5P-02

文書間類似度を用いたWebページからの応答文収集

竹内　慶介†　　杉本　徹‡

芝浦工業大学大学院　理工学研究科†　　芝浦工業大学　工学部‡

1. 研究背景と目的

データベースから適切な応答を選択する対話システムの性能は、システムが持つデータベースの質に依存する。このようなデータベースを拡張する研究として、収集したユーザ発話から未知の質問例を検出し、その質問例に対する応答文を生成する研究[1]や元々データベースに存在する質問例とテンプレートを用いて新しい応答文を生成する研究[2]などがある。しかし、近年盛んに研究されているニューラルネットワークを用いた応答文選択では、質問応答データベースに存在するような質問例は存在しないため、従来の手法を利用することはできない。

そこで、本研究では、人手で作成した応答文候補の集合をもとに文書間類似度を用いてWebサイトから新たな応答文候補を自動取得する手法を提案する。

2. 提案手法

2.1 処理の流れ

処理の流れを図1に示す。まず、人手で作成した応答文候補の集合とWebサイトの1ページに含まれている本文全体をそれぞれ一つの文書と見なし、文書ベクトルを計算する。その後、それらの文書ベクトルのコサイン類似度を計算し、既存の応答文候補の集合と内容が類似しているページを特定する。そして、そのページの各文に対して重みを考慮した文ベクトルの計算を行い、ページ全体の文書ベクトルとの類似度が高い文を応答文候補として抽出する。

2.2 文書ベクトルの計算

文書ベクトルの計算では、Word2vec[3]で学習した単語分散表現を各文書における単語の出現回数で重み付けした単語ベクトルを用いる。本研究では、以下の4つの手法で文書ベクトルの計算を行い、比較を行う。なお、手法3,4では大規模コーパス中の単語の出現回数としてGoogle 1-gram[4]を用いる。

手法1 : 各文書に出現する単語の単語ベクトルの平均ベクトルを文書ベクトルとする。

手法2 : 各文書に出現する単語の単語ベクトルの成分ごとの最大値から構成されるベクトルを文書ベクトルとする。

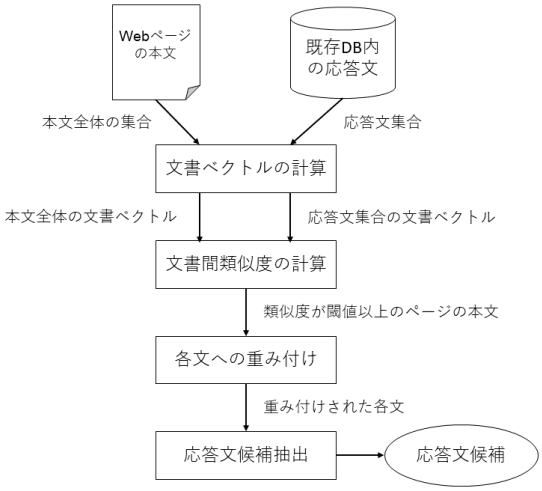

図1　処理の流れ

手法3 : 手法1で各単語の大規模コーパス中の出現回数を考慮に入れたもの。

手法4 :手法2で各単語の大規模コーパス中の出現回数を考慮に入れたもの。

文書ベクトルをvec(d)、単語分散表現をvec(w)、文書内の単語の出現回数をcount(w,d)、Google 1-gram中の単語の出現回数をcount(w,G)とすると、手法1~4は以下の式(1)~(4)のように表される。

$$\mathrm{vec(d)} = ave_w\ (vec(w) * count(w,d)) \quad \text{式(1)}$$

$$\mathrm{vec(d)} = max_w\ (vec(w) * count(w,d)) \quad \text{式(2)}$$

$$\mathrm{vec(d)} = ave_w\ (vec(w) * count(w,d) \,/\, \log count(w,G)) \quad \text{式(3)}$$

$$\mathrm{vec(d)} = max_w(vec(w) * count(w,d) \,/\, \log count(w,G)) \quad \text{式(4)}$$

本文全体の文書ベクトルと応答文集合の文書ベクトルを計算した後、それらの文書ベクトルのコサイン類似度を計算し、閾値以上の類似度を持つWebページを応答文抽出の対象とする。

2.3 応答文候補の抽出

前節の方法で選ばれたWebページ内のすべての文に対してスコア付けを行い、閾値以上の文を応答文候補として抽出する。文のスコアは、前節の手法1~4をWebページ内の各文に適用することで求めた

Collecting response sentences from Web pages using document similarity

†Keisuke Takeuchi ‡Toru Sugimoto

†Graduate School of Engineering and Science, Shibaura Institute of Technology

‡College of Engineering, Shibaura Institute of Technology

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

文ベクトルとその文が属すWebページの文書ベクトルの類似度と、その文が応答文として適切な表現や文の構造(例：名詞＋“とは”、名詞＋“では”)をしているかに応じた得点の和として定義する。

3. 対象Webページの選択に関する評価実験

3.1 評価方法

大学や研究室の紹介やそこで学べる専門用語の解説などを行う対話システムを想定し、そのような対話システムの応答文として適切な文を含んでいるWebページを特定できるかを評価する。実験対象として大学や学科のホームページとWikipediaから30ページを選び、各ページに適切な応答文が含まれているかを人手で判断したものを正解データとして、システムの出力の適合率と再現率を求めた。

3.2 評価結果と考察

手法別の評価結果を表1に示す。それぞれの手法における閾値には適合率が最大となるような値を用いている。

表1 対象Webページの選択に関する評価

手法	適合率	再現率
手法1	16/18(89%)	16/22(73%)
手法2	18/24(75%)	18/22(82%)
手法3	15/15(100%)	15/22(68%)
手法4	17/24(71%)	17/22(77%)

適合率は抽出した15ページ中15ページが正解データであった手法3が、再現率は正解データ22ページ中18ページを抽出することができた手法2が最も高いという結果になった。

4. 応答文抽出に関する評価実験

4.1 評価方法

3節の実験で正例と判断されたWebページに含まれる全文に対して、応答文抽出の処理を行い、抽出された文からランダムに100文を選び評価した。評価基準は、「応答文として出力したい内容が含まれているか」と「出力したい内容が含まれている場合、応答文として文の構造や表現が適切であるか」の2点である。また、本研究ではベースラインとして既存の応答文集合に含まれる単語をキーワードとし、Webページの本文集合からキーワードを含む文を抽出する手法と比較をしている。

4.2 評価結果及び考察

手法別の評価結果を表2、実際に抽出した文の例を表3に示す。

手法1では54%、手法3では58%とベースラインの手法よりも高い評価を得た。一方で、手法2は35%、手法4は37%とベースラインを下回る結果となった。これは、3節の実験で正解以外のページも抽出してしまい、それらのページ全体の内容と類似している文を応答文として多く抽出してしまっているためであると考えられる。また、手法1と手法3の結果の比較より、大規模コーパス中の単語の出現回数を考慮して単語ベクトルを計算することで、既存データベースの特徴をより反映した応答文抽出ができることがわかる。

表2 応答文抽出に関する評価

手法	内容が適切	構造が適切
ベースライン	48/100(48%)	34/48(71%)
手法1	54/100(54%)	46/54(85%)
手法2	35/100(35%)	28/35(80%)
手法3	58/100(58%)	45/58(78%)
手法4	37/100(37%)	29/37(78%)

表3 抽出した応答文の例

文1	工学部では、多くの工学系大学で行われてきた知識偏重教育を見直し、「ものづくり」に力点を置いた教育を行なっています。
文2	強化学習とは、結果に報酬や罰を与えることにより行動を選択するルールの強化や抑制を行い、徐々に最適な行動選択法を作り上げて行く学習方式です。

5. まとめと今後の課題

本研究では、人手で作成した応答文候補の集合をもとに文書間類似度を用いてWebサイトから新たな応答文候補を自動取得する手法を試みた。評価実験の結果、手法3ではベースラインを上回る58%の精度で既存データベースの特徴を反映した応答文を収集することができた。

今後は、抽出された正しい応答文の特徴を調査し、さらなる精度の向上を目指していきたい。

参考文献

[1] 竹内翔太他，“音声情報案内システムにおける質問応答データベース構築コスト削減の検討”，情報処理学会研究報告 Vol.2009-SLP-77 No.15，2009

[2] 森本高弘他，“質問応答データベースの自動作成に基づく音声対話システムの評価”，情報処理学会研究報告 Vol.2008-SLP-74(46) pp.267-272，2008

[3] Mikolov，T. 他，“Distributed representations of words and phrases and their compositionality”，Advances in neural information processing systems，pp.3111-3119，2013

[4] 工藤拓，賀沢秀人，“Web 日本語Nグラム第1版”，言語資源協会，2007

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.