機械翻訳のドメイン適応とカスタマイズの事例

内山将夫　隅田英一郎
情報通信研究機構多言語翻訳研究室

1. はじめに

統計的機械翻訳（SMT）の性能向上は著しい。SMT は、当初は、英語・フランス語などの構造が類似した言語間の機械翻訳のみで実用レベルの翻訳性能であったが、現在では、英語・日本語や中国語・日本語などの言語構造が大きく異なる言語間においても、対訳コーパス量が十分であれば、実用レベルの翻訳が可能となった[1][2]。なお、ここでいう「実用レベル」とは、実際に、それを使っている不特定多数のユーザーがいることとする。

たとえば、情報通信研究機構（NICT）においては、多言語音声翻訳アプリ「VoiceTra」を公開しており、ダウンロード数は、１００万以上である（http://voicetra.nict.go.jp/）。また、特許や汎用の機械翻訳エンジンをお試しいただくためのサイトとして、「みんなの自動翻訳＠TexTra®」（https://mt-auto-minhon-mlt.ucri.jgn-x.jp/）を公開しており、１０００ユーザー以上が利用している。

このように実用化されている SMT であるが、さらなる精度向上のためには課題も多い。本稿では、二つの課題について「みんなの自動翻訳」での取り扱いを述べる。

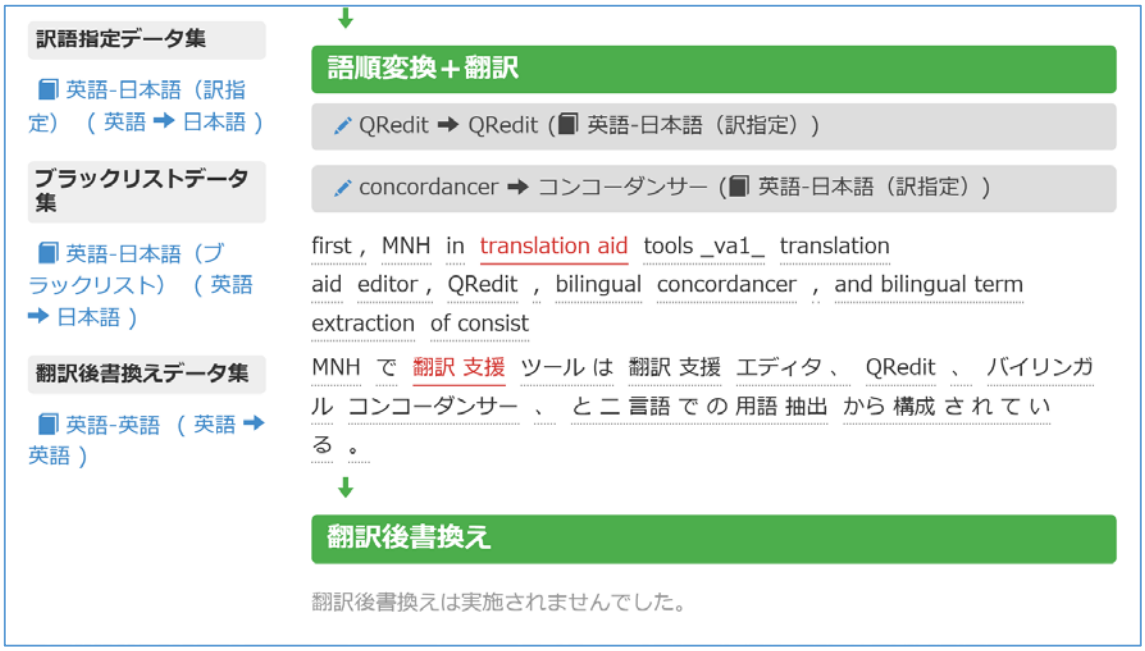
2. 対訳コーパスに起因する誤り等への対処法（カスタマイズ）

課題の一つは、SMT の知識源である対訳コーパスに起因する誤りである。すなわち、対訳コーパス中の対訳文が直訳でなかったり、対訳文自体に湧き出しや削除があったり、実は対訳になっていなかったりするために、機械翻訳が間違うというものである。もちろん、単に、単語アライメントエラー等により、フレーズ対応が間違う事も多い。

たとえば、新聞記事から抽出した対訳文[3]においては、日本語と英語の新聞記事のスタイルの違いにより、対訳文の一部のフレーズが非対応のことが多い。たとえば、日本語の記事では「３１日」というところが、「on Monday」などのようになっている。こうした場合には、誤ったフレーズ対応を学習してしまう。また、対訳文として、たとえば、「おはようございます」と「Good morning, Mr. Utiyama」が登録されているとすれば、単に、「おはようございます」と入力しただけであっても、「Good morning, Mr. Utiyama」と翻訳される可能性がある。さらに、「彼」「彼女」「弊社」「当社」「首相」のような代名詞的なものが、英語においては、固有名詞に訳されていることも多い。

これらのフレーズ対応や対訳文対応自体のノイズについて、あるクラスのノイズを一度

に削除する技術や、それができなくても、発見したノイズを即座に除去したり、修正する技術が必要になる。このような技術の一つとしてカスタマイズがある。



「みんなの自動翻訳」では SMT エンジンのカスタマイズとして以下が可能であり、それらの適用状況が上図のように確認できる。（上図では、訳語指定により、「QRedit」が「QRedit」に翻訳され「concordancer」が「コンコーダンサー」に翻訳されている。）

- SMT を適用する前に原文を書き換える。
- 翻訳メモリ中に原文と一致する対訳があれば、その訳文を出力する。
- SMT で訳語指定をする。
- SMT でブラックリスト（出てはいけないフレーズペア）を指定する。
- SMT の出力結果の訳文を書き換える。

これらを適切に使うことにより、SMT をカスタマイズ可能である。

3. ドメイン適応

別の課題は、ドメイン適応である。ドメイン適応としては、適応に利用できる対訳文が非常に少ない場合（たとえば１－１０００文）と、ある程度大きい場合（１０－２０万文）の２通りを考えている。

対訳文が非常に少ない場合のドメイン適応

ごく少量の対訳文からの適応については、たとえば、文献[4]がある。少量の対訳文から

翻訳モデルを作るときのボトルネックは、単語アライメントの正確性である。文献[4]においては、大規模データからの単語アライメントモデルを適用することにより、少量対訳文においても単語アライメントが正確にできるようにしている。

本稿においては、単語アライメントにおける対訳単語間の確率推定において、当該の対訳データから推定した確率と大規模データから推定した確率を補完した確率を活用することにより、小規模対訳データからであっても単語アライメントが正確にできるようにしている[5]。

単語アライメントができれば、あとは通常の方法で小規模翻訳モデル構築が可能である。この小規模モデルと大規模モデルとを線形補完した翻訳モデルを活用することにより、少量対訳文中の文と似た文については、大規模モデルのみよりも精度よく翻訳が可能になり、それ以外の文については、大規模モデルと同等精度での翻訳が可能になる。

対訳文がある程度大きい場合のドメイン適応

対訳文がある程度大きい場合のドメイン適応には様々な研究がある。たとえば、大規模コーパスからドメインに類似する文を選択したり、確率モデルを適応したり、フレーズテーブルを組み合わせたりである[6][7][8]。

本稿では、pre-ordering に基づく翻訳エンジン[9]を利用して、医療分野からの１０万文規模日英対訳データと２０００万文規模の日英対訳データ（前記１０万文を含む）から翻訳モデル（フレーズテーブル）を作成した。比較した方法としては、それぞれから翻訳モデルを作成して線形補完する方法と、フレーズペアの出現回数をそれぞれに求めた後でそれら回数を合計したフレーズペアから翻訳モデルを作成する方法である。なお、大規模翻訳モデルと小規模翻訳モデルについて、それぞれのフレーズの最大長は、７と１００としている。これにより、小規模モデルに特有の言い回しを優先できると考えた。また、言語モデルは線形補完した。なお、線形補完においては、大規模モデルの重みを０．９、小規模モデルの重みを０．１とした。

下記は、５００文のテストデータに対する BLEU 値[10]である。

	小規模のみ	大規模のみ	線形補完	回数合計
日英翻訳	14.17	26.31	26.60	27.40
英日翻訳	21.56	35.26	34.91	36.70

これからわかるように、フレーズペアの出現回数を合計したものから翻訳モデルを作成する方法の BLEU が一番高かった。理由としては次が考えられる。

- 大規模データにはすでに小規模データが含まれているので、出現回数を合計する方法は、そこにさらに小規模データを足しこむことになるので、重みを調整せずとも、ある程度のドメイン適応効果が得られた。
- 出現回数を合計する方法は、フレーズペアの確率の推定値が、線形補完に比べて適切な可能性がある。

- 線形補完について、補完の重みを加減することにより、さらに性能が向上する可能性がある。

4. おわりに

SMT の翻訳性能は近年向上したが、更に向上するためには、解くべき課題がたくさんある。本稿では、ドメイン適応とカスタマイズの事例を紹介した。

謝辞

本研究の一部は、総務省の情報通信技術の研究開発「グローバルコミュニケーション計画の推進-多言語音声翻訳技術の研究開発及び社会実証- I. 多言語音声翻訳技術の研究開発」の支援を受けた。

5. 参考文献

[1] 内山将夫, 隅田英一郎(2014)「AAMT 長尾賞記念講演」http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/pdf/AAMT2014.pdf

[2] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, Kevin Duh. (2010) *Head Finalization: A Simple Reordering Rule for SOV Languages.* WSMT and MetricsMATR

[3] Masao Utiyama and Hitoshi Isahara. (2003) *Reliable Measures for Aligning Japanese-English News Articles and Sentences*. ACL-2003, pp. 72--79.

[4] Joern Wuebker, Spence Green, and John DeNero. 2015. *Hierarchical Incremental Adaptation for Statistical Machine Translation.* EMNLP.

[5] 特願 2015-174465（単語アライメントモデル構築装置、機械翻訳装置、単語アライメントモデルの生産方法、およびプログラム）

[6] Kevin Duh, Graham Neubig, Katsuhito Sudoh, Hajime Tsukada. 2013. *Adaptation data selection using neural language models: Experiments in machine translation.* ACL.

[7] Rico Sennrich. 2012. *Perplexity minimization for translation model domain adaptation in statistical machine translation.* EACL.

[8] Arianna Bisazza, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2011. *Fill-up versus interpolation methods for phrase-based SMT adaptation.* IWSLT.

[9] Masaru Fuji, A. Fujita, M. Utiyama, E. Sumita, Y. Matsumoto. 2015. *Patent Claim Translation based on Sublanguage-specific Sentence Structure.* MT Summit.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: A method for automatic evaluation of machine translation.* ACL.