日英ニュース記事を用いた用例ベース翻訳システム

荒牧 英治 †‡　黒橋 禎夫 †‡　柏岡 秀紀 ‡　田中 英輝 ‡
†東京大学大学院情報理工学系研究科
‡ATR 音声言語コミュニケーション研究所

{aramaki, kuro}@kc.t.u-tokyo.ac.jp
{hideki.kashioka, hideki.tanaka}@atr.co.jp

1 はじめに

用例ベース翻訳 [5] とは，入力文と類似した用例をデータベースから探しだし，その用例を組み合わせて翻訳を行う手法である．この方式で実用的なシステムを構築するためには，大規模な対訳コーパスと高精度な構文解析が必要となり，それらによって構造的情報をもった対訳用例を大量に構築する必要がある．近年，大規模な対訳コーパスと高精度な構文解析は徐々に利用可能となってきている．しかし，対訳コーパスからスタートして最終的に翻訳文を生成するまでの一連の過程を実現し，十分に議論した研究はまだ数少ない．

本研究では，このような一連の過程を NHK の日英対訳記事コーパスをもとに実現した．まず，対訳記事の文対応付け，句対応付けを行い，その中の確信度の高い部分だけを対訳用例データベースとした．次に，日本語入力文に対して，日本語表現間の類似度に基づき翻訳に使える用例断片を取り出し，それらの英語表現を組み合わせることによって翻訳文を生成した．

本稿では，この一連の過程の中で，用例の作成方法と選択方法を中心に述べる．

2 TMの作成

用例ベース翻訳を行うために，まず，対訳コーパスから用例 (以降，**TM** とよぶ) を作成する必要がある．ここでいう TM とは，日本語と英語の句の対応関係が明らかとなっている対訳文である．

これを作成するために用いた NHK 日英対訳記事は 4 万記事ペア (5 年分) からなる (平均文数は日本語部分 5.2 文，英語部分 7.4 文)．表 1 に例を示す．日本語原稿では女子バレーボールについて，英語原稿では男女両方のバレーボールについて述べられている．このように内容全体として記事同士は対応しているものの直訳されて作成されているわけではなく，コーパスのすべてを TM として利用するのは困難である．そこで，小規模な実験で精度を調べ，アライメント結果が高精度なものだけを抽出することを目標とする．

2.1 文アライメント

2.1.1 文アライメント手法

本手法では，DP マッチングを用いた文アライメント手法を採用する．扱う文対応は，実験セット中で文対応全体の 84%を占めた (日本語文数:英語文数)=(1:1), (1:2), (1,3), (2,1), (2:2) の 5 種類の文対応とする．

対応の類似度 (以降，**文対応スコア**とよぶ) は，両言語の文対応内に含まれる内容語のうち翻訳辞書で対応する内容語の割合として，次のように定める．

$$\text{文対応スコア} = \frac{W_d \times 100}{W_j + W_e}$$

ただし，W_j を日本語内容語数，W_e を英語内容語数，W_d を辞書で対応する両言語の内容語数 とする．

利用した翻訳辞書は，EDR 日英対訳辞書，EDICT(一般的な日英対訳辞書)，ENAMDICT(固有名詞の日英対訳辞書)，アンカー日英対訳辞書，英辞郎である．これらの辞書にはのべ約 200 万語 (句) の対応が記載されている．

2.1.2 文アライメント結果

96 記事ペアで正解文対応を作成し，適合率 (文献 [7] の定義による) を調べた結果，60.7%であった．これは，十分な精度とは言えないので，1:1 文対応 (適合率 77.5%) のみに絞って，次節で述べる句アライメントを行った．

2.2 句アライメント

2.2.1 句アライメント手法

次に対訳文の句アライメントをとる．本稿では，文献 [1] をもとにした手法を用いた．本手法は次の 3step

表 1: NHK ニュース記事の例

バレーボールの女子チームのアトランタオリンピック出場が決まりました。女子バレーボールのアトランタオリンピック出場権をかけた世界最終予選で、今夜日本はクロアチアをセットカウント三対〇のストレートで破って三位以内を確定し、アトランタオリンピックの出場を決めました。日本の女子は、オリンピックにバレーボールが採用された一九六四年の東京大会以来、連続で出場権を獲得しました。

Japan, the Netherlands and Ukraine have qualified for the women's volleyball event in the Summer Olympic Games in Atlanta. The three countries secured the top three spots in the final qualifying tournament in Osaka on Saturday. Japan defeated Croatia three to zero, and together with the Netherlands and Ukraine had a record of five wins and one loss in the eight-country round-robin tournament. The Japanese women's team has qualified for every Olympic volleyball competition since the sport became an Olympic event in 1964. The Japanese men's team failed to qualify for the Atlanta Olympics.

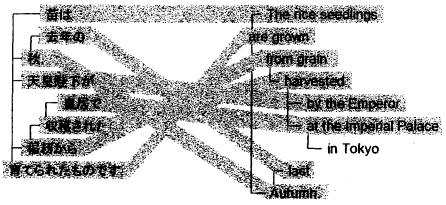

図 1: 基本句を単位とした依存構造と句アライメント (TM)

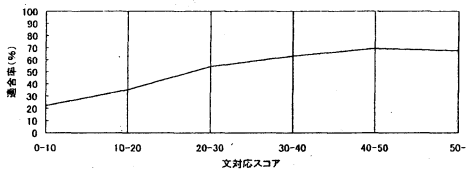

図 2: 文対応スコアと句アライメント精度

からなる.

Step1 基本句を単位とした依存構造への変換: ここでいう基本句とは，内容語を中心として機能語をまとめたもので，複合名詞や，名詞と前・後置詞，主動詞と助動詞などをひとまとまりとしたものである．日本語の文については，KNP[4] を用いて統語解析を行ない，その文節を基本句とする．英語の文については，Charniak の統語解析システム [2] を用い，日本語の文節と同様な単位に語をまとめる.

Step2 辞書の情報による対応関係の推定：次に翻訳辞書を用いて日英の内容語間に対応をつける．基本句に含まれる内容語が辞書に載っていれば，両言語の基本句同士を対応付ける.

ある語の訳語が相手側言語に複数存在すれば，基本句の対応は曖昧性を持つ．この曖昧性解消については，文献 [1] の手法による.

Step3 未対応句の処理：次に，対応付けられていない基本句 (以降，**未対応句**とよぶ) を規則によって処理し，対応付けを行う．規則に適合する未対応句が存在しなくなった時点で，句アライメント処理は終了し，図 1 のような TM を得る.

2.2.2 句アライメント結果

文アライメント実験に用いた記事から抽出した 1:1 文対応に，正解の対応を内容語単位で作成し，句アライメントの適合率と文対応スコアとの関係を調べた.

$$\text{適合率} = \frac{\text{内容語対応を含んでいる基本句対応の数}}{\text{システムが推定した基本句対応の数}}$$

図 2 のように，文対応スコアが高くなるほど，適合率は高くなる傾向にある.

表 2: TM の種類と数

TM	文対応スコア	数
NHK ニュース	30〜40	18290
	40〜50	6975
	50〜	2314
SENSEVAL	-	6920
白書	-	2225

2.3 TM のデータベース化

句アライメントの適合率が 60%を超える文対応スコア 30 以上の句アライメント結果を TM としてデータベース化した．また，ニュースコーパスの対訳文に加え，日英両言語で作成されている科学技術白書および経済白書，SENSEVAL[3] の用例についても同様にデータベース化した（白書と SENSEVAL の用例については，すでに文対応がとれている）．データベース化された TM の個数は表 2 のようになる.

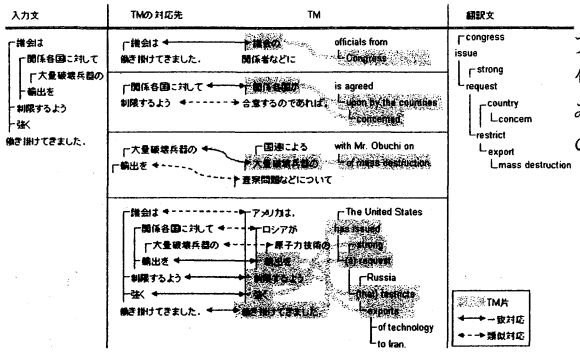

図 3: 手法のながれ

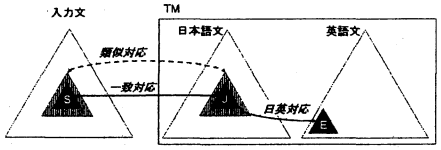

図 4: TM の選択

3 用例ベース翻訳システム

3.1 入力文の解析

入力文を構文解析し，**基本句**を単位とした依存構造にする．これは，句アライメント手法と同じ手法を用いる．この結果，図 3 左のような入力文の構造が得られる．

3.2 TM の選択

用例ベース翻訳では，多くの場合，一文を翻訳するのに複数の TM から翻訳に利用する部分木対を取り出し，それらを適切に結合する必要がある．入力文のある部分の翻訳に利用できる TM は次の条件を満たすものである (図 4)．

1. 翻訳対象となる入力文の部分木 (以降，部分木 S とよぶ) を翻訳するためには，部分木 S と TM の日本語部分木 (部分木 J) の表現が一致する必要がある．ここでいう一致とは，部分木内の内容語の一致とする．

2. 部分木 J と対応する英語部分木 (部分木 E) のアライメントされ日英対応が明らかとなっている．

3. それぞれの部分木 S,J,E は連続する部分木とする．これは，後の木構造の結合処理の曖昧さをなくすための制約である．

ここで，このような条件を満たす部分木 J，E のペアを TM 片とよぶことにする．これを入力文の基本句ごとにデータベースで検索する．通常，この条件をみたす TM 片は，TM データベース中に複数存在するので，もっとも適切な TM 片を選択する必要がある．

本手法では，適切な TM 片を次のように考える．

1. 入力文と TM 片の日本語側で一致部分が大きい．

2. 入力文と TM 片の日本語側で，一致部分に隣接している基本句同士が類似している．

3. TM 片内の日英対応の確信度が高い．

TM 片の選択は，これらを考慮した以下の TM 片スコアを用いて行う．

$$TM\text{ 片スコア} = \sum_{\text{一致対応}} \text{基本句類似度} \times \sum_{\text{日英対応}} \text{日英対応確信度} + \sum_{\text{類似対応}} \text{基本句類似度},$$

ただし，式の 一致対応 は入力文と TM 片の基本句対応の集合，類似対応 は 一致対応 に隣接する入力文と TM の基本句対応の集合，日英対応 は TM 片内の日英対応の集合である．

基本句類似度 の定義は以下のとおり．

$$\text{基本句類似度} = \frac{\text{内容語一致度} \times 2}{\text{対応内の内容語数}} + 0.2 \times \frac{\text{機能語一致度} \times 2}{\text{対応内の機能語数}}$$

$$\text{内容語一致度} = \begin{cases} 1.1 & \text{活用形も含めて一致} \\ 1.0 & \text{原型が一致} \\ 0.5 \times S_{ntt} + 0.3 & \text{シソーラスで類似} \\ 0.3 & \text{品詞が一致} \\ 0 & \text{その他} \end{cases}$$

$$\text{機能語一致度} = \begin{cases} 1.1 & \text{活用形も含めて一致} \\ 1.0 & \text{原型が一致} \\ 0 & \text{その他} \end{cases}$$

S_{ntt} とは，NTT の日本語語彙大系 [6] を用いて計算した類似度とする (最大 1.0)．品詞とは KNP が出力する体言と用言の分類とする．

日英対応確信度 の定義は以下のとおり．

$$\text{日英対応確信度} = \frac{\text{文対応スコア}}{60} \times \begin{cases} 1.0 & \text{内容語がすべて辞書で対応する} \\ 0.9 & \text{一部の内容語が辞書で対応する} \\ 0.5 & \text{その他} \end{cases}$$

以上のようなスコアによって，入力文の基本句ごとにもっともスコアの高い TM 片を選択する．図 3 中央に，選択された TM と使用される TM 片の例をあげる．図 3 の一番下の TM 片のように，日英間で構造が異なる表現の場合でも，それが一つの TM 片によって

表 3: 実験結果

	本手法	ベースライン
正解	169(159)	136
不正解	30(14)	63
精度	84.9%(91.9%)	68.3%

扱われることにより構造を変換する翻訳が自然に実現されることになる.

入力文の基本句と一致するTM片がデータベース中に存在しない場合は,入力文基本句の内容語を翻訳辞書で辞書引きし,得られた語をTMと同様に扱う.辞書引きは,入力文基本句の前方から最長一致法で行う.この際,翻訳辞書に複数の訳語が記載されていた場合は,ニュースコーパスでの頻度の高い語を採用する.

入力文の基本句ごとにTMを選択すると複数のTM片が同じ入力文基本句をカバーする場合があるが,この場合はTM片スコアの高い方を翻訳に用いる.

3.3 翻訳文の生成

次に,選択されたTM片の英語基本句同士を結合して依存構造を作成する.これは,次の2つの規則を用いて行う.(1) TM片内の基本句の依存構造は保存する.(2) TM片間は,対応先の入力文基本句の親子関係にもとづいて結合する.

次に,これを直列化し翻訳文とする.この処理は,TM片内の順序は保存し,TM片間の順序は規則によって決定する.

最後に活用,冠詞,単数-複数を決定し,翻訳文を得る.

図3右にTM片を結合した依存構造の例をあげる.

4 実験と考察

訳語選択という観点から手法を検証した.NHKコーパスで文対応スコアの高い(60以上)30文を無作為に抽出し,対訳文の日本語側を入力とし,対訳文の英語側を正解翻訳文として実験を行った(表3).評価は入力文の基本句ごとに正解翻訳文を参考にして行い,正解と不正解に分類した.この際,接続詞や日付・数字表現("三日","150人が"など)は評価の対象としなかった.

精度は,正解 / (正解+不正解)であり,ベースラインは翻訳辞書を用いた場合である.括弧内は,本手法でTM片が見つからず翻訳辞書が使用された場合を除いた値である.

結果は,表3のようになり,ベースラインを上回った.また,TM片が得られた場合の精度が高いことから,本手法の精度の低下は,TMが存在しないことが大きな原因と考えられる.

誤りを調べると,データベース中からTM片が少数しか得られず,日英対応の確信の低い対応を採用してしまう場合が多い.例えば,図3の例では,"大量破壊兵器の"の訳語が不正解である.基本的には,この問題は対訳コーパスの量が増えることにより自然に解消されると考えられる.また,現在は基本句を単位としてTM片を探索しているが,その日英対応の確信度が低い場合には,より小さい粒度で用例を利用することも考えられる.

5 まとめ

本稿では,日英対訳記事コーパスからスタートし翻訳文を生成するまでの一連の過程を実現した.訳語選択と実験の結果ではベースラインを大きく上回り,現在の対訳コーパス量で,用例ベース翻訳の実証的研究が可能であることを示した.

参考文献

[1] Eiji Aramaki, Satoshi Sato Sadao Kurohashi, and Hideo Watanabe. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of MT Summit VIII*, pp. 27–32, 2001.

[2] Eugene Charniak. A maximum-entropy-inspired parser. In *In Proceedings of NAACL 2000*, pp. 132–139, 2000.

[3] Sadao Kurohashi. Senseval2 japanese translation task. In *Proceedings of SENSEVAL2*, pp. 37–40, 2001.

[4] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, 1994.

[5] Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. In *In Artificial and Human Intelligence*, pp. 173–180., 1984.

[6] NTTコミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[7] 内山将夫, 井佐原均. 日英新聞記事の対応付けと精度評価. 情報処理学会研究報告 2002-NL-151, pp. 15–22, 2002.