テキストデータからの特徴抽出　佐々木先生による解説
～ニュースからの単語による特徴表現～

2019年 2月4日

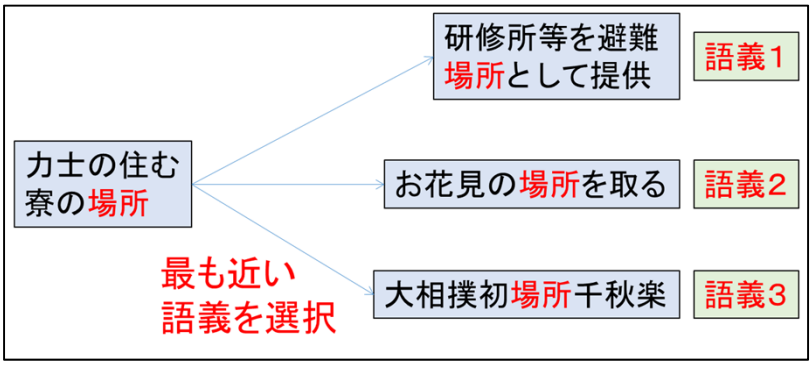
お伝えしたいポイント

- 佐々木先生は自然言語処理を専門に１８年の研究
- テキストを分析するための自然言語処理技術
- 分析の第一歩はテキストの単語分割
- 単語以外で分析に使える特徴「係り受け情報」と「類義語」
- 単語の特徴を使ってテキストの特徴を捉える

＜佐々木先生は自然言語処理を専門に１８年の研究＞

当社は、茨城大学との共同研究を進めてますが、今回は鈴木教授とともに共同研究の指導をして頂いています佐々木稔先生にテキストデータからの特徴抽出について解説して頂きます。アルファ碁が囲碁の世界戦優勝経験のあるプロ棋士を破るなど人工知能に一般の人たちの脚光が浴び始めたのは約３年前ですが、佐々木先生は１８年前からテキストデータについての研究を続けられており、これまでの蓄積によるアドバンテージはとても大きいと感じています。

佐々木先生の専門は自然言語処理で、ディープラーニングなどの機械学習を利用した語義曖昧性解消の研究を多数発表しています。語義曖昧性解消は、複数の語義を持つ単語（多義語）を含む用例文が与えられたときに、前後に出現する単語を特徴として多義語の語義を辞書の項目から選択する研究テーマです。辞書に記述された各意味区分について、周辺に出現する単語の特徴などを分析し、多義語の識別性能向上に取り組んでいます。



出所：茨城大学

図は「場所」の例で、語義１は「何かが存在する所」、語義2は「人がいる所」、語義３は「相撲を興行する所」を示し、入力文の「場所」は寮という建物が存在する所を表すため、語義１を選択します。

＜テキストを分析するための自然言語処理技術＞

有価証券報告書、決算短信などの企業開示資料やアナリスト・レポートといった大量のテキストデータと市場変動の関係性を発見し、市場の動向を分析することへの期待が年々高まっています。このようなテキストデータには、事業における業績の概要、経営戦略の方針、事業へのリスクや企業の不祥事に関する情報など、投資判断において数値で指標化されていない重要な情報が含まれています。このような自由記述で書かれた大量のテキストから有益な情報を効率良く取り出すために、テキストマイニング技術や人工知能といった機械学習技術を金融市場における分析に取り入れる研究が盛んに行われています。その背景として、自然言語処理の技術が進歩し、形態素解析や係り受け解析といった技術が実用化されるようになったことで、テキストデータも数値データとして指標化できるようになったことが挙げられます。今回は自然言語処理技術を用いたテキストデータからの特徴抽出について解説します。

我々人間が日常的に使っている言語をコンピュータで処理する技術を自然言語処理と言います。日常的に使う言葉を「自然言語」という聞き慣れない名称を使っているのは、コンピュータの分野では言語といえばCやPythonといった「人工言語」が一般的で、プログラミング言語と区別するためです。この自然言語処理には大きく４つのステップがあります(表参照)。この中の形態素解析技術により、テキストから単語出現頻度などの情報を抽出し、数値化できるようになりました。これらの数値データをデータマイニング手法に適用することで有用な知識や情報が得られます。例えば、 大手ショッピングサイトのAmazonでよく見る「よく一緒に購入されている商品」を発見するバスケット解析や、同じAmazonで「この商品を買った人はこんな商品も買っています」とあるようなルール(包含関係)を発見する相関ルールの分析などに使えます。

表：自然言語処理の解析ステップ

形態素解析	テキストを単語に分割し、単語の語形変化や品詞を決定する。
構文解析	単語や文節のつながりを求めて、文の構造を決定する。
意味解析	単語や文の意味を解析する。
文脈解析	複数文の系列を対象として、構文解析や意味解析を行う。

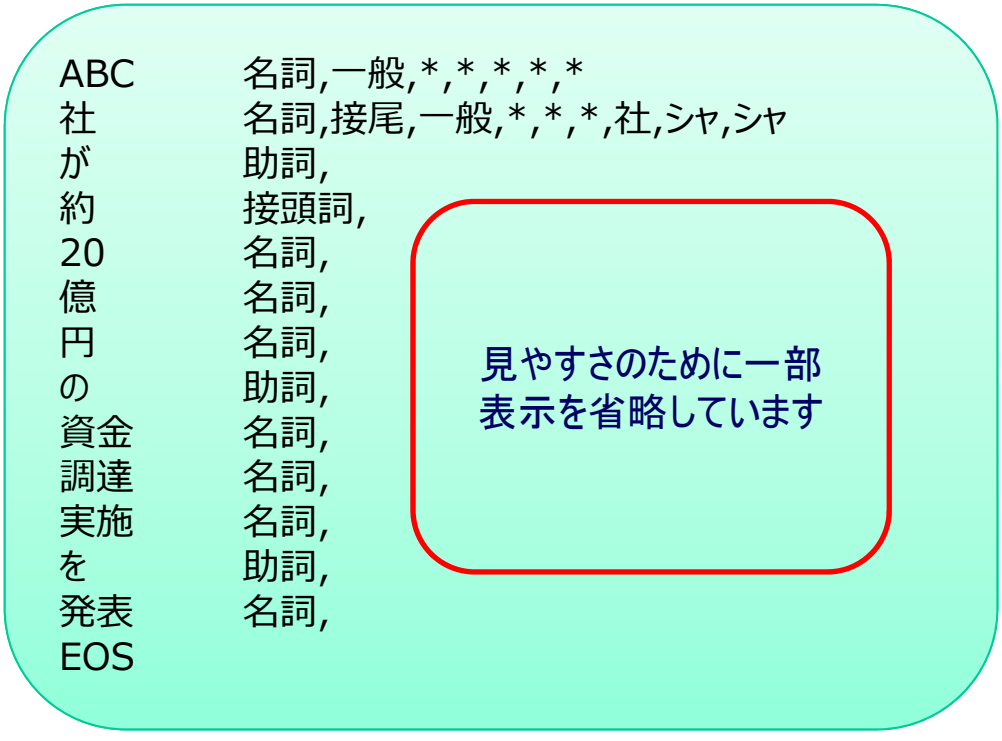
出所：茨城大学

＜分析の第一歩はテキストの単語分割＞

先ほど紹介した形態素解析は大量のテキストデータを統一的なルールの下で、自動的に品詞情報付きの単語列に分割してくれます。この形態素解析を行うフリーで利用可能なツールにはMeCab、Juman++、KyTea、Sudachiなどがあります。今回は広く使われているMeCabを利用して説明を進めます。例えば、架空の企業ABC社について以下のニュースがあったとします。

（ニュースタイトル１）ABC社が約20億円の資金調達実施を発表

このニュースタイトル１を、MeCabに入力して形態素解析を行います。このとき、MeCabの辞書には標準的な辞書であるipadicを用います。その解析された出力結果は以下のようになります。



出所：茨城大学

出力結果の各行は左側が分割された単語、空白を挟んで右側がカンマ区切りで「品詞情報、品詞細分類１、品詞細分類２、品詞細分類３、活用形、活用型、原形、読み、発音」の各情報となります。EOSはEnd-of-Sentenceの略で、文の終了を表します。

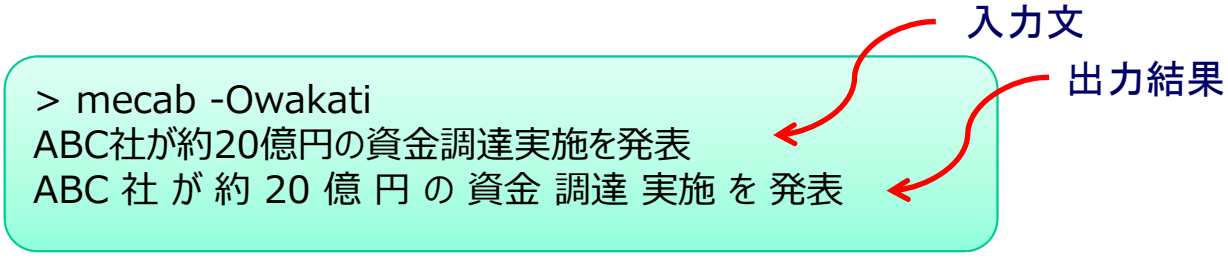
これらの情報を利用して、テキストの内容を表す特徴抽出を行います。内容を表す特徴として一般的に用いるのは、内容語と呼ばれる意味を持つ単語です。単語は内容語と機能語に分けることができます。内容語は単語自体が具体的な意味を持ちますが、文法的な役割は持っていません。機能語は逆に意味を持ちませんが、文法的な役割を持っています。内容語と機能語を品詞で分類すると、内容語は名詞、形容詞、動詞、副詞に相当し、機能語は助詞・助動詞・接続詞などが該当します。そのため、形態素解析の結果で品詞が名詞、形容詞、動詞、副詞と判定された単語を抽出することで、内容を表す特徴を取り出すことができます。上の例で、内容語を抽出すると、

「ABC、社、20、億、円、資金、調達、実施、発表」

出所：茨城大学

という特徴が得られます。

品詞で絞り込みをせず、テキストを単純に単語分割（分かち書き）することも可能です。MeCabでは"-Owakati"オプションを指定することで、品詞情報を出力せず、単語を分かち書きした単語列として出力することも可能です。



出所：茨城大学

分かち書きをした結果、単語の区切り文字として空白を入れて単語列が出力されます。すべての単語を使った単語列は、「AなどのB」で表現される上位下位関係などといった単語間の共起性や「ABC社」などのフレーズを抽出するために使われます。また、ニューラルネットワークを用いた単語間の意味的関係を特徴として持つ単語ベクトルの生成※、機械翻訳や自動対話などのシステムを構築するためのデータとしても使われます。

※本レター最終章＜単語の特徴を使ってテキストの特徴を捉える＞で具体的に説明します。

でも、MeCabの例のように単語の集合を特徴として抽出したときに疑問が生じます。それは「なぜ単語集合はテキストの内容を表す特徴とみなせるのか？」ということです。その理由は分布仮説と呼ばれるアイデアが根底にあるためです。分布仮説は1954年に提案されたもので、「同じ文脈で出現する単語は類似した意味を持つ傾向にある」ということを表します。例えば、以下の例文を考えます。

日立市の和菓子屋でお土産に「モーター最中」を買った。

この文にある「モーター最中」が何であるか分からないとしても、これまでの経験と「和菓子屋」や「お土産」の共起単語からどのようなものか推定することができると思います。この分布仮説に基づいて、テキストの内容を表す特徴はその中に含まれる単語の集合で表現することが可能だということになります。

このようにテキストを形態素解析することによって、得られた単語集合が内容的な特徴を表しますが、形態素解析は使用する辞書の違いで単語の分割結果が異なります。MeCabで利用可能な辞書には上述のipadicの他に、unidic、mecab-ipadic-NEologd および Comainu があります．これらの辞書の概要は以下のようになります。

表：MeCabで利用可能な辞書

ipadic	IPA品詞体系に基づき構築されたMeCabの標準的な辞書で、MeCabの辞書として広く使われる。
unidic	国立国語研究所が開発した、斉一な言語単位である短単位で書き言葉文書を自動解析するための辞書。
mecab-ipadic-NEologd (NEologd)	IPA辞書を拡張した新語、企業名や地名などの固有表現を効果的に抽出できる辞書。
Comainu	複数の短単位単語を結合した長単位解析を行うための辞書。イベント名や組織名など複数の名詞連続を効果的に抽出できる。

出所：茨城大学

実際に単語分割の結果がどれほど違うのか、以下の例文を用いて形態素解析結果の違いを示します。

（ニュースタイトル２）ドル・円下落、株安進行でリスク回避の動き優勢

このニュースタイトル２を、最も標準的なipadicを用いて形態素解析を行うと、単語分割の結果は以下のようになります。

(ipadic)
ドル ・ 円 下落 、 株 安 進行 で リスク 回避 の 動き 優勢

出所：茨城大学

「株安」の部分が一般名詞の「株」と接尾辞の「安」に分かれて出力されています。経済ニュースで、「安」や「高」といった接尾辞は、株価だけではなく為替相場でも出現します。「円安」をipadicで形態素解析を行っても「円」と「安」に分割されるため、「安」だけでは株価か為替相場か判断できず、「株」と「円」の出現数を比較する必要があります。

次に、unidicを用いて形態素解析を行います※。同じ例文を入力して形態素解析を行うと、以下のような単語分割になります。

(unidic)
ドル ・ 円 下落 、 株安 進行 で リスク 回避 の 動き 優勢

出所：茨城大学

分割結果がipadicと異なる部分は「株安」の部分です。「株安」はニュースなどで頻出するために、unidicでは辞書に登録され、普通名詞の一単語として抽出できます。

※unidicを使ったときの標準的な出力結果は、ipadicを用いた場合と比べて、発音と品詞情報の順序が異なっていますので、品詞情報を自動的に抽出する際には注意が必要となります。

今度はNEologdを用いて形態素解析を行います。

(Neologd)
ドル ・ 円 下落 、 株安 進行 で リスク回避 の 動き 優勢

出所：茨城大学

この場合は、「株安」だけではなく「リスク回避」がひとつの単語となっています。経済ニュースでは「株安」も「リスク回避」も頻繁に出現するため、一単語として扱う方が効果的ではないでしょうか。

最後に、長単位の単語を抽出できるComainuを用いて形態素解析を行います。

(Comainu)
ドル・円下落　、　株安進行　で　リスク回避　の　動き優勢

出所：茨城大学

Comainuの場合は、「ドル・円下落」「株安進行」「リスク回避」「動き優勢」が内容語として抽出されます。「ドル・円下落」や「株安進行」は有効なフレーズとして使えそうですが、「動き優勢」は他のニュースタイトルで出現する可能性が低いかもしれません。

それぞれの辞書に長所・短所があり、どの辞書を利用すればいいのか決めるのは悩ましい問題です。分析目的によって辞書を使い分け、有効な特徴を抽出できる辞書を選ぶことが大切だと思います。

＜単語以外で分析に使える特徴「係り受け情報」と「類義語」＞

単語以外で分析に使える特徴として、係り受け情報や類義語があります。係り受け情報は構文解析を行った結果として得られる文節間の修飾・被修飾の関係を表現したものです。この係り受け情報を抽出するためのツールにCabochaがあります。CabochaはMeCabで使用した辞書のうち、ipadic、unidic、NEologdが使用可能で、文を形態素解析をした後に単語間の修飾関係を解析します。Cabochaの解析例として、上と同じニュースタイトル２を入力して係り受け解析を行うと以下（次ページ）のようになります。

```
ドル・円下落、-D
        株安進行で---D
        リスク回避の-D
                動き優勢
EOS

* 0 1D 3/3 0.504642
ドル            名詞
・              名詞,
円              名詞,
下落            名詞,
、              記号,
* 1 3D
株安            名詞,
進行            名詞,
で              助詞,
* 2 3D
リスク回避      名詞,
の              助詞,
* 3 -1D
動き            名詞,
優勢            名詞,
EOS
```

見やすさのために一部
表示を省略しています

出所：茨城大学

出力結果は上に示すツリー構造と下に示すラティス構造などがあり、出力結果を選択することができます。ツリー構造は「ドル・円下落、」が「株安進行で」に係るように、係り先が視覚的に分かりやすく表示されています。一方、ラティス構造はアスタリスクが先頭にある行に文節に関する情報があり、文節番号と係り先の文節番号（**太字部分**）が表示されています※。

※3/3 0.504642などの係数についてはここでは説明を省略しました。詳しくはCaboChaの公式サイトなどをご参照ください。

類義語は意味の類似した異なる単語を表します。先ほどのニュースタイトル２において、「ドル・円下落、」の「下落」は類似した表現である「値下がり」を使う場合があります。この場合、「下落」も「値下がり」も同じ概念として扱うことができます。このような類義語の情報はシソーラスと呼ばれる類義語辞書を用いることで得られ、出現単語と同じ概念の単語を簡単に検索することができます。日本語では国立国語研究所が作成した分類語彙表、英語ではWordNetというシソーラスを使うことができます。

＜単語の特徴を使ってテキストの特徴を捉える＞

テキストを単語分割し、単語集合として表現することができれば、単語を特徴としてテキストを数値化することができます。その代表的な方法はベクトル空間モデルで、テキスト中の単語の重みを要素とするベクトルでテキストを表現するものです。単語の重みは単語の出現や頻度など様々な重み付け方法がありますが、今回は単語が出現すればそれに対応する要素を１、出現しない場合は０を与えてベクトル化を行います。例として、以下の２つのニュースタイトルをベクトルで表現してみましょう。

（ニュースタイトル1）ABC社が約20億円の資金調達実施を発表
（ニュースタイトル3）DEF社、製造業の調達購買業務を高度化

各ニュースタイトルをNEologdで形態素解析を行い、内容語を抽出すると以下のようになります。

（ニュースタイトル1）ABC 社 20 億 円 資金 調達 実施 発表
（ニュースタイトル3）DEF 社 製造業 調達 購買 業務 高度 化

「ABC」を１番目の要素、「社」を２番目の要素というように出現順にベクトルの要素を割り当ててベクトル化を行うと、各ニュースタイトルは以下のようなベクトルで表現することができます。

（ニュースタイトル1）{ 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0 }
（ニュースタイトル3）{ 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1 }

2番目の「社」と7番目の「調達」が重複していますが、それ以外はタイトル間で単語の重複がないため、内容は類似していないことが分かります。

【佐々木先生　プロフィール】

佐々木稔（ささきみのる）
徳島県徳島市生まれ．平成13年徳島大学大学院博士後期課程修了．博士(工学)．
平成13年茨城大学工学部情報工学科助手を経て，平成17年より同専任講師．
研究分野は，機械学習や統計的手法による情報検索，自然言語処理等に従事．情報処理学会，言語処理学会，計量国語学会，電子情報通信学会各会員．