人工知能研究のためのデータ共有

Data Sharing for AI Research

豊橋技術科学大学情報メディア基盤センター　センター長・教授
井佐原　均

通商産業省工業技術院電子技術総合研究所、郵政省通信総合研究所、独立行政法人情報通信研究機構を経て、現職。
産業日本語研究会世話人会代表。

1　はじめに

人間の知的活動の多くは言語によって行われており、AI・機械学習における言語データの重要性は言うまでもない。しかし言葉は人が作り出すものであるため、著作権等の権利の制約があり、研究開発に自由に使うことができ、かつ実用に直結した研究を可能とする大規模データは存在しない。

機械翻訳をはじめとする自然言語処理、さらには広く人工知能の研究開発にはビッグデータが必要である。AI・機械学習を活用したシステムの開発の重要性が高まっており、実社会のテキストを大規模に蓄積し、誰もがシステムの開発と実証を行える環境（プレイグラウンド）を整備することが必要である。産官学の様々なプレーヤーが連携し、この環境を用いてデータの蓄積・共有、システムの開発・実証を行い、協働ユーザの増加と実用化を推進することや、この枠組みを用いて、ビッグデータを用いたAI・機械学習教育による人材育成を行うことは重要な意味を持つ。

このようなデータの活用をとおして、新しいビジネスを作り出すことが可能となろう。誰もが実際のデータでサービスを開発することができ、それをビジネスとして実証できる。大企業や大学等での実用的研究開発から中小企業や学生によるユニークな研究開発まで、幅広い研究開発の実施が可能となる。このような考えのもと、我々はビジネス利用も可能な公開データを蓄積する枠組みの検討を進めている。

ニューラル機械翻訳など最近の研究は大量のデータが必要である。ニューラル機械翻訳は既にかなりの高精度になっているが、誤訳や訳抜けの問題がある。この解消には対象分野へのチューニングが有効と思われるが、チューニング用のデータが存在しない分野も多い。できるだけ多くの分野で研究用の対訳データを収集し、公開する枠組みが必要である。また、情報処理の分野ではアカデミックとビジネスの切り分けが難しいことから、現在検討中の枠組みではアカデミックだけではなく、ビジネスユースも可能なデータを公開することを目標とする。なお、ここでいう公開は無償・無制限ということに限らず、様々な条件での公開を想定している。

企業においては、たとえ公開を目的とした文書のデータであっても、公開を許諾しない場合が多い。しかし自社文書の提供（公開）が、より大きなビジネスチャンスに結び付き、非競争領域で協力することにより、競争領域での個々の競争力の強化に集中できるというメリットを理解してもらう必要があろう。

2　データ収集と公開の試み

我々は以下のような手段で、小規模ではあるが公開用のデータを収集することを目指した。

（1）大学や自治体、企業等が翻訳した対訳データを収集し、公開する。

本学で人手翻訳した大学のウェブページや、賛同してくれた地方自治体の対訳データを公開する。現時点では、地方自治体からは観光分野の対訳データの提供を受けて

いるが、その量は少ない。データの提供を受けた場合にも、冊子による提供の場合は公開データにまで加工するにはコストがかかる。自治体等からの翻訳の発注時に対訳形式のデータの納品も求めることが一般化することが望まれる。実際、本学のウェブページの翻訳では、その条件を付けたがコストは増加しなかった。

(2) 企業から公開可能なデータを提供してもらう。

上記の方法では、集められるデータの量は少ない。企業からは広報用の文書などが大量に翻訳会社に発注されている。企業の内部文書と異なり、プレスリリースなどの広報用の文書は公開されているものであり、そのデータを研究用に公開しても、企業側のデメリットはない。長期的に見れば、その分野の機械翻訳の性能向上につながり、企業側のメリットにつながる。このようなことを理解してもらい、対訳データの提供を受けることを目指している。

(3) ニューラル機械翻訳による対訳データを公開する。

上記の枠組みが実際に動き始めれば、日々のプレスリリースの対訳が蓄積されていき、大規模な対訳データベースとなる。しかしながら、賛同する企業を得るためにはデータの規模が翻訳精度の向上につながり、各企業や社会全体に対する大きなメリットとなることを示して、企業の理解を得る必要があるが、そのメリットを示すに足る対訳データがないことが問題である。このため、日本語テキストについて、所有者の許諾を得、ニューラル機械翻訳を用いて、それを英語に翻訳することで大規模な対訳データを作ることを考えた。

3 ニューラル機械翻訳による対訳データの作成

翻訳会社等の知見では、高性能のニューラル機械翻訳の出力が翻訳プロセスの効率化に役立つことが分かっている。機械翻訳システムの学習データとしてニューラル機械翻訳の出力を用いることが、対訳のない分野への機械翻訳の適用にどの程度有効かを検証し、それを示すことで、データ共有への理解を深めることを目指している。

高性能の実用ニューラル機械翻訳システムの実現のために、大規模な対訳データを用いた学習が行われている。そのようなシステムにおいても、専門用語や固有名詞は苦手で、誤訳を生じることがある。この問題を解決するには既存のニューラル翻訳システムに対し、それぞれの分野のテキストで学習させる、いわゆる転移学習の手法があるが、元のシステムが膨大なデータで学習している状態で、各分野の小規模なテキストを用いて有効な学習を行う手法は未発達である。一方で小規模ではあっても、その分野の対訳テキストだけで学習をさせて、その分野に特化した機械翻訳システムを実現するという考え方がある。

100 万文規模の ASPEC コーパス（他分野テキストとして使用）と 10 万文規模の自動車整備マニュアル（対象分野テキストとして使用）の対訳データを用いた実験では、他分野のテキストで学習したシステムでは高精度の翻訳は実現できなかった。転移学習もあまり効果的ではなかった。文書量は少なくても、同じ分野の文での学習が最も高い bleu 値を示した。10 万文規模の対訳データが準備できる場合は、ニューラル機械翻訳システムを分野チューニングできる可能性が示された。

では、対訳データがほとんど、あるいは全くない分野ではどのような方法が考えられるであろうか。元の文を人手翻訳してコーパスを作ることはコストの面で実用的ではない。流暢度が高いニューラル機械翻訳システムで翻訳し、対訳を作り、それで学習することで対象分野の機械翻訳システムを実現することは可能であろうか。

実験用に使ったデータは、約 14 万文のソフトウェアのマニュアルの日英対訳である。この日本語文を google 翻訳した。

100 万文規模の学習データ（ASPEC 対訳データ）で機械翻訳をトレーニングして、その分野の文を翻訳すると（試行 1）それなりの bleu 値がでる。しかしその翻訳システムで、別の分野の文（ソフトウェアのマニュアル）を翻訳させると性能は極めて低い。これに対して、ソフトウェアのマニュアルの日本語文を Google 翻訳して、できた対訳を学習データとしてトレーニングした機械翻訳システムで、ソフトウェアのマニュアルの文を翻訳すると、試行 1 と同等以上の性能が出る。

これをさらに向上させるためには Google 翻訳の出力から、精度の低そうな文を取り除くことが有効であろう。残った精度の高そうな文だけを使って学習させる。また、精度の低そうな文を見つける手法が実現できれば、

ニューラル機械翻訳の出力を後修正する場合に、後修正の必要性が高そうな文から作業を行うといったことが可能となる。

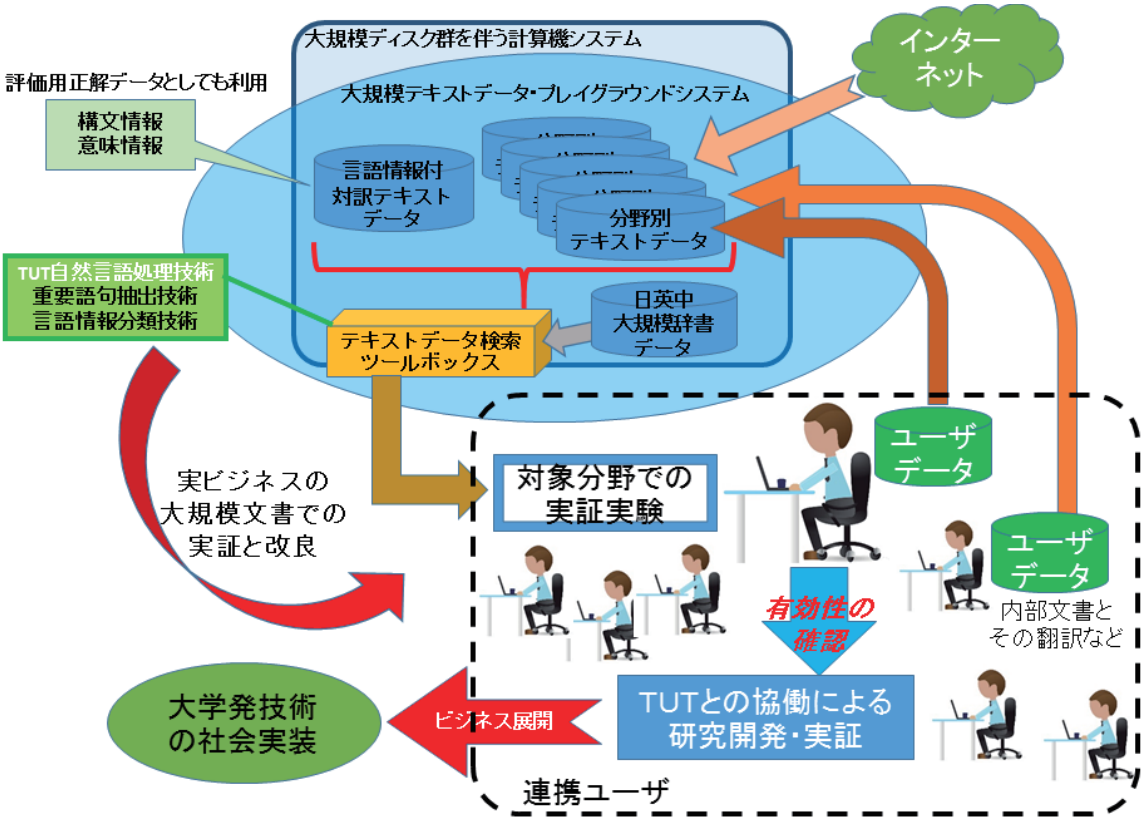
4 大規模テキストデータ・プレイグラウンドシステム

集められたテキストデータを用いて、誰もが研究を行える大規模テキストデータ・プレイグラウンドシステムは AI·機械学習分野において、様々な効果が期待できる。設置された辞書・大規模テキストデータ・ツールを用いて、ユーザは対象分野の精度の高い翻訳などの言語情報を容易に得ることができる。これにより、各ユーザはそれぞれの対象分野において、大規模データと AI・機械学習技術によって得られるメリットを確認し、実用化に進む。

利用できるデータ量（規模）が増え、高品質になることにより、得られるメリットは急激に増加する。ビジネス上は競争相手であるユーザたちが、文書作成や翻訳といった非競争領域では協力（データ共有）することが双方の（さらには我が国の）利益につながる。企業ユーザがプレイグラウンド上での実証で、これを認識することにより、従来クローズドであったデータ（内部文書）のオープン化が進展する。内容の保証のある高品質なデータを連携ユーザに提供することにより、雑多な大量データを利用する研究開発では得られない実用的な言語情報処理システムを実現する。

システムの構成は以下のようなものが想定される。

1）実用に足る規模の大規模テキストデータを迅速に利活用できるようにするため、大量のテキストデータを蓄積し、データを高速に検索し、実証実験を行える大規模ディスク群を伴う計算機システム。
2）実用化したい分野のテキストデータをインターネット上から収集したり、収集したデータを分類したり、データベース上のテキストを解析・翻訳するためには、言語に関する情報（辞書）が必須である。これを実現する日本語辞書・英語辞書・概念辞書等からなる大規模電子化辞書群。
3）ユーザが実際の場面でのテキストを大量に収集し研究開発を行う前に、フィジビリティスタディとして中規模のデータで実験を行うことを可能とするための、構文情報・意味情報等の言語情報（正解情報）



大規模テキストデータ・プレイグラウンドシステム

が付与された日英対訳テキストデータ。

4）計算機システム上で稼働し、大規模辞書を用いて、本システムに蓄積された大規模テキストデータから、ユーザの研究開発の条件に合ったテキスト群を的確に抽出する検索ツールボックス。

本システムは、情報系の研究者に限らず、企業も含めた幅広いユーザに大規模データの利用体験から実ビジネスに向けたシステムの研究開発までの環境を提供するものである。このため、学内利用、共同研究ユーザの利用はもちろん、他大学・企業等のユーザにも研究利用を可能とする。この場合、テキストデータの著作権・利用許諾が課題となるが、自治体からのデータなどについては、既に一般公開が可能との感触を得ている。

5 おわりに

システムを公開し、蓄積された技術やデータをデータごとに定められた条件で、誰もが容易に利用できるようにすることで、自然言語処理に関係する研究開発が急速に拡大・加速することが期待される。特に、これまで言語関連研究の経験のないユーザにおいても、データや技術の有効性を容易に確認することが出来、大規模データ・AI・機械学習のビジネス活用への発想を高めることが可能となる。これにより機械翻訳などの自然言語処理技術とデータを活用したシステムが様々な分野で社会実装される。

興味を持った企業が大規模な開発に乗り出す前にプレイグラウンドにおいて適切な分野及び規模のデータで検証することが可能になることによって、幅広い地域・分野・業種からの研究開発提案が期待できる。作成に高いコストがかかる辞書やデータを連携ユーザが持ち寄る、あるいは共同で作成することにより、個別開発と比べて、大幅なコストダウンが見込まれる。

参考文献

Hitoshi Isahara, Kyoko Kanzaki, Shigeru Nemoto, Akio Yoshida and Kozo Moriguchi, Can Neural Machine Translation System Create Training Data?, The 4th Asia Pacific Corpus Linguistics Conference（APCLC 2018）Takamatsu, Japan, September 17-19, 2018