第3章 記事の要約や新たな価値の導出に

文書を構造化するための単語オブジェクトの生成法

高橋 和弘

図1　単語Bが頻出する
この文書では無視せずに何らかの評価をすべき単語（結果的に無視することもあるが）と推定される

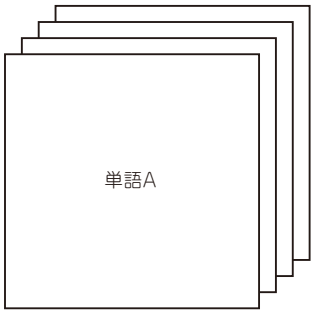

図2　単語Aが単独で出現する
他文書では出現しなければレア度の高い単語と推定される

自然言語処理は，コンピュータと人間をつなぐ機能として，さまざまな取り組みが行われてきました．今ではスマートフォンの普及によって音声認識など自然言語処理の機能を意識せずに日常的に使っています．しかし，ビジネス領域での自然言語処理の利用は，統計的な手法や文法的な手法の領域を超えられていません．コンピュータで日本語が使えるようになってから半世紀の時間を経て，自然言語処理は大規模言語モデルによる生成系AIの登場によって，その限界突破が可能となってきています．本稿では，この生成系AIを使用した自然言語処理の新たなトライアルについて紹介します．

これまでの自然言語処理

● 日本は構造化された文書でないケースが多い

ビジネスの現場で，日々多くの文書を処理するためには，その文書を理解する必要があります．それには与えられた時間や専門性など限定されたリソースの範囲で，これを行う必要があります．

文書の理解の効率を高めるには，その文書の中心となる部分を探し出す必要があります．論文などのように，パラグラフ・ライティングで意味構造をデザインしながら執筆された文書であれば，中心となる部分を比較的容易に探し出すことができるでしょう．

しかし，“行間を読む”ことが求められるハイコンテクスト言語の日本語の場合，一般的なビジネス文章は構造化されていないケースが多く，文書理解の効率が低下しているのが実態です．

● 文書内の単語の重要性の評価

日本語自然言語処理の1つの取り組みとして，文書中の重要単語を分析・探索し，文書中の単語の出現頻度や単語間の関係に着目した評価が行われています．

▶文書内の出現頻度（単語頻度，Term frequency）

単語の重要度の評価として，1つの文書に頻出する単語に着目する手法です（**図1**）．重要な単語は文書中に数多く出現するという仮説に基づく評価方法です．ただし，単語の出現頻度の多い単語の中には，重要ではない単語も含まれるので，これらを除去する必要があります．また，書き手側で意図的に出現頻度をコントロールされてしまう可能性もあります．

▶文書間の出現頻度（逆文書頻度，Inverse document frequency）

複数の文書の中で唯一出現する単語を評価する方法で，レア度の高い単語という評価になります（**図2**）．しかし，同じ意味にもかかわらず使われる単語が異なるケースも出てきます．また，意図的に違う単語が使われる場合もあるため，単純評価が難しくなってきます．