6Q-03

発表スライド内の用語に関する質問の自動生成

Jung Dawoon† 岡本 康佑‡ 松原 茂樹‡ 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 大学院情報科学研究科‡

1 はじめに

大学のセミナーなどの発表を伴う会議で活発な議論が展開されるためには、参加者が発表の内容を正しく理解する必要がある。しかし、発表スライド内には説明が不足している用語が含まれることがある。このような場合でも会議の雰囲気や立場によっては参加者自身が用語の意味について質問しにくいことがある。

本稿では、発表スライド内に出現する用語のうち、説明が不足している用語を検出する手法について述べる。検出された用語の意味を問い合わせる質問文を自動生成することにより、コミュニケーションロボットなどが代行して問い合わせることも可能となる。

2 スライド内の用語に対する質問の生成

本研究では、スライド内のある用語に対して、その意味が分からない議論参加者が存在する用語を**説明が不足している用語**とする。説明が不足している用語を含むスライドの例を図１に示す。この例では「オドメータ」が説明が不足している用語である。

図 1：説明が不足している用語を含むスライド

発表スライド内には説明が不足している用語が含まれる場合があり、このような場で会議に加わるコミュニケーションロボットがその用語の意味を自律的に質問することは、参加者による議論の理解に有効である。

コミュニケーションロボットが自律的に質問するためには、発表スライドから説明が不足している用語を検出し、その用語に適した質問文を生成する必要がある。著者らの研究室では、ディスカッションマイニングという会議記録システム[1]を運用し、議事録コンテンツを蓄積しており、その議事録を分析したところ、用語を質問する発言においては、用語が分からない理由が述べられる傾向があること、及び、その理由によって発表者の回答が異なる場合があること、を観察している。用語が分からない理由として以下の４つが挙げられる。

1. 用語の説明はあるが、それが理解できない
2. 用語を初めて聞いた
3. 用語を以前聞いたが、その意味を失念した
4. 用語は既知だが、スライドでの意味が曖昧

説明が不足している用語に対して該当する理由を抽出できれば、用語とその理由に基づき、あらかじめ用意されたテンプレートを選択することによって質問文を生成できる。質問テンプレートの例を図２に示す。

1.	説明がよく理解できませんので、 ・・・という用語について説明してください
2.	これまで聞いたことがない用語なので、 ・・・について教えてください
3.	自分が忘れているだけかもしれませんが、 ・・・という用語はどういう意味ですか
4.	・・・という表現の意味が曖昧だと思いますので 説明して頂けますか

図 2：質問文テンプレートの例

本稿では、コミュニケーションロボットが発表スライド内の説明が不足している用語について質問を自動生成するために、対象となる用語の検出と参加者が分からない理由の推定手法について述べる。

3 説明不足の用語の検出とその理由の判定

3.1 説明が不足している用語の検出手法

本稿で提案する検出手法では、発表スライド内のあらゆる用語に対して、説明が不足しているか否かを二値分類することにより検出する。分

Automatic generation of questions about terms in presentation slides
†JUNG, Dawoon (jung@nagao.nuie.nagoya-u.ac.jp)
‡OKAMOTO, Kosuke (okamoto@nagao.nuie.nagoya-u.ac.jp)
‡MATSUBARA, Shigeki (matubara@nagoya-u.jp)
‡NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
Department of Information Engineering, Nagoya University†
Graduate School of Information Science, Nagoya University‡

類には SVM を用いる。SVM で使用する素性として、用語の言語的な特徴に加えて、用語が出現するスライドの構造や過去のスライドなどの非言語的な特徴を用いる。本手法で使用する素性を以下に示す。

- 用語の言語的な特徴による素性
  - 単一語か複合語か
  - アルファベット・カタカナを含むか否か
  - 固有名詞を含むか否か
  - 一般用語か否か
  - 固有表現か否か
  - 固有表現のとき、その固有表現の種類
  - 情報処理分野の専門用語か否か
- 非非言語的な特徴による素性
  - 何枚目のスライドに登場するか
  - スライド内の階層の深さ
  - 用語を含む要素に対する子ノードの個数
  - スライド中の出現回数
  - 発表者の過去スライド中の出現回数
  - 発表者の過去スライドからの経過回数
  - 他人のスライドに出現する出現回数

Wikipedia の説明文に閾値以上出現する用語を一般用語とした。また、Wikipedia の説明文ではある閾値以下の出現に留まるものの、情報処理学会第 78 回全国大会論文集には出現する用語を情報処理分野の専門用語とした。固有表現抽出には CaboCha[2]を利用した。

3.2 説明が不足している理由の推定

本研究では、説明が不足している用語に対する理由は、2 節に示した、用語が分からない理由のいずれかであるとした。本手法では、3.1 節の手法により説明が不足していると検出された用語に対して、説明が不足している 4 つの理由それぞれについてそれに該当するか否かを二値分類により判定する。各判定は SVM によって行う。推定に使用する素性としては、3.1 と同様の特徴を用いる。

4 評価実験

4.1 説明が不足している用語の検出実験

発表スライド内の説明が不足している用語の検出実験を行った。分類には R 言語の kernlab[3]を利用した。実験データとして著者らの研究室のゼミで使われた、12 回分の発表スライドに含まれる 6034 個の用語を用いた。学部生及び大学院生 11 人が実験データ中の分からない用語をマーキングしたものを正解データとした。実験は 10 分割交差検定により行った。非言語的な特徴による素性を除去した場合を比較手法とし、適合率、再現率、及び、その調和平均である F 値を用いて検出性能を算出した。

$$\text{適合率} = \frac{\text{正しく判別された用語数}}{\text{説明が不足していると判別された用語数}}$$

$$\text{再現率} = \frac{\text{正しく判別された用語数}}{\text{説明が不足している用語数}}$$

実験結果を表 1 に示す。提案手法は比較手法と比べ F 値が高く、非言語的な特徴の利用により検出性能が上がることを確認した。

表 1：評価実験の結果

	適合率	再現率	F値
提案手法	54.2%	40.8%	45.1%
比較手法	14.1%	73.3%	23.5%

4.2 説明が不足している理由の推定実験

説明が不足している用語に対し、その理由を推定する実験を行った。分類ツール及び実験データとその分割法は 3.1 節と同様である。説明が不足している用語 249 個に対し、学生が回答した分からない理由を正解データとした。非言語的な特徴による素性を除去した場合を比較手法として、正解率を算出した。

$$\text{正解率} = \frac{\text{正しく予測された用語の数}}{\text{全説明が不足している用語の数}}$$

実験結果を表 2 に示す。提案手法は比較手法と比べ正解率が高く、非言語的な特徴を利用することの有効性が示された。

表 2：評価実験の結果

	正解率
提案手法	41.5%
比較手法	11.7%

5 おわりに

本稿では、会議での質問文の自動生成とその利用を目的に、発表スライド内の説明が不足している用語を検出し、その理由を推定する手法を提案した。今後は、提案手法の有効性及び、コミュニケーションロボットによる問い合わせの有効性について評価する予定である。

参考文献

[1] 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. TL, Vol. 112, No. 339, pp.59-64, 2012.
[2] 山田, 工藤, 松本, Support Vector Machine を用いた日本語固有表現抽出, 情報処理学会論文誌, Vol.43, No.1, 2002.
[3] Package Kernlab, https://cran.r-project.org/web/packages/kernlab/kernlab.pdf