平成23年度　国立情報学研究所　市民講座　第7回
「コンピュータで言葉を理解する―言葉の意味を処理するとは？―」

講師：宮尾　祐介
（国立情報学研究所　コンテンツ科学研究系　准教授）

◆　講　義　◆

お忙しいところたくさんの方にお集まりいただき、ありがとうございました。
今日は「コンピュータで言葉を理解する」というタイトルで、われわれが普段使っている言葉をコンピュータが理解するためには何をしなくてはならないかという話をします。
主に言葉が表している意味とはどういうことなのかを考えていきたいと思っています。

・スライド2「自然言語処理」

私は普段、自然言語処理の研究をしています。
自然言語処理とは、皆さんが普段使っている日本語や、私が今使っているような人間の言葉を理解するコンピュータを作ることを目指している学問です。
それができると、われわれが普段行っている言葉による情報交換やコミュニケーションを助けることができます。例えば、応用アプリケーションとして、かな漢字変換はおなじみのものです。
昔からの典型的な自然言語処理のアプリケーションと言えます。
最近では、皆さんも検索や自動翻訳を使ったことがあるかもしれません。
これも自然言語処理の応用の一つです。

また、まだそれほど一般的ではありませんが、対話システムというものもあります。
例えば買い物帰りにおなかがすいたので、「おいしいイタリアン知らない？」と質問します。
そうすると、ロボットが出てきて「いいところを知っていますよ。高級店とカジュアル店のどちらがいいですか？」と聞いてくれます。このような対話をしながら、イタリア料理店を探すというある目的を達成します。これが対話システムと呼ばれるもので、昔から研究が進められています。

・スライド（追加）「自然言語処理」

以下のような例もあります。
パソコンがハードディスクを認識せず起動しなくなってしまったとします。
こうなると大変な騒ぎになりますが、こんなときにも「ハードディスクのケーブルが外れているのかもしれません。では見てみましょう」と言ってくれると、非常に助かります。
これは実話で、今朝、研究所に来てパソコンの電源を入れると、「ハードディスクエラー」と出て全く起動しなくなってしまいました。
今日の発表資料がそのパソコンにしかなく、バックアップも取っていなかったので、頭が真っ白になりどこかに逃亡しようかと思いました。インターネットで一生懸命検索しても、どうしたらいいか分かりません。仕方がないのでサポートセンターに電話すると、持ち込み窓口に行ってくださいと言われ、急遽タクシーで秋葉原に行きました。顔面真っ青な状態で行って、データだけでも何とかしてくれと言い

ました。この場合はコンピュータではありませんでしたが、人間が「ハードディスクのケーブルが外れているのかもしれませんね」と対応してくれて、10 分後に元通りになり、今日はキャンセルすることなく無事発表できるようになりました。

日常生活では、無意識のうちに言葉によるコミュニケーションをしていますが、特にトラブルがあったときは、対話して問題解決することは本当に重要だと思いました。
今日の結論の一つは、今日帰ったら皆さんバックアップを取ってくださいということです（笑）。
このことを非常に痛感したので、講演の内容は半分ぐらい忘れてもいいので、これだけは覚えておいていただきたいと思います。

・スライド 3「自然言語処理の応用」
かな漢字変換、自動翻訳、対話システムも、最近スマートフォンが使われるようになってだんだん現実味を帯びています。このような応用があります。

・スライド 4「自然言語処理の応用」
また、実際に私は生命科学論文の検索システムを開発しました。
生命科学分野では毎日何千件もの論文が発表されるので、自分に関係ある研究を探すだけでも一苦労なのだそうです。それを探してくれる会社もあるくらい大変な分野です。
そこでは検索は非常に重要な技術です。
単にキーワードで検索するだけではなく、書かれている内容について検索したいこともしょっちゅうあります。例えば、MAPK1 という蛋白質が活性化されることについて書かれている論文がほしいときに、その意味に基づいて実際に論文を取ってきてくれるというシステムを開発していました。

・スライド 5「ロボットは東大に入れるか」
最近、国立情報学研究所が中心となって、人工知能技術を総動員して東大入試を突破することを目指すというプロジェクトを行っています。
人工知能にはいろいろな分野がありますが、自然言語処理もそのうちの一つだと考えられます。
特に試験問題はほぼ自然言語で書かれています。
ですから、ここでも自然言語の理解が非常に重要な役割を果たすので、今、研究を進めています。
このようにいろいろな応用があり、自然言語がコンピュータによって理解できるようになると、非常にいろいろなことができる可能性があります。
ただ、まだそれほど対話システムなどが実現しているわけではありません。
今日は主に、そこにある難しさについてお話しします。

・スライド 6「自然言語処理の考え方」
始めに、自然言語処理の問題にどのようにアプローチするかということについてお話しさせていただきます。まず、入力や出力が自然言語であると仮定します。
先ほどの例では、「おいしいイタリアン知らない？」という言葉が自然言語の入力として与えられます。

そして、コンピュータがそれを受け取ります。
コンピュータの中で何か処理をして、「いいところを知っていますよ。高級店とカジュアル店のどちらがいいですか？」という答えが返ってきます。
これを入力・出力といっています。この間に何かが行われています。
人間にとっては当たり前のプロセスなので普段は意識していませんが、このときに何かが頭の中で起きているはずです。それが「理解」であるとわれわれはまず考えます。
そして、理解は何か計算を行っているということがポイントです。
コンピュータは基本的に計算する道具です。
その計算の仕組みが明らかになれば、コンピュータで実行することにより入力・出力を作り出し、自然言語が処理できます。この自然言語の理解の仕組みを知りたいのですが、ここを研究するのは非常に難しいと言えます。
第一に、人間はこれを無意識に行っているので、そもそもどういう仕組みで動いているかを誰も知りません。例えば物理現象や生物現象がどういう仕組みで動いているのかが分からないのと同じように、言葉を理解することもある種の自然現象なので、その法則はまだ誰も知りません。
まずはそれを知らなければなりません。

・スライド7「自然言語処理の難しさ」
もう一つ難しさがあります。理解する仕組みを直接観察することができないことです。
脳みその中で何かをやっているのはまず間違いないのですが、脳みそを開けてみても、何か電気信号が走っているのが見えるだけで、それがどういう仕組みで動いて、最終的に言葉の理解が起きているのかを研究することができないのです。従って、直接脳みそを見ても難しいのです。

では、われわれはどういうアプローチを取るのでしょうか。
入力と出力は観察できます。こういう入力があって、こういう出力があったということは分かるので、その二つを観察し、その間に何が起きているかを推測します。
そして、その計算の仕方をプログラムとして実装して、うまくいっているか、うまくいっていないかを判断します。
このようなアプローチを取っているという前提で、言葉の理解はどういうことか、少し考えてみます。

・スライド8「言葉の理解とは？」
こういう出力と入力があったときに何が起きているかをもう少し考えると、まず一つ言えることは、われわれは言葉を使って情報をやり取りしているということです。
例えば「おいしいイタリアン知らない？」という文があると、「イタリア料理を探していますよ」という情報がコンピュータやロボットに伝わります。
ここで情報のやり取りが行われています。
その後、ロボットがイタリアンレストランを探して、その結果、高級店とカジュアル店があるから、どちらかを選んでほしいと思って、その出力を作ります。
そのときに、相手（ショッピング帰りのお姉さん）に対して、高級店とカジュアル店があるという情報

を渡しています。
この「情報」が、われわれが普段「意味」と呼んでいるものだろうと考えられます。
今日は、「言葉が表す情報＝『意味』」とは何かを主にお話ししようと思っています。

・スライド 9「『意味』とは何か？」
「意味」とは何か、あらためて考えてみます。
よく学生が自然言語処理をしたいと言ってくるのですが、その場合は大体何か意味を解析して、翻訳したり質問に答えるシステムを作ることだと思っています。
そこで、僕は「あなたが言う『意味』とはどういうことですか」と聞きます。そうすると、ほぼ誰も答えることができません。皆さんも普段、「意味」とは何かということなど、考えたことはないと思います。
例えばコンピュータが処理するという観点でいうと、自然言語の文はよく考えるとただの文字列、データですから、データとしてそのまま処理すればいいではないか、なぜ「意味」を考えなければいけないのかという疑問が当然わいてきます。
ここで重要なのが、文字列として処理することと、われわれがやりたいことが少し違うということです。
どうしてかというと、文字列が近いということと意味が近いということにはズレがあると、われわれは直感的にいつも思っているからです。

・スライド 10「文字列と意味の不一致」
幾つか実際の文を載せました。
「友達からケーキをもらった」と「友達からケヤキをもらった」は、文字列としては 1 字違うだけですが、われわれは全然違うことが起きていると認識します。
一方、「友達からモンブランをもらった」という文は、文字列としてはケヤキよりもだいぶ遠くなって変わっているようですが、われわれは「ケーキをもらった」という文と近いと感じます。
1 例目と 2 例目は文字列として近いけれども意味が遠く、1 例目と 3 例目は文字列としては遠いけれども意味は近いのです。
「友達がケーキをもらった」という文は、1 例目の「から」と「が」が違うだけですが、意味ががらりと変わった気がします。ここから、文字列の近さとわれわれが直感的に感じている近さは違うということが分かります。ここが「意味」を考えるときの一番のポイントです。

・スライド 11「『意味』とは何か？」
大辞林で「意味」を引いてみると、「1. 言葉・記号などで表現され、また理解される一定の内容」「2. ある表現・作品・行為にこめられた内容・意図・理由・目的・気持ちなど」「3. 物事がある脈絡の中でもつ価値。重要性。意義」とあります。
「意味」という言葉にも三つの意味があり、あいまい性があるのですが、今日お伝えするのは 1 番目の一番素直な「意味」です。

文があると、われわれはその文が何か意味を表していると考えます。
その意味を理解して何らかの行動をしているので、1 番の定義は間違っていませんが、この定義を聞い

たからといって、コンピュータにどうやって実装したらいいかが分かるかというと、残念ながらそうではなさそうです。
辞書的な定義から「意味」にアプローチしても、どうやら解決しなさそうだということが分かります。
われわれはあくまでもコンピュータの上で「意味」を実現したいのです。
恐らく人間の脳みその中で、「意味」は存在しているのでしょう。
電気信号なのか、どういう形かは分かりませんが、頭の中にきっと存在する何かでしょう。
しかし、それが何かは観察できません。
そこで、「意味」の性質をよく見て、「意味」がどういう形をしているのかを考えていこうと思います。

・スライド 12「自然言語処理における『意味』とは何か？」
自然言語処理において、「意味」とはどういう性質を持っているのでしょうか。
ここでは、先ほど述べた、意味が同じか違うかがポイントになってきます。
われわれは「意味」を考えるとき、意味が同じ、意味が違うという言い方をします。
ポイントは、異なる文字列が同じ意味を表すことがあったり、同じ文字列が異なる意味を表すという現象があることです。
例えば、「私が猫にエサをあげた」と「猫が私にエサをもらった」は同じような状況を表していますが、文字列としては全然違います。

「かわいい瞳の大きな女の子を見た」という文は自然言語処理の授業でよく出てきますが、瞳がかわいいのか、女の子がかわいいのか、瞳が大きいのか、女の子が大きいのか、瞳の大きな女の人の子どもを見たのか、いろいろな解釈があります。このように、同じ文字列でもいろいろな意味を表すことがあるのです。われわれは意味が同じなのか、異なるのかという直感を持っています。
これは、専門的には「意味の同値性」「意味の差異」といわれます。
このような直感がわれわれにはあるので、この直感を再現するように何かをすればいいのです。
自然言語処理において意味を処理するとは、意味が同じ、意味が異なるという直感をコンピュータで再現できるようにうまく表現したり、計算する仕組みを考えることになります。
これで、自然言語処理において「意味」を考える理由や、何を考えなくてはいけないかということが理解されたかと思います。

・スライド 13「文字列から意味へ」
具体例をお見せしながら、実際にどのように意味を表現して計算すればよいかという話をします。
写真は私の実家で飼っているメイちゃんというビーグル犬で、ビーグル犬といってもかなりかわいい部類に入り、普通のビーグル犬とは違います（笑）。ビーグル犬を飼っている人はご存じでしょうが、この犬種は相当意地汚くて、何でも食べます。散歩していると、道ばたにあるものを何でも食べるのでしょっちゅうおなかを壊します。もしインターネットのページに「犬に風邪薬を飲ませると貧血状態に陥ります」と書いてあったら、「これは大変だ、うちのビーグルが風邪薬を食べたら病気になってしまう」と思います。ここでそう思う気持ちを分かっていただけますか？
これが分からないとこの先の話は全く意味を成さないのですが。

ここでのポイントは、この二つの文は何となく同じことを言っていますが、文字列としてよく見ると、共通しているのは「風邪薬」という単語だけで、ほかの単語は全部違うことです。
「犬」と「ビーグル」、「飲ませる」と「食べる」、「貧血状態」と「病気」、「陥ります」と「なる」という対応関係があるということは分かると思いますが、文字列としては全く異なります。
従って、異なる文字列なのですが、同じようなことを言っているようだと考えられます。
もう少し言うと、助詞が違います。
上の文では「に」「を」「と」になっていますが、下の文では「が」「を」「たら」になっています。
これが違うと分かった人は全くいないと思います。
人間は無意識のうちに助詞を解釈して同じことを言っていると計算できますが、今のコンピュータにとっては違う文字だから違うとしか言えません。

・スライド 14「意味の計算」
どういう対応関係があれば、実際に意味が同じだと言えるのでしょうか。この意味の同値性をどう計算するかを考えます。「犬に風邪薬を飲ませると貧血状態に陥る」と「ビーグルが風邪薬を食べたら病気になる」は同じことを表現していますが、この間にどういう関係があるか、この二つの文をどうすればつなぐことができるかを考えることが、意味を計算するということです。
そのために使うのが「意味表現」というもので、意味を何らかの形で表して、それを使って意味の計算を行います。

・スライド 15「意味の 2 つの側面」
では、意味をどう表現したらいいか、どう計算するかを考えます。重要なのが、「意味」には二つの側面があるということです。
一つはわれわれが「構成的意味」と呼んでいるもので、文の中で単語が組み合わされて表現される意味、つまり文が表している意味です。
「ビーグルが風邪薬を飲んだ」と「ビーグルに風邪薬を飲ませた」は大体同じようなことを言っています。「に」や「が」などの助詞が変わっているので、本来は意味的関係が変わっているのですが、同じようなことを言っています。
このように、文が変わっても同じようなことを言えるものが構成的意味です。

一方、「語彙的意味」とは、文を考えなくてもそもそも単語が持っている意味のことです。
例えば、「犬」「ビーグル」「柴犬」は何か関係があることは分かります。全く独立の単語ではなく、何らかの関係があります。
「顔」「目」「鼻」もそうですし、「あげた」「もらった」も何か関係があることが分かると思います。
これは文が与えられなくても、常に成り立つ関係です。
自然言語の単語はこういう関係を持っています。

・スライド16「つまり・・・」

つまり、単語を並べることによって表される構成的意味と、単語がもともと持つ語彙的意味の二つを組み合わせて、われわれは「意味」というものを考えています。
この二つを何らかの形で表現して、それによって意味が同じ、意味が違うというわれわれが持っている直感を計算する仕組みを考えればよいということになります。
コンピュータサイエンスに詳しい方は、意味のデータ構造とアルゴリズムと言えば、何をしたいかが分かるかもしれません。データ構造とは表現の仕方で、アルゴリズムとは計算の方法です。
意味の表現の仕方と計算の方法を考えることが、自然言語処理の研究者が行っていることです。

・スライド17「構成的意味」

さて、実際には構成的意味と語彙的意味をどのように表現すればよいのでしょうか。
構成的意味とは、単語が並べられることによって作られる意味です。
例えば「ビーグルが食べた風邪薬を私も飲んだ」という文では、誰が食べたのか、何を食べたのか、誰が飲んだのか、何を飲んだのかが書かれています。
「私が風邪薬を飲んだ」という文で表現する意味は、「ビーグルが食べた風邪薬を私も飲んだ」という文に含まれています。
直接的に「私が風邪薬を飲んだ」ではなく、「風邪薬を私も飲んだ」という形で書かれており、文字列としては少し変わっていますが、「私が風邪薬を飲んだ」という意味も含まれています。
こういうことをうまく表現したいのです。
つまり、文の中で単語と単語がどうつながっているかを表現したいということです。
主語や目的語という言い方は、恐らくご存じかと思います。
要は、「誰が食べた」の主語は何か、「何を食べた」の目的語は何かという関係を表現したいと思っています。

・スライド18「構成的意味の表現方法」

構成的意味を表現するときにグラフ構造というものを使います。
グラフ構造とは、「食べる」「ビーグル」など単語で表されているところが点に相当し、点と点の間をつなぐ矢印を辺といいます。
点と辺で表現するデータ構造のことをグラフ構造といいます。
この場合は点が単語に相当し、辺は単語間のつながりを表現しています。
例えば、図では「食べる」という動詞の主語は「ビーグル」、目的語は「風邪薬」で、「飲む」の主語は「私」で、目的語は「風邪薬」などと表現されています。
これは「ビーグルが食べた風邪薬を私も飲んだ」という文に対する意味表現ですが、これをよく見ると、「私が風邪薬を飲む」という部分が表現されています。
このように、文字列ではなく意味表現の方法に変換すると、意味の同値性が見えてきます。

・スライド19「構文解析」

では、構成的意味の表現をどのように計算したらいいのでしょうか。

文の意味表現をどう計算したらいいかということなのですが、それを行う技術のことを「構文解析」といいます。
構文解析とは文の構造を解析することで、構文構造を計算する自然言語処理の技術です。
文の構造が分かると、構成的意味が計算できます。
例えば「犬に風邪薬を飲ませると貧血状態に陥ります」という文があるとすると、構文解析でどの単語とどの単語がくっつくかを計算します。
図の赤い木構造で表現しているものが、構文構造と呼ばれるものです。
S や NP が何を表現しているかということは今日はお話ししませんが、各単語がどういう文法的性質を持っているか、それがどう組み合わさるのかを計算していくと、文の構造が計算できます。
この計算ができると、「犬が風邪薬を飲むと貧血状態に陥ります」という意味表現が計算できるという仕組みになっています。

・スライド 20「構文解析の研究」
実は私は、構文解析を主に英語でずっと研究しています。
そして、Enju（エンジュ）というシステムを公開しています。
このシステムは 90%の精度で構成的意味が計算できます。

・スライド 21「意味の計算」
構成的意味の計算、構文解析は、この 10 年ぐらいで飛躍的に発展して、新聞などの限られた分野の文はだいぶ高精度に解析できるようになっています。
この構文解析を使うと、「犬に風邪薬を飲ませると貧血状態に陥ります」という文が、このように意味表現として表現できます。
ただ、この意味表現だけでは、まだ「ビーグルが風邪薬を食べたら病気になる」という文にはつながりません。

・スライド 22「語彙的意味」
ここで必要になってくるのが、もう一つの意味の側面である語彙的意味です。
語彙的意味とは、単語がもともと持っている意味のことなのですが、一口に語彙的意味といってもいろいろな種類があります。例えば、同義語・反義語とわれわれは言うと思うのですが、小学校のテストでも「これの反義語は何ですか」と出ますよね。あるいは、上位・下位関係や全体・部分関係もあります。

・スライド 23「語彙的意味を表す方法」
こういういろいろな種類の関係がありますが、それをどうやって表現するかというと、ここでもわれわれはグラフ構造を使います。
図は意味ネットワークといわれているものですが、先ほどと同じで、点が単語に相当して、辺が意味的な関係を表します。例えば、右上の図は上位・下位関係を表したもので、哺乳類の下位に猫、犬、猿があるというように表現します。
いろいろな種類の関係がありますが、その種類ごとにグラフ構造を使って語彙的意味を表現します。

ただ、先ほどの構成的意味との違いがポイントで、構成的意味は文中の単語の意味的関係なのですが、語彙的意味は文中の単語ではなく、文とは独立の単語を表しているので、どちらかというと辞書に相当するものと考えていただければいいでしょう。

・スライド 24「同義（類義）関係」

例えば同義関係とは何かというと、われわれは普段あまり意識することなく、車と自動車は同じものを指しているとか、二酸化炭素や炭酸ガスや CO_2 は同じ意味だと言いますが、同じ意味を表す単語のことを同義語といいます。しかし、同義語の定義をあらためて考えると、実はそれほど簡単ではありません。
同義語の定義として、同じ意味を表す単語だと言っても定義にはなりません。
今は「意味」とは何かを考えているので、こういう定義ではただの循環定義になってしまいます。
では、どのように定義すればいいのでしょうか。
われわれは、「ほとんどすべての文脈で置き換え可能な単語」と定義しています。例えば「冬休みに X で草津まで行く」「駅前の道はたくさん X が走っている」という文の X の部分は、「車」と言ってもいいし「自動車」と言ってもいいのです。
このように、ある単語が出てくる文はたくさんありますが、そのほとんどにおいて置き換えることができるものを同義語や類義語といいます。

・スライド 25「反義関係」

反義関係は、われわれが思っているほど簡単な概念ではありません。なぜかというと、全く無関係な単語を反義語とは言わないし、否定したから反義かというと、そうでもないからです。
「大きい」の反対を「大きくない」としては、幼稚園のテストでバツになります。
「大きい」の反義語は「小さい」です。
考えてみると意外と難しいのですが、自然言語処理の分野では、反義ではなく排他性を考えることが多いです。排他性とは、同時に成り立たないということです。
ある人が大人であったら、子どもであるということはないですよね。
また、ある国がアメリカであったら、そこは日本や中国ではありません。
何かが成り立っていると、それ以外のものは成り立たなくなる性質を排他性といいます。
「大きい」「小さい」という反義語もそういう性質の一つとして考えることができます。

・スライド 26「上位・下位関係」

恐らく語彙的意味の中で一番重要なのが、上位・下位関係と呼ばれるものです。
これは定義からいうと、単語 A のすべての性質を B が持っているなら、A は B の上位語ということになります。
「A が指す集合の方が B が指す集合よりも大きい」という定義が一番分かりやすいかもしれません。
「犬」という単語が指すものが図の水色の部分だったとすると、「ビーグル」が指す集合は図の青い部分です。水色と青色が包含関係にある場合、つまり青い部分が完全に水色に含まれている場合に、上位・下位関係といいます。
「犬」が上位、「ビーグル」が下位です。

もっと直感的にいうと、「B は A の一種である」「B is a A」と言えるかどうか。
そういう関係を専門的には「IS-A 関係」といいます。
左下の図が IS-A 関係、上位・下位関係を表していて、「哺乳類」が上位語で、その下位語に「猫」「犬」「猿」があり、「犬」の下位語が「ビーグル」や「柴犬」になります。

・スライド 27「動詞の意味関係」
今までお話ししたのは主に名詞に関する意味的関係でしたが、動詞についても似たような関係があります。例えば「勉強する」と「学ぶ」は大体同じ意味なので、同義語と考えることができます。
ただし、動詞の場合は若干ややこしくて、項の対応関係を考える必要があります。
項とは主語や目的語のことです。例えば「あげる」と「もらう」はある種の同義語で、同じようなことを表現できます。
つまり、「私が友達にケーキをあげる」と「友達が私にケーキをもらう」が同じ意味になりますが、ここで「が」と「に」が入れ替わります。
ここまできちんと見ないと、これが同義であるとは言えないので、そういう関係も考えた上で同義を定義しなければいけません。

それから、動詞特有の関係もあります。面白いのが含意関係というものです。
これは、ある動詞で表されていることが成り立つなら、もう一つも必ず成り立つという関係で、分かりやすいのは「後悔する」という例です。
「〇〇を後悔する」という場合、「〇〇」は必ず起こっていると言えます。
「勉強しなかったことを後悔する」というなら、「勉強しなかった」ことは絶対に起きています。
そういう関係があり、それを含意関係といいます。

・スライド 28「フレーズの意味的関係」
幾つかの単語が並んだフレーズについても、同じように意味的関係があります。
これぐらいまできちんと分からないと、文の意味の同値性が計算できません。
例えば、「強い雨が降る」と「大雨になる」は大体同じ意味を表しているので同義語であるとか、「いびきをかく」といえば必ず「寝ている」ということになるので、含意関係が成り立ちます。
「いびきをかく」と「寝ている」の関係など普段はあまり意識しないかもしれませんが、「あの人、いびきをかいているよ」といえば、ぱっと「寝ている」と思いますよね。
そういうことが人間がやっている非常に難しいことで、われわれはそれを再現するためにこのような意味的関係を考えています。
さて、このように上位･下位関係、同義･反義関係、含意関係が考えられます。
今日は全体･部分関係は省略しました。

・スライド 29「シソーラス、オントロジー」
意味的関係、語彙的意味を収録した辞書のことをシソーラスやオントロジーといいます。
シソーラスは普通に辞書として売っているので、もしかしたら使ったことがある人がいるかもしれませ

ん。
われわれは自然言語処理のためにこういうデータを使います。
世界的に有名なのはWordNetで、プリンストン大学が昔から作っているものです。
日本語では分類語彙表や日本語語彙体系がよく使われています。

・スライド30「語彙的意味の計算」

では、文を解析したときにそのシソーラスやオントロジーをどう使うかということが次の問題です。
それが語彙的意味の計算に相当するのですが、何をするかというと、先ほど計算した構成的意味の中の単語と、シソーラスやオントロジーの中の単語をつなぎます。
例えば「犬」がつながったり、「陥る」「貧血」「飲む」がつながったり、図のように文中の単語がシソーラスの中の単語に結び付けられます。
これで何が起きるかというと、例えば「陥る」という単語が「なる」という単語に置き換えられます。
これが「置き換え可能性」で、同義語でちらりと言いましたが、ある文の中にある単語を別の単語に置き換えるという性質をシソーラスやオントロジーが表現しています。

・スライド31「シソーラス・オントロジーをどうやって作る？」

シソーラスやオントロジーがあると話しましたが、実はそれを作ること自体が大変なのです。
世の中には無数の単語があります。1万〜2万ではなく、100万以上の単語がありますし、フレーズの間の意味的関係も考えなくてはいけません。
フレーズまで考えだすと、恐らく億単位では収まらないぐらいの表現があると思います。
しかも常に新しい単語が生まれています。最近だとスマホ、少し古いとスーパークールビズという単語も出てきます。その同義語を人間がいちいち登録していたら大変です。
そこが自然言語処理のボトルネックになっていて、それがうまくできないから自然言語が理解できないのだといわれていたのですが、最近は少し状況が変わってきました。
例えば、インターネット上に無数にあるテキストをうまく使うと、同義語や上位・下位語を自動的に獲得できるのです。
最近そういう技術の研究が進んでいて、かなり大規模なシソーラスやオントロジーを自動的に作ることができつつあります。

詳しい話はしませんが、一つだけポイントを述べると、同義語で言ったことと同じようなことですが、同じような意味の単語は同じような文脈に出てくるという性質があるので、それを使うと同義語が分かります。
例えば、図に「ぽげら」という架空の単語を入れました。人間であれば、「朝見たらぽげらが真っ赤に熟していた。おいしそうだったので、またぽげらを食べてしまった。塩をちょっとかけたぽげらは激ウマだね」という文章を見ると、トマトのようなものを想像するのではないかと思います。
それが分かるのは、人間は文脈を見て単語の意味を推測できるからです。
それをうまく使うと同義語や上位・下位語を自動的に獲得できる技術が今はあります。

・スライド 32「意味の計算」

ようやくこれでシソーラス・オントロジーがつながります。図で青い矢印で示したところが、オントロジーやシソーラスにつながってきます。

ここまで計算できると、「ビーグルが風邪薬を食べたら病気になる」という形で、めでたく最初に示した文が導き出せます。

構文解析によって構成的意味を計算して、シソーラスやオントロジーを使って語彙的意味を表現すると、全く違う文字列でも意味が同じだと言える文ができます。

・スライド 33「構成的意味と語彙的意味の相互作用」

しかし、構成的意味と語彙的意味には、かなり複雑な相互作用があると言われています。

語彙的意味を使うと単語の置き換え可能性を表せる、文の中のこの単語はこちらに置き換えても意味は変わらないと言えると言いましたが、「犬」と「ビーグル」の例を見てみると、「ビーグル」が下位語です。

逆に「貧血状態」と「病気」では、「病気」が上位語です。

「犬」を下位語「ビーグル」に置き換えても意味は通じるし、「貧血状態」を上位語「病気」に置き換えても意味は通じます。

ただ、逆を考えて「犬」を「哺乳類」という上位語に換えてみると、「哺乳類が風邪薬を食べたら病気になる」は必ずしも成り立ちません。

今は犬について言っているので、「ビーグル」なら成り立つけれども、猫が風邪薬を食べたらどうなるのかは誰も分かりません。「病気」についても同じことが言えます。

このように、構成的意味がどこにあるかによって、どういう語彙的意味に置き換えられるかが決まってきます。

これは皆さんが頭の中で無意識にやっていることですが、コンピュータにやらせようと思うと、本当に難しい問題です。

・スライド 34「あいまい性」

さて、これで大体意味が計算できるのですが、自然言語処理で必ず出てくる「あいまい性」という重要な問題があります。

文があって意味が計算できるという話をさらりとしましたが、言葉と意味の関係は 1 対 1 ではなく、あらゆるところにあいまい性があります。

例えば、「友達とケーキを食べた」と「せんべいとケーキを食べた」では意味が変わっているのが分かりますか。「友達とケーキを食べた」は友達と一緒にという意味ですが、「せんべいとケーキを食べた」は二つとも食べたという意味になります。

下の図にはパックマンとアカベイとケーキが描いてありますが、何を言おうとしているか分かりますか。

鋭い人なら分かると思うのですが、図のような状況なら「友達とケーキを食べた」という文でも、「友達（アカベイ）」と「ケーキ」を両方食べたということが成り立ちます。

従って、もちろん文脈によってはその意味が変わってしまいます。

「今日はネットにつなげない」「ボールがネットに引っ掛かった」では「ネット」に二つの意味があるの

で、文脈によって意味が変わります。
このようなあいまい性は至るところにあります。

・スライド 35「あいまい性の問題」
「友達とケーキを食べた」と「せんべいとケーキを食べた」では「一緒に」という意味なのか「両方食べた」という意味なのかが変わってくるし、「ネットで検索しよう」「ネットで掃除しよう」という場合も「ネット」の意味が変わってきます。
「大きな黒い瞳の女の子を見かけた」という文も、黒いのは瞳なのか、女の子が日焼けしているのか、いろいろな解釈があって、それも文脈がないと分からないというあいまい性があります。

・スライド 36「あいまい性解消」
このあいまい性をうまく解決して、人間が解釈しているように解釈させることが、自然言語処理の中で一番難しい問題でした。
90 年代はそれが全然できなくて、翻訳してもめちゃくちゃな答えが出てきた時代でした。
その問題を打ち破ったのが「機械学習」と呼ばれている技術です。人間が作った学習データをコンピュータに与えると、その中から規則性や傾向を自動的に学習して、知らないデータに対しても規則性を適用してうまくあいまい性を解決しました。
学習データは右の図のようなものです。
例えば「食べた」という動詞に関する文がたくさんあります。
「せんべいとケーキを食べた」は「両方食べた」という意味なので、「ケーキ」にかかります。
「先生とケーキを食べた」は、「一緒に食べた」ことになります。
「クッキーとケーキを食べた」なら両方食べたという意味になります。
人間はぱっと直感で分かるけれど、どうしてそうなるのかということは分かりません。
しかし、どちらが正しいかは分かるので、それを人間にデータとして作ってもらって、そのデータを機械学習というものに渡します。
そうすると、「友達とケーキを食べた」という学習データにないような文があったときも、この文に対して、友達は「一緒に食べた」のか「両方食べた」のかを自動的に判断することができます。
この機械学習という技術によって、あいまい性の問題は完全にではないのですが、だいぶ解決されてきました。

・スライド 37「意味の計算」
文からどうやって意味表現を計算するかは、シソーラス・オントロジーの語彙的意味の計算でもそうだし、構文解析でもそうなのですが、あいまい性解消という問題が実はあって、そこは機械学習という技術をうまく使って、今は解決されてきています。

・スライド 38「含意関係認識」
今、紹介した意味の計算は、自然言語処理の分野では「含意関係認識」といわれ、最近いろいろ研究が進められています。

動詞のときにも含意関係といいましたが、それとは少し意味が違います。
この場合は文が二つあるとき、その二つの間に含意関係が成り立つかどうかをコンピュータが自動的に判断するという問題です。
含意関係とは、「犬に風邪薬を飲ませると貧血状態に陥る」という文が正しいと仮定すると、「ビーグル犬が風邪薬を食べたら病気になる」という文も正しいと言えるという関係です。
これはある種の意味の同値性、もっと言うと下の意味が上の意味に含まれているという関係です。
「川端康成は『雪国』などの作品でノーベル文学賞を受賞した」という文があると、「川端康成は『雪国』の著者である」ということが分かります。
これもある種の意味の同値性です。
こういう判断をコンピュータにさせるという研究が行われています。

・スライド 39「含意関係認識で大学入試問題を解く」

なぜこの話を持ち出したかというと、最初に紹介した大学入試を突破するプロジェクトのとっかかりとして、知識を問う問題をコンピュータに解かせてみました。
知識を問う問題とは、教科書や参考書を見れば答えられる問題のことです。つまり、どれだけ教科書や参考書を記憶しているかを問われるような問題です。
コンピュータは無尽蔵に記憶できるので、今のハードディスクだと教科書何万冊分も保存できます。
従って、コンピュータにとっては知識を問う問題は簡単と思われがちですが、実は違うということがここでのポイントです。
単に記憶していればいいのではなく、記憶している内容と、今問われている内容が同じ意味であるかを判断しなければいけません。
これがまさに含意関係認識に相当しています。

・スライド 40「含意関係認識で大学入試問題を解く」

実際の例をお見せします。
2009 年度センター試験世界史 B で、「兵制や兵士について述べた文として最も適切なものを、次のうちから一つ選べ。①イェニチェリは、オスマン帝国の常備軍であった。②フランク王国では、テマ制（軍管区制）の下で屯田兵制が行われた」という問題があります。
最近いろいろなところでこの例を出しているのですが、出すたびに自信を持って答えられるかどうかを質問しています。
当てたりしませんから、正直に答えてください。
これに「答えられます」という方は手を挙げていただいていいですか。
一人いらっしゃいますね。
1 カ月ほど前にこのプロジェクトのシンポジウムをやったときも一人ぐらいしかいなかったのです。
しかし、受験生は 8～9 割この問題を解くのです。
みんな通ってきたはずの道なのに、どこで忘れるのでしょうか。
かく言う私も、「イェニチェリ」なんて言葉は初めて聞いた感じだったのですが（笑）。

教科書では、「イェニチェリ軍団は、軍楽隊、工兵隊、大砲隊、鉄砲隊などをそなえた皇帝直属の常備軍である」と書いてあります。
ビザンツ帝国のところにも「テマ制を採用した」と書いてあります。
図の赤い線と青い線が対応していることが分かります。
この文を読んだらこの問題の答えが分かると思いますが、どちらが正解か分かりますか。
これが分からないと困るのですが（笑）。これはどちらかというと、日本語力の問題です。
教科書には、イェニチェリ軍団は、オスマン帝国の皇帝直属の常備軍であったと書いてあるので、①は正解です。
②はビザンツ帝国という国があって、ここでテマ制を採用して屯田兵制が行われたので、フランク王国ではないので②は間違いとなります。
大丈夫ですか（笑）。
ちょっと疲れているから分からないのかもしれませんが。

例えば、①では「イェニチェリ」と言っていて、教科書の文では「イェニチェリ軍団」と言っています。
「イェニチェリ」と「イェニチェリ軍団」は同じということは自明ではありませんね。
これが当たり前だと思うと、コンピュータはなぜ言葉が理解できないのかということが分からないと思います。
「たけし」と「たけし軍団」は全く別のものですよね。
「○○」と「○○軍団」が同じとは限りません。
「たけし」と「たけし軍団」は違うけれど、「イェニチェリ」と「イェニチェリ軍団」は同じであることが分からないといけません。
もう少し難しいのが、「皇帝直属の常備軍」といった場合、それが「オスマン帝国の皇帝直属の常備軍」であれば、「オスマン帝国の常備軍」だと分かりますよね。
これも全く自明なことではありません。
これが分からないと、知識を問う問題といえどもコンピュータは答えられません。

・スライド 41「NTCIR」
これがどれくらいできるかを、この前やりました。
国立情報学研究所で NTCIR と呼ばれるワークショップを 1 年半に 1 回ずつのペースで開いています。
そこでは検索の技術などいろいろなことをみんなで一緒に評価しようと、共有の評価データを提供しており、みんなでそれを使って競争しつつ知見を共有しています。
ここに含意関係認識をするというテーマがたまたまあったので、今回そこに参加して、センター試験を使って含意関係認識のデータを作り、実際にそのシステムを評価してもらいました。

・スライド 42「大学入試にチャレンジ」
簡単に紹介しますが、センター試験の選択肢と Wikipedia を使って含意関係認識のデータを作って提供しています。
対象科目は世界史、日本史、政治経済、現代社会です。

最終的には6チームが参加して、現在最先端の含意関係認識技術がどこまでできるかを試しました。

・スライド43「データの作り方」

データとしては、センター試験から文を取ってきて、それに関係する文をWikipediaから取ってきました。この文の間に含意関係が成り立つかどうかを判定してもらいました。
歴史的事実として正しければ当然含意関係が成り立つし、うそであれば成り立たないと判断しなくてはなりません。

・スライド44「実際のデータ」

実際に含意関係のない例として、「パルテノン神殿は、ドーリア式神殿の最高傑作と言える作品である」とありますが、ドーリア式とヘレニズムに排他性があることが分かると、含意関係がなく、矛盾していることが分かる仕組みになっています。
このデータはまだ一般公開されていませんが、春ぐらいには利用可能になる予定です。

・スライド45「評価結果（試験の正答率）」

こういうデータを作って実際に試験を解かせた結果がこの表です。
試験の正答率は表のとおりですが、一番いいシステムがIBM-1で、57.7%の正答率が出ています。
ほかでも、いいチームだと5割ぐらいの正答率です。4択問題なので、25%は誰でも取れます（笑）。
従って、5〜6割はそれほど悪くないけれど、いいというほどでもありません。
今はそれくらいの意味処理ができると言えます。

・スライド46「意味に関わるその他の問題」

構成的意味と語彙的意味の話をしましたが、実は意味に関する問題はそれだけではなく、ほかにもたくさんあります。
例えば「お風呂が沸きました」「お風呂が沸いています」「お風呂を沸かしました」「お風呂を沸かしています」という文があります。始めの三つを聞いたときは、もうお風呂に入れると判断できますが、最後の文ではまだ入れないと分かります。
なぜでしょうか。
「沸く」と「沸かす」という動詞と、「た（過去形）」と「ている（助動詞）」の関係が複雑に絡み合ってこういうことが起きています。
それをアスペクトや様相といいます。

さらりと流しましたが、先ほどのWikipediaには、「オスマン帝国の」とは書いてありませんでした。
文脈からこの皇帝はオスマン帝国の皇帝であると分かるのですが、その処理のことを「参照関係」といいます。

メタファー、メトニミーについてです。
「つらい時期を乗り越えた」という文がありますが、「乗り越えた」は、もともとは物の上を越えるとい

う意味なのですが、「時期」という抽象名詞に適用したときに、今はつらい時期ではないと分からなければいけません。しかし、なぜそれが分かるかということが難しいのです。
このようなさまざまな問題があって、まだ完全な意味処理はできないのですが、こういうことができると人間と同等の意味理解ができると考えています。

・スライド 47「おわりに」

以上、構文解析、シソーラス・オントロジー、あいまい性解消などの技術を紹介しました。
そういう技術を総動員して、われわれが無意識にしている意味解析、意味処理ができるようになります。
センター試験の問題を解くということに適用すると、そこそこできつつあるけれども、まだまだ遠いと言えます。
今日は意味の話がメーンでしたが、意味を理解することと言葉を理解することは、全然違ったレベルの問題です。
人間は意味を理解して情報を受け取った後、それに対して何らかの反応をします。
反応して初めて言葉を理解したと言えるのですが、そこについてはまだ全然見えていないというのが正直なところです。
従って、このような意味処理は言葉の理解に向けた第一歩で、まだまだこれからたくさん研究することがあります。

・スライド 48「宿題」

最後に宿題を一つ出します。
「のび太はお風呂に入っています」と聞けば、のび太は風呂場にいることが分かります。
含意関係認識の一つの例ですが、なぜこう言えるのかを実際に考えてみてください。
「お風呂」と「風呂場」の関係や、「入る」と「いる」の関係を考えればいいと思います。
今日、急遽追加したもう一つの宿題は、パソコンのバックアップを取りましょうということでした(笑)。
これはぜひお忘れのないようにお願いします。
以上で終わります。
どうもありがとうございました（拍手)。