統計翻訳における対訳データ不足の問題について

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。

ushioda@jp.fujitsu.com　TEL

1 はじめに

多言語コーパス開発の進展と計算機パワーの向上により、既存の翻訳データを活用した翻訳作業の本格的な効率化に期待が寄せられている。その原動力となっているのが統計翻訳（SMT）、特に近年著しい発展をとげてきたフレーズベース統計翻訳の技術である。すでに翻訳業界に浸透している翻訳メモリに代わって、あるいは翻訳メモリと併用する形で SMT の成果を活用する研究が進められている [1,2]。また、WEB 上の翻訳サイトにおいても、従来のルールベース機械翻訳（RBMT）を用いた翻訳エンジンに代わって SMT ベースのエンジンが徐々に使われ始めている。

確かに SMT の潜在能力は非常に高く、未来の機械翻訳の中に SMT の要素技術の多くが取り込まれている可能性は非常に高い。しかし一方で、世界中の研究者がブレークスルーを目指して SMT の研究を推し進めていくうちに、これまで RBMT 研究者達が何十年もかけて解決策を模索してきた様々な障壁に同じように突き当たるようになってきているのも事実である。すでに様々な問題が指摘されているが、中でも大きな問題として、言語構造が大きく異なる言語間の翻訳において特に重要となる文の構造を SMT の枠組みの中で正しく捉えることの難しさや、動詞、形容詞などの活用や語尾変化を多く伴う言語の文を SMT でいかにして正しく生成するかなどが挙げられる。ここでは、もう 1 つの大きな課題である、データスパースネスの問題について少し考えてみたい。

2 データスパースネスの問題

コーパスベースの自然言語処理においては、大量のテキストの中から課題処理に有益な言語現象や言語統計を抽出し学習を行い、それを基に課題の処理を行うが、抽出する言語現象が精緻な（きめの細かい）ものであればあるほど、その頻度が低くなり、学習データに一度も現れなかった（頻度ゼロの）言語現象をテスト時に扱わなければならなくなるケースが増えてくる。これがデータスパースネス問題である。SMT もコーパスベースの自然言語処理の一種であり、対訳コーパスおよび訳文側言語の単言語コーパスから様々な統計情報を抽出して翻訳を行なう。一般に統計値（統計的推定値）はサンプルの量が多ければ多いほどその確からしさが高まる。SMT においてもコーパスの量が増すほど翻訳の精度が高まるが、人手で翻訳した高品質の対訳コーパスを収集するには量的に自ずと限界があるため、SMT の実運用に際しては対訳コーパスほど翻訳精度向上には寄与しないが量は桁違いに多い単言語コーパスを出来る限り集めて活用しようと言う傾向にある。

しかし、単言語コーパスをいくら集めても、対訳コーパス中に全く存在しない訳語を SMT が作り出すことはできないので、対訳コーパスの量が SMT においては決定的に重要である。また単に量だけの問題ではなく、仮に対訳コーパス中に対象となる（訳したい）単語なり表現が出現していたとしても、全く異なる文脈で使われていた場合かえって誤訳の原因になる場合もある。たとえ

ば大量のスポーツ記事の対訳文で SMT を学習させたとしても、政治記事をうまく訳せるとは限らない。

3 例文にない表現は訳せない

フレーズベース統計翻訳では、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。原文側のフレーズを対象言語に翻訳するのに使われるのがフレーズテーブルと呼ばれる一種のフレーズ辞書で、対訳コーパスから自動で作成される。対訳コーパス中に存在しない訳語を SMT が作り出すことができないのは、訳語や訳語を合成するための部分訳語がフレーズテーブルに存在しないからである。

図 1 に SMT を用いている代表的な翻訳 WEB サイトで翻訳を行なった例を示す。各段には日本語の原文と SMT による英訳文、さらにその訳文の SMT による日本語訳を順に→で繋げて表示してある。

例 1-1 の翻訳は正しい。実際のフレーズテーブルの中身はもちろん分らないが、この文の場合、「皮の赤いジャガイモ」が “red skin potatoes” と訳されている対訳例文と、「～が好きだ」が “I like ～” と訳されている対訳例文が SMT のトレーニング文（対訳コーパス）の中にあれば翻訳可能である。一方、例 1-1 と似た文でありながら例 1-2 の文はおかしな訳になっている。 「ジャガイモ」も「大根」も同じ野菜であるから、例 1-1 が訳せて例 1-2 が全く訳せないのは腑に落ちないと思うかも知れない。例 1-3 を見て分るように「皮の赤い大根」が正しく訳せていない、すなわちフレーズテーブルに存在しないために誤訳を生んでいると考えられる。SMT ではある表現の訳がフレーズテーブルに存在しない場合は、その表現の部分表現の訳同士を繋げるか、あるいはフレーズの別の切り方を探して訳文を生成する。例 1-2 の場合、例 1-3、1-4 を見て分るように「皮の赤い大根」は訳せないが「赤い大根が好きだ」は訳せている（SMT が訳文に対してある程度以上の確率を付与できている）。そこで、まず「赤い大根が好きだ」の訳文 “ I like red radish” を生成し、残った「皮の」の訳語 “skin” を最後にくっ付けた結果例 1-2 の訳文が生成されていると解釈できる。このような訳例文不足による弊害は、英語訳文を更に日本語に訳す「折り返し翻訳」を行なうと一層顕著に現れる。機械翻訳の誤訳と同じような表現を含む文が対訳例文に現れる可能性は極めて低いからである。例 1-2 の折り返し訳「私は赤大根の皮膚のように」ももともとの英語訳文より一層不可思議な訳になっている。

例 2-1 も同様にトレーニング文中に適切な例文がないために生じている誤訳である。例 2-2 が示すように、この SMT のフレーズテーブルには「くれないの」の訳として “me” が登録されている。突拍子もないように思われるが、例 2-3 と合わせて考えると、「くれないの」は「紅の」の意味ではなく、「私に～くれないの？」と言った「呉れない」の意味で登録されていると考えると納得が行く。ちなみに固有名詞の「紅の豚」は正しく訳されている（例 2-4）。

例 1-1	皮の赤いジャガイモが好きだ→ I like red skin potatoes → 私は赤い皮のジャガイモが好き
例 1-2	皮の赤い大根が好きだ→ I like red radish skin → 私は赤大根の皮膚のように
例 1-3	皮の赤い大根→ red radish skin → 赤大根の皮膚
例 1-4	赤い大根が好きだ→ I like red radish → 私は赤大根が好き
例 2-1	くれないの豚→ Pig me →私は豚
例 2-2	くれないの→ me →私に
例 2-3	くれないの？→ For me? → 私にとっては？
例 2-4	紅の豚→ Porco Rosso → 紅の豚

図 1　WEB サイト上の SMT の翻訳例

4 字面のマッチングのみに頼るのには限界がある

前節の例はほんの 1 例だが、SMT には RBMT のような文法規則に基づいて文を解析したり生成したりする枠組みが備わっていないために、未知の表現に対して柔

軟な（あるいはシステマティックな）対応ができない。翻訳できる表現を増やすのに最も有効な手段は対訳コーパスを増やすことである。しかし言語は多様性に富み、表現には余りにもバリエーションがあるために、過去の例文（対訳コーパス）との字面のマッチングのみに頼った学習をしていたのでは、いくら対訳コーパスを集めて来ても限界があるのは明らかである。Callison-Burch [3] らによるスペイン語テキストを用いた実験によれば、1 万語のトレーニング文を用いた場合テスト文中の 90%の単語が未知語となる。10 万語のトレーニング文で未知語が 70%である。テスト文中の未知語を 10%以下にするには約 1000 万語のトレーニング文が必要となる。しかもこれは単語ユニグラム（単語 1 語）の場合の統計であり、たとえばテスト文中のトライグラム（単語 3 語の連なり）について見ると、トレーニング文が 1000 万語の場合、テスト文中の約半分のトライグラムが未知（一度もトレーニング文中に現れなかった表現）になる。

5 言語表現の一般化が鍵

このような未知語の問題に対処するために、いくつかの取り組みが行なわれている。たとえば単語の活用や語尾変化などにより同じ単語がいくつもの違った表層表現を取る言語については、対訳コーパス中の対訳関係と形態素解析情報を用いてこれらの表層表現を 1 つのクラスに纏め上げることで、翻訳精度を落とさずにトレーニングコーパスの分量を減らす試みが行われている [4]。また、対訳コーパスから、対訳関係を利用して予め言い換え表現集（パラフレーズ）を抽出しておき、原文を翻訳できない場合は原文の一部をパラフレーズを用いて置き換えて翻訳することでトレーニングコーパスの分量を減らす研究も行なわれている [3]。これらの手法に共通しているのは、対訳コーパスを用いて同じ意味を持つ単語や表現を 1 つのクラスに纏め上げることで表現の一般化を行い、全く未知の表現を減らしているという点である。対訳コーパスを用いたクラス化は信頼度が高く有効性も確認されている。しかし対訳コーパスに全く現れない表現に対処することはできない。

そこで Marton らは単言語コーパスを用いてフレーズ間の統計的な近さを測定し、対訳コーパスと合わせて近似的パラフレーズを獲得する手法を提案している [5]。対訳コーパスが少量の時には有効であることが示されている。対訳コーパス . に依存しない他の一般化手法としては、単言語コーパスを用いた単語・フレーズの統計的クラスタリング [6] の活用が考えられる。このクラスタリングは、単言語コーパスの中で互いに置換可能な単語やフレーズを統計的に抽出する手法であり、言語モデルの改良や形態素解析の未知語処理などにおいて効果が確認されている。たとえば日本語の特許明細書を用いてクラスタリングを行なうと「被膜」、「保護膜」、「皮膜」、「被覆層」、「保護層」などが同一クラスに分類されたりする。この統計的クラスタリングを SMT の対訳表現の一般化に適用できれば、対訳コーパスに比べて圧倒的に大量に存在する単言語コーパスが活用できる点でメリットが大きい。しかし、対訳データの量を単言語データで補うのは容易でないのも事実である。最後に、その可能性について触れておきたい。たとえば以下の対訳文が日英翻訳用のトレーニングコーパスに含まれていたとする。

1a) 「本手法によりサンプルを傷付けることなく洗浄できる」

1b) "This method enables a cleaning of a sample without damaging it"

ここで、2 つの動詞「傷付ける」と「洗浄できる」はともに「サンプル」を目的語にしていることに注目したい。次に翻訳の入力文として次文が与えられたとする。

2) 「本手法によりサンプルを歪めることなく洗浄できる」

2) において 1a) と異なるのは「歪めることなく」の部分である。

上述のパラフレーズによる原文の書き換えにおいては、書き換えた文と原文の意味が同じであるという仮定の下に原文の代わりに置き換え文をそのまま翻訳して出

力する。単言語コーパスによる統計的クラスタリングにおいては、全く同じ意味のフレーズのみを収集することはできないが、置き換えても文法的に構造が変わらず、かつ意味的にも近くなるような表現を 1 つのクラスに纏め上げることができる。従って、「傷付けることなく」と「歪めることなく」が同一クラスあるいは類似のクラスに分類されていれば、その訳文にも 1b) と同じ文型の文が使えることが分る。たとえば国内出願特許約 33 万件の抄録文の「課題」部分のテキストの集合を用いてフレーズの統計クラスタリングをある条件で行なうとクラスの 1 つとして図 2 のような要素を含むクラスが得られる。フレーズの左にある数字はテキスト内におけるフレーズの出現頻度を表す。クラスの意味は後付けによる解釈になるが、ここでは「何らかの形でダメージを与えることなく」と言った表現が集まっていると考えられる。

次に入力文として 3) を考えてみる。

3)「本手法により肌荒れを気にすることなく洗浄できる」

3) において 1a) と異なる部分は「肌荒れを」と「気にすることなく」の 2 箇所である。文法的な観点で 1a) と 3) を比較すると、3) においては、「洗浄できる」の目的語が「肌荒れ」でない点が大きく異なる。従って 3) の訳文を生成する際に 1b) をベースにして同じ文型の文を作ると

“This method enables a cleaning of a skin damage without minding it”

と言ったおかしな訳文が生成されてしまう。SMT において同様な語並びの文の文法構造の違いを検出するのは難しいが、「傷付けることなく」と「気にすることなく」は同様のコンテクストにおいて使われることがほぼない

85	傷つけることなく	24	狭めることなく
83	悪化させることなく	18	煩わせることなく
8	痛めることなく	1	歪めることなく
63	損ねることなく	1	泡立てることなく
53	妨げることなく	1	沈めることなく
53	傷付けることなく	1	死滅させることなく
52	傷めることなく	1	薄めることなく

図 2　フレーズクラスの一部の例

という単言語に閉じた知見が統計的に得られるならば、3) の翻訳の際に 1a)-1b) の対訳文から得られたフレーズ対訳を使うことを抑止できる可能性は高い。

[参考文献]

[1] Ana Guerberof. 2009. Productivity and quality in mt post-editing. In MT Summit XII -Workshop: Beyond Translation Memories:NewToolsforTranslatorsMT, Ottawa, Ontario, Canada.

[2] Yifan He, Yanjun Ma, Josef van Genabith, and Andy Way. 2010. Bridging TM and SMT with translation recommendation. In ACL 2010, Uppsala, Sweden.

[3] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006.Improved statistical machine translation using paraphrases. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, HLT-NAACL. The Association for Computational Linguistics.

[4] Sonja Niesen , Hermann Ney, Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information, Computational Linguistics, v.30 n.2, p.181-204, June 2004.

[5] Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 381?390, Singapore.

[6] Akira Ushioda. 1996. Hierarchical clustering of words and application to NLP tasks. In Proceedings of the Fourth Workshop on Very Large Corpora, pages 28--41, Copenhagen.