英日・中日特許SMTシステムの実用化と課題

Deployment and Obstacles for English-Japanese and Chinese-Japanese Patent Statistical Machine Translation Systems

独立行政法人情報通信研究機構　ユニバーサルコミュニケーション研究所多言語翻訳研究室主任研究員　**内山　将夫**

PROFILE　独立行政法人情報通信研究機構主任研究員。対訳データ構築と機械翻訳の研究開発に従事。

1　はじめに

統計的機械翻訳（SMT）は、1990年前後から研究されているが、英日・中日特許SMTが初めて実用化されたのは、2010年代になってからである[1][2]。SMTシステムの実用化のためには、大規模特許対訳コーパスの構築と、アルゴリズムの発展による精度の良い機械翻訳に加えて、実用に耐える翻訳速度や前処理・後処理による翻訳精度の改善も重要である。本稿では、（独）情報通信研究機構（NICT）における特許SMTの実用化における取組について解説する。

2　大規模特許対訳コーパスの構築

対訳コーパスとは、異なる言語間で同一の意味内容を表す対訳文からなるテキストデータベースのことである。大規模対訳コーパスの構築は、通常は多大なコストがかかるが、特許対訳コーパスについては、これらのコストが通常の対訳コーパス構築に比べて大幅に低い。その理由は、パテントファミリーから対訳コーパスを自動構築することが可能だからである[3]。

その手順は、日英特許対訳コーパスについては、次の3段階である。

（1）日本特許庁に出願された日本語特許と、米国特許庁等に出願された英語特許から、パテントファミリーを抽出する。

（2）パテントファミリーから、動的計画法を用いることにより、最適スコアの対訳文を抽出する。このスコアは、対訳辞書を利用して単語の重なりを求めることにより計算する。

（3）対訳文をスコアでソートして、上位の対訳文のみを対訳コーパスとして活用する。

NICTでは、3000万文規模の日英対訳コーパスを日米同時出願特許から構築し、日英SMTエンジンの作成に活用している。

なお、同じ技術を活用して作成された特許対訳コーパスがNTCIR-7,8,9,10における特許翻訳タスクに活用された[4]。また、本対訳構築技術は、特許庁「H24年度　中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査」における日中特許対訳コーパスの作成に利用された。さらに、NICTと特許庁は、多言語特許文献の高精度自動翻訳の実現に向けて協力を合意しており、その一部として、今後、高品質な特許対訳コーパスの普及を促進していく予定である[8]。

3　語順変換による言語構造の違いの克服

英日・中日SMTにおいては、英語や中国語と日本語の言語構造の違いを克服する必要がある。具体的には、英語や中国語はSVO言語であるが、日本語はSOV言語であるので、この違いを吸収しなければならない。

日本語から英語や中国語へのSMTについては、この構造上の違いはまだ克服されていないと言えるが、英語や中国語から日本語への翻訳については、SMTの前処理として、英語や中国語から日本語へ似た語順に変換す

る語順変換の処理が非常に有効である[5]。なお、最近は、前もっての語順変換を利用せずに、語順変換と訳語選択を同時に実行する手法においても、英日 SMT の翻訳精度が向上してきた[6]。

NICT においては、文献 [5] と同様に、語順変換の後に SMT により訳語選択をするという手法を採用している。ただし、文献 [5] とは異なり、語順変換の規則を自動で獲得することにより、多言語に同様な手法が適応可能である[7]。その方法は、

（1）入力文を構文解析して 2 分木にする
（2）2 分木の各ノードにおいて子供のノードを入れかえるべきノードを同定する

というものである。

上記の方法は単純なものだが、複雑な語順変換もサポート可能である。たとえば、「FIG. 3C is a graph illustrating a simulation that includes the effects of resonance, cyclic clocks, and a change in logic current.」について、上記の手法を適用すると、「FIG. 3C _va1_ resonance of effects, cyclic clocks , and logic current in change _va2_ includes that simulation _va2_ illustrating graph is . 」のように日本語風の語順に変換されるので、それを SMT への入力とすると、「図 3C は、共振による効果、環状のクロック、および論理電流の変化を含むシミュレーションを示すグラフである。」のように機械翻訳される。なお、「_va1_」や「_va2_」は、日本語の助詞に相当するものを語順変換の後で挿入したものである。

4 翻訳精度と速度の問題

長文翻訳の実用化には、対訳コーパスの構築と翻訳精度の向上に加えて、翻訳モデルのサイズと翻訳速度も重要である。

必要なスペックは、運用の形態にもよるが、現状 128GB メモリ程度の計算機であれば比較的簡単にレンタルできることから、その程度のメモリでストレスなく動くような翻訳モデルのサイズとする必要がある。また、平均 23 単語程度の文について、1 文当たり 1 秒未満で翻訳できる必要がある。また、翻訳モデルの構築時間の短縮も、開発のターンアラウンドの観点から重要である。

特許 SMT の場合には、対訳コーパスのサイズが数千万文になるため、これらの問題の解決は困難であるが、現状の SMT の方式の中では、フレーズベースの SMT であれば、これらの問題を解決できることから、NICT ではフレーズベースの SMT を翻訳エンジンとして採用している。

5 前処理と後処理

NICT で開発した現状の翻訳エンジンがサポートしていない言語事象としては、化学式や数式や表などがある。また、請求項などの超長文も十分にはサポートしていない。これらについては、翻訳エンジンの前処理や後処理で対応可能な部分が多いと考えている。

たとえば、入力文中の化学式や数式については、BBB などの特殊記号に置き換えることにより、SMT エンジンに翻訳されることなく出力されるので、それを再びもとの化学式や数式に置き換え可能である。また、請求項などの超長文については、前処理で節などに分割し、節ごとに翻訳することが有効である。

一般に、SMT エンジンでは、入力を少し修正しただけで正しい翻訳がされることも多いので、適切な前処理は重要である。

6 今後の課題

英日特許翻訳について、2010 年以前の翻訳精度と現在の翻訳精度を比べると、以前は、端的に言って、意味不明な翻訳が多かった。一方、現在は、対訳コーパスが十分にあれば、意味が分かる翻訳文を出力することが可能になった。今後は各モジュールの精度を向上することにより、全体の翻訳精度を向上することが当然必要である。

それ以外にも

- 訳抜けをなくす
- 否定・肯定等のモダリティを正しく訳す

- 訳語を統一する
- ゼロ代名詞の復元

など様々な課題がある。

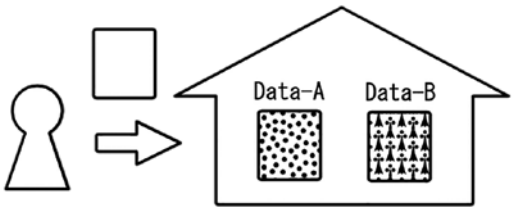
これらの課題を扱うときに注意しないといけないこととしては、それらの課題について個別に対処して特別な手だてを用意した方が良いのか、それとも、翻訳アルゴリズム全体が向上するなかで、いわば自然に解消するのを待つのが良いかというものがある。言い換えれば、疑似問題に時間を割かないように注意しないといけない。

7 みんなの自動翻訳 @TexTra®

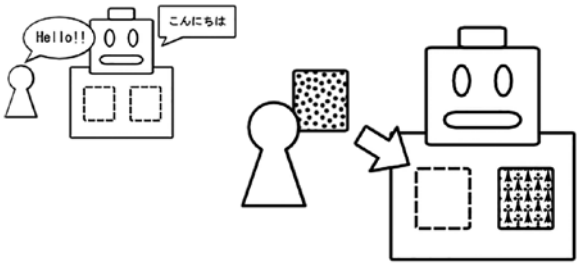
NICT で開発した SMT エンジンは「みんなの自動翻訳 @TexTra®」サイトにて一般公開されている。(https://mt-auto-minhon-mlt.ucri.jgn-x.jp/)

本サイトでは次のようにして NICT の SMT エンジンをカスタマイズできる。

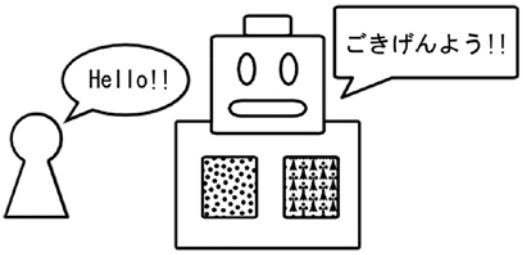
1．自分で用意した用語や対訳データを登録



2．登録したデータを用いて SMT をカスタマイズ



3．作成した SMT エンジンを利用



また、登録した対訳データからオリジナルの SMT エンジンを作成可能である。

本サイトおよび SMT エンジンについての、みなさまのご利用とフィードバックをお待ちしています。

参考文献

[1] 独立行政法人情報通信研究機構、一般財団法人日本特許情報機構.　NICT の高精度な中日自動翻訳ソフトウェアが Japio のサービスに (2013/3/28) http://www.nict.go.jp/press/2013/03/28-1.html

[2] 独立行政法人情報通信研究機構、日本発明資料株式会社.　“英語特許文” の高精度「自動翻訳ソフトウェア」を開発 (2013/3/21) http://www.nict.go.jp/press/2013/03/21-1.html

[3] Masao Utiyama and Hitoshi Isahara. (2007) A Japanese-English Patent Parallel Corpus. MT summit XI, pp. 475--482.

[4] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. Overview of the Patent Translation Task at the NTCIR-7 Workshop. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp.389-400, Dec. 2008.

[5] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head finalization: A simple reordering fule for

SOV languages. In Proc. of the Joint Fifth Workshop on Statistical Machine Translation.

[6] Graham Neubig and Kevin Duh. 2014. On the Elements of an Accurate Tree-to-String Machine Translation. ACL.

[7] 特開 2013-250605:　機械翻訳装置、機械翻訳方法、およびプログラム

[8] 独立行政法人情報通信研究機構、特許庁.「NICT と特許庁が多言語特許文献の高精度自動翻訳の実現に向けて協力合意」(2014/7/28)
http://www.nict.go.jp/press/2014/07/28-1.html