

# 要求仕様書中の文の分類手法

2020SE035 宮川翔海

指導教員：佐伯元司

## 1 背景と研究課題

システム開発を行うにあたり、発注者の要求を正確に理解し獲得することは、重要なことの1つである、機能要求や非機能要求だけでなく、様々な文を内包する。そこで、これらの文を分類することにより、要求仕様書の理解の助けになり、システム開発の効率化につながる。

近年では、仕様書中の要求の機械学習による自動分類が注目されており、その手法は有望である。

これまで、機械学習を使った要求文の分類 [1][2] は非機能要求文の細分類が行われている。これらは何らかのやり方ですでに非機能要求文と分類された文に対してであり、要求仕様書中に含まれている様々な文の分類は行われていない。

本研究では、機械学習による機能要求や非機能要求、それら以外の文の分類に着目し、要求仕様書中の要求分類手法を提案し、その評価も行う。

## 2 本研究で使用する技術

Word2Vec[3] とは、T.Mikolov らによって開発され、文章中の単語からベクトル分散表現を獲得する、自然言語処理の技術である。本研究では、文章のベクトル化に用いる。MeCab [4] とは、京都大学で開発されたオープンソース形態素解析エンジンである。本研究では文章の形態素解析をするために用いる。Chainer とは、オープンソースのディープラーニングフレームワークであり、機械学習プログラムの開発に使用する。

## 3 提案手法

### 3.1 分類項目

要求仕様書を収集し、それらの内容を分析した結果、既存の分類（機能、非機能）だけでは理解促進には不十分と考えたため、分類項目を表1のように決定した。

### 3.2 使用するニューラルネット

本研究では、入力層、隠れ層、出力層を持つ3層の線形ニューラルネット（NN と呼ぶ）と、同じ3層の Long Short Term Memory（LSTM と呼ぶ）の2つを用いて比較を行う。LSTM は前の文の分類結果も考慮するため、要求仕様書中の文の系列の情報を分類に用いることができると考えたため今回使用する。

### 3.3 評価手法

ラベル付与データと自動分類を経て得られるデータの作成手順、それらを比較するまでの流れを述べる。

要求仕様書から表1で示した分類項目に従って人手で分

表1 分類項目

分類項目	詳細
機能要求	システム開発における、クライアントから求められる要求
データ	成果品などが羅列されている文
非機能要求	機能以外の要件
機能 + 非機能要求	機能要求と非機能要求が合わさった要求
理由	システム開発の理由などが書かれた文
開発プロセスに関する要求	開発プロセスについて書かれた文
参照・添付文書および文書構造に関する情報	文書構造について書かれた文

類し、分類項目のラベルを付与する。要求仕様書1つにつき3名の被験者を割り当て、多数決によって分類項目を決定し、ラベル付きデータを作成する。ラベル付きデータを MeCab を用いて形態素解析し、Chainer を用いて学習させ、学習済みモデルを作成するプログラム作成する。ニューラルネットへの入力には、要求文を形態素解析したものを、Word2Vec を用いてベクトル化したものである。形態素解析を行う際には、名詞、動詞、形容詞、副詞を取り出した。生成した学習済みモデルに評価用の要求文を入力し自動分類を行わせる。自動分類結果と手作業で分類したラベル付きデータを比較し、最終的に適合率、再現率で評価する。要求仕様書はシステム開発や業務委託仕様書などで100行前後の6つの要求仕様書を用いた。

比較するまでのフロー図を以下の図1に示す。

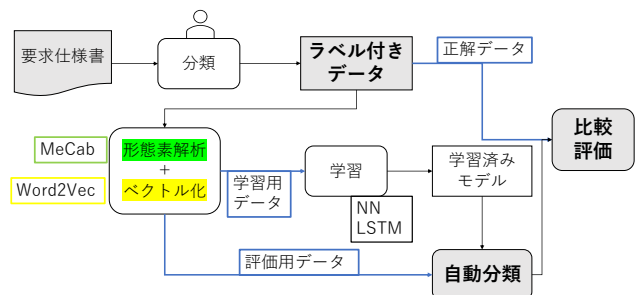


図1 評価の流れ

## 4 学習結果

6つの要求仕様書から1つを選び、それを評価用、残りの5つを学習用とし学習モデルを作った。これをすべての仕様書について行った。つまり1つのニューラルネットについて5回の学習と評価を行った。評価結果は5つ全部を合わせて、適合率、再現率を算出した。

学習は1つの要求仕様書からランダムに繋がった10文を選び、選べる回数だけ学習を行い、それを全部の要求仕様書に行き1つのepochとした。epochは20回繰り返した。NNは入力層、中間層、出力層の順にデータを流していく構造で、LSTMはNNが10個連なっており前の中間層の出力結果を次の中間層へ入力する構造になっている。

自動分類を行った結果の適合率、再現率を分類項目ごとに表2、表3に示す。また、「その他」は3人の被験者の意見が全く異なったものを指す。

表2 NNでの結果

分類項目	適合率	再現率
機能要求	0.158	0.106
データ	0.559	0.651
非機能要求	0.491	0.636
機能要求 + 非機能要求	-	-
理由	0.333	0.091
開発プロセスに関する要求	0.235	0.110
参照・添付文書および文書構造に関する情報	0.472	0.658
その他	0.172	0.125

表3 LSTMでの結果

分類項目	適合率	再現率
機能要求	0.154	0.094
データ	0.553	0.447
非機能要求	0.533	0.558
機能要求 + 非機能要求	-	-
理由	0.25	0.091
開発プロセスに関する要求	0.126	0.192
参照・添付文書および文書構造に関する情報	0.349	0.632
その他	0.192	0.159

## 5 考察

### (1) 学習結果の考察

「データ」「非機能要求」「文書構造」の分類は比較的順調であったがそれ以外の分類に関しては芳しくなかった。正解セットの分類数は機能要求：85(14)、データ：109(89)、非機能要求：258(39)、機能 + 非機能要求：7、理由：11(4)、開発プロセス：78(31)、文書構造：38(20)、その他：88で

あった。( )内は3人の被験者の意見が一致した文である。表2、表3に示すとおり機械学習によって機能 + 非機能要求に分類されたものはなかった。

「データ」と「非機能要求」は学習用のデータ量が十分で、「文書構造」は( )内の数が多いためデータの質がよかったことが原因ではないかと考えられる。

### (2) 学習モデルの考察

LSTMを用いた学習と通常の学習では後者の方が評価が良いという結果となった。LSTMは時系列を考慮するという特徴があるが、学習時に選んだ10文はランダムに選ばれたため、要求仕様書中の章をまたいで選んだものもあり、文の意味的なつながりが切れてしまったものもあるため低い結果になったと思われる。

## 6 今後の課題

### (1) 学習用データを増やす

データの質を保つため、3人に対して1つの仕様書を割り当て分類項目を決定することは維持したままで、データの量を増やすことが課題としてあげられる。

### (2) LSTMの学習方法

今回の学習では章をまたいでの学習で評価が落ちた。章構成も考慮して学習させることや、総数などのパラメータを増やしたり、他の構造をしたニューラルネットを使ったりして評価することも課題としてあげられる。

## 7 まとめ

近年では、要求の機械学習を用いた自動分類が注目を集めている。実現すれば、要求仕様の理解促進や保守性の向上に貢献し、システム開発の効率向上につながる。

本研究では、理解の助けになるであろう分類項目を決定した。また、自動分類器を用いて要求仕様書中の文の分類を行い、適合率、再現率で評価を行った。評価結果を踏まえて、考察や今後の課題を確認した。

## 参考文献

- [1] 古川実歩・松尾美穂・村瀬摩由加・佐伯元司：「機械学習を用いた非機能要求の分類法の評価」．(Academia.Sciences and Engineering.Journal of the Nanzan Academic Society)，22巻，198-202ページ，2022年．
- [2] 北川貴之・長岡武志：「深層学習による日本語で記述された要求仕様の自動分類の試み」．情報処理学会研究報告，Vol.2019-SE-201，No.17，1-8ページ，2019年．
- [3] T.Mikolov, K. Chen, G. Corrado and J. Dean: Efficient Estimation of Word Representations in Vector Space, Proc. International Conference on Learning Representations 2013 (ICLR) Workshop Track, arXiv:1301.3781, (2013)
- [4] Yet Another Part-of-speech and Morphological Analyzer : <https://taku910.github.io/mecab/> 2021 .