

# Rosette 形態素解析システム

## 形態素解析により、多言語文書の全文検索を行うことができます。

Rosette® 形態素解析システム (RBL) は、文節処理、基本形化、複合語分解などの必要不可欠な言語サービスを提供することで、情報検索およびテキストマイニングのアプリケーションで、多言語文書を処理できるようにします。

自然言語ごとに固有の問題が存在するため、RBLには複数のテクノロジーが搭載されています。東アジアの言語の場合、正確な検索結果を生成するためには、適切な文節処理が必要不可欠になります。RBL は、形態素解析を使用してこれを実現しており、句読点、接辞、活用語、単語の辞書形式など、特定言語の固有の特性を分析します。

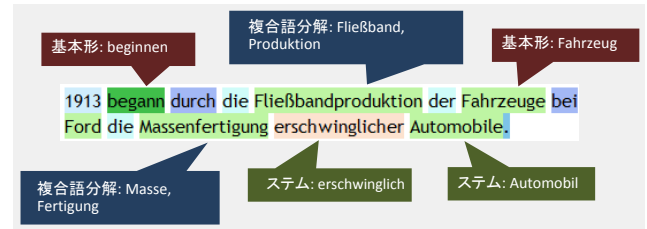
n グラムや基本形出力などの単純な方法より正確な結果を生成することができます。各言語を深く理解することで、定期的に辞書を更新して RBL を改善し、また学会で提唱された新しいアプローチを継続的に評価しています。

「Google は、日本語・中国語・韓国語版の最良サーチエンジンを作るべく、アジア言語処理技術を提供する Basis Technology を選びました。これは、世界中のインターネット利用者にもっとも好まれるサーチエンジンとしての Google の地位を築くための重要な一石を投ずることとなりました。」

— Google シニア・ヴァイス・プレジデント, Urs Hölzle 氏

### メリット

- 言語対応文節処理により、後続のテキスト処理の確かな基盤が得られます。
- 検索エンジンの機能には、基本形化、複合語分解、カスタマイズ可能な辞書などがあります。
- RBL は、スループットと拡張性が高いため、主要な Web 検索エンジンやエンタープライズ・サーチ・エンジンで採用されています。



### 新機能

- 文節処理は、中国語、日本語、朝鮮語などの単語間にスペースがない言語を自動分析する際に必要になります。
- 基本形化は、各単語の辞書形を生成します。これにより、検索の関連性を向上させたり、(すべての屈折形(「cruising」、「cruised」)ではなく見出し語(「cruise」)のみを索引付けすることで) 検索索引をスリム化したりします。
- 複合語分解は、複合語を構成要素に分割します。これにより、ドイツ語や朝鮮語の検索の関連性が向上します。
- 品詞タグ付けは、基本形化時に、曖昧な単語の正しい辞書形を選択するために使用されます(例えば、「spoke」が名詞か動詞か)。
- 文境界の検出は、文頭および文末を検出します。

### カスタマイズ可能な機能

ユーザーは、以下の機能を使用して、データのニーズに応じて Rosette をカスタマイズできます。

- ユーザー辞書。この辞書に、データ固有の単語規則や文節処理規則を指定することで、文節処理をカスタマイズできます。
- 中国語字体変換システム。これにより、汎中国語検索を行うことが可能になります。中国語話者は、簡体字と繁体字の両方の文書を単一のクエリーで検索して、結果を設定した字体で表示できます。このモジュールは、文字ベースではなく単語ベースで変換することで、高い精度を実現しています。
- 日本語表記ゆれアナライザー。辞書ベースの正規化モジュールであり、非標準の旧漢字を現在の形式に変換し、またカタカナの綴り字のゆれを正規化します。ユーザーは、必要に応じて、正規化辞書を拡張できます。

## 英語の場合

言語解析は、非英語だけではなく、英語の場合にもメリットがあります。基本形化 — 単語の辞書形を検出して、**関連したクエリー**を追加することで検索を拡張します。これは、従来型の基本形出力では実現しません。

| 検索クエリー   | 従来型基本形出力 | RBLを使用した基本形化                        | 比較  |
|----------|----------|-------------------------------------|---|
| animals  | anim     | animal                              | 2つの無関係な単語が同じ語幹(この例では「anim」)を共有する可能性があります。 |
| animated | anim     | animate                             |   |
| several  | sever    | several                             | 基本形出力が意図しない結果になる可能性があります。                 |
| children | children | child                               | 不規則動詞および名詞が原因で、基本形出力モジュールがうまく機能しません。      |
| spoke    | spoke    | speaks (動詞の過去形の場合)<br>spoke (名詞の場合) |   |

## 中国語、日本語、朝鮮語の場合

これらの話者人口の多い言語は、単語間にスペースを入れることなく記述されます。形態素解析には、バイグラムやn-グラムの手法と比較して、索引サイズの低減、検索関連性の向上など、多くの利点があります。「**北京大学生物系**」(北京**大学**生物**学部**)を索引付けする場合の問題を考えてみます。

従来型の文節処理では、以下のように、索引に6つのバイグラムが追加されます。

| term position | 1  | 2  | 3  | 4  | 5  | 6  |
|---------------|----|----|----|----|----|----|
| bigrams       | 北京 | 京大 | 大学 | 学生 | 生物 | 物系 |

Bigram segmentation produces **non-words** and words which are **incorrect in this context**

「学生」を検索すると、「北京大学生物学部」がヒットしますが、これは正しくない結果です。

形態素解析では、以下のように、索引に2つの正しく文節処理された単語が追加されます。

| term position | 1    | 2   |
|---------------|------|-----|
| tokens        | 北京大学 | 生物系 |

「学生」を検索すると、「北京大学生物学部」にはヒットせず、これは正しい結果です。

## アラビア語の場合

アラビア語は、高度に屈折した言語です。単語の先頭、中、および末尾に接辞が付加されるため、文字列による完全一致検索では、多くの関連するヒットが検出できなくなってしまいます。RBLは、以下の処理により、検索の再現率と適合率を向上させます。

- 正規化。スタイルの違いや綴り字の間違いのためにゆれているアラビア語の単語の綴り字を標準化します。
- 基本形化。関連した一致を検出します(「book」を検索した場合の「two books」や「my books」など)。



## ドイツ語および朝鮮語の場合

デンマーク語、オランダ語、ドイツ語、朝鮮語、ノルウェー語、およびスウェーデン語では、複合語を自由に作成することができます。索引付けのために、この複合語を分解する必要があります。

例えば、ドイツ語で、「**Samstagmorgen**」(「土曜日の朝」)は、「**Samstag**」(「土曜日」)と「**Morgen**」(「朝」)から成る複合語です。「**Samstagmorgen**」を複合語分解することで、「**Samstag**」(「土曜日」)の検索時に、この複合語に一致させることができます。

## 対応言語

RBLは現在、以下の言語に対応しており、他の言語についても現在開発中です。

|          |         |         |
|----------|---------|---------|
| 日本語      | ギリシャ語   | ハンガリー語  |
| 朝鮮語      | クロアチア語  | フィンランド語 |
| 中国語(簡体字) | スウェーデン語 | フランス語   |
| 中国語(繁体字) | スペイン語   | ブルガリア語  |
| 英語       | スロバキア語  | ヘブライ語   |
| アラビア語    | スロベニア語  | ペルシア語   |
| アルバニア語   | セルビア語   | ポルトガル語  |
| イタリア語    | タイ語     | ポーランド語  |
| インドネシア語  | チェコ語    | マレー語    |
| ウクライナ語   | デンマーク語  | ラトビア語   |
| ウルドゥ語    | ドイツ語    | ルーマニア語  |
| エストニア語   | トルコ語    | ロシア語    |
| オランダ語    | ノルウェー語  |         |
| カタロニア語   | パシュトゥ語  |         |

詳細 [www.basistech.jp](http://www.basistech.jp) お問い合わせ [info@basistech.jp](mailto:info@basistech.jp) 電話 03-3511-2947

〒102-0084  
東京都千代田区二番町9-6

One Alewife Center  
Cambridge, MA 02140

2553 Dulles View Drive  
Herndon, VA 20171

171 Second Street  
San Francisco, CA 94105

