ふしぎの国のスウガク使い

確率と統計の科学でヒトのことばの謎を解く

内村直之

▶最近のコンピュータは，私たち人間のことばをずいぶん理解するようになりました．スマホに「今日の天気は？」と聞けば，天気情報が書いてあるホームページを見て答えてくれます．Googleでわからないことがらを調べれば，「ここに説明してあるでしょう」とばかりにたくさんの文書を提示してくれます．私たちのことばをコンピュータで扱うのに，実は確率統計的な考え方がとても役に立っています．今回は，統計数理研究所の持橋大地准教授にことばを確率的統計的に扱う最先端の言語研究について，お話を聞きました．「数学でことばを？」とちょっとびっくりですが，機械による翻訳や情報探索など，これからの私たちの情報生活について不可欠なものとなりそうです．

持橋大地さん．
統計数理研究所（東京・立川）のロビー壁に刻まれた「数」の字を前にして．

「ことばというものには，子供のときから興味があったようです．幼稚園のときに『宇宙戦艦ヤマト』と漢字で書いて両親をびっくりさせた」と持橋さん．もちろん，高校でも英語と国語は大の得意科目で，当然，大学は東京大学の文系を受験しました．同時に数学も好きで，「大数へ数学」誌の熱心な読者だったといいます．

「言語学の研究をやりたかったんですが，大学へ入って知った文学部の言語学の内容は，○○語の音韻規則はどうしたこうしたとか，細かいマニアックな話ばかり．もっと抽象的・一般的な『ことばの問題』の研究に憧れていたので，こりゃいかんと……」．

そこで決めたのが，抽象的な学問のできる理科に進むことでした．理科から文科へ鞍替えする「文転」は珍しくないけれど，文科から理科への「理転」はめったにいません．一年留年し猛勉強の末に「基礎科学科第二」という理系学問なら何でもやれるところへの進学に成功しました．

ことばの勉強は「センスが勝負」と考えがちだけれど，そうではなく，実はたくさん触れてパターン認識をうまくできるようになることがポイント，と高校時代から考えていたそうです．「書き換えとか要約とか問題演習をいっぱいしないと英語はできるようにならない．なんか，語学って機械的だな，と思っていた」．それが今の持橋さんの研究の原点のようです．

コンピュータでヒトのことばを理解したい

コンピュータが開発されて，「ヒトのことばがわかる電子頭脳」があるといいなあ，と思った研究者は多かったようです．もちろん，コンピュータはコンピュータ用に設計された人工の言語（たとえば今ならC言語とかJavaとかのプログラム用言語）を「理解」することができます．その命令に従って「プログラム」通りに処理を進めます．ヒトのことばも同じように処理できるだろうか，と機械翻訳などを含めたいろいろな試みが1950年代からなされました．ヒトがきっかけの文を入力すると，あたかもそれに答えるような「会話ソフト」も作られましたが，これは，あらかじめたくさん用意された文を相手の入力文に合わせて出力するだけでとても「理解」というわけにはいきませんでした（これは，後に「人工無脳」というあだ名がついたこともあります）．

ヒトのことばには，たくさんの語彙に加え，それを組み合わせて文を作るルール「文法」があります．これにそってきちんと組み立てればヒトのことばをコンピュータで扱えるのではないか？……こういう試み「自然言語処理」の研究が始まったのですが，これはとてつもなく難しかったのです．

大学時代に持橋さんはコンピュータになじみました．2年生のときから，プログラムの相談を受ける指導員を務めていたほどです．ことばへの興味もあって，当然，その眼は自然言語処理に向かったのでした．

「ルールを基にして僕たちの使う言語をコンピュータ内で組み立てようとすると，例外があまりにもたくさんあったり，一つの単語を知らないだけで処理が破綻した

りしてうまくいかなかった」と持橋さんは説明し，いろいろそんな例をあげてくれました．

たとえば，「走る」なら「急いで歩く」と同じ，というような言い換えルールを作れますが，「たたずまい」となると，国語辞典的な定義を基にしてもさっぱりうまく使えない．たくさんの語彙があるのが大事だとはわかりますが，それをどう組み合わせたら，文や文章になるのかはコンピュータには理解が難しすぎました．

私たち日本人が利用する「かな漢字変換」でもそうです．「へんなじがでる」というのを変換すると「変な字が出る」「変な地が出る」「経んな辞が出る」「変な自我出る」のどれが正しいのか，コンピュータにはわかりません．

そこで，1990年代後半から使われだしたのが確率の利用です．「かな漢字変換」を例にすると，「変な」の方が「経んな」や「偏な」よりも頻繁に使われます．さらに二つのことばの組み合わせの頻度を見ると，「地が出る」「痔が出る」よりも「字が出る」の方が頻度は多い．まして「自我出る」というような使い方はほとんどありません．漢字も含めたそれぞれのことばの出現確率，あるいは二つのことばの組み合わせの出現確率から，一番実現の確率の高い変換結果を求めると「変な字が出る」にたどり着く．これがことばを確率統計的に考える基本の方法です．

「文法」よりたくさんの単語の相互関係が大事

「ルールよりも，語彙が大切．実現される文や文章はあることばと別のことばの『関係』の強さ，つまり確率で決まっていくのです」と持橋さん．たとえば，「たたずまい」という単語は「おだやか」とか「静けさ」とかいう単語と連携することが多く，「暑苦しさ」などという単語と一緒に使われることはない．その関係の強さを確率として数値化していけば，そういうネットワークとしてことばの世界が成立するのではないか……言語を確率と統計的にみてこういうふうに考えるようになった1990年代後半，持橋さんは，大学を卒業して奈良先端科学技術大学院大学（NAIST）の松本裕治教授の研究室に入ります．ここは，自然言語処理の日本の最先端を行く研究室の一つでした．

そこで研究したのは，たとえば，長いいくつもの文字が並ぶ文や文章を入力してそれを単語に区切り，各単語の役割（品詞：名詞，動詞，助動詞，形容詞などですね）を決めることでした．こういう作業を「形態素解析」といいます．確率統計的に言語を考えれば，これをなんと数学的な式であらわし，解くことができる問題になるというのです．ちょっと抽象的で面倒ですが，その式をチラ見してみましょう．

N字からなる文章$S=c_1c_2c_3\cdots\cdots c_N$があるとします．問題は，それを$n$個の単語が並んだ列
$W=w_1\cdots\cdots w_n$（その単語の品詞の並び方は
$T=t_1\cdots\cdots t_n$だとします）に分けるとき，単語の出現確率とそれが並ぶ確率から計算される文章の出現確率
$P(W, T)$が最大になるような単語列Wと品詞列Tの組み合わせを求めなさい，というのが問題になります．$P(W, T)$が最大になれば，それが実際に現れる文章になるはずというわけです．

文字の分け方によって現れる単語の出現確率は変わります（先ほどの「変な自我出る」という切り方を思い出してください）．単語を並べる確率は，多くの単語を考えると複雑すぎるので，二つ並んだ品詞の出現確率だけを考えることにします．これはたとえば名詞－名詞，形容詞－名詞と並ぶ確率は大きいけれど，名詞－形容詞と並ぶ確率は小さいということを実際に調べておくわけです．品詞別の単語の出現確率を$p(w_i|t_i)$，品詞が二つ並んで出現する確率を$p(t_i|t_{i-1})$とすると，文章の出現確率$P(W, T)$は，それらを全部掛け合わせて

$$P(W, T)=\prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1})$$

と書き表すことができます．

たくさんの単語の出現率を調べ，N字からなるSを入力すれば，この式を最大にするようなSの単語への分け方を求めることのできる計算方法が開発されています．

どんな言語でも使える強力な方法

これは一番簡単な言語の模型ですが，さらに詳しい確率統計的な理論で，誰にも教わらずに見えている文字列だけから自然に単語に分けることに持橋さんは挑みました．ここで開発された「ベイズ階層言語モデル」という方法はちょっと難しいので，説明は省略しますが，その威力を見てみましょう．

どれが単語かわからない「切れ目なし」の文章を入力し，どう単語に分割されるかをみましょう．

日本語（新聞記事の一節）

国連による対イラク制裁解除に向け，関係の深い仏に一層の協力を求めるのが狙いとみられる．

→国連 による 対 イラク 制裁解除 に 向け，関係 の深い 仏 に 一層 の 協力 を 求め る の が 狙い とみられる．

日本語はかなと漢字があるため，ちゃんと単語に分け

る前と分けた後を読んでも違いはわかりにくいかもしれません．ところが英語になると，その威力がわかります．

英語（ルイス・キャロル「Alice in Wonderland」の一節）

first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspeduponherknee,andthebrighteagereyeswerelookingupintohers

→first, she dream ed of little alice herself, and once again the tiny hand s were clasped upon her knee, and the bright eager eyes were looking up into hers

単語がくっついてしまうととても読みにくいのが，単語に分けられてすっきりと読めますね（-ed や -s が1語になっているというところもありますが）．さらにアラビア語ではどうでしょう．

アラビア語（AFP のアラビア語ニュースから）

الفلسطينيبسببتظاهرةلانصارحركةالمقاومةالاسلاميةحماس.

↓

الفلسطيني بسبب تظاهرة ل انصار حركة المقاومة الاسلامية حماس .

これを Google による自動翻訳にかけると違いははっきりわかります．単語に分かれていない前者では，

“Filstinebsbptazahrplansarhrkpalmquaompalaslamiphamas.”

となんだかわからないのですが，単語に分かれていると，

“Bcause of a demonstration for Palestinian supporters of the Islamic resistance movement Hamas.”

と明確に意味がわかります．

この方法は，単語がくっついてしまうことが多いドイツ語のような言語，辞書が十分にない古文，さらには未知の言語で書かれた文章にも使えます．持橋さんは「いづれの御時にか，女御更衣あまたさぶらひたまいける……」から始まる『源氏物語』の解析に使ってみてうまくいったことを嬉しそうに話してくれました．

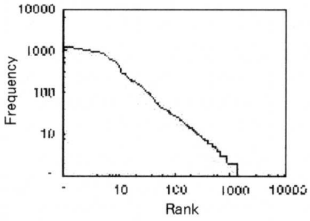
確率と統計の力を借りると，ことばの不思議な性質も見えてきます．たとえば，宮沢賢治の代表作『銀河鉄道の夜』に出てくる単語を出現頻度順にならべ，その出現確率を調べてみるということをしてみます．結果の一部は右上の表になります．この物語は 1307 種類の単語（「。」や「、」も一つの単語と見ています）からできているのですが，助詞の「の」や助動詞の「た」のようにたくさん使われるものもあります．物語のキーワードである「天の川」などは中くらいの頻度です．一方，「燈火」「天蚕」「鶴嘴（つるはし）」などごく少数しか使われていない単語もいくつもあります．

順位	単語	頻度	出現確率
1	の	1266	0.055005
2	。	1120	0.048662
3	、	988	0.042927
4	た	951	0.041319
5	て	884	0.038408
18	ジョバンニ	189	0.008212
34	カムパネルラ	101	0.004388
104	風	26	0.001130
104	天の川	26	0.001130
482	橙	5	0.000217
482	ボート	5	0.000217
482	ステーション	5	0.000217
1307	燈火	1	0.000043
1307	天蚕	1	0.000043
1307	鶴嘴	1	0.000043

『銀河鉄道の夜』に出てくる単語の頻度と出現確率．

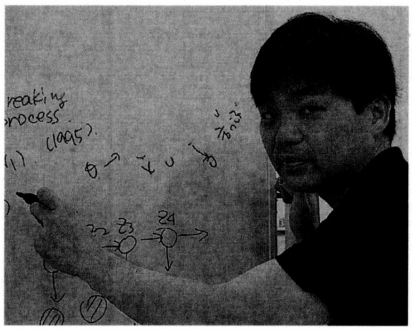
少数の高頻度のことばがある一方で，低い頻度のことばがいくつもあることがわかる．

単語の出現回数の順位 r と頻度 f を掛け合わせるとだいたい一定になります（つまり反比例関係）．両対数でプロットしたグラフはこんなぐあいです．



順位 r（横軸）と頻度 f（縦軸）を両対数でプロットした．
$\log(r)+\log(f)\fallingdotseq$ ほぼ一定，つまり
$r\times f\fallingdotseq$ ほぼ一定，という関係がある．

これは以前からジップ（G. K. Zipf はアメリカの言語学者）の法則として知られています．「ことばは，よく使うものはホントにたくさん使う『Rich gets richer.（富めるものはますます富む）』という性質がある一方で，頻度の少ないことばもたくさんある，いわゆるロング・テール現象もあるんですね」と持橋さん．複雑な現実を表現するには，莫大な語彙数がどうしても必要になる，といいます．現実に合わせて，新しい単語が生まれる一方，古い単語は忘れられていくという筋道を「ポアソン・ディリクレ過程」という確率の生まれ方を導入して調べると，確かにこのジップの法則を再現できます．「まるでことばの進化をみているよう．集団遺伝学で使われる数学と重なるところもあるんですよ」．



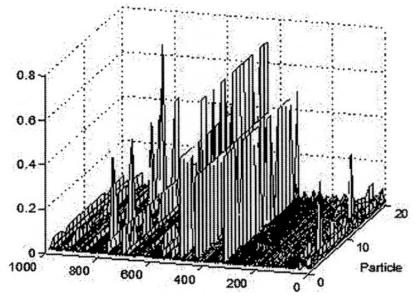
夢ふくらむ自然言語処理の世界

90年代，インターネットの世界にあらゆる文書を結びつけることのできるWWW（World Wide Web）が登場，人々の情報収集手段を革命的に変えました．そこで注目されてきたのはGoogleやYahoo!などが提供する「検索エンジン」です．ここでも，人間が書いたり話したりする言語「自然言語」をコンピュータで扱う技術が中心となっています．自然言語処理が取り組むべき目的は，単語や構文の解析を超え，文書から欲しい情報を抽出して要約したり，異言語間で自動翻訳したり，さらには，音声認識入力や，コンピュータによる質疑応答にいたるまで広がっています．そこでは，持橋さんがずっと取り組んできた「統計的自然言語処理」という方法が今や主流になっています．

「文章の『意味』も機械で判別することはできないだろうか？」と持橋さんは考えます．単語とその並びの後ろに隠れた「意味」を，私たちは読み取ることができます．朝，届けられた新聞を見て「大逆転 笑顔が呼んだ甲子園」という見出しを読めば，「これは高校野球の記事だな」とわかります．こういう文章の背景にある「話題」をトピックと呼びましょう．文章は普通，いくつかのトピックにまたがっている存在ですね（たとえば，この文章も，「ことば」というトピックと「数学・確率統計」というトピックが混ざっていると考えられます）．混ざったトピックから単語の並びが生まれてくることになりますが，先ほどの単語の出現率からどういうトピックが確率的に混ざっているかを判断できれば，「文章の意味」がコンピュータにわかるということになります．これを実現するのが「潜在的ディリクレ配分法（LDA）」という理論ですが，詳細は大学入学以後に学んでください．

この方法を応用すると，一つの文書の中で，文脈がどう展開されているか，という問題を扱うこともできます．たとえば，ある中国に関する文書で「香港の政治問題」「中国の議会問題」「中国の内政問題」「香港の経済問題」などと話題が変化するポイントを確率的に捉えることもできたそうです．こういう方法は，Webを使って商品を購入するのを見ていて，購買者の興味がどう変わっていくかを判断するというようなことにも使えるといいます．

今，世の中のあらゆる情報が文字データ化され，Webという世界に積み重なりつつあるいわゆる「ビッグデータ」の時代です．それはとても有用な情報をたくさん含んでいるはずですが，とても人力で探索するわけにはいきません．そういうときには，「自然言語処理」という技術がカギになるはずです．統計・確率という数学でそれに挑む持橋さんの夢は大きくなるばかりでしょう．



中国問題に関する文書の文脈をコンピュータで追った．グラフが飛び上がった点で話題が変化した確率が高まっている．

持橋大地（もちはし・だいち）　1973年横浜生まれ．都立小石川高校を経て東京大学文科三類入学．96年，同大教養学部基礎科学科第二進学．98年奈良先端科学技術大学院大学情報科学研究科へ入学し2005年博士（理学）を取得．ATR音声言語コミュニケーション研究所専任研究員，NTTコミュニケーション科学基礎研究所リサーチスペシャリストを経て11年，統計数理研究所准教授．専門は統計的自然言語処理と機械学習．

○もっと知りたい人のために

・この分野は始まって20年経ったかどうかという新しい分野です．専門的な情報が多く，あまり高校生が読めるものはなさそうです．Googleなどで「自然言語処理」を探すか，持橋さんのWeb
http://chasen.org/~daiti-m/
で読めそうな文献を探してください．

・この分野を研究しているのは，言語処理学会（http://www.anlp.jp/）や情報処理学会の自然言語処理研究会（http://www.nl-ipsj.or.jp）などです．

・ことばとコンピュータ，Webという問題を考える際，Googleという存在の動向は見逃せません．その理念
https://www.google.co.jp/about/company/philosophy/
と，していること
https://www.google.co.jp/about/company/products/
を見ておきましょう．

（うちむら なおゆき，科学ジャーナリスト）