10年を迎えた産業日本語と自然言語処理技術の課題

10th anniversary of Technical Japanese Association and NLP

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 産業日本語と産業日本語研究会

「産業日本語」とは「産業・技術情報を人に理解しやすく、かつ、コンピュータ（機械）にも処理しやすく表現するための日本語」と定義されている。これにより、言語処理技術を活用することによって、明瞭な日本語文を作成することや高品質な翻訳文を低コストで作成することが可能になる。産業日本語研究会は自然言語処理や言語学の研究者、産業・技術情報を持つ様々な産業、言語サービス·言語ビジネスなど､広く日本語に関わる人々が、このような目標に向けて集い、議論し、研究·開発·普及活動を推進することを目指して設立された。

2007 年に特許版・明晰日本語策定委員会が設置され､日本特許情報機構（JAPIO）による「技術用日本語」､「明晰日本語」という活動が開始された。2008 年には名称を「産業日本語」に統一し、2009 年に産業日本語研究会を発足し、特許版・産業日本語委員会を設置した。2016 年に特許版・産業日本語委員会を産業日本語研究会に統合した。

産業日本語は産業·技術情報を表現するものであるが、コンピュータ処理だけではなく、「人に理解しやすい」ということも重要である。このような観点から、産業日本語研究会のもとにライティング分科会、文書作成支援分科会、特許文書分科会を設置して活動を行っている。また、コンピュータによる特許ライティング支援を目指し、的確な「言い換えルール」の抽出を行い、2013 年に特許ライティングマニュアル（初版）を作成した。その後、内容の見直しを行い、7 つのカテゴリー、27 のルールに再構成し、併せて、例文の追加・修正を行った第 2 版を発行した。

2 データ（テキスト）の重要性

活動を開始した 10 年前と今を比較すると、機械翻訳においては 10 年前には統計翻訳や用例翻訳といったデータに基づく機械翻訳システムが実用化されていたが、翻訳精度は不十分であった。2016 年以降ニューラル機械翻訳が実用化されたことにより、翻訳の精度が大幅に向上した。他の自然言語処理技術応用分野においても、深層学習を利用することにより高性能のサービスが可能になっている。

自然言語処理の手法は規則による手法から統計や機械学習による手法を経て、現在は深層学習によるものが中心となっている。これらは学習に基づく手法である限り、質の高いデータが大量に必要となる。大量データによる深層学習によるシステムは処理の過程が分からないブラックボックス的なシステムであり、正確性の保証には、分野と応用に合った質の高いデータが重要である。

データの重要性にはいくつかの観点がある。一つは量の問題である。機械学習によるシステムではデータの量がその性能に直結する。次は対象の問題である。話し言葉を対象とするのか、書き言葉を対象とするのか。取扱説明書などの手順書か、新聞記事か、小説か。さらにデータの質の問題がある。データの質には元となるデータの信頼性とデータに付与された情報の信頼性の両者がある。

機械翻訳システムであれば、統計翻訳であれ、ニューラル翻訳であれ、学習に用いた対訳データに沿った翻訳を出力する。そのため、システムの翻訳精度は、対訳の品質すなわち元の文を適切に目標言語に翻訳した対訳であるかどうかに依存する。このようなデータの品質とシステムの精度との関わりは、翻訳システムと対訳データとの関係だけではなく、機械学習による自然言語処理システムに共通する問題となる。形態素解析システムの学習データに付与された形態素情報が不適切であれば、学習後の形態素解析システムは不適切な解析結果を出力する。高精度のシステムを実現するためには、適切な学習データを利用することが重要である。

このようにデータの品質の重要性がますます高まっているが、ここには大きな課題がある。機械翻訳システムを例にとれば、人間が作った規則や、統計情報を用いた変換であれば、処理はいくつかのステップに分けられ、途中の段階で人間のチェックが入ることがあり、ある程度の保証はある。しかし、最近の人工知能システムは深層学習という手法を用いており、入力をニューラルネットに入れて、出力を得るということになり、途中の過程はブラックボックス化している。

日本語と英語の対訳データであれば、人手チェックもある程度可能であろうが、他の言語との対訳となるとチェックは困難である。対訳が機械翻訳システムで作られたかどうかすら、わからないことがあろう。内容に抜けがないか、誤訳がないか、適切な表現になっているかなど、十分には確認できない場合があろう。

このようなデータの問題は、データを用いて学習するシステムだけの問題ではなく、データから取り出した結果をもとに議論をする場合にも同様であり、データ自身やそこへの情報付与が信頼できるかどうかは注意しなくてはならない。ある結論をデータから導き出したとして、その推論過程はもちろんであるが、用いたデータが適切であるかも重要である。データの分野、対象、時期、集め方、情報付与のルール、付与の精度などは常に気を付けるべきものであろう。

3 言葉と人工知能

ボードゲームで人間に「勝つ」人工知能システムが出現したことが話題になった。チェスにおいては 1988 年にすでに人間のチャンピオンに勝つコンピュータシステムが開発された。将棋や囲碁は盤面の大きさなどの条件から人間に勝つコンピュータシステムの開発は困難だと言われてきたが、、将棋は 2013 年に、非常に困難であろうと予想されていた囲碁でも 2016 年には人間に勝つシステムが登場した。

コンピュータが人間に勝つということであれば、最初から計算では人間よりも高速で実行できたのであり、機械が人間に勝つということでは自動車は人間よりも走るのが速い。ボードゲームに勝つシステムが話題になったのは、ボードゲームが単純な計算などより、ずっと「知的」な活動であると思われているからであろう。とはいえ、ボードゲームでは、ある時点の状況から次に実行できる手順の数は有限であり、実行結果は確定している。また各状況はコンピュータ上で表現することが可能である。そのような空間の探索がシステム実現の課題であった。一方、言葉の処理においては、同じ言葉が別の文脈では別の意味に解釈される。さらには文脈をコンピュータ上で表現する方法が未確定である。コンピュータは人間の脳よりも演算速度が速く、記憶容量が大きいにもかかわらず、人間よりコミュニケーションが上手な人工知能というのは現時点では想像しがたい。人と人とのコミュニケーションのような柔軟さを実現するのは解決すべき要素が多々存在する。

人間のコミュニケーションの基盤は会話参加者の関係性であり、知識や常識の共有であろう。さらにコミュニケーション時においては、完璧な目的志向の発話だけではコミュニケーションを人間的とは感じづらいと思われる。人間同士の対話では、「ゆらぎ」が発生する。このゆらぎがコミュニケーションを人間的にする重要な要因ではないだろうか。ゆらぎには以下に示すようないくつかの種類が考えられる。

言語的ゆらぎ　：語の省略、単語選択の間違い、言い間違いなどによるゆらぎ
感情・感覚のゆらぎ：感情や感覚からくるコミュニケーションのゆらぎ
対話におけるゆらぎ：膨らみのある会話、冗談、独創的な連想などを許容し、紡いでいく会話

このようなゆらぎをコンピュータシステム上に実現することにより、人と同じように理解し、場合によっては人と同じように誤解するシステムが実現でき、「不気味の谷」を感じさせない対話システムが実現できよう。

産業やビジネスで用いる言葉と、個人が用いる言葉には差がある。ビジネスにおいては、正確な命令や正確な意味伝達が必要とされる。機械に命令する場合やコンピュータから知識を得る場合などで、単一方向のコミュニケーションである。一方、個人の場合は「人間的な」対応や言語表現が重視される。人間とコンピュータの双方向のコミュニケーションが必要となる。

4 おわりに

産業日本語はコンピュータ処理に適するとともに、人間の理解にも寄与する言語であり、ビジネスでの再利用可能性と加工性を持ち、コンピュータシステムの学習データとなりうる正しい言語・情報となることを目指している。従来の自然言語処理技術においては、システムの性能を向上するための言語の制約は、文の長さの制約など、比較的明確であった。ニューラル時代の正しい言語を作るための制限がどのようなものになるかは、今後の課題である。


123456789
Dddddd
Ccccc