不要語は不要か

根津　裕太†　　三浦　孝夫†
†　法政大学理工学部創生科学科

1. 前書き

情報検索において欠かせない前処理の一つに、不要語の消去が挙げられる。不要語とは文字通り「情報検索において不要な語」であり、SlothLib が一般公開している不要語リストが代表例である[1]。不要語の定義は厳密に規定されておらず、一般的には括弧や演算子、数字などといった「記号」、単体では意味を待たない「付属語」、さらには、前記の条件には含まれないが文面での出現頻度が多く、文書を特徴づけうる語には到底なりえない「非特徴語」が不要語として扱われている可能性がある。

「非特徴語」の選定には明確な定義がなく、極めて抽象的な概念となっている。SlothLib の不要語リストに含まれる単語はまさしく「非特徴語」であり、なぜこの単語群が不要語と判断されているのかは不明である。不要語は果たして本当に不要なのか？　本研究では、この疑問を解消するために、単語出現頻度(tf)、逆文書頻度(idf)を用いて、不要語の特性を明確化し、SlothLib の不要語リスト内の有効性を検証する。

2. 不要語指標

本稿では、不要語の判定基準として不要語指標 tf-tfidf を用いる。この値が高い単語を不要語と定義し、SlothLib 不要語リストと比較して、この条件の有効性を検証する。具体的な数式は①である。

$$\log tf_{\omega} - tfidf_{\omega} \cdots ①$$

$$tfidf_{\omega} = \frac{tf_{\omega}}{df_{\omega}}$$

ここで、tf_{ω}は全文書に含まれる単語ωの数であり、df_{ω}は単語ωを含む文書の数である。不要語とその他の単語との閾値については、全文書に含まれる単語のうち、不要語指標の上位 10%の値とする。

3. 実験

本研究で対象となるコーパスは、毎日新聞の 2016 年度版のデータ 1 年分（記事数：83669 件、総単語数 12698459 語、244470 種類）である。また、不要語リストとして SlothLib が公開している 301 語を用いる。

表 1 条件を満たす単語と満たさない単語

条件を満たす単語	日,こと,年,人,ため,後,中さん,月,円,万,第,回,今,時,前,者,的,歳,目,これ,以上,もの,たち,時間,関係,国,自分,それ,私,分,何,上,元,ぶり,点,今回,うち,億,ほか,力,多く,間,以降,次,ところ,手,一つ,場合,新た,先,面,方,同じ,内,度,部,会,カ月,他,その後,化,ごろ,近く,際,ここ,下,感,まま,各,気,性,別,話,市,数,それぞれ,一,形,場,家,法,どこ,体,台,口,誰,はず,男,名,そこ,地,わけ,金さまざま,以下,外,二,道,方法,町,すべて,ごと,三,とき,左,夏,県,高,子,店,系,名前,右,水,みんな,兆,前回,用,五,七,等,六,八,頃,四,例,時点,あと,式,所,毎日,違い,個,彼,以前,校,春,たくさん,係,文,書,秒,区,はじめ,作,我々,課,席,もと,室,首,感じ,秋,界,線,いつ,ちゃん,確か,なか,どれ,これら,あなた,どちら,彼女,未満,週,木,員,女,みたい,論,手段,火,土,土,自体,校,たび,都,いま,結局,輪,冬,こちら,村,いろいろ,品,あれ,十,字,略,あたり,ほう,半ば,俺,玉,ひと,楽,いくつ,ひとつ,九,連いや,本当,よう,屋,扱い,歴,全部,なん,誌,千,府,事,はるか,もん,通,わたし,類,かく,列,だめ,向こう,なに,集,みなさん,段,百,行,以後,器,毎,匹,ふく,様,レ,観,こっち,境,おれ,やつ,かたち,紀,箇所,あそこ,おまえ,それなり,よそ,情,どっち,がい,まとも,達,下記,とおり,みつ,喜,束,まし,伸,くせ,幾つ,へん,簿,そちら,しかた,ハイ,奴,そっち,様々,から,あちら
条件を満たさない単語	てん,すね,あっち,上記,ちゃ,べつ,さらい,そう,がら,すか,どっか,ごっちゃ,そで,なかば,ぜんぶ,おおまか,哀,婦,ずつ,べん,ヶ月
コーパスに生じない単語	など,見,怒,ヶ所,ヵ所,ヵ月,箇月

コーパスに含まれる総単語のうち、不要語指標値が 24447 番目に高い値は 2.43 であり、これを閾値とする。コーパス中に登場した 294 語のリスト語のうち、273 語が不要語指標の上位 10%の条件を満たす（92.9%、表 1 参照）。

SlothLib の不要語リスト内で条件を満たさなかった 21 語には「すね」「あっち」「どっか」「ごっちゃ」など、新聞では好まれない砕けた表現が多い。こういった単語は tf の値が小さくなり、不要語の条件を満たさない。これらは、コーパスの特徴を反映した結果である。

4. 結論

本研究では不要語指標 tf-tfidf を提案し、これを用いて不要語の判別を行い、SlothLib が公開している不要語リストの単語を検証した。この結果、92.9%(273/294)が不要語であると判断できた。また、リスト外で不要語指標条件を満たした単語の多くは、コーパスに依存している。

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]　徳永 健伸: 情報検索と言語処理（言語と計算）, 東京大学出版会, 1999

Copyright © 2018 IEICE