統計数理研究所シリーズ〈データの時代〉③

言葉は数字

統計数理研究所
数理・推論研究系　准教授
理学博士　持橋大地氏

持橋先生の研究は、教えた言葉を話すロボットではなく、
教えなくても言葉を分析する機械だというから驚きだ。

持橋大地氏

1973年生まれ。横浜市出身。義務教育も高等学校も公立校。都立小石川高校から東京大学文科三類に進学。言語系に進むも、個別言語ではなく一般的な言語学を研究するために、年間2名だけという狭き門を突破、文系から理系に移行し、統計的に言語を研究し続け、2011年から統計数理研究所。数理・推論研究系 准教授。高校時代合唱部で指揮者に抜擢され、指揮の勉強をするために、高校のある巣鴨から立川に通っていたことがあるという。

持橋　国語研の田窪所長のインタビュー、読みました。おもしろかったですよ。

――ご存知なのですか?

持橋　僕は言葉の研究をしているので、特に最近は国語研と一緒に研究しているんです。

――言葉は文系ではないのですか?

持橋　言葉は文系的なものだと思われていますが、言葉の使われる頻度という観点から見ると、途端に数字が入ってきます。数字を扱うということは理系です。昔から、言語学は数字を無視してきていましたが、そうするとすべての言葉が等価になってしまいます。どういうことかと言うと、例えばとても有名な言葉〈Time flies like an arrow.〉の場合。普通に考えたら〈時間は矢のように過ぎ去る〉です。でも、flyはハエという意味もありますから、〈時バエは矢が好きだ〉とも解釈できるわけです。言語学の研究者は「こういう可能性もある」と言いますが、そもそも〈時バエ〉なんていないし、文脈を考えればfliesは〈飛び去る〉でしょう。

――はい。

持橋　人間が普通に考える解釈はこれだということを論理的に示すことはできません。そこにはやはり頻度が必要になってきます。言語学は理系の頭で考えていかなければ、きちんとした研究はできないということになります。

――今までの言葉の認識とは違う気がします。

持橋　普段その人が使っている言葉から、その人の属性のようなものが見えてきます。例えばインタビューをされていて、普通の人が滅多に使わない言葉に出会い、「あ、この人こういう人だったのか」と思う、そんな経験はありませんか?

――キャンベル館長がまさにそうです!

持橋　どんな日本語が出てきたのですか?

――「僥倖（ぎょうこう）」です。藤井聡太6段が使われて以来皆の知る所になりましたが、それ以前、インタビュー中に「僥倖」に出会いました。

持橋　日本語力、すごいなあ。

――先生、「香箱」はご存じですか?「花冷えの朝。猫は慎重で香箱を崩しません」って。

持橋　聞いたことはありますが、どういう意味ですか?

――文字通り薫香などを入れておく茶道具、香道具のことですが、猫の姿勢のひとつを香箱座りと言って、その姿勢をすることを「香箱を組む」と言うんです。これもキャンベル先生です。

持橋　やっぱり専門家ですねえ。他の例として、例えば「開腹する」。一般人は「開腹」とはあまり言わないので、この人は医療関係者だとわかる。大江健三郎の『恢復する家族』、あの『恢復』を使ったら、相当小説を読んでいる人だとわかります。というように、言葉からその人がどういう人かということがわかってきます。

――なるほど。

持橋　人文系の研究者は言葉を調べる時、言葉を1つ決めて研究します。統計学では1つではなく全部を調べる。たくさん調べることで、平均とちょっと違うという証拠を集めると、この人はこうなのだということが見えてきます。たくさん調べる時、人力でタグ付けしていたら根負けしてしまいますが、コンピュータならできます。これは言語処理の中の大きな技術なんです。

――具体例を…。

持橋　これはトピックモデルと言われる技術について10年くらい前に書いたものですが、例えば『銀河鉄道の夜』の中に出てくる言葉の頻度を調べました。

――作品の最初から最後まで?

持橋　はい。これは自動的に数えられるんです。言葉1つひとつにその言葉が持つ意味のようなものを自動的に割り当てます。38番とか58番とか振ってありますね。その番号、数字が意味なんです。58番に割り当てられた言葉を見ると、「ライラック」「雪上」「登山」とあり、ハイキングのようなイメージのものが集まっています。これは「自然」というような意味を持つ言葉の集まりになっています。6番を見ると、いかにも鉄道好きが喜びそうな言葉の集まり。これらは人間が割り当てたわけではなく、人間が教えなくても勝手に振り分けてくれた結果です。それを集めて17番はいくつ、6番はいくつと並べるとヒストグラム、棒グラフができます。するとその人が医者っぽいとか教育者だとか、数学に強いとか化学が入っているとか、そんなことがわかって来るんです。

――すごい！ 企業が放っておかないでしょう?

持橋　私はいくつも企業との共同研究をやっているのですが、例えばデンソーとの共同研究は会話のわかる人工知能の開発などですね。今若い人が使う「～じゃね」という言葉。「じゃね」は辞書にないですよね。だから「じゃね」の意味を人工知能に学習させよう、みたいな研究です（笑）。

――へええ。

持橋　これ、意外と難しくて。言葉はもともと音なんですが、音を文字に起こした時に、「とか言ったらわかるんじゃね」と言う場合。今の日本人が見ればすぐわかりますが、もともと「じゃね」が単語だという認識はないわけですよ。なので、「じゃね」が単語だとわかるように区切りを入れる。でも、これを人がいちいち教えていたのではきりがないので、これを自動的にやるための開発です。

――宇宙人が来ても何分か話していれば意味がわかるようになるという研究ですね。

持橋　そうです。それを目指しているということです。「教師あり」というのは基本的に例を教えてあげてその通りにやっていく人工知能のことで、「教師なし」というのはデータだけ与えて、そこから言葉の裏側にある隠れた構造まで学習するというものなんです。ただ、僕の中ではかなり有名な研究なので、今回は別のお話をしようと思っていました。ロボティクス（ロボット工学）というものです。

――なんだかまた難しくなってきた…。

持橋　どこから説明したらいいか考えてしまいますが、僕の研究とのつながりで言うと、ダンス譜ってご存知ですか?

――楽譜みたいな?

持橋　そうです。ダンスの一連の動きの中で、ここからここまでがひとつの流れというのを示すものです。バレエのように多くの人が知っている踊りではなく、民俗舞踊のようなものが芸術分野の研究対象です。今までは動画を録ってそこにひとつずつ人がタグ付けしていたそうなんです。でも民族舞踊などはタグ付けそのものが大変で、録画しても何もしないで、というかできなくて放ってあることが多い。それを自動認識させるという技術の開発です。これができることで舞踊学のデータ解析が格段に進歩して、踊っているビデオを見ただけでその舞踊がどういうものかと解析できるようになると思います。

――はい。

持橋　これは単に分析するだけではなくて、生成するモデルなので、統計的には乱数を振ると勝手に動いてくれる。躍らせることもできるわけです。

――はあ～。すごいですね～。

持橋　ロボット工学も共同研究ですが、僕個人の研究は別にあって、例えば「岩波データサイエンス」という、最近出て立川のオリオン書房にもある本に書いてあります。言語処理についてですね。例えば「国連」という言葉のIDは45701番目の単語とすると、文章はIDを表す数字が並んでいるように見えるわけです。人間が見れば、これは名詞、これは動詞と品詞がわかりますが、それを自動的に見つけさせる。昔はむずかしいと言われていた技術ですが、2007年あたりからはかなりきれいにできるようになりました。『不思議の国のアリス』のテキストを見せたら、もののみごとに振り分けて、1番に振り分けられた言葉は名詞ばかり、2番には形容詞、3番には動詞が集まってきている。5番も名詞なのですが、1番の抽象名詞に対して、ねずみとか女王といった単語が集まってきている。しかも、この後、その品詞が何個でてくるかということも推定してくれて、僕自身が感動しました。

――なぜ? なぜそんなことができるの?

持橋　なぜって、原理的には品詞は無限にあるのですが、よく出てくる品詞とめったに出てこない品詞ってあるじゃないですか。名詞や動詞はよく出てくる、それらの統計モデルを立てて、そこにテキストを与える。

――プログラムを書くわけですね。

持橋　書きます。7000行くらいのプログラムを3か月とか4か月とかかけて書いて、最初は間違ったりしますから、どこが間違ったかなと一生懸命考えて、間違いを直す。そういうのを何回もやって、バグを直すだけでも半年かかりますし、まだパーフェクトじゃない。言葉はデータ量がものすごく多いので、計算だけでも何十時間もかかる。僕がどんなに気を遣ってプログラムを書いても、計算に10時間くらいかかります。プログラムをうまく書くと10時間で終わるものが、下手な人が書くと、1週間とか2週間とかかかるようになりますから、プログラムの技術も求められていますね。

――大変でも、結果がこんなにうまくいけばうれしいでしょう?

持橋　楽しかったですね。だって勝手に学習してくれるんですから。本来は国語研の人などにこういう手法をもっと広げないといけないと思っています。英語とかフランス語とかなら人間にもできますが、全然知らない言葉だとできませんよね。仏教用語などもそうですよね、サンスクリットとかパーリ語とか。でも僕の技術を使うとそれができるのではと思っています。