用例利用型日英機械翻訳の基本設計

高橋大和　　白井　諭
ＮＴＴ　コミュニケーション科学研究所
立花正敏　　西垣万亀子
ＮＴＴ　アドバンステクノロジ
池原　悟
鳥取大学

1.はじめに

機械翻訳の方式はルール型と用例型に大きく分けることができる。また、機械翻訳システムは翻訳したいドメインを限定しなければ、実際に運用するのは難しいことも良く知られている。しかし、ドメインへの対応と翻訳精度の向上を図る場合、ルール型においては辞書とルール作成が難しいこと、各処理の精度の向上が難しいことが挙げられ、また、用例型においては、大量のタグ付コーパスをどう作成するか、といった問題が挙げられる。この他、いずれにも未解決の課題が多く、特に日英翻訳では、言語類型が大きく異なることもあり、実用段階とは言い難く[1]、研究途上であると考えられる。そこで本稿では、ルール型及び用例型の特徴を生かし、統計的手法を併用することにより、現状の技術レベルで構成可能な機械翻訳方式を提案する。

2.各種翻訳方式の得失

本節では、各方式の得失について検討を行う。

2.1.ルール型翻訳方式

現在市販されている機械翻訳ソフトの多くはこの方式によると思われる。いずれも、入力文を解析する処理(形態素解析、構文解析、意味解析など)と、構造変換を施した後の内部構造または中間言語から出力文を生成する処理を直列に配置する。各処理の動作は辞書とルールにより制御する。

システムの翻訳精度は各処理の精度の積で効いてくるため、辞書やルールを大規模化、高精度化する必要がある[2]が、それには多大な工数を要するという問題がある。また、深く解析すると詳細な関係情報が使えるようになる反面、単語間の関係を喪失しがちになる。このような要素合成的手法により訳文を生成すると文全体としての体裁が整えにくい事もこの方式の持つ問題といえる。

2.2.用例型翻訳方式

ルール型翻訳方式における辞書やルールの問題を克服するため、類推により翻訳する手法[3]が提案された。これは、あらかじめ対訳例文集を用意しておき、翻訳対象文と類似した翻訳例を真似ることにより翻訳するもので、翻訳例があれば整った訳文が生成されるという利点がある。また、言語現象を個別に分析して辞書やルールを作る必要がなく、対訳例文を追加するだけで翻訳能力の向上が期待できる。

この方式では、１対１に厳密に対応する対訳例文集の存在と、例文への単語やフレーズなどのタグ情報の付与を前提とすることが多い。しかし、現実には１対１の対訳例文は多くは存在せず、シソーラスを用いた類似判定も確実ではないため、マニュアルなどの改版に伴う前版からの流用が効果をあげているに過ぎない。また、大量の対訳例文を確保できたとしても、それらへ均質かつ正確にタグ情報を付与するのも容易ではない。

2.3.融合型翻訳方式

ルール型と用例型を併用するタイプとそれらを積極的に統合しようとするタイプがある。併用するタイプは、各方式の得失を引き継いでいるほか、最適な結果を選択するにはどうするかという新たな問題を生じる。統合するタイプは、解析結果と対訳例文集のタグとの整合が問題になる。

3.用例利用型翻訳方式の提案

ルール型の利点は解析により言語情報が得られる点、用例型の利点は翻訳例により整った訳文が生成される点にあると思われる。そこで、これらの利点を生かし、各処理の精度の積で全体の精度が決まる解析型の欠点と、対訳例文集やタグ付与といった用例型の欠点をカバーする方式として、図 1に示すような用例利用型翻訳システムを提案する。

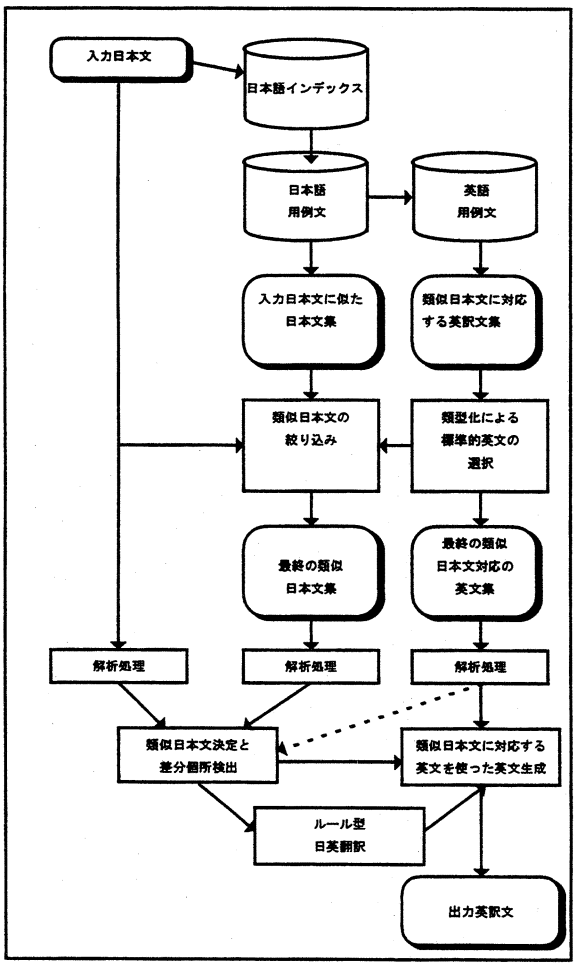

図1　用例利用型翻訳システムの基本構成

3.1.対訳例文の収集

本稿では、日経新聞の市況記事を対訳例文のソースとして利用する。市況記事の場合、日英の記事は1対1の対訳ではないが、半数の英文に直訳的に対応する日本文が存在している[4]ため、対訳用例として適用可能であると考えられる。大量の対訳例文を容易に収集するためには、対訳例文集の作成には人手の介入を不要とし、自動的に抽出、日英対応付けを行う必要がある。そこで、対訳例文を作成する際、再現率よりも適合率を重視して日英文対応付けを行う。この方針に従った緩い対応付け(厳密な対応ではないが、ある程度直訳的な例文対の収集)は可能であると考えられる[5]。また、新聞記事を対象にする事により、大量の対訳例文集が構築可能となる。ただし、対訳例文集には、3.3.節の理由により例文へのタグの付与は行わない。

3.2.対訳例文集の利用

利用にあたっては緩い文対応である事を念頭に置く必要がある。そこで、対訳例文を以下の手順で選択する。

1. 入力文と同じ文字を含む例文を複数取り出す。
2. それらに対応付けられた訳文を統計的に類型化する。
3. もっとも代表的な訳文を選択する。
4. その文と対応付けられた例文を入力文の類似文として利用する。

この方法により、妥当な訳文を与えるような類似文と対訳表現を選択する。

3.3.解析処理の適用

入力文と類似文に対して、平行して解析処理を適用する。解析によって得られた各種言語情報を利用して差分個所を検出する。平行して解析する事により、入力文と用例文の同一のある表現で解析誤りが生じた場合においても、同じ解析結果を得るため、解析誤りを打ち消しあう事が期待できる。また、例文へのタグの付与が不要になり、タグと解析結果のミスマッチが防止される。

3.4.訳文生成の基本的な考え方

類似文に対する訳文を解析し、差分個所を決定する。類似文との対応部分に応じて、訳出の単位を単語からフレーズへ変更する。差分個所の翻訳には、当初はルール型翻訳の利用を考えている。

4.用例利用型翻訳システムの試作

3.節で提案した用例利用型翻訳方式を基にプロトタイプの試作を行った。

4.1.対訳コーパスの構成

対訳コーパス(文対応のみ)は、和文データベースと英文データベースに分けて構築される。英文データベースは和文データベースのリンク情報により検索することができる。日本語インデックスは、和文データベースを基に、N-gram[6]を用いて対訳例文に3回以上出現する表現を抽出し、高速な検索を行うために、トライ構造[7]を用いたデータベースとして構築している。

4.2.候補文の検索

N-gram インデックスにより入力文の形態素がある割合以上含まれる対訳コーパスの和文を候補文として抽出する。

4.3.候補文の評価

候補文と入力文の動詞の部分や各文の形態素の

並び順(を格、に格など)、また、パターン対辞書利用による英訳文などを利用して、類似度を計算し、もっとも値の大きいものを類似文として選択する。ただし、プロトタイプでは、英文の類型化は行っていない。

4.3.1.類似の定義

本稿では、入力文と候補文の文間距離を、候補文の形態素を入力文の形態素の並びに極力一致するようバブルソートした際のスワップ回数を基準とした。

4.3.2.類似度の計算式

類似度は以下の式で計算する。

類似度＝(1－スワップ回数／最大スワップ回数)
×(一致した形態素数／候補文の形態素数)
×(一致した形態素数／入力文の形態素数)
×((一致した助詞の直前の形態素数＋1)
／(候補文中の助詞の数＋1))
×((一致した助詞の直前の形態素数＋1)
／(入力文中の助詞の数＋1))

第一項は、入力文と候補文を形態素単位に切り分け、その共通項が入力文と同じような並びになるまでのバブルソート回数から算出する。

第二項は、候補文が入力文をよりも長い場合の補正項である。

第三項は、入力文の方が長い場合の補正項である。

第四項は、候補文から見た文法構造の類似度の補正項である。

第五項は、入力文から見た文法構造の類似度の補正項である。

第二項・第三項は、入力文と類似文を比較する際、冗長か、不足しているかを類似度に反映させる。また、第四項・第五項により、文法構造を類似度に反映させて、評価を行う。

4.4.英文の生成

入力文と類似文での差異を識別し、該当部分に対して日英辞書による翻訳により英訳を取得し、置換修正を行う。プロトタイプでは、名詞の置き換えが可能になっている。

5.本システムの特徴

試作した用例利用型翻訳システムの利用環境を図 2に、実行画面を図 3に示す。

主な特徴

- クライアント・サーバ方式により、多数のユーザーでの利用を可能とした。サーバ・アプリケーションは UNIX、クライアント・アプリケーションは、windows-NT で動作する。
- 個人用と共通の対訳コーパスデータベースに分割しているため、ユーザ独自の対訳コーパスを登録できる。
- 対訳データの入力インターフェイスにより、大規模対応の対訳コーパスが実現できる。
- 形態素解析システムとして ALTJAWS を利用している。
- 融合型翻訳インターフェイスにより、既存の翻訳システムと併用することにより、より良い翻訳環境が実現できる。
- GUI の活用により、操作性の容易化と高度化を計る。

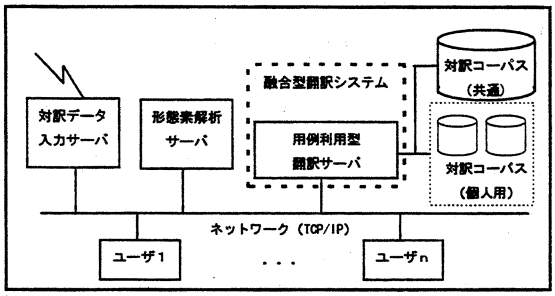

図2　試作した用例利用型翻訳システムの構成図

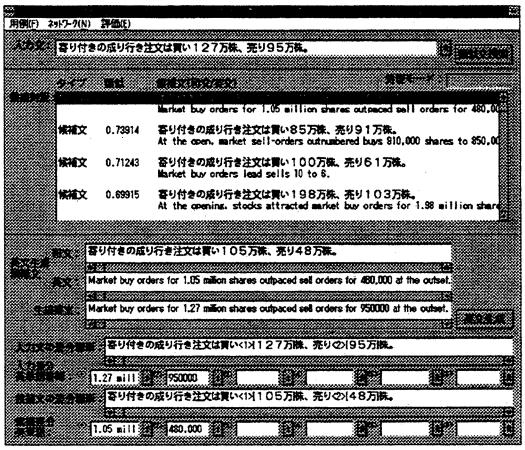

図3　実行画面

6.類似文検索実験

試作した日英用例利用型翻訳システムに対して、95/8 から 95/11 までの市況記事文(約 23000 文)を用例データベースとして用い、この期間内の 100

文(ウィンドウテスト文)と期間外の100文(ブラインドテスト文)との類似文の検索率を確認した。結果として、ウィンドウテスト文では、第一候補に入力文と同一の文が検索されたのが80文、ブラインドテスト文では、同一文が2文検索され、他はすべて候補類似文を検索することができた。図4にブラインド文における類似度の分布を示す。

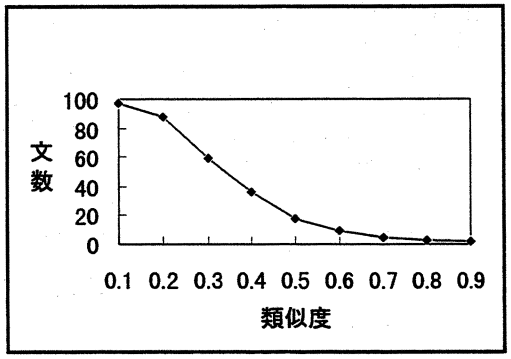

図4　類似候補文の類似度(ブラインド文)

図4を見ると、類似度が0.5以上の類似文は18文であるが、表1により実際に検索された候補文と入力文を比較すると、類似度が0.4～0.5であっても類似している。対訳データベースの充実、類似度の計算式、評価の改良、また、英文生成において、部分対応している類似文から合成する方法などを検討することにより、性能を向上できると考えられる。

表1　類似度と類似文(ブラインド文)

類似度	入力文(上)／類似文(下)
1	譲渡性預金（ＣＤ）は取り引きが成立していない。
	譲渡性預金（ＣＤ）は取り引きが成立していない。
1	株価指数先物・オプション・前引け
	株価指数先物・オプション・前引け
0.804949	株価指数先物・オプション・大引け
	株価指数先物・オプション・大引け――買い戻しで高値引け。
中略	
0.584716	ＴＯＰＩＸ先物６月物は同１０ポイント安の１６６８ポイント、日経３００先物６月物は同１．０ポイント安の３０９．０ポイントで前場を終えた。
	ＴＯＰＩＸ先物12月物は同８ポイント安の１４０３ポイント、日経３００先物12月物は同１．８ポイント安の２６２．７ポイントで前場を終えた。
0.569083	14時現在、前週末比52銭円安・ドル高の1ドル＝105円32－35銭で取引されている。
	14時現在、前週末比24銭円安ドル高の1ドル＝101円55－58銭で取引されている。
中略	
0.467692	一方、ＮＴＴデータ、ソニーミュが安く、邦チタも軟調
	一方、ＮＴＴデータは小高い。
0.463644	無担保コール翌日物は前日比０．０１％高の０．５１％で若干ながら出合っているようだ
	無担保コール翌日物は前日比横ばいの０．５０％程度で推移。

7.おわりに

本稿では、緩い対訳データで動作可能な日英用例利用型機械翻訳システム方式の提案とその試作を報告した。現在、プロトタイプの改良および評価を行っている。プロトタイプでは、入力文の解析は形態素解析のみであるが、今後は類似判定の適正化、構文解析や意味解析などの導入を考えている。また、評価式、英文生成の改良もあわせて進める予定である。

謝辞　本システムの実現にご協力くださったNTTアドバンステクノロジ(株)の田邊俊明氏に感謝いたします。

参考文献

[1] 成田:言語類型と機械翻訳,情報処理学会研究報告,Vol.96,No.65,96-NL-114-21,pp.143-150

[2] 池原,宮崎,横尾:日英機械翻訳のための意味解析用の知識とその分解能,情報処理学会論文誌, Vol. 34,No.8,pp.1692-1704

[3] Nagao, M. : A framework of a mechanical translation between Japanese and English by analogy principle, *in Artificial and Human Intelligence*, Elithorn & Banerji, eds., pp.173-180

[4] 白井,藤波,池原,上田,井上:新聞記事日英対訳コーパスの構築(1)―基本構想と検討課題--, 平成七年度(第四十八回)電気関係学会九州支部連合大会, 1373, p.855

[5] 高橋,白井,藤波,池原,上田,松島:ＤＢから抽出した日英新聞記事の自動対応付け, 言語処理学会第2回年次大会, B3-3, pp.201-204

[6] 池原,白井,河岡: 大規模日本語コーパスからの連鎖型及び離散型の共起表現の自動抽出法, 情報処理学会論文誌, Vol. 36, No.11, pp.2584-2596 (1995)

[7] 青江順一:ダブル配列による高速ディジタル検索アルゴリズム,信学論(D), Vol. J71-D, No. 9, pp.1592-1600(1988)