2U-04

seq2seq モデルベース RNN による 会議中の発言からの重要単語抽出

川瀬 卓也† 大平 茂輝‡ 長尾 確†

†名古屋大学 工学部電気電子・情報工学科 ‡名古屋大学 情報基盤センター

1. はじめに

大学の研究室や企業でディスカッションを行う会議は頻繁に開催され，その価値を高めるためにはその会議の内容を記した議事録を残す必要がある．

会議中に効率よく議事録を作成するための研究も行われている[1]が，議事録作成にかかる人的コストは少なくなく，また，全ての発言内容を記述するのは非常に困難である．

そこで本稿では，会議中の音声とその会議の議事録を学習データとして，seq2seq モデルベース RNN による重要単語抽出手法を提案する．

2. 学習データの取得

著書らの研究室では，ディスカッションマイニングという会議記録システムを提案・運用している[2]．参加者はリモコンとタブレットを所持しており，発言する際はリモコンを用いて発言を登録し，それが話題を提起する発言(導入発言)であるか話題を展開する発言(継続発言)であるかを表明する．発言が終わったら発言の終了をリモコンで表明し，発言区間として登録する．システムは会議の様子を映像データとしても記録しており，書記による発言内容の要約テキスト（以下、発言テキスト）と合わせて議事録コンテンツとして記録している．本研究では，発言テキストに含まれる単語を重要性の高いものとみなし，その単語の自動抽出を試みる．

本研究の重要単語抽出にはディスカッションマイニングシステムで取得した過去 8 年分の映像データと過去 13 年分の発言テキストを用いた．

3. 学習データの前処理

本研究における学習データの前処理の流れを図 1 に示す．

3.1 議事録コンテンツにおける発言音声処理

ディスカッションマイニングが記録している会議映像データから音声データを取得し，発言区間ごとの音声ファイルに分割した．音声ファイルは，音声認識システム Julius[3]により音素列へと変換した．この音素列から，無音単語を示す「sp」とそれぞれ文頭・文末の無音単語を示す「silB」「silE」を除去した．

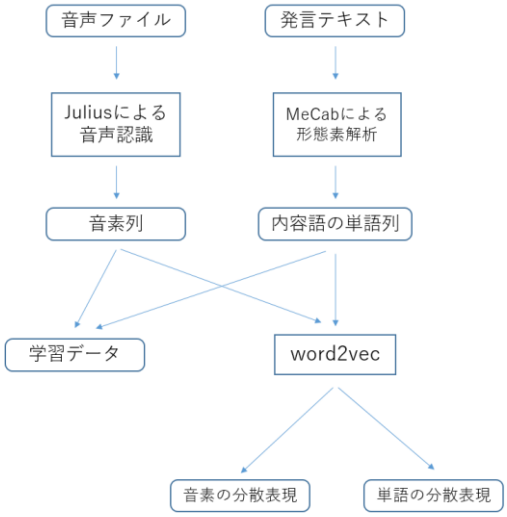

図 1. 学習データの前処理の流れ

本研究では 60 秒未満の発言の音声ファイルを使用した．また，音素の個人性による違いを考慮し，会議の参加者のうち日本人成人男性の発言音声ファイルから取得した音素列を使用した．

3.2 議事録コンテンツにおける発言テキスト処理

書記が記録した発言テキストは，形態素解析エンジンである MeCab による前処理を行った．MeCab にはデフォルトで IPA コーパスに基づく IPA 辞書が公開されている．本研究では，新語に対応して IPA 辞書からの差分を提供している mecab-ipadic-NEologd という辞書を併用した．

これらの辞書を用いて発言テキストを形態素に分割し，その形態素を基本形に統一した．さらにその品詞が「名詞」「動詞」「形容詞」である，いわゆる内容語のみを抽出することで，会議の内容を損なわず無駄な単語を省略した．

3.3 word2vec による分散表現の生成

本研究において，重要単語を抽出する過程で単語や音素をベクトル化する必要がある．本研究では word2vec[4]による分散表現を用いた．word2vec には CBOW(Continuous Bag-of-Words)と Skip-gram の 2 つのアーキテクチャがある。本研究では CBOW を用いて分散表現を作成した．その際のパラメータの設定は以下の通りである．ウ

Extracting Keywords from Voice Data in Discussion using seq2seq Model RNN
†KAWASE,Takuya (kawase@nagao.nagoya-u.ac.jp)
‡OHIRA, Shigeki (ohira@nagoya-u.jp)
†NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)

インドウサイズは最大5単語，出現回数による単語の無視は行わない．階層的ソフトマックスを使用し，学習の繰り返し回数は100回とした．

音素コーパスには「sp」「silB」「silE」を除去する前の音素列を使用し，単語コーパスには3.2節で記述した処理と同様の処理をした単語列を用いた。それぞれのコーパスから，43 種類の音素を10次元，17750種類の単語を300次元のベクトルに変換するモデルを生成した。

表1．各コーパスと分散表現取得モデルの説明

	データサイズ	データ数	モデルサイズ
音素コーパス	5.67GB	14995	43×10
単語コーパス	11.1GB	99879	17750×300

4. seq2seq モデル RNN を用いた重要単語予測

以下では，本研究で実装したモデルと実験状況について述べる．

4.1 seq2seq 単語予測モデル

seq2seq 単語予測モデルは会議中の発言の音素列を入力とし，Encoder と呼ばれる RNN によってその音素列を順次読み込み，発言内容を表現する中間ベクトルを生成する．続いて，生成された中間ベクトルを Decoder と呼ばれるもう一つ RNN の初期状態とし，Decoder の出力と Decoder の内部状態に基づいて単語を入出力する．Encoder と Decoder の各 RNN 内部に音素または単語をベクトルと対応させる層がある。その部分での計算に3.3節で述べた分散表現を利用した．

また，分散表現取得時のコーパスに文章の末尾を示す「eos」トークンを追加しており，「eos」トークンが出現するタイミングを予測することが出来る．本研究では RNN として Long-Short term Memory (LSTM)[5]を採用し，ミニバッチ学習を用いて学習させた．

4.2 実験設定

データセットには3.1節，3.2節で述べた音素列と単語列を用いる．詳細は以下の表2に記述する．

表2．データセット

	データサイズ 音素/単語	データペア数
学習データ	3.86GB/1.14GB	11728
テストデータ	0.95GB/0.28GB	2932

ミニバッチ学習におけるミニバッチサイズは64とし，学習の繰り返し回数は710回とした．

4.3 評価基準

単語予測モデルの出力は，最初の「eos」トークンが出力されるまでとする．再現率の計算に「eos」トークンは含まない．本研究では，以下の式で表される再現率を評価基準とし，再現率の計算に「eos」トークンは含まない．

$$再現率 = \frac{正解単語数}{予測単語数} \times 100$$

4.4 実験結果

再現率は全て小数点以下を切り捨てて計算した．予測された単語列が「eos」トークンのみの場合など，正常に再現率を計算できない場合などを除いて2915ペアの予測結果を得た．再現率の平均は9.61%であった。図2は再現率を棒グラフに表したものである．横軸は再現率を表し，縦軸はそれに該当する予測結果数である。

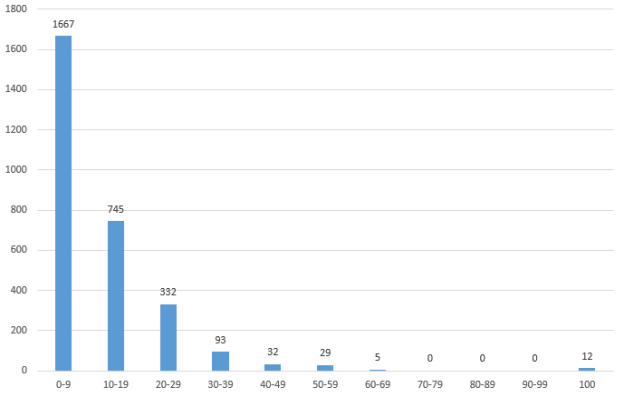

図2．実験結果

5. おわりに

本稿では，会議中の発言から重要な単語を抽出する手法を提案した．今後は，単語予測の性能向上を目指し，議事録における単語の出現頻度や議論構造を考慮した手法の検討，単語の重要性のより厳密な定義などを行う予定である．

参考文献

[1] 三浦，平田，議事録生成技術に関するサーベイ，言語・音声理解と対話処理研究会，2017.
[2] 長尾，ディスカッションマイニング ： 対面式会議での議論からの知識発見，信学技報.2012.
[3] 李，大語彙連続音声認識エンジン Julius, http://julius.osdn.jp/,(2018年12月参照).
[4] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space, CoRR, 2013.
[5] Sepp Hochreiter, Jurgen Schmidhuber, LONG SHORT-TERM MEMORY, Neural computation, 1997.