ハイブリッド翻訳のための フレーズアラインメント

株式会社富士通研究所
ソフトウェア&ソリューション研究所
主管研究員
潮田　明

PROFILE 1983年東京大学理学部物理学科卒業。同年株式会社富士通研究所入社．表面磁気光学効果、空間光変調器、統計自然言語処理の研究などに従事．博士（理学）

ushioda@jp.fujitsu.com

1 はじめに

日英・英日機械翻訳のように言語構造の大きく異なる2言語間の自動翻訳においては、従来より構文解析を含むある程度深い言語的解析が重要だと考えられ、実用システムは主にルールベース翻訳を中心に開発が行われて来た。一方仏英間などヨーロッパ言語間での適用から始まった統計翻訳（Statistical Machine Translation, 以下SMT）は、近年著しい発展をとげ、最近ではアラビア語英語間や中英間などの翻訳においても実用化を目指した開発が進められている。SMTでは現在フレーズベースの統計翻訳が最も有望視されているが、フレーズベースの統計翻訳における「フレーズ」とは一般に言語学的フレーズ、あるいはconstituentとは無縁の場合が多い。そのため、翻訳の単位として汎用性に欠ける、対訳コーパスの分野に過適合するため対訳フレーズの分野間移植性が低い、などの本質的問題を抱えている。

本稿では、日英翻訳を念頭に、従来のルールベース翻訳および用例翻訳と、フレーズベースの統計翻訳とを融合するハイブリッド翻訳について展望し、融合に際してのカギとなるハイブリッド翻訳のためのフレーズアラインメント方式について考察する。

2 ハイブリッド翻訳

ルールベース翻訳は汎化が比較的容易で、ルールによる細かい細工が可能であるなどの長所がある一方、辞書や文法には大量の人手コストが必要な上、ユーザによる調整が難しいなどの問題を抱えている。更には、最近多くの言語間で整備が進みつつある大量の対訳コーパスから、直接自動で（そのまま既存の翻訳エンジンに組み込める形で）文法規則を抽出する手立てが現時点では得られていない。用例翻訳はコーパスから自動で翻訳知識が構築できる点で優れているが、不特定分野の翻訳においては、大量の用例収集が必要であり、対訳コーパスを効率的に活用するためには、用例の汎用化を自動で行うメカニズムの開発が不可欠である。統計翻訳は対訳データさえあれば人手による辞書作成やルール作成が不要というメリットがある反面、学習用対訳データとは全く違う分野や文種の翻訳は苦手であるという欠点もあり、汎用的な翻訳システムとして見たときに従来手法に比べて実際どのくらいの翻訳品質が単独で得られるかはまだ未知な部分が多い。

そこで、少なくとも日英・英日翻訳においては、従来より商用システムとしても幅広く使われてきたルールベース翻訳および用例翻訳と、大量の対訳コーパスから翻訳に有用な情報を定量的に抽出できる統計翻訳の長所を組合わせたハイブリッド翻訳が次世代自動翻訳の有力候補として期待できる。しかし、単にハイブリッドといっ

ても、組合わせ方には様々な形態が考えられる。最も疎な組合わせ方としては、それぞれの翻訳結果からベストと思われる結果を抽出する方法、すなわち投票方式がある。またそれぞれの翻訳システムの中間結果を相互利用する方法も考えられる。たとえば、フレーズベース統計翻訳において、抽出されたフレーズテーブルの中から、パーサの出力と一致するフレーズ（すなわちconstituent）のみを選択、あるいは一致するフレーズを優先して使う方式などが提案され、その有効性についても報告されている。更に踏み込んだ融合型のハイブリッド方式では、統計情報と言語解析情報を結合しながら翻訳を進めて行く方法などが考えられる。

いずれのアプローチにせよ、従来のルールベース翻訳や用例翻訳の資産を最大限に活用しようと考えたときに必ず問題となる重要なポイントは、ルールベース翻訳におけるフレーズとフレーズベース統計翻訳におけるフレーズの間に整合性が全くないことである。前者は構文木におけるconstituentをフレーズの単位として解析を進めるのに対して、後者のフレーズは、言語学的意味とは無縁に、統計的にある意味で有意な単語の連なりをフレーズとして活用している。SMTの枠組みの中でconstituentを用いることの是非についてはここでは深くは論じないが、少なくともSMTの最大の問題の1つである異分野間移植性、すなわち、ある分野の対訳コーパスから学習したSMTを全く異なる分野での翻訳へ適用しようとしたときの適用可能性、を考えた場合、人間の言語知識を基に築かれた汎化性に裏打ちされたconstituentの概念が重要な役割を果たすことは十分考えられる。ましてや、ハイブリッド方式の中で従来のルールベース翻訳の開発過程で築かれた文法規則あるいは文法記述の枠組みや、言い換え可能性を基準に構築された用例翻訳用の用例の蓄積を有効に活用しようと考えたときには、効率よくconstituent　あるいはそれに準拠した対訳フレーズと統計翻訳の枠組みとを組み合わせる手立てを考える必要がある。

3 フレーズアラインメント

ここでは、ハイブリッド翻訳のためのフレーズアラインメント方式の一例として、統計情報と辞書情報を組み合わせながらボトムアップにバイリンガルパージングを進めることにより統計的最適化の枠組みの中に言語学的制約を組み込むことが可能な手法を紹介する。前述のようにconstituent を基準にしたフレーズアラインメントを抽出することが目的であるが、実際の対訳コーパスの中で意味的に過不足なく対応の付く部分（チャンク）を括り出して行ったときに、結果的に得られたチャンクが実際constituentになっているかと言うと、その補償はない。実際には一方の言語においてconstituentであっても、対応するもう一方の言語側のチャンクはconstituentでない、ということは多々ある。従って両言語側ともconstituentであるという制約は強すぎる可能性がある。またもちろんそもそも何がconstituentであるかも、もとになる文法のルールに依存するため、両言語側の文法の相性と言ったものも関係してくる。そこでこの手法では、日本語側のフレーズがconstituentになるべく近づくことを優先するアプローチを選択している。図1に本手法の構成図を示す。入力として対訳文（日本語の文と英語の文のペア）が与えられると、まず構文の解析が行われるが、導入したいと考える言語的解析の深さに応じて、形態素解析あるいは構文解析が施される。これらの解析の後、日英両文は名詞句や句動詞などのベースフレーズの単位に分割され、フレーズマージ処理部に送られる。フレーズマージ処理部では、ある評価指標（総合評価値）に従って、日本語の隣接するフレーズ同士、あるいは、英語の隣接するフレーズ同士を今度は順に繋ぎ合わせて（マージ）、より長いフレーズを組み立てて行く。このときどの日本語のフレーズとどの英語のフレーズが対応しているかは、図2に示すようなマトリックス表現によって常に記録されている。繋ぎ合わせの各段階において、日本語の隣接するフレーズ同士を繋ぐ

のか、あるいは、英語の隣接するフレーズ同士を繋ぐのかは、どちらの操作がより高い総合評価値を生むかによって判断される。図1に示すように、総合評価値は統計評価値と構文構造評価値の組み合わせにより計算される。統計評価値は、日本語のフレーズと英語のフレーズの対応の度合いを、それぞれの構成単語同士の対応（単語同士が互いの訳語として現れる確率）を基に求めたものである。単語同士の翻訳確率は単語ベースの統計翻訳モデルを用いて、大量の対訳文から統計的に求めることができる。構文構造評価値は、繋ぎ合わせの繰り返しによって得られる文の構造が、文法に照らし合わせてどのくらい自然かを基準に数値化されたものである。本手法では、対訳文の構文構造は一切考えずに、フレーズ同士の統計的な対応度のみをもとにフレーズアラインメントを行うこともできるが、その場合は図1において統計評価値をそのまま総合評価値として用いればよい。統計翻訳の枠組みの中で、構文情報を用いた方が良いのか、あるいはどのくらい用いれば良いのかは、まだ未解決の問題であるが、ここで紹介したようなフレーズアラインメントの手法を用いて検証することが可能であると考えられる。

図2に、本手法により日英対訳特許抄録の課題文15万文対から自動抽出されたフレーズアラインメントの結果の1例を示す。マトリックスの列方向に、得られた日本語のフレーズが、行方向に英語のフレーズが並べられている。マトリックス中の数字は日本語フレーズと英語フレーズの対応の度合いを示したものであるが、0より大きい数字は「対応有り」0は「対応なし」を表している。日本語のフレーズと英語のフレーズを比較すると、日本語のフレーズの方がより文法的な単位（constituent）に近いものになっているが、これは、総合評価値の算出において、日本語のフレーズがよりconstituentに近いものになるように設定した結果である。

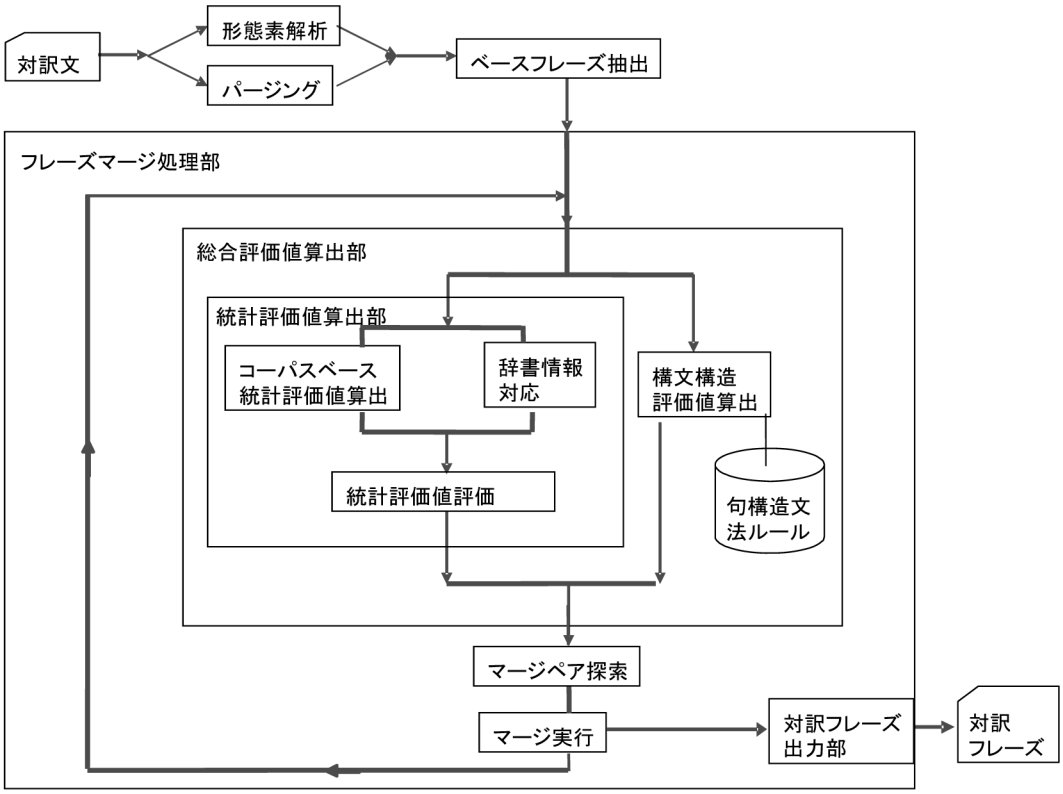

図1　フレーズアラインメントの構成図

なお図2の結果は1例であり、同じ入力文対からこれ以外にも様々な長さのフレーズ対応が抽出される。これらのアラインメント結果をフレーズベース統計翻訳システムに組み込むことにより翻訳精度が向上することが確認されているが、本手法によるフレーズアラインメントの本格的な活用はまだこれからの課題である。

参考資料

Franz-Josef Och and Hermann Ney(2004) "The alignment template approach to statistical machine translation." *Computational Linguistics*, 30(4),pp.417-450.

Kenji Yamada and Kevin Knight(2001) "A syntax-based statistical translation model." *Proceedings of the 39th Annual Meeting of the ACL*, pp.523-530.

入力文：

「ガス不透過性フィルムの一面に、特定物質を含む樹脂層を形成し、その上にガス不透過性フィルムを積層することにより、食品その他のかび発生を防止する包装材料として用い、防かび効果を発揮する」

“To be used as a packaging material for preventing mildew of food or the other and to perform a mildewproofing effect by forming a resin layer containing specific substance on one surface of a gas impermeable film, and laminating a gas impermeable film thereon.”

[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]		
0	0	0	0	0	0	0	0	31	0	0	:	To be used
0	0	0	0	0	0	0	137	0	0	0	:	as a packaging material
0	0	0	0	0	0	350	0	0	0	0	:	for preventing mildew of food or the other
0	0	0	0	0	0	0	0	0	0	1	:	and to perform
0	0	0	0	0	0	0	0	0	80	0	:	a mildewproofing effect
0	0	0	84	0	0	0	0	0	0	0	:	by forming
0	0	428	0	0	0	0	0	0	0	0	:	a resin layer containing specific substance
0	62	0	0	0	0	0	0	0	0	0	:	on one surface
215	0	0	0	0	0	0	0	0	0	0	:	of a gas impermeable film
0	0	0	0	0	88	0	0	0	0	0	:	, and laminating
0	0	0	0	307	0	0	0	0	0	0	:	a gas impermeable film thereon

[0]:ガス 不透過性 フィルム の	of a gas impermeable film
[1]:一面に ,	on one surface
[2]:特定 物質 を含む 樹脂 層 を	a resin layer containing specific substance
[3]:形成し ,	by forming
[4]: その上 に ガス 不透過性 フィルム を	a gas impermeable film thereon
[5]:積層 する ことにより ,	, and laminating
[6]: 食品 その他の かび 発生 を 防止 する	for preventing mildew of food or the other
[7]:包装材料 として	as a packaging material
[8]:用い ,	To be used
[9]:防 かび 効果 を	a mildewproofing effect
[10]:発揮 する .	and to perform

図2　フレーズアラインメント結果の例