対訳コーパスからの機能表現翻訳パターンの自動学習

京都大学大学院情報学研究科教授
黒橋　禎夫

PROFILE　1994年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006年4月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

kuro@i.kyoto-u.ac.jp　075-753-5344

京都大学大学院情報学研究科
中澤　敏明

PROFILE　2006年東京大学大学院情報理工学系研究科電子情報学専攻修士課程修了。現在京都大学大学院情報学研究科博士後期課程在学。機械翻訳の研究に従事。

1 はじめに

機械翻訳において翻訳文の滑らかさや文法的な正確さなどを向上し、翻訳をより高精度にするために必要な情報は、多くの場合、態や時制、主語と目的語の関係、モダリティなど、内容語ではなく機能語に含まれている。つまり、翻訳の精度向上を実現するためには、機能語の情報を十分に活用する必要があると言える。しかしながら近年の機械翻訳手法の主流である統計翻訳手法は、言語情報を利用しないため内容語と機能語の区別をしておらず、このため機能語の翻訳に弱いという問題がある。特に日本語は機能表現が多様であり、この問題の翻訳制度の影響は無視できないほどであるといえる。例えば「・・・流れる。」という文と「・・・流れ始める。」という文をそれぞれ統計翻訳システムで翻訳してみると、どちらも“flows”と訳され、正しい訳である“begins to flow”といった訳は出てこない。

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、本稿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。

2 対訳文アライメント

我々の手法では対訳文アライメント（用例の獲得）から翻訳に至るまで、一貫して文を依存構造木の形で扱っており、これにより言語構造の違いを柔軟に吸収することができる。機能表現パターンはアライメントされた対訳文から学習する。アライメントの例を【図1】に示す。

日本語文の解析には、日本語の形態素解析システム

JUMAN、依存構造解析システムKNPを用いる。これらは日本語文の構造を非常に高精度に解析することができ、新聞ドメインでは、形態素解析が99%、構文解析が90%の精度である。またこれらのシステムは、新聞以外のドメインの文に対しても、十分な精度での解析が可能である。日本語の依存構造の単位（ノード）は、各自立語が1ノードとなるもので、助詞、接辞、助動詞などは自立語のノードにまとめる。英文については、Charniakパーサを用いて句構造に変換し、そこからheadを定義するルールによって依存構造に変換する。英語の依存構造の単位（ノード）は、日本語と同じく、各自立語が1ノードとなるもので、前置詞や助動詞は自立語のノードにまとめた。また木構造のルートノード（文全体のヘッドノード）は最も左側に、各文節は上から下に順番に並んでいる。

日英間の語、語列の対応候補探索には、対訳辞書、Transliteration、数字のマッチングなどいくつかの手がかりを利用する。得られた対応候補の中には曖昧な対応や不適切な対応が含まれている可能性があるが、木構造全体が整合的に対応づくように考慮しつつ、対応候補の中から適切なもののみを選択する。

3 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1】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、さらに実線の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/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

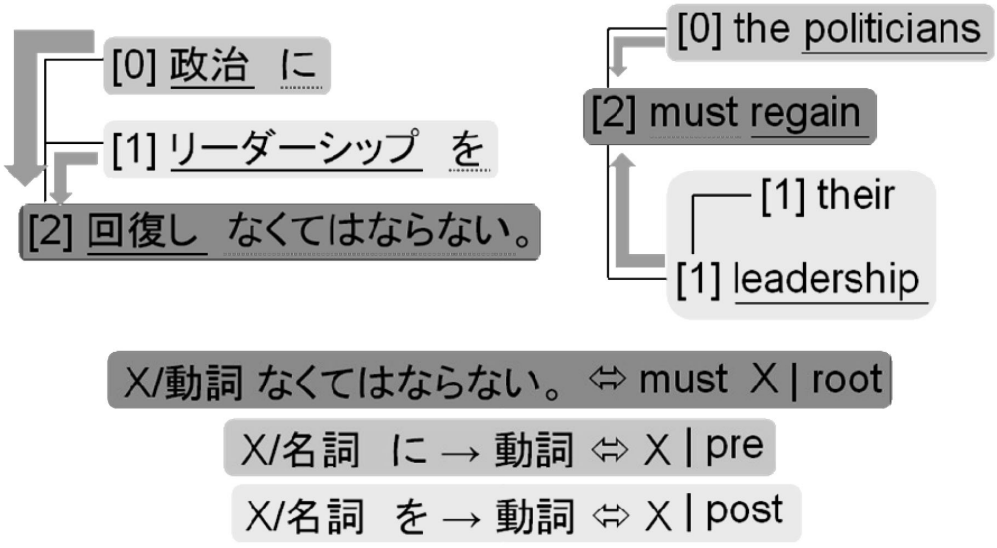

図1 対訳文アライメントおよび機能表現パターンの学習例

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。

さらに、原言語側については係り先の情報も利用する。これは係り先の違いによって翻訳パターンが変わることが考えられるからである。係り先の情報としては、句のタイプのみを利用する。日本語の句のタイプとしては「体言」「動詞」「形容詞」「判定詞」の4つを考える。これにより、最終的には[0]の対応からは“X/名詞 に → 動詞 ⇔ X | pre”、[1]の対応からは“X/名詞 を → 動詞 ⇔ X | post”というパターンが得られる。

4 実験

AAMTから提供された特許日英対訳文約4万対訳文を用い、機能表現パターンの学習を行なった。学習された機能表現パターンの例を【表1】および【表2】に示す。

日本語表現	英語表現	頻度
X/動詞 こと が できる 。	can be X | root	154
	can X | root	103
	to X | root	26
X/動詞 こと に よって → 動詞	by X | post	54
	by X | pre	12
	X | pre	3
X/名詞 は → 動詞	X | pre	1384
	this X | pre	52
	X | post	32
X/名詞 を → 動詞	X | post	18806
	X | pre	3119
	of X | post	552
X/名詞 に → 動詞	to X | post	2117
	in X | post	808
	X | post	494

図2 学習された機能表現パターン

日本語表現	英語表現	頻度
X/動詞 → 体言	X | post	2192
	X | pre	675
	for X | post	631
X/動詞 → 判定詞	for X | post	25
	X | post	20
	X | pre	7
X/動詞 → 動詞	X | pre	1596
	by X | post	847
	is X | pre	458
X/動詞 → 形容詞	by X | post	4
	X | pre	2

図3 係り先の違いによる機能表現パターンの変化

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、頻度の低いものは特殊な場合か、アライメントの誤りから来るものである可能性が高いため、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。例えば【図1】から学習された“X/名詞 に → 動詞 ⇔ 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。

【表1】より、日本語の機能表現の違いにより適切な英語表現（補助動詞など）が学習できていることがわかる。また日本語の「X/名詞 は」は主語になることが多いので英語では動詞の前から（pre）かかり、「X/名詞 を」や「X/名詞 に」は目的語になることが多いため英語では動詞の後から（post）かかるといった情報が学習できていることもわかる。【表2】は係り先の違いにより翻訳パターンがどのように変化するかを示したものである。

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。

5 まとめと今後の課題

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。またかかり先の情報などを用いてパターンをさらに細分化することも考えられる。

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] 中澤敏明, 黒橋禎夫：“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”, 言語処理学会 第14回年次大会 pp.37-40 （2008.3）

[2] Toshiaki Nakazawa, Kun Yu, Sadao Kurohashi：“Structural Phrase Alignment Based on Consistency Criteria”, In Proceedings of Machine Translation Summit XI, pp.337-344, Copenhagen, Denmark （2007.9）