1B-03

RNNを用いたテキスト二値分類による用語説明文抽出方法の提案

†李 天瑶　　†鹿糠 秀行　　†大島 敬志　　†前岡 淳
†株式会社 日立製作所 研究開発グループ

1　はじめに

文書の理解性や保守性を向上させるために、文書中に出現する用語と、用語を説明する文章（以降、説明文）をまとめた用語辞書の作成が重要である。用語辞書作成には工数がかかるため、用語辞書を自動構築する研究が行われている。自動構築では、用語の抽出と説明文の抽出が必要である。用語の抽出はTermExtractなどのツールが開発されている [1][2]。説明文の抽出は、予め定義した抽出ルールに則って抽出する方法が提案されている [3][4]。しかし、説明文の書き方が規定されていない場合には抽出ルールを定義することが難しく、その結果として抽出の網羅性が低くなることを問題と考える。

本研究では、用語辞書作成の支援を目的として、RNN (Recurrent Neural Network) を用いて、文書中のセンテンスを説明文と非説明文に二値分類することにより、説明文を自動抽出する方法を提案する。あるソフトウェアに関わる文書を題材に、説明文と非説明文からなる計50文の教師データと計30文のテストデータを用いて試行し、説明文抽出の精度と網羅率ともに約8割の結果を得た。

2　RNNを用いた用語説明文抽出方法

近年、様々な課題に深層学習が応用されており、深層学習は、ルールベースより柔軟に判定できるという特徴をもつことが知られている。自然言語処理分野では、文章を単語の時系列情報として扱うことができるRNNが特に活躍している。例えば、Sonらは、RNNを用いて、ベトナム語の法律文書の文を、仮定や措置の5種類に分類した [5]。

本研究では、文を分類できるというRNNの基本的な機能を用いて、説明文を抽出できると考える。すなわち、RNNを用いて文書内の文を「説明文」と「非説明文」とに二値分類し、「説明文」に分類された文を集めればよいと考える。

設計したRNNのネットワーク構造を図1に示す。中間層としては、埋め込み層及びLSTM (Long Short-Term Memory) ユニットからなる層を取り入れた。

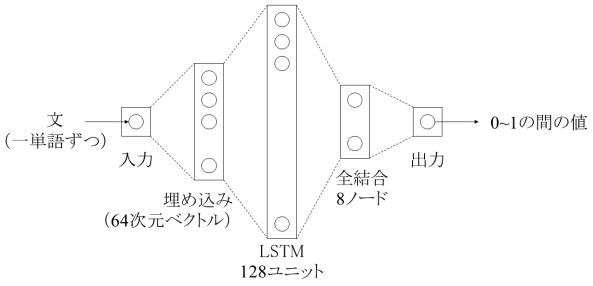

図1: 本研究で設計したネットワーク

埋め込み層では、単語をベクトルへ変換する。本研究では、ベクトルのサイズを64とした。対象文書全体を、記号や数字の除去および分かち書きをした上で、Word2vecを利用して得られる「単語-ベクトル対応表」を利用する。中間層を、128のLSTMユニットとした。LSTMユニットにより、文脈における長期依存を解析できるようになる。

比較的に「浅い」ネットワークを設計した理由は、説明文と非説明文の二値分類タスクは比較的に簡単であることと、実用において用意できる教師データは少ない（50〜100文程度）と予想されることである。

3　RNNの教師データとテストデータ

対象文書を調査したところ、大多数の説明文は、下記2種類の文型のいずれかとなっていることがわかった。

- ［用語］とは、［ほげほげ］です。
- ［ほげほげ］を、［用語］といいます。

本研究では、RNNがルールベースよりも柔軟に説明文と非説明文を分類できるかどうかを検証したい。すなわち、一般的な文型に適合せず、ルールベースによる抽出が困難な「レアケース」を、RNNは網羅可能かどうかを検証したい。そのため、教師データの説明文は、上記2種類の文型のものに絞った。一方、テストデータの説明文には、対象文書の中で見つかったレアケースの説明文3文（下記のとおり）を入れた。

- ［用語］は、［ほげほげ］です。
- ［ほげほげ］が、［用語］です。
- ［ほげほげ］を［用語］といい、… 。

なお、対象文書では、説明文は非説明文に比べて非常に少なかった。文書と同様な比率の教師データとテ

Extracting Term Definitions Using Recurrent Neural Network
†Tianyao LI, †Hideyuki KANUKA, †Keishi OOSHIMA, †Jun MAEOKA
†Hitachi, Ltd. Research & Development Group

ストデータだと、学習済みの RNN はすべての文を非説明文に分類するようになり、テストデータをすべて非説明文に分類することで、高い正解率を取りかねない。そのため、人為的に説明文と非説明文の比率を調整した。結果、表 1 に示す数の教師データとテストデータを用意した。

表 1: 教師データとテストデータ

	説明文	非説明文	計
教師データ	22	28	50
テストデータ	12	18	30

4 結果

テストデータに入れたレアケースの説明文 3 文は、すべて説明文に分類できた。全体の分類結果は表 2 のとおりである。表 2 により、二値分類の正解率は 87%である。説明文分類の精度は約 79%であり、網羅率は約 92%である。

表 2: 実験結果

		テストデータ	
		説明文	非説明文
分類結果	説明文	11	3
	非説明文	1	15

5 考察

5.1 RNN の柔軟性について

ルールベースでは対応できないレアケースの説明文 3 文を、RNN はすべて説明文に分類することができた。この結果から、RNN がルールベースより、説明文のバリエーションに対して柔軟性が高いことが分かる。

一方、ルールベースが正しく分類できる説明文を、RNN が誤って非説明文に分類しているケースも存在する。深層学習を含み、機械学習は全体的に正解率が 100%となることは非常に難しく、特に今回のように教師データが少量の場合には高い正解率を達成することは難しいと予想できるため、特に悪い結果ではないと言える。

また、この結果から、実用においては、RNN とルールベースの併用が効果的であることがわかった。すなわち、頻出パターンについてはルールベースで確実に対応し、ルールベースによる抽出が困難なレアケースを RNN で補うといったアプローチが良いと考える。

5.2 提案手法の実用性について

説明文分類の精度は約 79%であり、網羅率は約 92%であるが、十分実用レベルに達していると考える。さらに、前述したとおり、RNN で分類できなかった説明文は、ルールベースでは分類できる説明文を含んでいるため、両手法の組み合わせにより、漏れがより少ない説明文の分類を実現できる。また、50 文程度の教師データは、プロジェクト毎に用意するとしても、十分現実的な工数（本実験の場合は 1〜2 人日）で用意できるため、文書をすべて目検する作業に比べてより少ない工数で辞書を構築できると考える。

6 おわりに

RNN を含む機械学習を利用したアプローチの問題点として、教師データにおける偏りの影響を受けやすいことが挙げられる。例えば、教師データの説明文の大半に、説明文であることとは無関係に特定の名詞が偶然含まれており、逆に非説明文の大半にその名詞が偶然含まれていない場合、その名詞を含む文を説明文に分類してしまうモデルが生成されてしまう可能性が高い。

今後は、学習時・分類時共に、文中の名詞を全て同一の識別子（例えば【名詞】など）に置換することで、上述の問題に対応できるか検証すべきと考えている。これは、説明文と非説明文を区別する特徴は、名詞の種類よりも、名詞・動詞・助詞を含む各品詞がどのような順番・バランスで並んでいるか、といった点に現れるという仮説に基づく。名詞を始めとする各品詞をどの程度抽象化すると良いか、例えば、固有名詞と一般名詞を区別した場合としない場合の分類精度に差は生じるか、などについて検証していきたい。

参考文献

[1] 中川 裕志, 湯本 紘彰, 森 辰則: 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, 10 巻 1 号, pp.27-45 (2003)

[2] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

[3] 木田敦子, 乾裕子, 落谷亮, 西野文人: 新聞記事からの用語集作成のためのテキスト分析, 情報処理学会研究報告自然言語処理, No.95, pp.85-92 (1999)

[4] 佐々木靖弘, 佐藤理史, 宇津呂武仁: ウェブを利用した専門用語集の自動編集, 言語処理学会第 11 回年次大会論文集, 言語処理学会, pp.895-898 (2005)

[5] Nguyen Truong Son, Nguyen Le Minh, Ho Bao Quoc, Akira Shimazu : Recognizing logical parts in legal texts using neural architectures, 2016 Eighth International Conference on Knowledge and Systems Engineering, pp.252-257 (2016)