シンポジウム

英語コーパスと機械翻訳

永 田 昌 明
(NTTコミュニケーション科学基礎研究所)

本日はせっかくお招きに預かりながらシンポジウムに参加できず大変に申し訳ありません。実は股関節の病気のせいで歩くのが難しい状態になってしまい、車でなら移動できるのですが、京都から東京まで出張するのはちょっと無理ということで、講演要旨を代読して頂くことになりました。

私が皆様にお話する予定だったのは、

1．大量の対訳コーパスから統計的手法により機械翻訳を実現する統計的機械翻訳

という技術の紹介と

2．統計的機械翻訳は既に実用化のフェーズに入っており、今後の主流となるだろう

という技術動向の紹介です。

統計的機械翻訳の技術内容は、私本人がPowerPointのスライドを使いながら説明しなければ理解して頂けないと思いますので、大量の多言語テキストデータがあれば言語間の翻訳が可能であるということを例で示します。

大量の多言語テキストデータといえば、インターネットです。検索エンジンは強力な翻訳支援ツールです。例えば、英語のメールを書いていて、「～してもらえると有難い」と言いたいときに、“I would be grateful if you could …”だったかなあ“It would be grateful if you could …”だったかなあと迷ったと仮定します。Googleで“I would be grateful if you could”と二重引用符でこの７つの単語列を検索キーとして入力すると、この７つの単語列そのものズバリを含むページが161,000件検索できます。これに対して“It would be grateful if you could”は26,800件なので、結論は、どちらもOKらしいが、“I would be grateful”の方がより適切な表現らしいということが分かります。

すなわち、ある言語（日本語）から別の言語（英語）を翻訳するときに、翻訳先（英語）のコーパスが大量にあればあるほど、正しい翻訳ができる可能性が高まるのです。これが機械翻訳における英語コーパスの利用法です。

次に訳語を探すことが考えてみます。例えば、「股関節」をどう英語に訳すか分からないとします。この場合、「股関節」という日本語を含み、かつ、同じページに何らかの英語の文章を含むページを検索するために、「股関節」と“+of”の二つの単語をキーワードとして入力し、検索結果をじっと眺めます。“+of”を検索キーワードに加えるのは、ofが英語で最も出現頻度が高い単語の一つだからです。ただし、単純に“of”と表記すると、Googleはこれを無視してしまう（「ストップワード」）ので、無視されないために“+of”と入力します。

さて、「股関節+of」というキーワードの検索結果を眺めると、hipという単語が何度も出てくることに気づきます。「あれ？Hipは『おしり』や『腰』の意味じゃないのかな？」といぶかりながら、今度は「股関節」と“hip”の組合せで検索してみると、検索結果の

中に“hip joint”という単語列が大量に出てくるので、「股関節」の訳語は“hip joint”であることが分かります。

すなわち、ある言語（日本語）から別の言語（英語）へ翻訳された大量にテキストがあるとき、互いに翻訳になっていると思われる部分に何度も共起（同時に出現）する単語のペアを探すことにより、対訳辞書を作ることができるのです。また、互いに翻訳になっているテキストのペアが大量にあればあるほど、より正確かつ多くの語彙を持つ対訳辞書を作ることができます。これが機械翻訳における英語と日本語の対訳コーパスの利用法です。

「統計的機械翻訳」は、このような大量のコーパスを利用した対訳辞書の作成や翻訳先の言語の文の生成を、コンピュータの力を借りて大規模かつ系統的に行う手法に関する研究です。統計的機械翻訳の特徴は、アルゴリズムが言語に全く依存しないため、対訳コーパスさえあれば、どんな言語ペアの間の翻訳器でもすぐに作れることです。

互いに翻訳になっているテキストのペア（例えば英語と日本語）を大量に所有している企業の代表例はMicrosoftです。WindowsやOfficeのマニュアルは英語から世界中の言語へ翻訳されています。翻訳先の言語（例えば英語）のテキストを大量に所有している企業の代表例はGoogleです。Googleは世界中の言語のテキストをかき集めて、手元に所有しています。

話は少しそれますが、Googleは、“Web 1T 5-gram”という名前で、インターネット上の英語テキストの連続する5つの単語（5-gram）の出現頻度を6枚のDVDに収めたものを公開しています。これはLDC（Linguistic Data Consortium）というコーパスや辞書を扱うアメリカの非営利団体から購入できます。

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13

また、つい最近（2007.11.1）、Google Japanから「大規模日本語n-gramデータ」も公開されました。これは、Webから抽出した約200億文（約2550億単語）の日本語データから作成したn-gramデータ（1〜7gram）で、「言語資源協会」という日本の特定非営利活動法人から購入できます。

http://googlejapan.blogspot.com/2007/11/n-gram.html

話を統計的機械翻訳に戻すと、実は、統計的機械翻訳を最も熱心に研究している企業もGoogleとMicrosoftです。大規模なコーパスが自社に存在し、コンピュータマニュアルの翻訳やWebページの翻訳といった大きなニーズが自社に存在するのですから当然といえば当然です。

現在、Googleが提供する翻訳サービスは、すべて統計的機械翻訳に基づいています。昨日（2007.11.23）の時点で、アラビア語から英語、中国語から英語、英語から日本語など、15の言語ペア、29方向の翻訳サービスが提供しています。このように非常に多くの言語ペア間の機械翻訳を容易にサポートできるのが、統計的機械翻訳の特徴です。

Googleにおいて、何でもいいから英語のキーワードを入力して、「このページを訳すBETA」というリンクをクリックすると、統計的機械翻訳による翻訳結果を実感することができます。翻訳精度に関する評価は皆様にお任せしますが、一般的には、英語と日本語のような語順が大きく異なる言語ペアではあまり精度がよくありませんが、アラビア語と英語のような比較的語順が近い言語ペアではそれなりに使える翻訳になっていると言われています。

日本では、NTTやATRにおいて統計的機械翻訳の研究が行われています。まもなく市

場に出回るNTTドコモの905iシリーズには「じゃべって翻訳」という音声翻訳ソフト（音声で入力し翻訳結果を文字で表示）がiアプリとしてプリインストールされています。残念ながら「しゃべって翻訳」で使われている翻訳エンジンはATRが開発したものです。ATRは旅行会話に限定した日本語・英語・中国語・韓国語の大規模な対訳コーパスを所有しており、この対訳コーパスがこの音声翻訳ソフトで使われています。これに対して、NTTは、GoogleやMicrosoftと同じようなアプリケーションを想定して、テキストの翻訳を主に研究しています。

以上、（英語）コーパスを機械翻訳に応用する技術である「統計的機械翻訳」の概要、および、この技術がWebページの翻訳や携帯電話での音声翻訳で実用化されつつあるということをご紹介しました。

最後に「英語コーパスと機械翻訳」の今後について少しだけコメントします。工学的な立場では、翻訳先（英語）の言語において、翻訳したい分野や翻訳したい文体のテキストが大量にあればあるほどよいということにつきます。日本英語コミュニケーション学会に参加されている皆様が様々な英語コーパスを作って公開して頂ければ、これほど素晴らしいことはありません。

最近、Research Channelに代表されるような、講義や講演を収録した映像アーカイブがインターネット上で急速に普及しています。やがてこれらの英語の講演に対して英語や日本語の字幕を付与したい、あるいは英語を日本語に通訳したいというニーズが出てくると思われます。

http://www.researchchannel.org/

そのためには録音した音声と、その音声から英語の話し言葉を書き起こしたテキストを大量に用意する必要があるのですが、現状では人手により力づくで作る以外によい方法がありません。このような、音声とその書き起こしがペアになった大量のコーパスを作成することが、今後の課題の一つではないかと思います。