

# 自然言語処理とエンタープライズ・サーチ

Benjamin Douglas\*

自然言語処理（NLP）技術がエンタープライズ・サーチの質に与える影響は少なくない。この技術は、単なるキーワード検索を越えた、より自然な検索を実現するため、言語検出、分かち書き、名詞句抽出などの言語学的な機能を提供する。ローエンドと見なされることがあるが、これらの技術を活用することにより、ランキングやファジー・マッチングなど高度な検索機能が可能になる。本記事は自然言語処理技術の紹介とエンタープライズ・サーチとの関係を説明する。最新の研究を紹介し、将来的なエンタープライズ・サーチへの影響を推測する。

キーワード：自然言語処理、ランキング、形態素解析、統計モデル、固有表現抽出

## 1. エンタープライズ・サーチとは

Google や Yahoo などが提供する、ウェブ検索の基本サービスは多くの人にとって身近なものだろう。一方、エンタープライズ・サーチは、組織内のドキュメントやデータベースにおける情報だけを検索・抽出するものである。

社内情報検索を活用する企業の事例はたくさんある。Amazon などの消費者向けサイトは、顧客に製品カタログを検索するサービスを提供する。新聞社などのメディア企業のサイトでは、読者が現在/過去の記事やその他のコンテンツを閲覧可能である。このような企業に限らず、FAQ（よくある質問）の提供や、社内情報の共有などのために検索機能を構築する場合もある。情報量が多く、人の手による整理が困難な状態において検索機能が役立つ。

エンタープライズ・サーチとウェブ検索は技術的に近いが、違いもある。ウェブ検索エンジンは静的なデータであるウェブページをインデックスするのに対し、エンタープライズ・サーチは、データベース内のコンテンツなどの動的な情報も検索できる。情報量が比較的少ないので、ウェブ検索エンジンと比べると、カスタマイズおよび先端的なデータ処理が可能。自然言語処理（NLP：Natural Language Processing）は、その先端的なデータ処理の一例で、言語学的な処理である。本記事では自然言語とエンタープライズ・サーチの関係を説明する。

## 2. エンタープライズ・サーチの代表機能

基本的にエンタープライズ・サーチ・エンジンは、多数のデータセットの中から特定のキーワードを含むドキュメントを探すことができる。検索エンジンで必要な情報を探

す方法は多くあるが、本セクションでは、主要検索エンジンが提供する標準検索機能の概要について述べる。

比較的小さいデータセットにおいても、検索結果は、効率的に順序付けられていることが重要である。通常は「関連性」にもとづいて順序付けられる。この場合、関連性とはドキュメント内にユーザの知りたいことが含まれる度合いを意味する。一般に、質の高い情報を含むドキュメントが、リストの上位にあることが望ましい。関連性を数学的に定義するのは難しいが、数理モデルを使えば実用的な近似値を求めることが可能である。

関連ドキュメントに表示されるキーワードが不明な場合、試行錯誤しながらキーワードを見つける。その際、代替クエリーを提供することでキーワードを探す手助けになることがある。入力や綴りを誤ったために、適合するドキュメント数（いわゆるヒット数）が少ない場合は、綴りが似ていて、使用頻度の多い言葉を候補として検索エンジンが提案するという機能がある。逆にキーワードが一般的すぎるため、ヒット数が多すぎる場合は、更なるキーワードを提案し、検索結果の絞込みをサポートする。

膨大なデータセットでは、必然的に情報の類似したドキュメントが複数存在する。同じ文章でも保存形式が異なっていたり（例：html と pdf など）、新聞社と通信社のように二社が同じ情報を公開していることがある。これらのケースでは、新しい情報が追加されるわけではないので、複数のソースの表示は検索結果を混乱させてしまう。その場合、すでに表示された類似ドキュメントを結果から取り除く、「duplicate removal」と呼ばれる技術を用いて、複数の類似ドキュメントの表示を避けることができる。

自然言語はとても奥行きが深いので、同じことがらを表現するのに、微妙にニュアンスの違う表現方法が複数存在する。例えば、「車」と「自動車」のような単語である。同じ意味を表す場合でも、文書によって異なる単語を使用することがあるため、本来の意味を含む完全なデータセットを見つけるためには、可能性のある単語をそれぞれ検索する必要があるが、ユーザは、「車」「自動車」などの同義語

\*ベンジャミン・ダグラス Basis Technology Corp. (サンフランシスコ支社)

ベイシス・テクノロジー (株) 〒102-0084 東京都千代田区二番町 9-6 パウ・エプタ 3F

Tel.03-3511-2947

(原稿受領 2009.6.16)

をわざわざすべて入力する必要がなく、検索エンジンが、クエリーに入力されたキーワードの同義語も自動的にマッチしてくれる。類似の事象として、「home」と「homes」のように微妙に違う単語であっても、同じコンセプトについて言及していることがある。これらのクエリー語は、可能性のある表記にも自動的にマッチできる。意味的に類似する単語マッチングは、「ファジーマッチング」と呼ばれる。

企業情報に関連がなくとも、スパムやアダルト向けコンテンツのフィルタリングが必要なケースがある。「スパムフィルタリング」や「アダルトコンテンツフィルタリング」と呼ばれ、ユーザ作成型インプットを許可するウェブサイトが主流である中、不適切な情報を検索結果から削除することができるのが特徴である。これは、質の高いコンテンツとそうでないものを区別するために、重要性を増している。

「自然言語クエリー」は他のアプリケーションでも有用である。例えば、話し言葉や書き言葉がクエリーとして使われる場合があり、標準的なキーワード検索の、「車 仙台」ではなく、「仙台では、どこで車を購入できるのか?」と文章で検索する。これは、特に、コンピュータ知識の十分でないユーザに焦点を置くウェブサイトでは有効な手段である。

### 3. エンタープライズ・サーチで使われる最新自然言語処理技術（基本概念説明）

- 言語検出
- 統計的分かち書き
- 基本化
- 名詞句抽出
- 固有表現抽出
- 構造解析
- 構文解析
- 品詞タグ

自然言語処理とは、人間の話し言葉をコンピュータに取り込む際に起こる問題を解決しようとする、人工知能の分野の一つであり、検索エンジンで使われる自然言語処理技術は、話し言葉ではなくテキストに基づくものである。このセクションでは、エンタープライズ・サーチに関連する様々な自然言語処理のコンセプトについて概説する。

まず、言語モデルのコンセプトを簡単に説明する。言語モデルとは、言語の統計的表示であり、いろいろな形で機能し、シンプルであったり、また複雑なものもある。シンプルな例として、単語のリストと相対度数がある。例えば、日本語のモデルでは、「は」の出現頻度が高く、「灰皿」の出現頻度は低い。このモデルは文章全体の出現頻度の把握ができ、テキスト解析に使われる。言語は曖昧なことが多く、こうしたモデルは二つの可能性のうち、どちらがふさわしいか判断するのに使われる。モデルは手動で生成されるが、機械の学習能力からアルゴリズムを借りることで自動的に生成(トレーニング)されることもある。大きいデー

タセットはこういった機械学習のアルゴリズムで解析すると、単語数だけでなく数多くの特徴を自動的に抽出することができる。いくつかの言語モデルについて以下で説明する。

世界の電子ドキュメントは、様々な言語と文字コードで書かれている。これらのドキュメントは、コンテンツの言語やコードでラベル付けされていなかったり、ラベル付けの情報が間違っていることもある。「言語検出」は電子テキストの文字コードを決定付けるプロセスである。問題への主なアプローチは、辞書ベースのアプローチと統計的アプローチの二種類がある。辞書ベースのアプローチでは、テキストの単語はそれぞれの、可能性のある言語がよく現れる特有の単語リストと比較される。例えば、「the」は英語で一般的な単語だけでなく、世界中の言語で極めて特有のものである。しかし、これら特徴的な単語を特に同系言語で見つけるのは難しい。一方、統計的アプローチは最新のもので、世界言語の統計的モデルの生成に基づいている。各言語の多数のドキュメントを解析し、統計的フットプリントを作成する。同様にフットプリントはサンプルテキストを取り込み、各言語の代表的フットプリントと比較する。このような統計的アプローチでは、より大きい可能性のある言語セットを考慮し、あらかじめ言語の語彙や文法を知る必要がない。

西洋言語では、表記の規則にもとづいて単語間にスペースが入るので、文章を単語ごとに分かち書きするのは難しくなく、スペースとスペースの間にあるテキストを一つの単語として定義できる。一方、単語間にスペースがない多くのアジア言語ではこのような単語定義は不可能である。「統計的分かち書き」とは、単語の始まりや終わりがどなのか、表記法の手がかりがない文章において、単語の切れ目を見つけ出す処理のことである。言語モデルのもう一つのアプリケーションでは、与えられた言語の文章の多くは、手動で分かち書きされたもので、単語の切れ目のパターン抽出に使われる。これら統計的パターンは、ブレンテキストで単語の切れ目を適合させるのに使われる。言語辞書はこれらモデルの有効性を増すために使われ、モデルのトレーニングでは見られなかった単語の切れ目を見つけることを可能にする。

単語は文法という一連のルールに従うことで文章形成が可能となる。表面上、文法は文章中で単語の置かれる位置を制限するが、より一般には、どのような種類の単語がどこに位置するかを制限する。これらの単語区分を品詞といい、名詞、動詞、形容詞などがそれにあたる。単語の品詞を知れば、文章中でどのように使われるかについてのヒントが得られる。また、体系化されていないテキストにおける文章構造を把握する際に非常に重要である。「品詞タグ」は自然言語処理の手法であり、文章中でマークされない単語の品詞を決定する。文法規則に基づくアプローチがこれまで用いられることが多かったが、一般的には統計的アプローチがより優れていると考えられる。統計的アプローチでは、付加的な品詞情報のある長い文章が、品詞ラベルを

元のテキストに割り当てることができるモデルのトレーニングに使われる。

先に述べたように、同じ単語であっても、本来の特性に基づいて異なった綴りでの表示が可能である。例えば、「青い」の連用形は「青く」で、また、「考える」の過去形は「考えた」となる。基本形とは、単語グループのベースとなる語形（辞書形）で、上記の「青い」や「考える」にあたる。ベースの語形（辞書形）でない単語を基本形に変換する処理のことを「基本化」と言う。英語のように表記の限定された言語では、基本化は簡単にできる。しかし、日本語やフィンランド語など形態素学的に複雑な言語では、語形が多すぎて、それらを辞書にリストアップすることは実質的にできない。この場合、単語の基本化には、部分的語形辞書と形態素学的組み合わせの規則を使われなければならない。

名詞は、言葉で意思相通しようとする「もの」を表すので、文章の意味を理解するうえで非常に重要である。しかし、名詞そのものがかなり曖昧なこともよくある。大抵、名詞の前後にある 2~3 語がその名詞の説明をし、文脈を与える。これらを名詞句という。例えば、「車」という単語を含む文章からは、「私の車」「盗難車」「中古車」とほぼ同じ文脈は見られない。これらの名詞句は、瞬時に文章の意味解釈をする短い文脈として機能する。そのため、品詞タグ付けとは別の解析、「名詞句抽出」という、自然言語処理の別の技術がある。これは文中の名詞句を識別するものである。これら名詞句は、何も加えられていない単語と比べ、それほど曖昧さがないので、非常に効果的な情報単位であることが多い。

固有名詞は、日常生活に結び付く名称を持つものである。最も分かりやすい例は人名で、そのほか、場所、企業、地域名などもある。テキスト中の固有名詞を識別する処理を「固有表現抽出」という。固有名詞は、会話中のトピックや中心人物であることが多く、ドキュメントの意味を理解するのに重要である。固有名詞は、名前の付けられていない名詞にくらべ、曖昧さが少ない。たとえば「岩」は一般的であり、「ジブラルタル」は固有のものである。名称のリスト（ガゼティアと呼ばれる）は文章中で固有名詞の識別に役立つが、十分なリストとなっていないことが多い。名称のパターンはしばしば一様でなく、他の語句に比べ、形態素学的ルールと異なる。そして、常に新しい名称が作り出されているので、静的リストの弱点を克服するために、統計的手法が求められる。

ドキュメント内には、全体的な意味把握に重要な部分と、そうでない部分がある。例えば、ヘッダーとフッターはサイトのすべてのドキュメントでまったく同じ形で、そのページの意味情報を伝えることはない。ドキュメントのテーマを判定する際は、こういう点は無視し、意味を持つ部分を重視したほうが良い。どの文章が重要かの判断は容易ではないが、テキストの位置とフォントの大きさ、文章の長さなどを見ることで推定できる。普通はフォントの大きさなどの規則を手動で書くにもかかわらず、データ量が

十分多ければ統計モデルを学習することも可能である。

構文解析をすることでドキュメントを構造化することができる。これにより、文章中のそれぞれの言葉が、どのように文法規則に準拠しているかを認識し、文章中のそれぞれの語の関係が明らかになる。そしてこの構造を理解することにより、さらに高いレベルの解析が可能となる。文法規則を使用することは可能だが、規則が多くて複雑な上、自然文は規則に合わないことが多い。よって、より柔軟な統計モデルが使用されることが多い。

#### 4. 検索機能と自然言語処理の関係

前述の自然言語処理の概念は一般的な言語学のもので、技術はそれぞれの問題を数学的に解決しようとする。本セクションでは、このような概念をエンタープライズ・サーチに応用して説明する。エンタープライズ・サーチは Google のようなウェブ検索に似ているが、エンタープライズ・サーチが持つ特徴ゆえに自然言語処理の技術をうまく利用できる。その主な特徴はデータの規模。前述の自然言語処理のアルゴリズムは高い性能が要求されるため、データ量が多過ぎるとコンピュータのリソースが不足してしまうが、少ない量のデータを扱う場合には、柔軟に処理を実行し、サーチ機能を向上することができる。また、エンタープライズ・サーチはデータのテーマが限られていることが多いので、自然言語処理技術のカスタマイズにより、より高い性能を出すことができる。

関連性順の並び替えは、検索エンジンの重要な機能の一つである。自然言語技術は、言語学的志向に関係なく、検索における高レベルの関連性処理を助ける。ドキュメントの言語や文字コードが判別されなければ、どの処理も始まらない。文字コードが分からないとドキュメントを一様に保存することができず、また言語が分からないと言語別の処理がおこなえない。ユーザの立場からみると、使用言語が読めなければドキュメントが理解できないので、言語判別は関連性に不可欠である。

検索処理の基本単位はトークンである。トークンは言語の最も小さい単位で、スペースがある西洋言語では抽出が容易であるが、日本語のように文章の単語の間にスペースが入っていない言語では、トークンは文字、あるいは文字列となり、言葉の真の意味が薄れてしまいがちで、関連性のある結果が出にくくなる。しかし自然言語処理の言語モデルを使うと、上質のトークンを自動的に抽出でき、関連性処理を向上させることができる。

自然言語処理技術を利用することにより、関連キーワードを抽出でき、絞り込み検索などに利用できる。複数のニュアンスを持つ言葉で検索すると、結果リストが混乱する可能性があるからである。例えば、「ボルト」で検索すると、電気や映画、五輪選手などを含むドキュメントが出てくる。自然言語処理で抽出された名詞句を集めることで、「ボルト」という言葉の付近にある文脈を表す言葉がわかる。例えば、あるデータセットでは、「ボルトとアンペア」「ウサインボルト」などという「ボルト」を含む名詞句の頻度が

多いだろう。このような言葉をクエリーに追加すると検索結果から無関係な情報を削除できる。また、「ボルト」の場合、五輪選手の「ウサイン・ボルト」や記者の「アンドリュー・ボルト」のように固有名詞の場合も追加検索の実施を推奨する。

自然言語クエリーは、品詞タグ付けのように、自然言語処理技術の利用により、その処理が向上できる。あらかじめ用意された質問パターンに対するクエリーを比較するのが、自然言語のクエリー検索インターフェースを作成するうえでは標準的な手法である。基準は、「Xはどこですか？」または、「Xについてもっと知りたい」のようなものである。ユーザが基準に沿った正確な表現方法でクエリーを入力すれば、キーワードは抽出され（上記の例の場合は「X」）、そして、クエリーエンジンに送信される。

## 5. 日本語特有の問題

- 複数の文字コード（Shift-JIS, EUC-JP, など）
- 分かち書き
- 複合語解析
- 非常に屈折した言語学的形態素論
- 複数の文字体系（平仮名, 片仮名, 漢字）
- 複数の表記体系（全角/半角, 漢数字/アラビア数字）

日本語は、検索に対して数多くの言語学的課題を投げかける。これらの課題は、日本語テキストのアルゴリズム処理特有の難しさに由来するものである。さらに、検索技術は、本来西洋言語を話す国々で設計、研究されたものであるため、他言語で使用するには、そのモデルを大幅に修正する必要がある。この課題とそれらを解決する自然言語処理技術を説明する。

テキスト表現にいまだに広く使われている旧来の文字コードが、日本語処理を複雑にしている。プラットフォームまたはアプリケーションによって、テキストは Shift-JIS, EUC-JP, ISO-2022-JP などの文字コードで保存される。その結果、文章を日本語と判別するためには、文字コードも判別しなければならない。この情報はしばしば、ドキュメントそのものから消えていたり、間違っていたりすることがある。自然言語処理技術を使う統計的手法は、正確な識別を確実におこなう最善の方法といえよう。

最新の検索エンジンはすべて、意味のあるトークンが容易に入手できるという仮定のもと、基本的検索単位としてトークンを使う。これは単語間にスペースがあり、表記文字が限定されている西洋言語には有用だが、日本語には当てはまらない。統計的分かち書きは、標準に従って（もしくはおおよそ従って）トークンを識別するのに使われるが、矛盾する標準が存在する場合がある。日本語では、信頼性のある表記に統一するという習慣がないので、単語の定義とは何かというトピックは極めて議論の余地がある。「印象的」という語句の「的」は、それ自身で一つの語句なのか、それとも、前の語句の一部なのか？「動的」はどうだろうか？これらの例は、テキストの自動的分かち書きを困難に

するだけでなく、「正しい」回答が不明なことさえある。検索に関して、不可分の検索単位を持つという習慣は日本語を話す者にはない。例えば、「東京都内」という語句の場合、日本語を解す読者ならば、「東京」「東京都」「都内」これらすべての語句が「東京都内」に当てはまると考える。しかし、全ての単語を抽出するために、重複させずにフレーズを分割するのは不可能である。これらの問題を軽減するために、日本語検索エンジンでは、一つ目は分かち書きで、二つ目は N-gram でというように、データを二重に索引付けすることがある。分かち書きによって、同義語や名詞句抽出などの高レベルな処理が可能となる。一方、N-gram は、分かち書きが必然的に間違えるケースを見つけるのに役立つ。

分かち書きに関する課題に、長い語句または固有名詞の複合語分割がある。例えば、「世界遺産」という語句を見てみると、二つの総称的名詞で構成されているが、組み立てると、二つのそれぞれの意味よりもさらに具体的な意味のある語句になる。これは具体的な概念なので、「世界遺産」を一つの単語と捉える。しかし、一般的に「遺産」を検索して、「世界遺産」が見つからないのは都合が悪いだろう。この場合、ユーザは語句の関連性と同時に、単語単位の再現率も求める。高レベルの処理においては複合語がトークンとして扱われ、また基本検索では単語がトークンとして使用されるようになっていけば、適切な処理ができる。

日本語の屈折形態変化は、単語の標準定義の欠如と相まって、日本語テキストの分かち書き特有の課題をもたらす。大抵の人は、「行った」は「行く」を基本形とする一つの単語であるとするが、他の形になると意見が分かれる。「行っている」は、一語（基本形「行く」）だろうか？または二語（基本形は「行く」と「いる」）なのか？「行ってもらった」はどうか？言語的観点から語句の形式だけでなく、ユーザが期待する標準とは何かという検索的観点からも、これは議論の余地がある。「行っている」を二語であるとみなすように、単語をより短い断片へと分割するならば、ユーザにより多くの検索結果をコンピュータ計算して示すのは容易である。

日本語の特異点として「表記ゆれ」があり、たとえば「皮膚」という単語は、「皮ふ」あるいは「ヒフ」と書かれることがある。「異体字」の場合も表記が異なり、それらを正しく扱うための特別な処理が必要なことが多い。このような表記ゆれをリストアップし、そのリストを利用・管理することが肝要である。

## 6. エンタープライズ・サーチの将来

ここまでの、エンタープライズ・サーチエンジンの標準機能を紹介し、自然言語処理とどのような関係があるかを説明してきた。将来はどのような機能が標準になるだろうか。今期待されるテクノロジーや先端技術研究がヒントとなる。

Google などの検索エンジンは、検索語の前と後ろにある語句をドキュメントのタイトルと同時に表示するが、将来

的にはその代わりに意味を持つ実際のコンテキストを表示するようになるだろう。自動要約という研究分野は、テキストの全体的な意味を簡潔な文章にまとめようとするものである。ドキュメントの要約を読むことでユーザが関連性をいち早く判断することができる。

ウェブで人や製品に関しての一般的な意見を探すことが良くある。自動要約の下位分野は感情解析といい、ドキュメントの執筆者の意見が良いものか悪いものかなどを判別し、レストランや製品の評価に役立つ。現在でもテキストが明確で強い意見を表すと自動的に識別できる。近いうちに、曖昧な表現を使っても、矛盾している意見を挙げても、正しく解析されるようになることが期待される。

「コンテキスト・クラウド (Context cloud)」という検索結果表示法も使われ始めている。検索語に関係する言葉を表示する方法だが、リストという形ではなく、不定形なクラウド (cloud) の形で表示される。関連性が強いほどフォントが大きく表示され、強い関係を持つ言葉はより目立つ。リスト形と比べると瞬時に読める情報量が多くなり、フォントの大きい言葉がユーザの求める言葉と合わない場合は、フォント大きさを順に目を通すことになる。ユーザが、クラウド表示の方が検索結果を読む効率が上がると判断すれば、今後さらに多くの検索エンジンがこの表示法を採用するだろう。

エンタープライズ・サーチにおける将来的傾向として、自然言語処理への依存があげられる。自然言語処理技術の利用により、さらに高レベルな処理が可能になる。自動要約と感情解析は自然言語処理研究の最新のテーマで、これら技術の検索エンジンへの採用はその研究の成功に負うことであろう。コンテキスト・クラウドには検索結果表示のために関連性の概念が利用され、自然言語処理技術の改善

とともに関連性が向上し有用なものとなると思われる。以上のような考察から、将来の検索エンジンの不可欠な機能は自然言語処理と密接に関連すると言っても過言ではないであろう。

学術的なプロジェクトだけでなく、商業的に自然言語処理技術を専門とする企業が多数ある。Basis Technologyはその分野のトップ企業であり、10年以上にわたり自然言語処理関連ソフトウェアを開発・提供している。Google, Amazonなどの顧客にとって検索機能は欠かせないものであり、Basis Technologyの言語処理ソフトウェアを使用することで、その検索に付加価値を付けることができる。

#### 参 考 文 献

- 1) Langville, A., Meyer, D. Google(TM)'s PageRank and Beyond: The Science of Search Engine Rankings. Princeton, NJ: Oxford University Press. 2006.
- 2) White, M. Making Search Work: Implementing Web, Intranet and Enterprise Search. Medford, NJ: Information Today. 2007.
- 3) Spink, A., Lerner, A., Zimmer, M. Web Search: Multidisciplinary Perspectives (Information Science and Knowledge Management). New York, NY: Springer. 2008.
- 4) Croft, B., Metzler, D., Strohman, T. Search Engines: Information Retrieval in Practice. Reading, MA: Addison Wesley. 2009.

ベンジャミン・ダグラス (Benjamin Douglas)

現在、自然言語処理ソフトウェア開発の専門企業、ベイス・テクノロジー (サンフランシスコ支社) のプリンシパル・エンジニア。エンタープライズ・サーチ開発の経験もあり、検索エンジンで日本語処理を可能にするソフトウェアの設計と実装に携わる。2年間、日本でソフトウェア会社勤務の経験がある。

**Special feature:** Enterprise search. Natural Language Processing in enterprise search. Benjamin Douglas (Basis Technology Corp. c/o Basis Technology K.K. 9-6 Nibancho, Chiyodaku-ku Tokyo 102-0084 JAPAN)

**Abstract:** Natural Language Processing (NLP) has a potentially large influence on the quality of Enterprise Search systems. These technologies enable going beyond simple keyword search and providing a more natural search experience. NLP is a set of technologies that implement linguistic features such as document language identification, tokenization and noun phrase extraction. Integrating these low-level functions into a larger search application enables higher-level functions such as relevance ranking and fuzzy matching. This article provides an introduction to NLP technologies and their relationship to Enterprise Search. Current research in the field will be introduced and its likely effect on the future of Enterprise Search will be inferred.

**Keywords:** Natural Language Processing (NLP) / ranking / morphological analysis / statistical models / entity extraction