潜在意味解析による推移的特徴語の発見と応用

中山　瑛李†　　三浦　孝夫†
† 法政大学理工学部創生科学科

1．前書き

特徴語は時代や扱う文章によって変化する．不要語リストは固定ではないため，本研究では流動的に入れ替え続ける方法を提案する．潜在意味解析を用いて文脈を考慮した語を抽出する．さらに，情報検索で議論されている TF-IDF で高速かつ効率よく判定し，単語に重みづけをする．

2．潜在意味解析と TF-IDF

まず，文書ごとの単語出現頻度から単語-文書行列Dを作成する．次に特異値分解$D = U\Sigma V^t$(Uは単語の固有行列，Vは文書の固有行列，Σは対角行列)を行う．そして，より安定した語共起を得るために次元削減を行う．これにより新たに得られたD'に対して単語相関行列，文書相関行列を計算する．

TF は単語の出現頻度を表し，IDF は逆文書頻度を表す．高速かつ効率よく特徴語判定をするには，情報検索で議論されている TF-IDF を利用することが効果的である．

3．実験

本研究で対象となるコーパスは、CD-毎日新聞 2017 年版に採録されている 1 月 1 日から 1 月 14 日までの 2 週間分のデータ（2443 文書，名詞 547 語）から，文書サイズが 10 未満の文書，そして代名詞を除いた 2068 文書，537 語を用いる．また，不要語リストは SlothLib が公開している語を利用する．

本研究では 4 つの実験を行う．まず 1 つ目に不要語がどのように推移するのか，2 つ目に重要語がどのように推移するのかを実験する．また 3 つ目に，類似文書がどのように推移するのか，4 つ目に、独立文書がどれほど特徴的であるか実験する．

推移した特徴語は，今回使用したコーパスにおける特徴語である．表 1 と 2 は主な結果を示す．

表 1　不要語推移（一部例）

スタート不要語	推移→	
話	思い	父
男	捜査	逮捕
新た	国	必要
目	言葉	声

表 2　重要語推移（一部例）

スタート重要語	推移→	
町	復興	災害
龍	海	黒
放送	作家	毎日
契約	介護	サービス

捜査や逮捕といった単語は新聞では頻出するため不要語と判断されたといえる．「黒」は将棋の先手の意味で使われるように，色とは別の意味を持つ「黒」が表れ重要語と判断された．一方，色での「黒」を扱っている文書もあり，完全に重要語かどうかを判断するのは難しい．

6．結論

本研究において，不要語が TF 3.55 以上，IDF 1.56 以下を満たし，重要語が TF 3.55 未満，IDF 1.48 以上を満たした．このため，潜在意味解析と TF-IDF を用いた推移的特徴語の抽出ができたといえる．しかし同じ語が使われていたとしても，文脈が異なる場合もある．潜在意味解析と TF-IDF だけでは完全に特徴語かどうかを特定するのは難しく，更なる条件が必要である．

参考文献

[1]SlothLib　不要語リスト
http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt
[2]猪原敬介(2009)「LSA(Latent Semantic Analysis)の概要と実行」
http://cogpsy.educ.kyoto-u.ac.jp/personal/Kusumi/datasem09/inohara.pdf