自動学習された機能語の翻訳パターンを用いた用例ベース機械翻訳

中澤 敏明　　黒橋 禎夫
京都大学大学院情報学研究科
nakazawa@nlp.kuee.kyoto-u.ac.jp　kuro@i.kyoto-u.ac.jp

1　はじめに

近年の機械翻訳手法の主流は、統計的な情報から単語対応を自動的に獲得し、その結果から単語列で連続な部分を句として抽出して翻訳知識とする prase base SMT [3] である。また句の抽出時や翻訳において構文解析を行うなど、文の構造を利用する手法が多く研究されており、構文情報の重要性が認識され始めている。しかしこれらの手法は、そのベースとして文の構造を一切利用しない単語アラインメント手法を利用しており、言語構造の違いを正確に扱っているとはいえない。日英間などの言語構造の大きく異なる言語対に対しては構造情報を利用する方が精度のよい翻訳を生成することが可能であると考えられる。このため我々は翻訳対象である言語対の言語的な違いを柔軟に扱い、より精度の高い翻訳を目指すために、言語構造を積極的に利用した用例ベース翻訳の研究を行っている。

そこでは内容語をノードとする木構造として文を表現しており、付属語は内容語に付随する形で扱われる。翻訳は、ノードごとに利用可能な用例を検索し、複数の用例を組み合わせることによって実現される。基本的には内容語が入力と一致していれば用例として利用可能であると判断する。この際、機能語まで一致する用例があれば機能語の翻訳まで正確に行えるが、機能語が一致しない場合や機能語の情報が無い場合にはこれまでは正確に扱えていなかった。高精度な翻訳を目指すときにもちろんこれでは不十分であり、機能語の翻訳も正確に扱う必要がある。

この問題を克服するため、本論文では対訳コーパスから「機能表現の翻訳パターン」を自動学習し、翻訳時に利用する手法を提案する。これにより入力と用例との間で機能語にズレが生じている場合においても、内容語と機能表現パターンとを組み合わせることにより正確な翻訳文を生成することが可能となる。

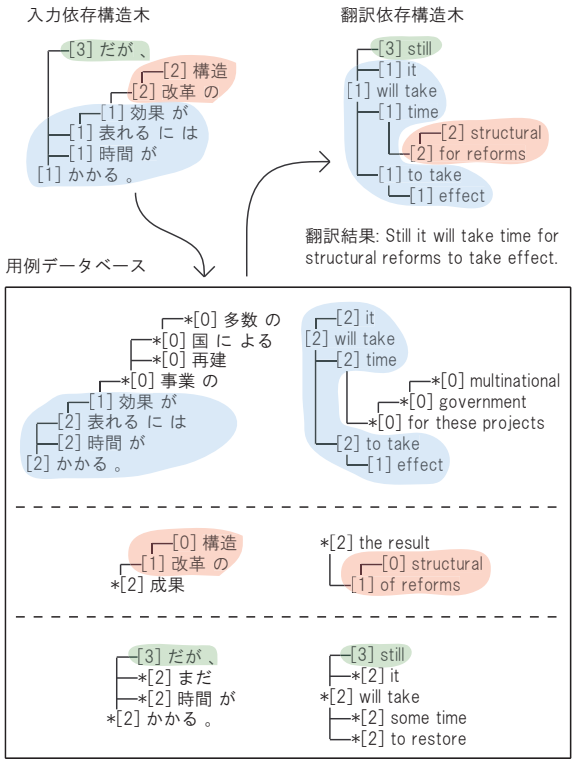

図 1: 用例ベース機械翻訳の例

2　用例ベース機械翻訳の概要

我々が開発している用例ベース機械翻訳システムでは構文解析器を利用して句を正確に扱っており、構造的な言語処理に基づく翻訳を行う。これにより、言語対間の構造的な違いを柔軟に吸収し、高精度な翻訳を実現することができる。ここで言う”句”とは、1 つの内容語と 0 個以上の機能語の組を指す。我々はこの句を単位として、用例のデータベースの構築及び翻訳を行う。

翻訳の流れは、まず入力文を依存構造木に変換し、あらゆる連続な部分木について利用できる用例を検索

する。利用可能かどうかの判断の基本は、内容語が一致しているかどうかである。次に検索された用例の中から、入力文をちょうどカバーできるような用例の組み合わせを探索し、実際に用いる用例を決定する。最後にそれらを組み合わせることにより、翻訳文を生成する。図 1 に翻訳の例を示す。図 1 では入力の依存構造木を翻訳するために、3 つの用例を組み合わせて出力の依存構造木を作りだし、最終的な翻訳文を生成している。

用例を組み合わせる際には、まず入力構造木のルートノードをカバーする用例をベースとし (図 1 の一つめの用例)、そこに他の用例を貼り付けていく。用例を貼り付ける際には、貼り付ける場所や貼り付ける方向が問題となるが、ここで「のりしろ」情報を利用する。のりしろは実際に翻訳で使う用例の部分木の外側の部分木 (図で*が付与されたノード) のことであり、のりしろの部分に他の用例を貼り付けることにより、出力構造木を生成する。のりしろ情報を利用することにより、文の構造や語順などを自然と表現することができる。

3 機能表現パターン

3.1 機能表現の翻訳

機能語の部分の翻訳を行う際には、次の 3 つの場合が考えられる。

1. 親の用例ののりしろ情報が利用できる
2. 子の用例ののりしろ情報が利用できる
3. のりしろ情報がない

図 2 に “方法 として 検討 した。” を翻訳する場合を例として挙げる。親の用例にのりしろがあり、のりしろの機能語が入力と一致している場合は、のりしろの機能語を残して子の用例を組み合わせることで、正しい訳が得られる。子の用例にのりしろがあり、子の用例の機能語が入力と一致している場合は、子の用例の係り受け情報を利用して、用例をそのまま組み合わせればよい。

問題となるのは親にも子にものりしろがなく、機能語の翻訳知識が得られない場合である。このときは機能語の翻訳だけではなく、子の用例を親の前に組み合わせるのか、後に組み合わせるのかも判断できない。ここで機能表現の翻訳パターンを利用することを考える。“X に” という入力に対しては、“to X” という訳を後ろから組み合わせればよいという知識があれば、2 つの用例と機能表現パターンの 3 つを組み合わせることにより、所望の訳を生成することができる。

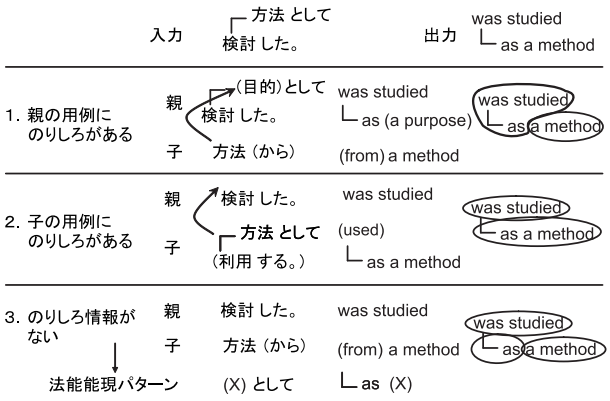

図 2: 機能表現の翻訳 1

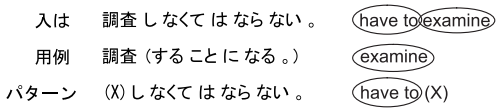

図 3: 機能表現の翻訳 2

また文のルートとなる句の機能語が一致していない場合も、機能表現パターンの利用が可能である。図 3 の例では、“調査 しなくてはならない。” を翻訳する際に、“調査 することになる。 ↔ examine” という機能表現が不一致な用例と、“X しなくてはならない ↔ must X” という機能表現パターンを組み合わせることにより、“must examine” という正しい訳を生成している。

3.2 機能表現パターンの学習

機能表現パターンは、対訳辞書などにより内容語同士の対応関係が得られた句のペアから行う。例えばアラインメントの結果、図 4 のような対応関係が得られたとする。同じ番号が振られている基本句同士が対応しており、さらに下線部は単語レベルでの対応を表している。ここでまず、両言語におけるルートノード ([2] のノード) に注目する。ルートノードの中で、“回復” と “regain” はどちらも内容語だということはすでにわかっており、かつそれらが対応していることがアラインメントによりわかっている。すると、残りの機能語の部分 (破線部) を翻訳パターンとして学習することができる。つまり、この機能表現パターンを利用することにより、“X なくて は なら ない 。” という文は、“must X” と訳せばよいことが学習できるのである。

内容語をすべて X として一般化してしまうと、X が名詞なのか動詞なのか区別がつかなくなるため、“X/

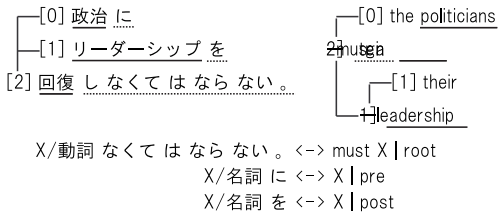

図 4: 機能表現パターンの学習例

動詞 なくて は なら ない 。↔ must X“というように品詞情報込みのパターンとして学習する。さらに同時に係る方向も学習する。この場合はルートノードにおける学習なので、係り先はない。そのため、最終的には “X/動詞 なくて は なら ない 。↔ must X|root“というパターンとなる。

同様にして、[0] の対応からは “X/名詞 に ↔ X|pre”、[1] の対応からは “X/名詞 を ↔ X|post” というパターンが学習される。“pre” と “post” はそれぞれ前からと後ろから係ることを表す。

この方法により、JST 日英抄録コーパス (100 万文対) から学習された機能表現パターンの例を図 5 に示す。JST 日英抄録コーパスは、科学技術振興機構所有の約 200 万件の日英抄録から、内山・井佐原の方法 [5] により、情報通信研究機構が作成したものである。図 5 に示すように、ある日本語の表現に対する英訳の候補は複数得られる場合がほとんどであるが、ここではそれぞれの表現に対して最も頻度の高い英訳を代表として利用する。図 4 から学習された “X/名詞 に ↔ X|pre” というパターンは実際には特殊なケースであり、このパターンを他の文に適用してしまうと、誤りとなる可能性が高い。しかし頻度の高いものを利用することによって、このような特殊なパターンを除外することが可能である。

このようにして学習された翻訳パターンを翻訳に利用することによって、用例の機能語にズレがあり、のりしろ情報が利用できない場合や、文末の機能表現が一致していない用例を用いる場合でも、正しい訳を生成することができる。

4 実験と考察

機能表現パターンの利用による翻訳精度向上を調べるための実験を行った。JST 日英抄録コーパス (100 万文対) [5] からランダムに抽出した 500 文をテストデータとし、残りの文から長すぎるものを除いた 96.6 万文対をトレーニングデータとして、用例の学習および機能表現パターンの学習に利用した。翻訳精度の評

日本語表現	英語表現	頻度
X/動詞 なくてはならない。	**must X|root**	6
X/名詞 に	**to X|post**	7987
	in X|post	7833
	X|pre	6720
X/名詞 を	**X|post**	60352
	X|pre	41705
	of X|post	11588
X/名詞 は	**X|pre**	36463
	X|post	2716
	of X|post	2105
X/動詞 ためには	**in order to X|post**	115
	to X|post	103
	to X|pre	95
X/名詞 より	**than X|post**	322
	from X|post	256
	of X|post	212

図 5: 学習された機能表現パターンの例

表 1: 機能表現パターンの有無による翻訳精度の変化

		BLEU4
テストデータ全て (500 文)	パターンなし	18.43
	パターンあり	**19.17**
機能表現パターンが適用された文 (340 文)	パターンなし	17.16
	パターンあり	**18.15**
翻訳結果が変化した文 (296 文)	パターンなし	16.52
	パターンあり	**17.65**

価には、1 リファレンスでの BLEU4 を用いた。表 1 に実験結果を示す。テストデータ 500 文のうち機能表現パターンが適用された文が 335 文あり、さらにそのうち翻訳結果が変化した文が 287 文あった。表 1 にはこれら 340 文と 296 文のみでの精度比較結果も示した。なおこの 296 文中で、機能表現パターンを利用することにより BLEU 値が向上したものは 174 文あり、逆に低下したものは 77 文であった。

表 1 を見ると、機能表現パターンを用いることによって翻訳精度の向上につながったことが明らかである。またスコアが低下した文の半分以上は、語順の変化によってたまたまリファレンスとの違いが大きくなっただけで、誤差程度のものであった。

表 2 は実際の翻訳結果例であり、日本語、機能表現パターンなしの出力、機能表現パターンを用いた出力の順で、それぞれの BLEU 値も示した。一つめの例では、“始める” の部分を “begins to” と訳すことに成功している。二つ目の例では” に対する” が”for” と訳され、さらに” については” は前から係るというパターンが適用され、正しい語順が得られている。

三つ目の例では、“に” を “to” と訳しているが、この場合は “by” と訳すのが正解である。もちろん “に” を “by” と訳すパターンの頻度は非常に低く、このままでは正しい翻訳が得られない。この問題を解決する

表 2: 翻訳結果例

入力	電界強度が２１．４ｋＶ／ｍｍを越えると分極反転電流が流れ**始める**。	BLEU
出力 1	Then the inverse current flows when an electric field strength exceeds 21.4 kV / mm .	26.66
出力 2	Then the inverse current **begins to flow** when an electric field strength exceeds 21.4 kV / mm .	**44.82**
入力	高齢者**に対する**セメントスＴＨＡの適応**については、**長期経過を十分に配慮し、適応性を判断する必要があると考えた。	BLEU
出力 1	It was considered deeply consider the long term progress that had to judge the adaptability on the adaptation of the cementless THA the elderly.	19.60
出力 2	On the adaptation of the cementless THA **for** the elderly it was considered deeply consider the long term progress that had to judge the adaptability .	**25.94**
入力	ダイオキシン**に**汚染された環境をいかにして治療するかは，環境科学の最も大切な問題の一つである。	BLEU
出力 1	How treatment for polluted dioxin environment is one of the most important problems of environmental science .	**27.23**
出力 2	How treatment for polluted **to** dioxin environment is one of the most important problems of environmental science .	26.07

ためには、注目している句だけではなく、その係り先の機能語の情報までみることなどが考えられる。

また注目する句の機能語自体に複数の意味がある場合もある。図 5 の最後の例では、“より” に複数の意味があり、“from” や “than”、ときには “above” や “since” といった様々なパターンが得られる。現在は最も頻度の高かったパターンのみを適用しているが、これらの様々なパターンからどれを選択するか、またそのときに必要な情報は何かを模索しなければならない。

5 関連研究

翻訳のパターンを学習する手法の一つとして Chiang[1] による手法がある。Chiang は一般的な単語列アラインメントツールである GIZA++の結果から文脈自由文法に似た形式の機能表現パターンを学習する方法を提案し、翻訳精度の向上を達成した。この手法は Phrase base SMT を拡張したものであり、ある程度の文の構造を扱うことが可能である。これにさらに構文解析結果を利用する研究として Quirk ら [4] や Galley らの手法 [2] があり、構文解析を利用するため Chiang の手法よりも構造的に意味のある機能表現パターンの学習を行うことができる。

このような発展的な翻訳知識学習の手法は、翻訳においてある程度の文の構造を用いることにつながるが、そのベースとなるアラインメント手法である IBM モデルは、文の構造情報は一切用いていない。このように単語列として文を扱う手法は、英語とヨーロッパ言語など言語構造に大きな違いがない言語対では精度よいアラインメント結果が得られるが、日英などのように言語構造が大きく異なる言語対に対しては不十分である。一方で我々が提案する手法はアラインメントにおいても各言語での文の構造を利用しており、そこから学習された機能表現パターンは言語構造の違いを吸収し、より正確で滑らかな翻訳を生成するために非常に有効であるといえる。

6 結論

本論文では機能表現パターンの自動学習及びこれを利用した翻訳手法を提案し、翻訳実験において提案手法の有効性を示し、翻訳精度の向上を達成した。今後の課題として、学習された機能表現パターンが複数ある場合にいかに正しいパターンを選択するかを考える必要がある。現在のパターンでは元の単語を品詞情報のみに汎化して学習しているが、この汎化がいささか強引であると考えられ、どの程度汎化すればよいかを深く考えることが重要である。今回は日英翻訳における適用例を示したが、逆方向での実験や他の言語対における実験などを行い、提案手法が言語対によらずロバストであることを示す必要がある。

参考文献

[1] David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, 2005.

[2] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Coling and 44th Annual Meeting of the ACL*, pages 961–968, 2006.

[3] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL 2003: Main Proceedings*, pages 127–133, 2003.

[4] Chris Quirk, Arul Menezes, and Colin Cherry. Dependency treelet translation: Syntactically informed phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 271–279, 2005.

[5] Masao Utiyama and Hitoshi Isahara. A japanese-english patent parallel corpus. In *MT summit XI*, pages 475–482, 2007.