E-002

# 薬と効能の関係獲得におけるニューラルネットワークの適用
# Acquisition of Relationship between the Medicine and Efficacy Using a Neural Network Model

鳥海 心† 宮崎 太郎‡ 後藤 淳‡ 山田 一郎‡ 八木 伸行†
Shin Toriumi Taro Miyazaki Jun Goto Ichiro Yamada Nobuyuki Yagi

## 1. はじめに

Web 上に存在する膨大な量の知識を活用するため、我々は Web テキストからの知識獲得の研究を行っている。その一環として、健康に関する医学的知識を獲得する研究を進めており、Web 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した[1]。

このペアを獲得する過程において、Web テキストから抽出した連体修飾節が「薬と効能」の関係にあるか判定する必要がある。従来手法[1]では、対象の連体修飾節に出現する単語の情報を足し合わせた特徴を利用し、Support Vector Machines(SVM)[2]を用いて判定したが、この手法では、単語の出現順序などを考慮していなかった。

そこで、単語の出現順序を考慮したモデルを作成できる再帰型ニューラルネットワーク(RNN)[3]を利用した判定を試みた。これにより、文脈を考慮した判定が可能となり、判定性能の向上が期待できる。SVM を用いた従来手法と比較実験を行った結果、RNN の優位性が明らかになった。

## 2. 提案手法

提案手法の処理の全体の流れを図 1 に示す。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、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。[1]で報告した手法との違いは③の処理で、それ以外は同一の処理を行っている。各処理の詳細を以下に説明する。

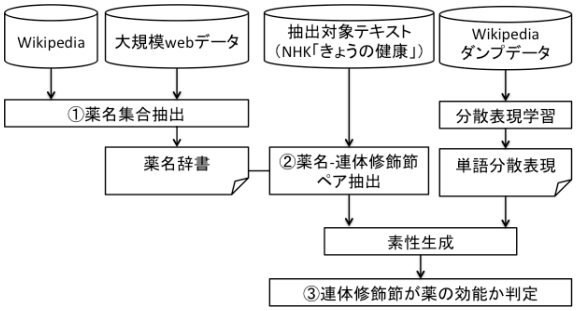


図 1 処理の流れ

### 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[4]と、Wikipedia を解析する手法[5][6]を利用する。大規模 Web データを対象とした手法[4]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[7]、このサービスを利用する。Wikipedia を解析する手法[5]は、ALAGIN フォーラム「上位下位関係抽出ツール」として公開されている[7]、このツールを利用し、上位下位関係を獲得する。さらに、山田らの手法[6]により、上位下位関係の誤りを除外し、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で「薬」に属する単語集合を生成する。

大規模 Web データを解析する手法と Wikipedia を解析する手法により獲得した単語集合の和集合を「薬名辞書」として用いる。

### 2.2 薬名-連帯修飾節ペアの抽出

解析対象のテキストデータから、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。

### 2.3 連体修飾節が薬の効能か判定

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 RNN 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使用する。図 2 のように、対象となる連体修飾節に出現する個々の単語の分散表現 $W_i$ を前から順に RNN に入力し、最後の単語が入力された時点での RNN の出力により、対象の連体修飾節が薬の効能を表すかどうかの判定を行う。入力に用いる素性に単語の分散表現 200 次元を用いる。

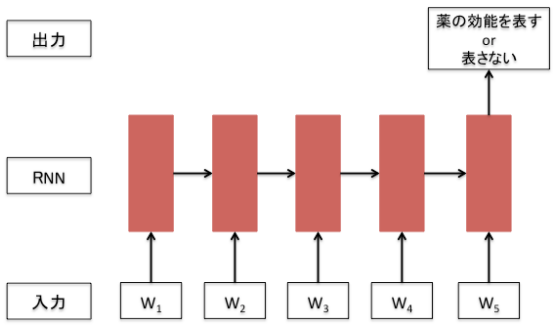


図 2 RNN による判定

† 東京都市大学 Tokyo City University
‡ NHK 放送技術研究所
NHK Science & Technology Research Laboratories



Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

## 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に評価実験を行った。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、薬名-連体修飾節ペアの抽出処理で使用する係り受け解析器には CaboCha[8]を用いた。提案手法に用いた RNN の実装には Chainer[9]を使用し、入力層、中間層、出力層の 3 層でネットワークを構築した。入力層と中間層は 200 次元、出力層は 2 次元に設定し、また中間層には LSTM を用いた。RNN の入力には対象となる連体修飾節に出現する単語の分散表現を用い、Skip-gram による単語の分散表現を使用した。計算には Word2Vec[10]を用いた。

評価実験の対象データには、NHK「きょうの健康」の 2009 年 3 月から 2015 年 6 月までの 5.5 年分の Web ページから抽出した 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-連体修飾節ペアを取得し、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を判別した。その結果、正例 146 個、負例 411 個が含まれていた。5-fold cross-validation で適合率と再現率を求め、調和平均である F 値で評価をした。

RNN では学習時にランダム要素を含むため、学習ごとに異なったモデルができる。そのため、同一条件で 3 回のモデル学習を行い、学習回数ごとの訓練誤差を比較し、誤差が最小となる学習時のモデルを用いて評価した。

### 3.2 ベースライン手法

ベースライン手法として、[1]の提案手法を用意した。これは 2.3 節の判定タスクに RNN ではなく SVM を用いた手法を利用する。SVM には SVM-Light[2]を使用し、多項式カーネルにより判定を行った。素性には、単語の分散表現を用い、図 3 のように判定の対象となる連体修飾節に出現する単語の分散表現の和をとり、単語数で正規化したものを用いる。単語の分散表現の計算には提案手法と同様に Word2Vec[10]を用いた。

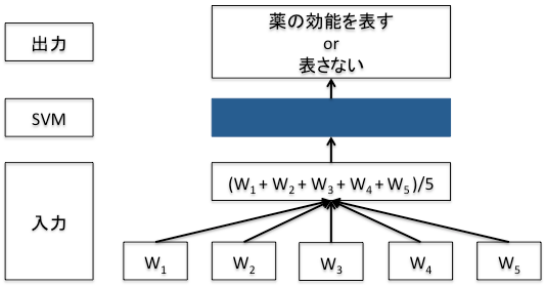


図 3 SVM による判定

### 3.3 結果

表 1 に実験結果を示す。ベースライン手法として用いた SVM の F 値は 0.685、提案手法である判定に RNN を用いた手法では 0.796 であった。提案手法はベースライン手法より 0.111 の向上が見られ、判定に RNN を利用する効果を確認することができた。

表 1　評価実験結果

| 手法 | 適合率 | 再現率 | F値 |
|---|---|---|---|
| SVM | 0.793 | 0.603 | 0.685 |
| RNN | 0.739 | 0.863 | 0.796 |

### 3.4 考察

判定に RNN を用いたことで、語順を情報として扱うことができ、より詳細な判定ができるようになった。このため、ベースライン手法を上回る結果になったと考えられる。

ベースライン手法に使用した SVM は、学習データ中の正例と負例の数に差がある場合、学習の重みを調整することで判定性能が向上する。この重みを、評価データを用いて事後的に最適なものに調整すると、F 値が 0.788 となり、今回の提案手法と同等の精度になる。これは SVM を用いた場合の性能の上限と考えることができるが、RNN ではこれとほぼ同等の性能を、事後的なパラメータ調整をせずに実現できた。このことからも、今回のタスクにおいては提案手法が従来手法と比べて適していることがわかる。

## 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、RNN で判定をした。実験の結果、RNN を用いた提案手法の F 値が 0.796、SVM を用いた手法の F 値が 0.685 という値であり、判定に RNN を適用した提案手法が有効であることがわかった。

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である。今後、他分野知識での適用を確認したい。

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] 鳥海ほか, “Web テキストからの薬と効能の関係獲得” ,第 22 回言語処理学会年次大会(NLP2016), pp. 1045-1058 (2016).
[2] Thorsten Joachims, “Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning,” MIT-Press, (1999).
[3] Thomas Mikolov, et al, “Recurrent Neural Network Based Language model”, in Proceedings of Interspeech, (2010).
[4] 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).
[5] Asuka Sumida, et al., “Hacking Wikipedia for Hyponymy Relation Acquisition,” In Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP2008), pp.883-888, (2008).
[6] 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).
[7] https://alaginrc.nict.go.jp/
[8] 工藤ほか, “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).
[9] http://chainer.org/
[10] Tomas Mikolov, et al., “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, (2013).



Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.