる。」

出力：「… are consumed, is become」

「になる」を動詞のモダリティと同じように扱う予定であったが，ルールの不足からそのように扱われなかった．

4　おわりに

入力文の格構造と同じ格構造を持つ直訳調の訳を生成する翻訳手法を提案した．入力文を自動解析した結果を用いて，特許文で中程度の長さの文を対象とした日英翻訳で階層フレーズベース統計翻訳と同程度の結果が得られた．特許文の形態素・構文解析精度は改善の余地があり今後の向上が期待できるため，入力文の解析精度に依存する提案手法の精度も向上が期待できる．

提案手法は，名詞間の修飾の翻訳精度が高くなかったが，このような表現は語順が大きく変わらないため統計翻訳の翻訳精度が高いと思われる．そこで今後は，名詞句など語順が大きく変わらない表現の翻訳に統計翻訳を利用して，それらの訳を格構造にもとづいて組み合わせて全体の文構造を生成することで，文構造としては直訳調を保ちながら，各部分の訳の精度を向上させたい．

参考文献

[1] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *CL*, 1993.
[2] D. Chiang. Hierarchical phrase-based translation. *CL*, 2007.
[3] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro. Overview of the patent translation task at the ntcir-7 workshop. In *Proc. NTCIR-7 Workshop*, 2008.
[4] H. Hoang, P. Koehn, and A. Lopez. A unified framework for phrase-based, hierarchical, and syntax-based statistical machine translation. In *Proc. IWSLT*, 2009.
[5] D. Karp, Y. Schabes, M. Zaidel, and D. Egedi. A freely available wide coverage morphological analyzer for english. In *Proc. Coling*, 1992.
[6] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 2003.
[7] M. Tonoike, M. Kida, T. Takagi, Y. Sasaki, T. Utsuro, and S. Sato. A comparative study on compositional translation estimation using a domain/topic-specific corpus collected from the web. In *Proc. the 2nd International Workshop on Web as Corpus*, 2006.
[8] K. Uchimoto, M. Murata, Q. Ma, S. Sekine, and H. Isahara. Word order acquisition from corpora. In *Proc. Coling*, 2000.