Webテキストからの薬と効能の関係獲得

鳥海心† 宮﨑太郎‡ 山田一郎‡ 八木伸行†
†東京都市大学 ‡NHK放送技術研究所
g1232144@tcu.ac.jp miyazaki.t-jw@nhk.or.jp yamada.i-hy@nhk.or.jp yagi@tcu.ac.jp

1. はじめに

近年、Webの発達やSNSの普及によりWeb上には膨大な量のテキストデータが存在するようになった。この中には重要な知識が含まれており、Webからコンピュータの扱える知識を獲得する取り組みが、多く報告されている[1][2]。獲得した知識は質問応答システムなどの構築に応用されている[3]。

我々はWeb上のテキストからの知識獲得の試みとして、健康に関する医学的知識を獲得する研究を進めている。本稿では、テキストから「薬と効能」の関係を持つ名詞と節ペアの獲得手法について報告する。例えば、「熱を下げる解熱剤」という文からは、「解熱剤」に「熱を下げる」という効果があることがわかる。このように、名詞と連体修飾節から有用な知識を獲得することを目的とする。

従来、名詞と連体修飾節を4種類の関係に分類する手法が提案されている[4]。この手法では、連体修飾節に含まれる単語表記を特徴量として利用するため、学習データに出現しない単語の特徴を使うことが出来ず、分類には一定量の学習データが必要であった。この問題を解決するために、今回、単語を多次元のベクトルで表した分散表現を利用する。

提案手法では、まずWeb上のテキストから、薬名を表す集合を抽出し、薬と連体修飾節ペアを抽出する。次に、抽出したペアが「薬と効能」の関係にあるかを機械学習により判定する。この際、連体修飾節に含まれる単語の分散表現を利用する。実験では、分散表現を用いる手法の有効性を示した。さらに、分散表現とIDFを併用した特徴量と比較したところ、IDFの効果が少ないことを確認した。

2. 提案手法

NHK「きょうの健康」のWebページを対象とし、健康に関する医学的知識を獲得する。提案手法は、図1のように、①薬名集合の抽出、②薬名-連体修飾節ペアの抽出、③連体修飾節が薬の効能を示しているかの判定、の3段階の処理を行う。各処理の詳細を以下に説明する。

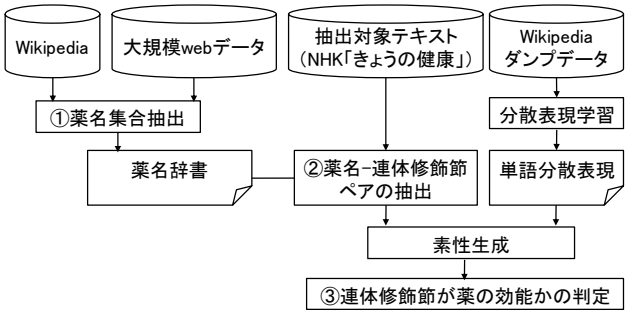

図1 提案手法の概要

2.1 薬名集合抽出

薬名と連体修飾節のペアを取得する際に用いる「薬名辞書」を作成するために、薬名の集合を抽出する。抽出には、大規模Webデータを解析する手法[5]と、Wikipediaを解析する手法[6][7]を利用する。

大規模Webデータを対象とした手法[5]では、まず、数十個程度の集めたい単語集合を人手で作成する。次に、単語集合に含まれる単語と、Webテキストに出現する単語の文脈類似度が高いものを数百個程度抽出する。最後に、この数百程度の単語を正例、ランダム抽出した単語集合を負例として学習したSVMにより、Webテキスト中の任意の単語が集めたい単語集合に属するか否かを判定することで、数千個程度の単語が収集される。これらの処理により、「薬」に属する単語集合を獲得する。この手法はALAGINフォーラム「カスタム単語集合作成サポートサービス」として公開されており[8]、本稿ではこのサービスを利用する。

Wikipediaを解析する手法は、まずWikipediaの定義文（最初の1文）、カテゴリー、見出し語の階層構造を利用して上位下位

関係となる単語ペアの候補を大量に抽出し、SVM により単語ペアが上位下位関係であるかを判定する[6]。この手法は ALAGIN フォーラム「上位下位関係抽出ツール」として公開されており[8]、本稿ではこのツールを利用する。さらに、山田らの手法[7]によって、獲得した上位下位関係の上位語から見た下位語らしさと、下位語から見た下位語らしさを判定し、上位下位関係の誤りを除外する。Web テキストから獲得した「薬」の下位に属する単語を収集することにより、大量の「薬」に属する単語集合を生成する。抽出した薬名の例を表 1 に示す。

Web テキストから獲得した「薬」に属する単語集合と Wikipedia から獲得した単語集合の和集合を、「薬名辞書」として用いる。

表 1　獲得した薬名の例

ACE 阻害薬、アクテージ、頭痛薬、イルベサルタン、バファリン、タミフル、パブロン、ニトログリセリン、タイレノール、抗血小板薬

2.2 薬名-連体修飾節ペアの抽出

解析対象のテキストデータ（NHK「きょうの健康」Web ページ）から、薬名とそれに係る連体修飾節のペアを抽出する。

テキストデータを係り受け解析し、名詞節に該当する部分が「薬名辞書」に含まれる場合に、その名詞節と連体修飾節をペアとして抽出する。複数の連体修飾節が一つの薬名に係る場合には、それぞれを別のペアとする。表 2 に、抽出したペアの例を示す。

表 2 抽出した薬名と修飾節ペア

修飾節	薬名
血栓ができないようにする	抗凝固薬
多くの人に用いられている	トラスツズマブ
ウイルスの増殖を抑える	プロテアーゼ阻害薬
よく使われるのは	吸入ステロイド薬

2.3 連体修飾節が薬の効能かの判定

2.2 節で抽出した薬名と連体修飾節のペアについて、連体修飾節が薬の効能を表すものであるか判定する。判定には SVM を用いる。学習には「薬名と連体修飾節のペアが薬とその効能の関係にあるか」の正解ラベルを人手により付与したデータを用いる。SVM で用いる素性は、以下の A と B の 2 種類を用意し、性能を比較する。

A. 単語の分散表現による素性

単語の分散表現は、単語の表層を用いず、単語の意味により多次元の特徴を抽出することができることから、近年注目を浴びている。本稿では、Skip-gram[9]による単語の分散表現を用いて素性を作成する。

単語の分散表現による素性は、判定の対象となる連体修飾節に出現するすべての単語の分散表現の和をとり、正規化したものを用いる。

$$v^{(i)} = \frac{\sum_{w \in W} v(w)^{(i)}}{|W|}$$

ここで、$v^{(i)}$は i 番目の素性を、W は連体修飾節に出現する単語の集合を、$v(w)^{(i)}$は単語 w の分散表現の i 番目の要素をそれぞれ表す。

Skip-gram により作成されたベクトルの和は、意味的な足しあわせを表現できることが報告されている[10]。連体修飾節に出現するすべての単語の分散表現の和は、連体修飾節全体の意味を足し合わせた素性になっていると考えられる。

B. 単語の分散表現*IDF による素性

A の分散表現による素性ではすべての単語について同じ重みで分散表現を足し合わせた。一方、本素性では単語ごとに IDF を掛け合わせることにより単語の重みを与える。

$$v^{(i)} = \frac{\sum_{w \in W} v(w)^{(i)} \cdot IDF(w)}{\sum_{w \in W} IDF(w)}$$

$$IDF(w) = \log \frac{|D|}{|\{d : w \in d\}|}$$

ここで、$|D|$は全文書数、$|\{d : w \in d\}|$は単語 w が出現する文書数をそれぞれ表す。

単語の分散表現*IDF による素性は、連体修飾節中で特徴的な単語の寄与率が高くなるため、より連体修飾節の特徴を表した素性となることが期待できる。

3. 評価実験

3.1 実験条件

提案手法の効果を確認するための評価実験を行った。

薬名と連体修飾節のペアを抽出する際に使用する係り受け解析器には CaboCha[11]を用いた。ペアが「薬名とその効能」を表すか否かの判定には SVM-Light[12]を使用し、多項式カーネルにより判定を行った。SVM の素性に用いる Skip-gram による分散表現の計算には Word2Vec[9]を用いた。薬名辞書は、2.1 節の処理により抽出した 7,628 個の単語を用い、分散表現の学習と IDF の計算には、Wikipedia の 2015 年 4 月のダンプデータを用いた。

評価実験の対象データには、NHK「きょうの健康」の Web ページ、2009 年 3 月から 2015 年 6 月までの 5.5 年分 1,159 記事を使用した。ここから 2.2 節の手法により、557 個の薬名と連体修飾節ペアを取得した。取得したペアに対して、1 名の作業者により人手で薬名とその効能の組み合わせであるか否かを判別した。その結果、正例は 146 個、負例は 411 個が含まれていた。評価データ例を表 3 に示す。5-fold cross-validation で適合率と再現率を求め、この調和平均である F 値で評価をした。この際、全体の F 値が最大となるように SVM-Light のパラメータを調整した。

表 3 評価データ例

正例
血栓ができないようにする 胃酸の分泌を抑える ウイルスの増殖を抑える 気管支を広げる作用がある 心臓の働きを抑える
負例
働くのを抑える 多くの人に用いられている 市販の よく使われるのは 作用の弱い

ベースライン手法として、単語表記をベクトルの要素として連体修飾節を特徴付ける手法を利用した。ベクトルの要素の値として、検索処理における単語への重み付けなどで使われる TF-IDF を用いた。連体修飾節に出現しない単語に対応する要素には 0 が入る。このベクトルを素性として用い、SVM によって連体修飾節が薬の効能を表すかを判定する。

3.2 結果

表 4 に実験結果を示す。ベースライン手法の F 値は 0.595、提案手法となる分散表現を用いた手法では 0.788、分散表現*IDF を用いた手法の F 値は 0.752 であった。分散表現を用いた手法はベースライン手法より 0.193 の向上が見られ、分散表現を利用する効果を確認することができた。

表 4 実験結果

手法	再現率	適合率	F 値
ベースライン	0.934	0.435	0.595
分散表現	0.92	0.689	0.788
分散表現* IDF	0.855	0.671	0.752

3.3 考察

3.3.1 分散表現*IDF について

分散表現*IDF による素性では、分散表現をそのまま使う場合と比較して、F 値が低下した。人手による正解は、多くの場合に「血栓ができるのを防ぐ」のように、「病気の症状や原因」と「作用」が書かれているものを正例としている。このうち、「病気の症状や原因」を表す部分は、専門的な名詞で IDF が高い場合が多い（表 5）。それに対し、「作用」を表す部分は IDF が低い動詞で表現される場合が多い（表 6）。このため、単語の分散表現*IDF の素性は、IDF が高い「病気の症状や原因」を表す単語の分散表現の影響が大きくなり、連体修飾節全体の特徴を適切に表現することができなかったと考えられる。

性能を向上させるためには、素性を足し合わせる際に、名詞のみを足し合わせた素性と動詞のみを足し合わせた素性をそれぞれ用意するなど、品詞ごとに素性を分けることが考えられる。

3.3.2 提案手法で判定が難しい連体修飾節

提案手法で判定が難しい連体修飾節についての分析を行った。それらの多くは「薬の効能」を表現しそうな単語を含むが、全体の単

表 5「病気の症状や原因」を表す単語の IDF

単語	IDF
血栓	9.13
胃酸	9.76
発作	7.26

表 6 「作用」を表す単語の IDF

単語	IDF
防ぐ	5.13
促す	5.73
抑える	5.11

語数が少ないという特徴がみられた。例えば「素早く効く」のように、「薬の効能」を表現する単語「効く」の存在により、誤判定したと考えられる。単語数の少ない連体修飾節にはペナルティを与えることで、さらなる精度の向上が期待できる。

4. おわりに

本稿では、Web から健康に関する医学的知識の獲得を目指し、「薬と効能」の関係にある名詞と連体修飾節ペアの獲得手法を提案した。

Web 上にあるテキストから抽出した名詞と連体修飾節ペアが「薬と効能」の関係にあるか、分散表現手法と分散表現*IDF 手法のそれぞれの素性により SVM で判定をした。実験の結果、TF-IDF を用いたベースライン手法の F 値が 0.595、分散表現を用いた手法の F 値が 0.788、分散表現*IDF を用いた手法が 0.752 という値が得られ、分散表現を用いる手法が有効であることがわかった。また、IDF を併用しても、ほとんど効果が無いことがわかった。

提案した手法は、薬の効能に特化した素性を用いていないので、「薬と効能」以外の知識の獲得も可能と考えられる。今後、他の知識獲得の評価実験を行い、他分野にも適用可能か検討していく予定である。

なお、本研究の一部は、JSPS 科研費 25280036 の助成を受けたものです。

参考文献

[1] István Varga, et al., “Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster,” *In Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.1619-1629 (2013).

[2] Johannes Hoffart, et al., "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Journal of Artificial Intelligence*, Vol.194, pp.28-61, 2013.

[3] Masahiro Tanaka, et al., "WISDOM2013: A large-scale web information analysis system," In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013.

[4] 山田ほか, "Web を情報源とした Q&A システムの検討," 第 9 回言語処理学会年次大会発表論文集, C7-5, 2003.

[5] Stijn De Saeger, et al., "A Web Service for Automatic Word Class Acquisition," In *Proceedings of the 3rd International Universal Communication Symposium* (IUCS'09), pp.132-138, 2009.

[6] Asuka Sumida and Kentaro Torisawa, "Hacking Wikipedia for Hyponymy Relation Acquisition," In *Proceedings of the Third International Joint Conference on Natural Language Processing* (IJCNLP2008), pp.883-888, 2008.

[7] 山田ほか, "上位下位関係からのインスタンス集合の獲得," 電子情報通信学会技術報告 vol.114, no.444, NLC2014-44, pp.1-6, 2015.

[8] https://alaginrc.nict.go.jp/

[9] Tomas Mikolov, et al., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[10] Tomas Mikolov, et al., "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, 2013.

[11] 工藤, 松本, “チャンキングの段階適用による日本語係り受け解析,” 情報処理学会論文誌, Vol.43-6, pp.1834-1842, 2002.

[12] Thorsten Joachims, "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," *MIT-Press*, 1999.