第 8 回
コロキューム講演録

Cognitive Computing
未来をつくる人工知能技術

Topic

第 1 部　講演：P1-4

- Cognitive Computing とは
- 自然言語処理の可能性
- 人のセンサーとしての SNS
- オープンデータの活用
- スマートな都市づくりのための
  ソーシャルメディア活用

第 2 部　質疑応答：P5

講演者：村上　明子

日本アイ・ビー・エム株式会社
東京基礎研究所　専任研究員

講演日 2014/12/18

招聘者
澤谷由里子/白井裕子

アーカイブ担当
寺田翔太/出来寛祥/白井裕子

早稲田大学実体情報学博士プログラムのコロキューム、第 8 回は日本アイ・ビー・エム株式会社基礎研究所で人工知能やテキストマイニングによる情報技術に携わる村上明子さんをお招きして「スマーターシティ実現のためのソーシャルメディア分析」と題して講演をいただきました。村上さんが実際に研究された数々の事例をもとに、今後の展望を聞かせていただくことができました。講演後の質疑応答では、内容に関することだけでなく、日本 IBM の社風などリーディング生から多数の質問が寄せられました。

第 1 部　講演

講演者紹介

村上 明子(むらかみ あきこ)

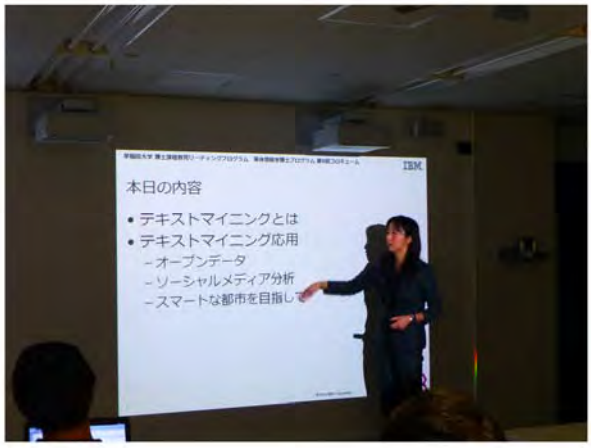
1999 年日本アイ・ビー・エム（株）入社。以後、同社東京基礎研究所において自然言語処理の研究に従事。テキスト・マイニング・ツール IBM TAKMI の研究開発において、品詞管理や辞書作成などを担当した。昨今では、電子メールや掲示板など人と人のコミュニケーションの文書を対象としたコミュニケーション分析などを行っている。近年では IT を活用した災害からの復興や減災、リスク管理を実現する「レジリエント工学」の分野にも関わっている。このほかに著書として「チャンス発見の情報技術（東京電機大学出版）（共著）」、訳書として「情報検索の基礎（共立出版）（共訳）」「Google Hacks 第 2 版、第 3 版（オライリージャパン）（共訳）」がある。

■Cognitive Computing とは

人間に近い思考過程を持つ人工知能 Watson

そもそも人工知能は人間と同じ能力のものをシステムで実現しよう、という研究です。歩いて喋って目を見て話す、という人間の様々なインタラクションを再現しようという見方もできます。IBM が開発した人工知能システムの一環である Watson も、人工知能のひとつであると言えます。この Watson はアメリカのクイズ番組「ジョパディ！」に出演し、2 人のクイズ王と互角に戦ったという実績を持っています。この人工知能でつくられたシステムは、出題されたクイズに対して、「この問題は何について聞いているのか？」、「答えの候補としていくつかありうるが、何が一番解答になるか？」ということを導くために処理を行います。問題の内容を判断し、それに対する最適な解答を過去の知識源の中から探すといった点で人間に似た思考プロセスを持っていると言えます。

IBM はこのようなシステムを人工知能とは呼ばず、Cognitive Computing と呼んでいます。Cognitive Computing は人間そのものを再現しようとしていません。Cognitive Computing は機械にできることと人間にできることを区別し、それぞれの能力をそれぞれの分野で最大限活かすことを目指しています。例えば、「目の前の物体を椅子と判別すること」について考えてみてください。普通の 4 本足の形状をした椅子であれば機械にとってもそれを椅子と判断することは可能でしょう。しかしその椅子の形状が、前衛的な形をしていて、いわゆる 4 本足の椅子でなければ、機械にとって椅子の判別を行うことはとても難しいタスクとなります。しかし人間は目の前の物体が見たことのないものであってもそれを椅子であると推測する力に長けています。人間と機械、それぞれに長けているところを補完しあう、それが私たちの考える Cognitive Computing の基本的な発想になります。

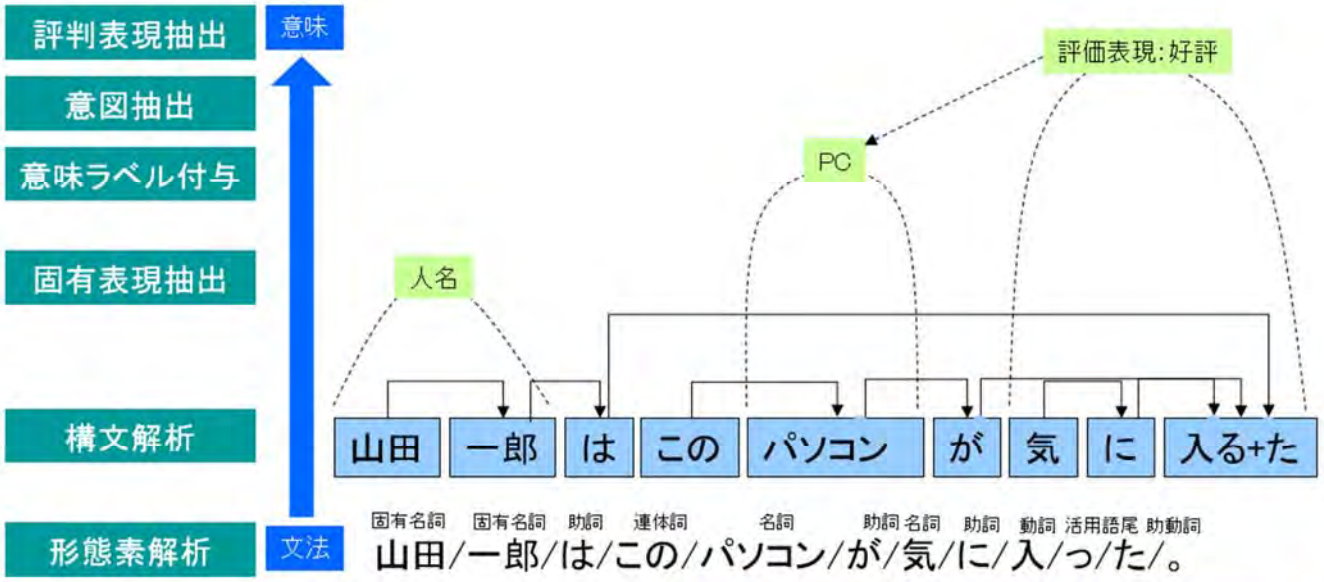


■自然言語処理の可能性

まず、私が現在携わっている仕事の基本的なことについて話したいと思います。私たちは色々なデータを仕事柄扱いますが、そもそもデータにはどのようなものがあり、そうしたデータを処理することでどのようなことが可能になるのでしょうか？

そもそもデータにはどんなものがあるか？

データには大きく分けて２種類あります。一つは構造化データ、もう一つは非構造化データです。構造化データとは端的に言えば、スキーマの決まったデータです。例えば人に関するデータベースがあり、その中で名前・年齢・性別など決まった項目が埋められているデータベースといったものがあったとすればそれは構造化データにあたります。もう一方の非構造化データはスキーマが決まっていません。例えば Twitter 内のつぶやきなどそれぞれの単語に意味付け、ラベル付けがされていない文章は非構造化データといえます。他にも、新聞の文章や画像データ・音声データなども完全に構造定義をもたないため非構造化データにあたります。一般に非構造化データしか手元にない場合、特にそれがテキストデータであった場合には形態素解析や構文解析といった自然言語処理の手法を用いてラベル付けを行い、実際に意味を取り出す作業が必要になります。以下の図は自然言語処理の例になります。文章内の単語に対して固有表現の抽出や意味ラベルの付与を行い、さらにはその文章がどのような評価表現をおこなっているのかを解析します。



自然言語処理の応用

私のメインとなる仕事は、自然言語処理によって以上のような解析をし、非構造化データを構造化すること、そしてそれを利用するになります。こうした自然言語処理を応用すれば、Twitter 内のつぶやきもマーケティングのためのデータとして活用することができます。

例えば、Twitter の口コミ情報から競合他社の弱みや自社の強みを分析することで戦略をたてたり、自社の製品の品質管理などが可能になります。

■人のセンサーとしての SNS

以前、分析対象として使えたテキストデータは新聞のテキストデータなどであり、今あるデータと比べれば数も少なく、とても高価なものでした。しかし、最近ではインターネットの普及とともに SNS など一個人が情報を発信する機会が増え、数多くのデータがインターネット上に溢れる世の中になっています。例えば Twitter 上には世界各地の人間がその場で得たリアルタイム性のあるつぶやきがデータとして流れています。こうしていわば高機能なセンサーとしての人間が生み出したデータはうまく処理・解析することでそこから今までにない価値ある情報を生み出せると考えられます。これはまさに冒頭に申し上げた Cognitive Computing のひとつの実践方法であるといえます。

Twitter によって浮き彫りになる様々な現象

例えば、2011 年 3 月 11 日に日本を襲った東北地方太平洋沖地震の際には Twitter 上のつぶやきから店頭で買えるもの買えないもの、手に入るもの手に入らないものに関する情報を分析しました。この分析によると、ガソリンは地震後すぐに購入できなくなったのに対して、水は同月 20 日頃までは比較的手に入りやすく、その後いきなり購入できなくなったことが判明しました。この現象について検証を進めてみるとこの頃にメディアを通して原子力発電所からの放射能漏れの影響に影響による水道水汚染に関するニュースが報道され、水道水に対する安全性・信頼性が危ぶまれたためであることがわかりました。変わったところでは、タバコが購入できなくなってしまう現象が Twitter 情報を分析することで明らかになりました。これはタバコの外部フィルムを作る工場が東北地方に集中していたために、供給が一時的に滞ってしまったことが原因であることがわかりました。このように Twitter 上に投稿される人々の声を収集し解析することで様々な現象を浮き彫りにされます。

利用できるのは Twitter のデータだけではありません。いまや政府や公共団体などによって様々なデータベースが無償で利用可能な形でウェブ上に開設されており、これを利用することで分析できることが近年ますます広がっています。一般にオープンデータと呼ばれるこうした無料公開データベースは企業における経営戦略、保険料の設定、はたまた株価予測など実に多種多様なシーンで活用されています。

■オープンデータの活用

インターネットの普及に伴い誰もがオープンデータを利用できるようになり、今では多くの企業がそうしたデータを利用することで企業の経営やマーケティングに活用しています。ここからは広い意味でのオープンデータ（＊通常はオープンデータというと LOD をさすことが一般的ですが、ここでは政府等が公共で利用できるという広い意味でオープンデータと言っています）の活用方法について「分析対象としての活用」と「知識源としての活用」という２つの側面から見ていきたいと思います。

知識源としてのデータ

みなさんは GeoNLP というオープンデータ共有サイトをご存知でしょうか？ GeoNLP では地名に付随してその地域の緯度経度情報が含まれるデータがアーカイブされています。この GeoNLP で公開されているデータを活用することでユーザがつぶやいたテキスト情報に地名が含まれていた場合、そこから緯度経度情報も割り出すことが可能になります。しかし、実際に誰かがつぶやいたテキスト情報に含まれる地名は多くのものに曖昧性があります。例えば「中央公園」というテキスト情報があったとしても、それは特定の場所を指示していないため、それが新宿の中央公園なのかどこか別の地域に属する中央公園なのかテキスト情報からだけでは判断できません。そういった場合には情報発信者の住所や職場の住所などそのユーザが有する個人情報も含めて分析していくことでより正確な地理情報を推定することが可能になります。

また PubMed という医療関係の論文アーカイブがあります。ここにアーカイブされている論文はそこに含まれるテキスト情報を分析することで医療分野における専門用語同士の関係性を解析し、遺伝子相互関係のサマリーを生成することができます。

このように知識源としてのデータは、テキストデータと緯度経度情報を紐付ける等スキーマの決まっていないデータや表現方法が統一されていないデータを一つの決まった表現に統一するためのヒントとなるようなデータであるといえます。

分析対象としてのデータ

次に、分析対象としてのデータ活用についてお話しします。例えば、National Highway Traffic Safety Administration（NHTSA）という日本における国土交通省にあたる機関が運営するサイトでは交通事故に関する記録がアーカイブされ公開されています。このデータを活用して、例えば、「何歳の人が事故を起こしたのか」、「どの車種で事故が多いのか」などといった情報から自動車の保険料を算出する事が可能となるでしょう。

コールセンターでの会話履歴情報も立派な分析対象データです。製品に関する不具合情報に関するご連絡を数件頂いた時点で修理の必要性を判断、早期にリコールを実施することでリコール遅延による損失額を低く抑えるなどといったことも可能となります。特に自動車をリコールすることを考えるとリコールの実施が１日遅れただけでもその損失は何億円単位といったものになります。このように、オープンデータやコールセンターなどから収集したデータをいち早く分析にかけることで素早い経営判断が可能になるでしょう。

このように分析対象としてのデータは、知識源としてのデータとは異なりそのまま分析することができるデータであるといえます。

■スマートな都市づくりのためのソーシャルメディア活用

ここまでオープンデータの解析をしていくことでどのようなことができるのか、実際に企業でおこなわれてきた事例を紹介する事でデータマイニングの可能性を実感していただけたかと思います。ここからは私自身がこのようなデータマイニング技術を使ってこれから何をしていこうかということについてお話ししたいと思います。

都市における災害早期発見

ある場所で火災が起こったとします。このとき例えば Twitter 内での火災に関する話題を分析することで火災の早期発見、ひいては火災現場の早期特定などが可能になると考えられます。

都市における交通整備システムの改善

ソーシャルメディアから得られる情報は他の技術、たとえば交通シミュレーションなどと組み合わせることによってより活用できるようになります。たとえば、従来の交通シミュレーションの結果に人間が発信した Tweet からの分析情報を付加することで、どの部分で渋滞による不満が多いのかなど、より有意な交通情報を得ることができるでしょう。こちらは京都の祇園祭りを例にしたデモですが、気象予測と事前のデータによる交通シミュレーションによって渋滞部分を予測し、それにもとづいたバスの増便、通行規制といったことが可能です。しかし実際にお祭りに参加している現地の人の声を分析してみると、混んでいても不満が無かったり，逆に混んでいなくても不満であったりという現象がみえてきます。実は混んでいる地域でもきちんと列を組んで並んでいる場合それほど不満は生まれず、そんなに混んでいなくても人ごみが雑然としていることで不満が生まれることがあり得るのです。このようにその場にいないとわからない情報をソーシャルメディアから取得し既存の交通シミュレーションと組み合わせることで、今まで以上に効率的な交通整備対策が可能になるのではないかと考えています。

■まとめ

本講演では「データとは」、「自然言語処理とは」何なのかといった基本的なことに始まり、そうした技術が実に幅広い分野で応用されてきているということを少しでもご紹介できたかと思います。弊社の掲げる Cognitive Computing はいまだ進化段階にあります。これからは私自身、Cognitive Computing によって未来の社会システムを改善するために、その可能性を信じて研究を進めていきたいと思っています。

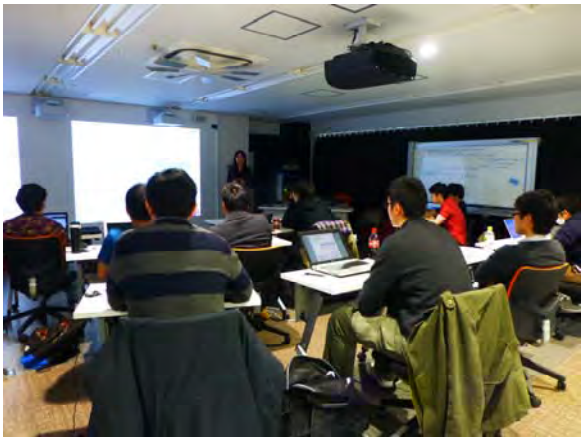
第 2 部　質疑応答

■ Q. オープンデータの活用において他の企業との差別化をいかに図るか？（高橋城志）

■ A. 企業が内部で持っているデータとオープンデータの組み合わせと独自の分析技術によって差別化を図ります。

オープンデータと顧客が持っているデータの組み合わせることで、事故の発見など受動的な分析のみならず、企業の売り上げ増加対策など積極的な経営戦略も分析によって実施できます。

また、テキスト分析やシミュレーションなど複数の側面から得られた結果を組み合わせることでより高度な分析を実施します。



■ Q. 例えば水が買えないなど社会的問題が起こった際問題の沈静化のタイミング判断はどうするのか？（林良彦）

■ A. 買えない人が減ったのであれば解決と判断しています。

基本的には買えない人の数が減っていくことを、沈静化判断の基準としています。しかし，現場には物資が届いているのに物資不足に関する Retweet は拡散し続けていってしまうという現実問題があり，現場と Twitter の流通情報に齟齬が生じる場合があります。その場合は、先ほどお話した拡散経路分析を行うことでオリジナルの情報発信者を特定し，Retweet 情報の信頼性を判断します。

■ Q. ユーザが複数のユーザ層に属することも考えられますが、こういったことを加味していますか？（佐々木一磨）

■ A. 分析の結果、複数の層ができてしまう場合はユーザを複数の層に帰属させます。

ユーザをクラスタリングする際にいかにクラスタリングするかは分析の目的に依存しますが、ユーザに複数の特徴が認められれば複数の層に帰属させます。マーケティング分野の人はひとつの層に決めたがる傾向にあります。例えば、F1 層などがその例ですが、こういったあるユーザをひとつのユーザ層に固定化して分析していくこと自体には問題があります。

■ Q. ユーザ分析はクラスタありきで行うのですか？（佐々木一磨）

■ A. クラスタはあらかじめ決めるものではなく分析の結果、みえてくるものです。

クラスタは決めつけではなく、あるユーザの分析を通して何かしらの特徴が見出された時にはじめてクラスタが現れます。例えば、政治に関してあるトピックについて興味を持っている人にどのようなクラスタの人が含まれているか調べてみたところ、子供を持つ方や専業主婦の方が多いということがわかるということがありました。

※質問者のうち、本プログラム所属の学生のみ、氏名を記した。

実体情報学博士プログラム

http://www.leading―sn.waseda.ac.jp/