2S-04

異分野テキストデータを対象とした同一単語の分野間における意味の差異抽出方式

田丸 翔大† 岡田 龍太郎† 中西崇文†
武蔵野大学データサイエンス学部データサイエンス学科†

1. はじめに

現代の知識社会において，自身のアイデアを整理，創出，発信することは非常に重要である．一般的に，新たなアイデアを創出する際に自身が持つ知識だけではなく，異種異分野の知識を組み合わせることが有効とされている．異種異分野の知識を組み合わせるのに必要なことは，異種異分野の人とのコミュニケーションによる相互理解が必要となる．しかしながら，異種異分野の人とのコミュニケーションの場合，背景知識の違いによって，使う用語が異なっていたり，同じ用語でも違う意味で用いていたりといった分野依存での用語の用法の違いが存在しており，これらを明確することが，異種異分野の人とのコミュニケーションにおいて重要と考えられる．

本稿では，異分野テキストデータを対象とした同一単語の分野間における意味の差異抽出方式について示す．本方式は，異種異分野ごとのテキストコーパスを対象として，各コーパスのどちらにも用いられている共通の単語について，分野ごとのその単語に関する類似単語を比較することで，全く異なる意味で用いられている単語を判別し抽出することが可能となる．本稿では，Wikipedia におけるカテゴリを分野として捉え，生物学と情報学の 2 分野において，両方の分野において使われている単語でかつ意味に違いのある単語を抽出することを実現した．本方式は，同じ単語であっても別の意味で用いられる場合をあらかじめ把握でき，コミュニケーションギャップを防ぐことが可能になる．

2. 関連研究

瀬端[1]らは，シラバスの文書を対象として，各文書の単語の出現頻度の様相から，異文書間に出現する同一単語の機械的な意味の差の抽出を実現している．

Extraction Method of Semantic Differences between Disciplines for Text Data
Shota Tamaru† Okada Ryotaro† Nakanishi Takafumi†
† Musashino University, Department of Data Science

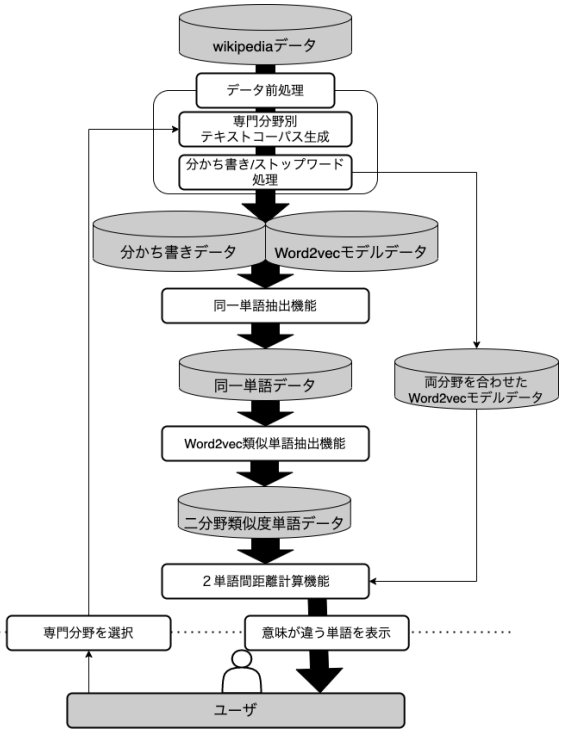

図 1：同一単語の分野間における意味の差異抽出方式

森永ら[2]は，Word2vec の分散表現を用いたカテゴリの推定による解釈機能を有する逆引き辞書の実現方式を示している．

本方式は，異分野テキストデータを対象とした同一単語の分野間における意味の差異抽出を実現することにより，同じ単語であっても別の意味で用いられる語を明らかにすることを可能としている．

3. 同一単語の分野間における意味の差異抽出方式

本章では，図 1 を用いた同一単語の分野間における意味の差異抽出方式について示す．本方式は大きく分けて，データ前処理，同一単語抽出機能，Word2vec 類似単語抽出機能，2 単語間距離計算機能で構成される．データ前処理は，専門分野別テキストコーパス生成，わかち書きストップワード処理で構成される．これらの処理に

よって，ユーザに選択された2つの分野のテキストコーパスが生成された上で，テキストコーパスそれぞれにおいて，分かち書き，ストップワードによる不用語削除を行う．データ前処理によって作られたテキストコーパスを用いて，それぞれWord2vecモデルデータ化を行う．また，2単語距離計算機能に用いるため，両分野を合わせたWord2vecモデルデータも作成する．本稿では，テキストコーパスの元データとしてWikipediaデータを用いる．Wikipediaデータのカテゴリーデータから専門分野を検索することで専門分野別に分けることができる．

3.1 同一単語抽出機能

分野間における意味の差異を抽出するためには，両分野に出現する同一単語を取得する必要がある．2つの分野のコーパスからそれぞれ共通して出現する単語を抽出する．

3.2 Word2vecによる類似単語抽出機能

3.1節で得られた同一単語について，情報学，生物学それぞれについてWord2vecにより上位1つの類似単語を抽出する．つまり，同一単語について情報学における類似単語と生物学における類似単語1つずつが抽出されることになる．

3.3 2単語距離計算機能

2単語距離計算機能は，3.2節で得た2つの類似単語を両分野を合わせたWord2vecモデルデータを使い距離計算を行う．これを行うことにより，3.1節で得た同一単語は両分野で同じ意味として使われているのか，別の意味を持っているのかを数値で判断することが可能となる．

4. 実験結果

図1で示す方式を実現する実験システムを構築し，生物学と情報学との単語の違いを導出することを試みた．

生物学と情報学の二分野において3.3節の距離計算の結果，0.7以下の数値を出した同一単語のみを抽出し，類似度計算結果をソートした結果を表1に示す．

表1より，『食べ物』，『ヒント』，『コンピュータネットワーク』，『具合』が抽出された．本結果から，これらの単語の両分野における用途の違いを考察したところ，『食べ物』は，情報学において，食べ物そのものに関する情報や，栄養情報学と関連があることが分かった．それ対して，『食べ物』は，生物学において，味覚や好き嫌いが関連していることが分かった．『ヒント』は，両分野ともに，処理の仕方や考え方についてのヒントという文脈で用いられており，アプローチの仕方に対する違いであることが分かった．『コンピュータネットワーク』

表 1：同一単語類似度計算結果

	同一単語	生物学類似単語	情報学類似単語	類似度計算
0	食べ物	要素	瞬く	0.579836
1	ヒント	3番	NDC	0.571053
2	コンピュータネットワーク	モルフォゲン	扱え	0.468026
3	具合	体細胞分裂	掛け	-0.571765

は，情報学において，コンピュータネットワーク構成学といった学問を表しているのに対し，生物学において，『ゲノムネット』や生物化学，情報科学の融合分野の一つである『バイオインフォマティクス』が関連していることが分かった．『具合』は，両分野ともに染色具合や会話の盛り上がり具合といった，様々な言葉の後ろにつける使われ方をしており，特定の概念を指していないことから類似度が低くなっている．

これらのことから，食べ物やコンピュータネットワークといった同一単語で意味の異なる単語を抽出することが可能となった．一方，『ヒント』や『具合』といった文章や単語に後付けすることで成立する単語も抽出されやすいことが分かった．これにより，生物学と情報学でコミュニケーションギャップになりうる同一単語でかつ別の意味で用いられる単語を抽出することが可能になった．

5. おわりに

本稿では異分野テキストデータを対象とした同一単語の分野間における意味の差異抽出方式について示した．本方式により，Wikipediaにおけるカテゴリを分野として捉え，生物学と情報学の2分野における同一単語における意味の違いの抽出を実現した．本方式を用いることにより，同じ単語であっても別の意味で用いられる場合をあらかじめ把握でき，コミュニケーションギャップを防ぐことが可能になる．

今後の課題として，専門分野だけでなく，TwitterなどSNSデータを用いた個人間の意味の差異抽出．インタラクションシステムであるチャットボットと組み合わせ，異分野の知識を組み合わせたアイデア創出の手助けとなるシステムの開発などが挙げられる．

参考文献

[1] 瀬端 賢人,中島 克也,小林 亜樹,シラバス中の組織間での単語意味揺らぎの分析，第78回全国大会講演論文集，pp.871-872，2016.
[2] 森永雄也，山口和紀，カテゴリ情報を付与した文の分散表現による逆引き辞書の精度向上，SIG-AM，16(01),pp.1-8，2017.