Japio

ルールベース翻訳と統計翻訳の融合における syntax の役割

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 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、機械翻訳等の研究に従事。マサチューセッツ工科大学修士、カーネギー・メロン大学博士。2007 年度から特許版産業日本語委員会委員。
ushioda@jp.fujitsu.com TEL

1 はじめに

特許文書は専門用語が多く、文も長くて複雑な構造をしているため読みにくいとよく言われるが、これらは機械翻訳の対象として見た時も同じようにやっかいな問題である。それでも専門用語については、機械翻訳の場合は辞書を充実させることでかなりの程度まで対処可能である。多くの実用機械翻訳システムで今なお取り入れられているルールベース機械翻訳（RBMT）の方式では、長年地道に辞書の拡充が行なわれてきており、今では数百万語の辞書を備えたものも珍しくない。また、近年著しい発展をとげてきた統計翻訳（SMT）、特にフレーズベース統計翻訳（PBSMT）の技術を用いると、対訳コーパスがあれば自動で大量の対訳フレーズが収集できるようになってきている。たとえば数百万文対の対訳特許文書から数千万対～ 1 億対規模の対訳フレーズが自動収集できる。但しここで言う対訳フレーズというのは、従来 RBMT で用いられてきたような言語学的に意味のある単位で切り出されたものではないため、そのままでは RBMT で用いたり、あるいは人間が参照したりすることは難しい。しかし PBSMT の枠組みの中でうまく使えば、専門用語の訳出に関して言えば RBMT よりこなれた訳語が生成できる場合が多い。

一方長くて複雑な文の構成は機械翻訳にとっては本質的かつ最も困難な問題である。RBMT あるいは PBSMT 単独の枠組みで乗り越えることは非常に難しく、両者の融合が必須であると思われる。本稿では、RBMT と SMT を融合する上で鍵となると期待されている syntax の役割について考察してみたい。

2 日英間の語順のギャップ

複雑な文構造の特許文がうまく翻訳できるように RBMT を改良する上でのハードルはルールの拡充である。辞書は人手をかければかけた分だけ拡充できるが、ルールの拡充には、1）システムに精通した専門の技術者でないとルール作成できない、2）分野ごとに個別のルールを人手で作成するには膨大な時間とコストがかかる、3）複数人でかつ長期間にわたってルールを作成した場合ルール間の整合性を保つのが容易ではないなどと言った問題がある。従って、特許翻訳のように分野ごとにまとまった量の対訳コーパスが存在する場合には、自動あるいは半自動でルールが獲得できることが望ましい。その意味で SMT の活用に期待がかかる。

一方で従来の PBSMT の枠組みにおいて、長くて複雑な文を翻訳しようとしたときに遭遇する最も難しい問題は、語順の入れ替えである。PBSMT では、入力文（原文）をフレーズと呼ばれる単位に分割したのち、それぞれのフレーズを訳文側の言語に翻訳し、最後に翻訳されたフレーズを並び替えて訳文を生成するが、日本語－英語間のように文の構造や語順の大きく異なる言語間の翻訳においてはこのフレーズの並び替え (phrase reordering) が非常に難しい。図 1 に、日英間でフレーズの順序が極端に異なる特許文の例を示したが、殆ど順

To obtain an information carrying sheet in which an information pattern is scarcely visually observed by bare eyes by arranging an information pattern formed of infrared absorption ink containing infrared absorption substance represented by the specific structural formula on an upper surface of a substrate having infrared reflectivity.

赤外線反射性を有する基材の上面に，特定の構造式で示される赤外線吸収物質を含有する赤外線吸収インキによって形成した情報パターンを配設することにより，情報パターンが肉眼では目視されにくい情報担持シートを得る。

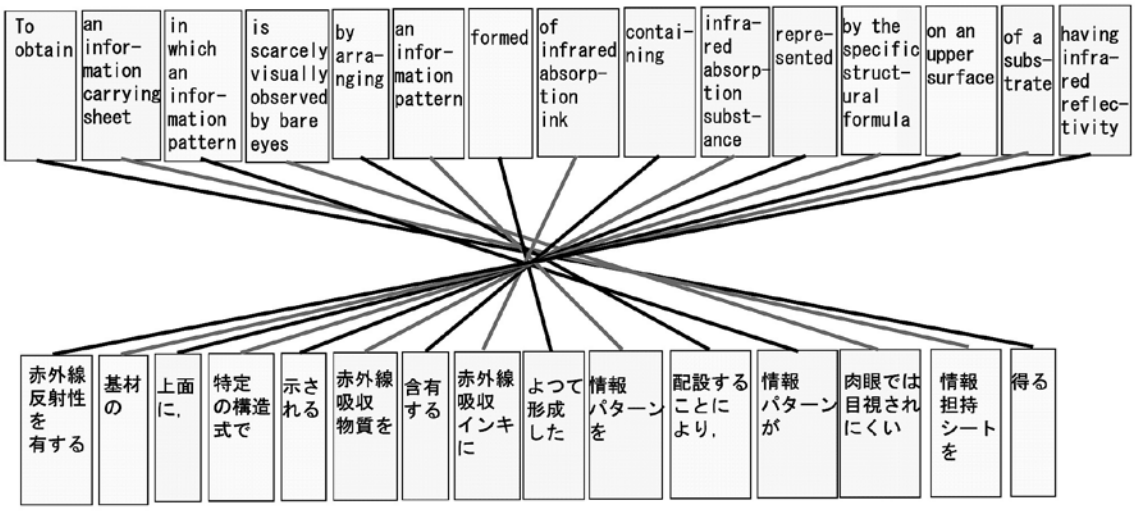

図１　両言語間で語順が極端に異なる例

序がさかさまになっているのが分かる。なおこの図におけるフレーズは一般の PBSMT におけるフレーズではなく、より言語学的に意味のある単位（constituent）に近づけるように切り出されたものである（Ushioda, 2007）。このフレーズ並び替えの困難さが故に、日英特許翻訳においては、局所的には RBMT よりもこなれた訳が作れるにも関わらず、文全体の訳質においては PBSMT はまだ RBMT には及ばないのが現状である。

3 syntax の役割

そこで、近年 RBMT と SMT の長所を組み合わせて新しい翻訳方式を作ろうという試みが盛んになされるようになってきた。まだこれと言った決定打は打ち出されていないが、ここでは RBMT のコアをなす syntax の概念を SMT に導入する手法について考えてみる。なお、SMT における syntax には日常使われる言語学的な意味での syntax の他に、統計的に自動学習される formal syntax も含まれるがここではただ syntax と言った場合には言語学的な意味での syntax を指すものとする。

PBSMT では、フレーズと呼ばれる単語列を単位として翻訳を行うが、このフレーズに相当する部分を木構造で置き換えることで Syntax-based SMT と呼ばれる翻訳モデルが構築できる。原文側のみ木構造を用いる tree-to-string モデル（Quirk, 2005; Huang, 2006）、訳文側のみ木構造を用いる string-to-tree モデル（Yamada and Knight, 2001）や、両言語側木構造を用いる tree-to-tree モデル（Melamed, 2004）などが提唱されている。ここで言う木構造とは厳密には文を構文解析器（パーサ）で解析してできた構文木の部分木を指す。PBSMT におけるフレーズに関しては、原文と訳文の間の単語の対応が保たれるかぎり

任意の単語列のペアがフレーズ対となり得るが、パーサによる木構造を導入したモデルでは、パーサが部分木をなすと認めた単語列（constituent）のみが翻訳の単位と成りうるため、一般に PBSMT よりも制約がきつくなり、その分翻訳文の候補が絞られる。これには一長一短がある。

Syntax の導入により候補が絞られることのメリットとしては

・より文法的に正しい翻訳文が得られる
・日本語－英語間のように文の構造や語順の大きく異なる言語間の翻訳においては、PBSMT では難しかった大局的なフレーズ（あるいは部分木）順序の入れ替えが可能になる
・訳文探索のサーチスペースが狭められるため効率的な探索が可能になる

などが挙げられる。

逆にデメリットとしては、

・Syntax による制約のせいで探索がより複雑になる
・PBSMT ほど柔軟なフレーズや訳文の形成ができない
・パーサにエラーがあると正しい訳文が得られない

などがある。

PBSMT 開発の初期のころには、Syntax による制約の導入によりかえって翻訳精度が落ちたというような報告もあり（Koehn 他，2003）、今でも Syntax による制約には懐疑的な意見も多く聞かれるが、Syntax により実際に翻訳精度が向上するかどうかについては、上に挙げたデメリットをどう克服するか等まだこれからの取り組み如何にかかっていると言える。

しかし現時点で明らかに言えることは、PBSMT の延長線上からは決して「意味」の世界に踏み込むことはできないということである。なぜなら、たとえ正確な翻訳文が得られたとしても、PBSMT からは構文に関する情報は全く得られないからである。文の意味を正確に把握するためには構文を把握することが不可欠である。一方 Syntax-based SMT では正確な翻訳文が得られた段階で、原文側か訳文側かあるいは両方の構文木が得られるため、意味解析への手がかりが得られるという点で実用的にメリットがあり、更に意味解析の結果から構文構造へフィードバックをかける等の新しい改良手法へも道が開かれていると言える。

4 文法的な正しさの意義

最後に、果たして解析や生成の過程における文法的な正しさが、翻訳の正しさに通じるかという問題を考えてみたい。Noam Chomsky はかつて統計的な文法モデルの不備を指摘する目的で以下の２文を例示した。

（１）Colorless green ideas sleep furiously.
（２）Furiously sleep ideas green colorless.

Chomsky の議論は、「（１）も（２）も実際の英語の会話の中には一度も出てきたことはないであろうから、文法的な正しさを計るいかなる統計モデルにおいても共に不適格文の烙印が押されるであろう。しかしそれにも関わらず、いかにばかげた文であろうと（１）は文法的に正しく、（２）は正しくないと判断できる」というものだった。つまり、文法的な正しさを統計モデルで推し量るには限界があるという主張である。それに対して Fernando Pereira は新聞記事のテキストから学習したクラスベースの統計言語モデルを用いて、（１）の文の出現確率は（２）の文の出現確率に比べて２０万倍大きいことを示して見せた。すなわち人間が直感的に文法的に正しいと感じる文は正しくないと感じる文に比べて統計的にも優位に現れるということが示されたわけである。

RBMT にせよ、SMT にせよ、あるいは他のどのような機械翻訳モデルでも、「一度も現れたことのない文や表現を、過去のデータにつき合わせてどのように処理するか」という一般化（generalization）の問題が究極の課題である。RBMT においては、人間の言語能力を駆使することで、文法ルールを通して一般化を実現することができるが、上述のように人手がかかり過ぎる。SMT においては学習データの量をただひたすら増や

し、全く未知のデータに対しても類似のデータから統計的に各種パラメータが推定できるようにすることで、一般化を果たす。しかし、データ量だけで一般化を行うにはまだ世の中のコーパスの量も計算機パワーも十分とは言えないのが現状である。従って現実的な解は、SMTの自動学習の枠組みの中にできる限り人間の持つ言語の一般化知識を埋め込むということになる。この一般化知識のある種凝縮されたものが、syntaxだという考え方がSyntax-based SMTの背景にある。Pereiraの実験で見たように、文法的な正しさが確率の高さに反映されると考えるならば、統計モデルにおいては翻訳文の確率の高さが翻訳の正しさの指標であるから、文法的な正しさを維持した解析や生成、すなわち言語的な意味でのsyntaxをベースにした解析や生成を統計的な枠組みの中で行うことにより、効率良く最適解すなわち極大確率の解に近づけるものと期待できる。

参考資料

Noam Chomsky (1957). Syntactic Structures. The Hague/Paris: Mouton.

Liang Huang, Kevin Knight, and Aravind Joshi (2006). “Statistical syntax-directed translation with extended domain of locality.” In Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA).

Philipp Koehn, Franz Josef Och, Daniel Marcu (2003). "Statistical Phrase-Based Translation" In Proceedings of the Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada.

I. Dan Melamed (2004). "Statistical Machine Translation by Parsing" In Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics (ACL).

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

Fernando Pereira (2000). "Formal grammar and information theory: together again?", Philosophical Transactions of the Royal Society vol.358, no.1769, pp.1239--1253.

Chris Quirk, Arul Menezes, and Colin Cherry (2005). “Dependency treelet translation: syntactically informed phrasal SMT.” In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Akira Ushioda (2007) “Phrase Alignment Based on Bilingual Parsing.” Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation, pp.241-250.

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