自然言語処理における知識と経験

knowledge and experience in natural language processing

長岡技術科学大学准教授
山本　和英

豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士（工学）。1996 年～ 2005 年株式会社国際電気通信基礎技術研究所（ATR）、2002 年～現在まで長岡技術科学大学、現在准教授。自然言語処理の研究に従事。言語処理学会理事、アジア太平洋機械翻訳協会理事。

yamamoto@jnlp.org

1 はじめに

私は 2017 年に「知識を書こう」という原稿を書いた（言語処理学会「自然言語処理」[1]）。本稿ではこの原稿で書いた内容について改めて議論したい。文献 [1] は学術雑誌の巻頭言として書いた雑感でいわばエッセイであるが、私自身は今後の自然言語処理の技術を進めていく上で重要な問題提起のつもりで執筆した。以下では、文献 [1] で紙面の関係で書ききれなかったことについて本稿において補足すると共に、少しでも議論を深めていきたい。

まず文献 [1] で述べた問題提起について、その概要を述べる。当該原稿では、コーパスを情報源とした自然言語処理への懸念を指摘した。近年では多くの自然言語処理研究が何らかのコーパスを入力として何らかの課題（タスク）を解くという問題設定になっている。これに対して、コーパスを用いない自然言語処理研究、すなわち事前作成可能な何らかの情報、規則、パタンなどを構築した上で行う自然言語処理研究が現状ではあまりに少なく、またあまりにそれら研究の扱いが軽視されていることを問題提起した。この上で、学術的にこの状態が今後も続いていくことに危機感を持っていることを述べた。

以上の指摘は、コーパス自然言語処理の一連の研究やその成果を否定することが意図ではない。むしろ、これらの研究成果を活かした上でさらに高度な処理を実現するためには、事前作成した何らかの知識が必要なのではないか、というのが私の主張である。このような観点から現状を見ると、少数の例外を除いて何らかの知識を構築しようといった動きは見られない。何にせよ知識の構築には年月がかかるので、世間が AI ブームと言われている今のうちに、水面下でこのような研究や構築を進めていかないと「次」がないのではないかと、私は大いに危機感を持って文献 [1] を執筆した。そして、その気持ちは今も何ら変わらない。

以下では、知識と経験という二つのキーワードを使って議論を進めていく。なお、本稿で言う「知識」とはコーパスを用いない、あるいはコーパスから人手で加工して作成した何らかの情報を指し、「経験」とはコーパスから自動で獲得した何らかの情報を指す。

2 自然言語処理で用いられる情報

まず、自然言語処理で用いられる情報について分類と検討を行う。一般論として、自然言語処理のある課題（タスク）を解く際には必ず何らかの根拠となる情報を用いて解く。この根拠には大きく、事前作成可能なもの（静的情報）と入力テキストから得られるもの（動的情報）に大別される。また、情報にはどのような単語がどのように並んでいるか、あるいはある単語の周囲にどのような単語が共起しているかといったテキスト情報とそれよりも高度な非テキスト情報とに分類することもできる。以上の 2 軸によって我々が自然言語処理を解く際にどのような情報を用いているかを整理すると表 1 になる。

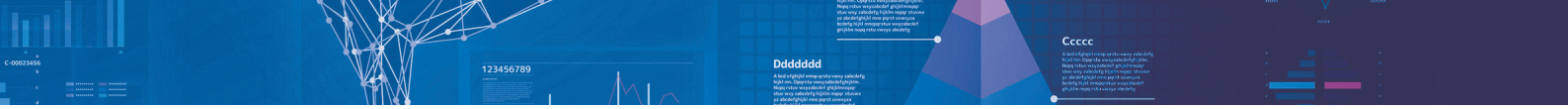


表１　自然言語処理における情報の分類

	静的情報 （事前作成可能）	動的情報 （入力文から獲得）
テキスト情報	経験 （言語モデルなど）	入力文 （文脈情報など）
非テキスト情報	知識（辞書、ヒューリスティックスなど）	解釈／理解

表１に示した情報のうち、当初から自然言語処理として使ってきたのが入力文情報である。入力文に含まれている単語や係り受け関係の情報を根拠としてタスクを解いてきた。この際に、形態素解析や構文解析など様々な解析技術が必要となるためこれら技術が進展してきた。

1990 年代ごろから始まり、現在の主流となっているコーパスを用いた言語処理は、入力文中の情報を用いるだけでなくコーパス（経験）を新たに情報として用いるようになった。これは問題を解くための情報源が入力テキストしかないのと比較すると、タスクを解くための情報源が大幅に増加したという意味で画期的な進展ではないかと思う。経験、すなわち多くのテキスト（コーパス）から得られる言語情報は、言語モデル（初期の頃は n-gram）という形で整理され、統計情報という形でタスクを解く根拠として用いられた。

以上のように、入力文そのものとコーパス（経験）、すなわち（静的及び動的）テキスト情報は自然言語処理における情報源としてごく普通に使われるようになってきた。この一方、表１に示す非テキスト情報の利用に対する検討はかなり遅れている。これまでの研究で使われた静的な非テキスト情報としては、形態素解析における単語辞書、あるいは形態素解析や構文解析における各種ヒューリスティックスが該当する。後者はコーパスが十分に整備されていない 1980 年代には盛んに用いられていた印象があるが、1990 年代のコーパス時代に入ってからはあまり見かけなくなった。考えてみれば、ヒューリスティックスすなわち「経験則」はコーパスを使って各種統計を取れば獲得することができるので、ヒューリスティックスを頭で考える必要がなくなったのは自然なことかもしれない。

この他に非テキスト情報として思い浮かぶものはシソーラス（is-a オントロジー）である。単語の同義関係や包含関係を木構造で表現したシソーラスは、２単語の同義関係や対義関係の認識や、語彙的換言処理の一部で利用される。以前は単語表現の汎化（抽象化）や２単語の類似性判定（類似度計算）にもシソーラスが用いられていたが、近年になって word2vec に代表される分散表現の技術がこれに取って代わられた。この結果現在の自然言語処理でシソーラスが使われるのは、前述した対義語の認識や同義語である表記ゆれの認識など、木構造とは関係のない局所的な語彙知識を使った一部のタスクに限定される。

最後に、本稿の主眼からは外れるが表１の残った部分、すなわち動的に生成される非テキスト情報についても簡単に触れる。この欄に入るべき情報がいわゆる言語解釈、あるいは入力文や入力テキストの「意味」ということになる。しかし、これまで多くの試みがなされているにも関わらず、この欄にどのような情報を入れるべきかはいまだにはっきりしていない。1980 年代ごろに提案されてきた各種の意味表現、対話処理で用いられる言語意図（発話意図）、あるいは近年の doc2vec などによって得られるベクトル形式の文分散表現はいずれもこの候補であるが、いずれも十分なものとは言えない。従って、極論すれば自然言語理解の技術はこの欄に埋めるべき項目を確立することが最終目標なのかもしれない。

3　知識とは何か

次に、知識とは具体的に何であろうかという点について、形態素解析を例に考える。

形態素解析では「外国人参政権」という表現の単語分割問題が時々引用される。通常は、この文字列は「外国人」「参政権」と単語分割すべきで「外国」「人参」「政権」とは分割しない。この根拠として、古典的なヒューリスティックスである最長一致法、あるいは分割数最少法でも説明がつくし、コーパスによる言語モデルで相対的に出現する可能性が高い単語の連接という根拠でも（おそらく）「外国人」「参政権」と分割される。これは多くの場合に正しい判断となるが、逆に言えば「外国」「人参」「政権」と判断される余地がないことを意味する。そこで、仮に「外国」「人参」「政権」と単語分割しなければならない状況があったとすればそれはどのような状況か、あるいはそのように単語分割するためにはどのような知識が事前に必要かを考えてみる [1]。すると、この単語の前後

1　例が悪いため、状況設定にやや無理があることはお許しいただきたい。

の文脈において野菜や果物などの話がされているとか、あるいは「人参」という名前の人がいるとか、こういった場合には「外国」「人参」「政権」と分割しなければならない可能性があることが分かる。これは表１における入力文の解釈の問題ではあるが、これら解釈を得るためには既存の情報源だけでは不十分で、何らかの知識の準備が事前に必要であるように感じる。

以上は形態素解析を例に説明したが、知識は形態素解析だけでなく他の解析でも同様に必要と考える。例えば、構文解析では係り先の曖昧さ解消にコーパスの統計的な情報を用いると、最も自然である（＝最も高頻度に出現する）解釈はおそらく統計的に得ることが可能であるが、少数解釈を統計的情報から得るのは難しい。正確に言えば、統計的な自然言語処理においても複数解釈を出力することは可能であり、少数解釈の把握は可能である。しかし、ある入力文においてこの少数解釈のほうが正しい解釈だと出力するための根拠が存在せず、第２位の解釈が第１位にはいつまでもなり得ない。同様に、意味解析においても、語義曖昧性解消を行うために文脈を用いて最尤の解釈を求めるが、この際に少数解釈が逆転して採用されることは基本的にない。

すなわち、自然言語処理のために必要な知識とは、経験（コーパス）から得られた共起や統計情報からでは判断できないような情報なのではないかと考える。そして、経験から得られた最頻出の解釈を状況によって上書きするために知識が必要になるのではないか。コーパス言語処理の進展によって我々は経験が手に入り、最頻出の解釈を得ることが可能になりつつあるが、状況によっては最頻出ではない解釈をしなければならないことが少なくない。そして、そのような場合に本稿で言うところの知識が必要になるのではないだろうか。

4 コーパスだけでは知識は網羅できない

以上述べてきたように、知識とは予め静的に持っている情報で、経験は自然言語処理の場合コーパスそのものである。しかし、知識と経験は排他的な関係ではなく、お互いに関係または重複があると考えることは自然だと思う。すなわち、多くの経験を積み重ねて、効率的な記憶や処理のためにこれらを抽象化したものが知識なのではないかという考え方である。もしこれが正しいとすれば、本稿で知識と呼んでいるものは実は経験そのものであり、情報源としてはコーパスだけあれば別途知識を構築しなくてもタスクを解くことができるということになる。

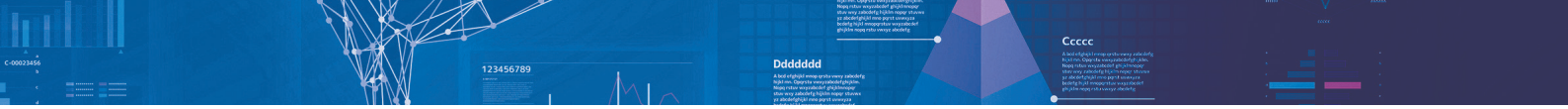
しかし、私は下記の二つの理由から、大量のコーパスから何らかの知識抽出処理を行うだけでは知識の構築は難しいと考える。

理由の一つは、コーパスの過疎性（スパースネス）の問題である。自然言語処理テキストにおいては、単語の出現頻度はジップの法則（Zipf's law）または冪乗則（power law）に従うことが経験的に知られている。この経験則に従えば、どのような分野のコーパスをどの程度集めようとも、そのコーパスにおいて高頻度に出現する単語は極めて少数であり、逆に頻度１の単語は頻度別で最も多い、という性質がある。今行いたいのはコーパスから知識の獲得であるが、簡単のため一定頻度以上の単語の出現そのものが知識だと仮定すれば、コーパスから集めることができるのは中頻度以上の知識であり、その総数は比較的少数であることが予想できる。また、知識の網羅性という観点から、今我々に必要な知識があるコーパスにまんべんなく含まれていることは期待できず、これをコーパスだけで網羅するのは難しい。以上の傾向は、構築しようとする知識の規模が大規模化すればするほど顕著になる。

もう一つの理由は、前述の過疎性とは無関係に、そもそも本質的にコーパスから獲得できない知識があると考えるからである。確かに、いかなる知識の断片もコーパス上に出現する可能性があり、その意味で仮に無限にコーパスを収集そして加工することで知識が構築できるようにも感じる。しかし、この知識の断片を集めて（人間の思うような形で）抽象化あるいは高度化することができるのは人間のみであり、計算機による処理では（そのような教示データを人手構築しない限り）実現できないと私は考える。ただし、この点については議論の余地があり、知識の抽象化も可能であるという考え方もありえると思う。

5 知識を自動で獲得する必要はない

コーパスを情報源として自然言語処理のタスク（課題）を解くという問題設定は、以下の２つの部分問題に分解



することができる。

1．（知識獲得）コーパスからタスクを解くために必要な何らかの知識を（自動で）獲得する
2．（知識適用）獲得した知識を用いて（自動で）課題を解く

ここで言う知識には、n-gram など様々な言語モデルも含まれる。モデルによっては知識獲得と知識適用が内部処理で明確に分離されている場合とこれらを同時に行う場合があるが、いずれにしても、何らかの形でこれら2つの処理をこの順に行っていることに変わりはない。このことから、コーパスによる自然言語処理は知識構築を自動で行っていると言っても差し支えない。

改めて言うまでもないが、自然言語処理（計算言語学）という研究領域は言語に対する様々な処理を計算機によって、すなわち自動で行う研究領域である。本稿で議論している知識の自動構築も理想的には自動で行うことが望ましく、これら研究が進展することが期待されているのは確かである。しかし、私は後述する理由から知識を自動で構築する現在の枠組みと並行して、人手によって知識を構築し、その知識によって高度な課題を解くというアプローチも必要ではないかと考えている。

このように考える最大の理由は、知識自身の検討が足りないためである。人工知能分野における議論も含め、知識とは具体的にどのような情報が必要で、どのように構築し、どのようなデータ形式で利用すべきかについて学術的に合意がないだけでなく、私の知る限り議論もあまり進んでいるように見えない。従って、知識とは何かについて十分に議論されていない現状において、（よく分からない対象を）自動で構築しようというのは時期尚早ではないだろうかと考えている。知識とは何か、どのような知識を構築すると高度な処理が実現できるのかについて、現段階は様々な提案や試みがなされるべき段階ではないかと思う。

また、自然言語処理の課題を知識獲得と知識適用の二つに分類できるとすれば、知識獲得と知識適用の両者に固有の課題が存在すると考えるのが自然である。従って知識獲得とは別に知識適用についても今から検討を進めるべきであり、現状では不十分な知識を用いて知識適用しているだけでは本当の課題が見えてこないのではないかと危惧する。

さらに、産業応用としての観点からも人手による知識構築は必要と考える。学術的には知識構築の自動化というのは興味深い課題であるが、それよりも目の前にある課題を部分的にでも解いて一刻も早く社会に貢献することが肝要である。厳しく言えば、現在のコーパス自然言語処理研究は知識を自動獲得することにこだわってばかりいて、社会の需要に応えるという責務を見失っていないだろうか、とさえ感じる。

知識構築は言葉にすれば簡単だが、実際は一朝一夕にできるような作業ではない。実のところ、何をどうやっていいのかもまだよく分からない。しかし、だからこそ今から取り組まなければいつまでもその先には進めないような気がしている。

6 まとめ

コーパスによる自然言語処理は経験（コーパス）から自動的に知識を構築し、この知識に基づいて各種処理を行うという考え方と捉えることができる。知識は静的情報であることから予め時間をかけて構築することもでき、必ずしも自動で構築する必要はない。コーパスを情報源として知識を自動構築するのは、低頻度の知識になるにつれ収集効率は悪化し、結果として効率が悪い。また、コーパスから本質的に得られない知識も存在するであろう。以上より、人手で知識を構築することが優位とまでは言えないが、研究の多様性の観点からもっと人手で知識を構築して各種タスクを解く研究が増えるべきで、そのような研究が増えることを期待する。

参考文献

[1] 山本 和英．知識を書こう．自然言語処理，Vol.24, No.4, pp.521-522, 言語処理学会（2017.9）