「人の言葉」を計算する研究

藤井　敦
図書館情報メディア研究科助教授

はじめに

私の研究内容を一言で言うと、「人の言葉を計算する」ことです。ここでの「計算する」は、「calculate」ではなく「compute」の意味です。こんな漠然とした言い方をする理由は、既存の分野名で説明すると、「自然言語処理」、「情報検索」、「音声言語処理」という3つの領域に渡ってしまい、一言で言えないからです。逆に、「計算機科学」や「情報学」といった総称的な分野名を使うと具体的なテーマが分かりません。

本稿では、私の研究内容を紹介するために、個別の研究領域について一般的な説明をして、その後で研究事例を紹介します。

自然言語処理

自然言語処理（Natural Language Processing）は人間の言葉を計算機で処理する研究です。プログラミング言語と対比して、人間が使う言葉を「自然言語」と呼びます。すなわち、自然言語処理は、「人の言葉を計算（compute）する」ことです。

しかし、情報検索や音声言語処理は自然言語処理に含めないため、「自然言語処理」というキーワードだけで私の研究内容を説明することはできません。

自然言語処理とは具体的にどのような研究でしょうか。「解析系」と「生成系」に分けて考えると分かりやすくなります。解析系は、人間が作成したテキストを入力として受け取り、著者が伝えたい内容を解読することが目的です。

生成系は、解析系とは入出力が正反対です。すなわち、「伝えたい内容」を形式的に表現した意味情報を受け取って、人間にとって分かりやすいテキストを生成します。よく考えて物を言うことを「言葉を計算する」と言うことがあります。生成系には、この能力が必要です。ただし、この場合の「計算する」は「compute」よりも「calculate」

に近いと思いますし、事実、私の研究内容を一部しか表していません。

解析系と生成系を使う応用例として「機械翻訳」があります。日本語から英語への翻訳であれば、まず解析系で日本語テキストの内容を解読し、形式的な意味表現に変換します。次に、生成系でその意味表現を英語テキストに変換します。

情報検索

情報検索（Information Retrieval）は、蓄積された文書集合から必要な文書だけを探し出す処理です。GoogleやYahoo!のような検索エンジンでお馴染の技術です。ただし、データベースのように定型化された情報ではなく、自然言語で書かれたテキストが対象です。そのため、ここでも「人の言葉を計算する」ことが必要になります。

情報を効率良く検索するためには、索引が必要です。そこで、テキストを解析して、その内容をよく表す索引語を抽出する必要があります。ユーザが入力した検索キーワードで不十分な場合は、関連する語句を自動的に追加して検索する必要があります。検索されたテキストが長い場合には、どこに必要な答えが書かれているのかを解析する必要があります。

音声言語処理

自然言語処理が「書き言葉」を対象とするのに対して、音声言語処理（Spoken Language Processing）は「話し言葉」を対象とします。具体的には、音声認識、音声合成、音声対話などの研究を含みます。音声認識は、人間の発話を音響や言語の観点から分析して、テキストに転記します。音声合成は、音声認識と入出力が全く反対です。音声対話は、人間とシステムが音声で情報をやり取りしながら問題を解決します。私の研究は音声認識が中心です。

人の言葉を計算する技術

自然言語処理、情報検索、音声言語処理は、それぞれが「人の言葉を計算する」研究です。しかし、うまく組み合わせることで、より高度な処理が可能になります。

例えば、自然言語処理の翻訳技術と情報検索を合わせると、外国語で書かれた情報の検索や内容理解が容易になります。

情報検索と音声言語処理を合わせると、録音された音声アーカイブから必要な情報だけを取り出すことができます。携帯電話に声で命令して必要な情報を手に入れることもできるようになります。

インターネットの普及によって、世界中から多種多様な情報が爆発的に発信され続け、また蓄積されています。情報の洪水に

飲み込まれることなく必要な情報を手に入れるためには、「人の言葉を計算する」技術が必要なのです。

具体的な研究テーマ

今までは基礎的な話をしてきたので、以降は、私が研究しているテーマのうち、実用性が高い事例を中心に紹介します。

事典検索システム

World Wide Webが無かった時代には自分で辞書を調べたり、他人に聞いていたようなことでも、最近はWebを調べるようになりました。しかし、既存の検索エンジンでは、知らない言葉の意味を調べたり、関連する言葉を調べるための機能が不十分です。

そこで、事典的な調べ物を目的とした検索サイト「Cyclone」を開発しました。Cycloneを使うと、言葉の意味、関連語、同義語などをWebから効率良く調べることができます。IPAの「未踏ソフトウェア創造事業」で研究開発しました。Cycloneは以下のURLで公開しており、誰でも自由に使うことができます。

http://cyclone.slis.tsukuba.ac.jp/

多言語情報検索（言語横断情報検索）

Webには、外国語でしか書かれていない情報が存在します。論文のような学術情報や特許のような技術情報についても同じです。多言語情報検索は、検索キーワードや検索された文書を機械翻訳して、外国語の運用能力が低いユーザに必要な情報を提供する技術です。言葉の壁を越えて情報を検索するので、「言語横断情報検索」と呼ぶこともあります。

しかし、ユーザが検索したいような新語や専門用語は辞書に載っていないことが多いため、辞書に依存する翻訳手法は有用性が低いです。辞書に載っていない言葉の多くは、外国語の発音をカタカナで表記した、いわゆるカタカナ語です。そこで、辞書にないカタカナ語でも発音に基づいて翻訳する「翻字」という技術を開発しました。この技術は、商用の多言語特許検索サービスで実用化されています。

質問応答

既存の検索エンジンを使うと、入力したキーワードを含む文書の一覧が検索されます。しかし、文書中のどこにどういう答えがあるのかは、ユーザが読んで考えなければいけません。そこで、質問文を入力すると、具体的な答えを返す質問応答システムについて研究しています。

ただし、あらゆる質問に答えるような万能なシステムは将来の話です。現在は、質問内容を5W1Hで分類し、Who（誰）、

Where（どこ）、When（いつ）、What（何）、How（どうやって）の観点を対象に研究しています。例えば、「印刷機を発明した人は誰ですか？」は、「Who」を尋ねる質問です。「Why（なぜ）」に答えることは依然として難しく、今後の研究課題です。

特許情報処理

特許には様々な発明に関する知識や技術が蓄積されています。これらを体系的に活用することができれば、学術研究や産業における価値が高いと考えています。

具体的には、特許検索システムの高精度化、テストコレクション（検索システムのベンチマークテストに用いるデータセット）の構築、特許情報を用いた言語資源の構築について研究を進めています。テストコレクションの構築は、国立情報学研究所が主催する「NTCIR」という国際ワークショップの一環として行っています。言語資源の構築は、NEDOの「産業技術研究助成事業」で行っています。

機械翻訳

「自然言語処理」の所で説明したように、機械翻訳は解析から生成まで行う巨大なシステムです。そこで、システム全体ではなく、機械翻訳に関する要素技術について、「韓国語」、「中国語」、「モンゴル語」といったアジア言語を中心に研究しています。具体的には、以下の研究をしています。

複数の言語で書かれたテキストを大量に集めると、「information」と「情報」のような対訳関係にある言葉の組が現れます。そこで、テキスト集合から対訳辞書を自動的に編集する研究をしています。

モンゴル語にはモンゴル文字表記とキリル文字表記の2種類があります。これらは音声言語としては似ているため、発音情報を用いて一方のモンゴル語からもう一方に翻訳するための研究をしています。

外国語を取り入れるときに、日本語ではカタカナで発音を表記するのに対して、中国語では漢字を用います。しかし、同音意義の漢字が複数あります。そこで、元の外国語と関連する意味を持つ漢字を選択し、機械翻訳する研究をしています。

おわりに

理工系の研究では、学術的な価値と産業上の価値をバランスよく追求することが重要です。私の研究においても、「人間が言葉を自由自在に操る仕組み」を解明するという学術的な探求と、検索や翻訳といった技術シーズを産み出すことの両方が求められています。どちらの精神も失うことなく、これからも研究を続けていきます。

（ふじい　あつし／計算機科学）