

対話システムにおける話題誘導のための 単語間の関係性を用いた応答文生成*

山内 祐輝, Graham Neubig, Sakriani Sakti, 戸田 智基, 中村 哲 (奈良先端大)

1 はじめに

対話システムにおける応答文生成は、自然で円滑な対話を実現する上で、重要な要素技術の一つである。システム構築の汎用性を高める手法として、テンプレートをを用いた手法 [1] やフレーズ単位に基づく統計的手法 [2] が提案されており、主にユーザ側の目的に沿った対話での応答を対象として、その有効性が確認されている。我々は、システム側が特定の目的を持ち対話を進める説得対話における応答に着目する。この種の対話では、ユーザの興味を誘導するという要素が含まれるため、現在の話題から意図した別の話題へと誘導するための応答文(話題誘導応答文)の生成が必要となる [3]。

本稿では、ある単語に関する話題から別の単語に関する話題へと誘導するために、テンプレートをを用いた話題誘導応答文生成法を提案する。人手で用意するテンプレートに対して、誘導対象となる単語データベースの中から埋め込み対象となる単語ペアを効率的に抽出するために、概念辞書と Web 検索という情報源を活用する。

2 対話システムにおける話題誘導

本研究では、対話のタスクとして、所属研究室を選定する学生に対して、特定の研究室へと勧誘する説得対話を対象とする。説得対話システムは、個々の研究室に関連する単語(研究分野等)をデータベースとして保持しており、各単語に関する話題で対話を進める。その際に、説得を行うために、ユーザが興味を持つ話題を推定しつつ、現状の話題(Current)から勧誘目標とする研究室に関する話題(Next)へと、ユーザの興味を誘導する。話題誘導応答文生成では、対話制御部から出力される応答指令(例: GUIDE(⟨Current⟩, ⟨Next⟩))に対して、自然に話題を誘導できる応答文を生成する。

3 テンプレートによる話題誘導応答文生成

現在の話題に関する単語と誘導する話題に関する単語の関係性を用いて、テンプレートにより応答文を生成する。単語の関係性を抽出するために、概念辞書に基づく手法と、Web 検索を用いる手法について述べる。

3.1 概念辞書に基づく手法

概念辞書とは、概念と概念間の関係を記述した辞書である。その例としては、WordNet[4]がある。本稿では、概念辞書の作成は人手で行い、辞書の内容として、概念を表す単語と概念間の上位/下位関係を記述する。辞書の構造例を Fig. 1 に示す。登録した単語は、奈良先端科学技術大学院大学情報科学研究科に属する単語である [5]。最上位概念に研究科名である情報科学を記述し、その下位概念には領域名、研究室名、研究テーマの単語を記述する。また、共通の単語が含まれる複合名詞に対して、その共通の単語(+α)を上位概念に追記する。

概念辞書に記述のある概念間の関係を基に応答文テンプレートを生成し、現在の話題と誘導する話題を入れることで応答文を生成する。例えば、上位語から下位語へ誘導する応答文テンプレートとして「⟨Current⟩の研究には⟨Next⟩もありますよね。」を作成する。

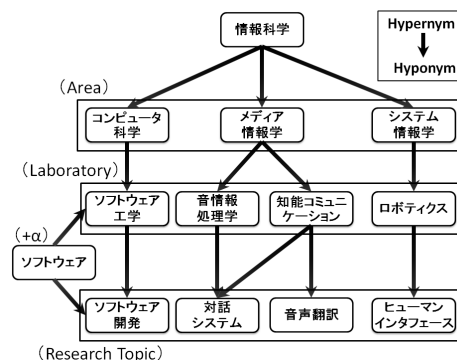


Fig. 1 Example of the concept dictionary structure

また、概念辞書の構造上で共通の子孫(上位/下位のどちらか一方を辿って繋がりがある単語)となる⟨Hypernym⟩/⟨Hyponym⟩を持つ話題間(例えば、Fig. 1の「対話システム」と「音声翻訳」で、この場合の共通の子孫は「知能コミュニケーション」である)のテンプレートも作成する。例えば、上位語に共通の子孫を持つ場合の応答文テンプレートとして「⟨Current⟩は⟨Hypernym⟩の研究ですよ。⟨Hypernym⟩の⟨Next⟩の研究には興味ないですか。」を作成する。

3.2 Web 検索を用いた手法

Web 情報は、単語間の関係性を取得する上で有効な情報源である [6], [7]。本稿では、Web 検索により取得した単語間の関係性を利用して、応答文テンプレートに対して埋め込み対象となる単語ペアを抽出する手法を提案する。提案法の手順を以下に示す。

1. 話題誘導応答文に利用できる可能性が高い関係を想定した応答文テンプレートを作成
(例) 応答文テンプレート:
「⟨Current⟩は⟨Next⟩にも使われていますよね。」
2. 応答文テンプレートを基に検索テンプレートを生成
(例) 検索テンプレート: 「⟨Current⟩を用いた⟨Next⟩」
3. 検索結果を基に応答文テンプレートに当てはまる関係のある単語ペアを決定
(ヒット件数が0の場合は関係がないと判断)

誘導に利用する単語ペアを決定する際の尺度として、ヒット件数と単語間の相互情報量を用いる。ページのヒット件数はヒット数が多いほど単語間の関係性が強いと判断する。相互情報量を求める式を以下に示す。

$$I(x, y) = \frac{P(x, y)}{P(x) * P(y)} = \frac{N * C(x, y)}{C(x) * C(y)}$$

$P(x, y), P(x), P(y)$ はそれぞれ、検索テンプレート、検索テンプレートで用いた ⟨Current⟩, ⟨Next⟩ がヒットする確率、 N は検索サイトの総ページ数、 $C(x, y), C(x), C(y)$ はそれぞれのヒット件数である。

4 実験的評価

4.1 実験条件

辞書に登録された単語から単語に話題を誘導することを想定した話題誘導応答文を生成する。評価基

* Answer Sentence Generation using Relationships between Words for Guiding Users to New Topics in Spoken Dialog Systems by YAMAUCHI, Yuki, NEUBIG, Graham, SAKTI, Sakriani, TODA, Tomoki, NAKAMURA, Satoshi (NAIST)

Table 1 Experimental conditions

Evaluation data	3400 (sentences) (20 sentences × 10 methods for each person)
Human evaluators	17
Subject	Words regarding the Graduate School of Information Science in NAIST (148 words)
Search site	CiNii (A search site for Japanese papers)

準は、辞書内の単語ペアのカバー率と生成された文章の内容の自然性(1. 不自然, 2. どちらとも言えない, 3. 自然)とする。カバー率とは、辞書に存在する単語ペアの内、各手法を用いた際に話題誘導応答文が生成できる割合である。

実験条件を Table 1 に示す。概念辞書に登録する単語の数は 148 個で、辞書の作成に要した時間は約 8 時間である。応答テンプレートは、上位語から下位語、下位語から上位語、上位に共通の子孫を持つ話題、下位に共通の子孫を持つ話題への話題誘導応答文の 4 種類である。また、Web 検索に用いた手法で利用した検索テンプレートは、(a) 「〈Current〉を用いた 〈Next〉」、(b) 「〈Current〉技術 〈Next〉」、(c) 「〈Current〉分野 〈Next〉」、(d) 「〈Current〉〈Next〉」の 4 種類である。4 種類の内、値が最も高い関係を用いて応答文を生成する。ただし、ヒット件数を尺度とする際、(a)、(b)、(c) の 3 種類を用いて検索し、3 種類ともヒットしなかった場合に (d) を用いる。

実験は、概念辞書を用いる手法 3 種類、Web 検索を用いる手法 6 種類、単語間の関係を用いない手法 1 種類で行う。具体的な手法については以下に示す。Web 検索の頻度計算には、日本語論文検索サイト「CiNii」[8] の検索結果を使用する。

- 概念辞書を用いる手法
 - 辞書に記述のある単語ペア (Dic1)
 - 辞書の構造上で子孫関係の単語ペア (Dic2)
 - 共通の子孫を持つ単語ペア (Dic3)
- Web 検索を用いる手法
 - 検索ヒット件数を選択尺度に用いる場合
ヒット件数の上位 10% (Hit1), 10 ~ 50% (Hit2), 50 ~ 100% (Hit3)
 - 相互情報量を選択尺度に用いる場合
相互情報量の上位 10% (MI1), 10 ~ 50% (MI2), 50 ~ 100% (MI3)
- 単語間の関係を用いない手法
 - 関係を考慮しないでテンプレートに単語ペアを入れて応答文生成 (ALL)

4.2 実験結果および考察

各手法の単語ペアのカバー率を Table 2 に示す。概念辞書を用いた手法全体でのカバー率は 92.5% で、Web 検索を用いた手法全体でのカバー率は 40.3%、関係を用いない手法でのカバー率は 100% である。概念辞書のカバー率が 100% でないのは、 $+\alpha$ の単語は上位概念がなく、応答文を生成できない単語ペアが存在するためである。

各手法における主観評価結果を Fig. 2 に示す。Fig. 2 より、概念辞書に記述のある単語ペアを用いた Dic1 の自然性が最も高く、関係を用いない ALL の自然性が最も低いことが確認できる。

概念辞書を用いた手法に関しては、Dic1 の自然性が最も高く、Dic3 の自然性が最も低いことが確認できる。Dic2 の自然性が Dic1 より低くなった原因としては、単語間の関係の距離(単語間にあるノードの

Table 2 Coverage [%]

Method	Coverage	Method	Coverage
Dic1	1.8	Hit/MI1	4.0
Dic1+2	4.6	Hit/MI1+2	20.2
Dic1+2+3	92.5	Hit/MI1+2+3	40.4
ALL	100		

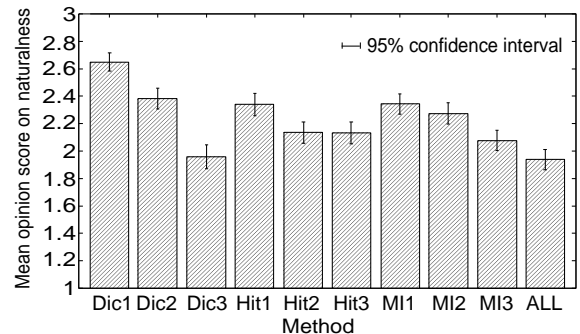


Fig. 2 Result of subjective evaluation

数)が離れたため単語間の関係性が薄くなり、自然性が低くなったと考える。Dic3 では、単語間の直接的な関係性が伝わらず、自然性が低くなったと考える。

Web 検索を用いた手法に関しては、ヒット件数と相互情報量共に、上位 10% (Hit/MI1) の自然性が最も高いことが確認できる。また、Hit2 と MI2 を比較すると、MI2 の方が自然と評価されていることが確認できるため、Web 検索での選択尺度において、相互情報量を用いた場合の方が自然性の高い単語ペアが上位に位置づけられたと考える。なお、概念辞書を用いた場合と Web 検索を用いた場合を比較すると、Dic2 と Hit1, MI1 が同等の自然性と評価されていることが確認できる。

以上より、自然性とカバー率を考慮すると、辞書に記述がある単語ペアに関しては手法 Dic1 を用いて応答文を生成し、関係の記述がない単語ペアに関しては Web 検索を用いて、選択尺度を相互情報量として応答文を生成するのが良いと考える。また、新たにシステムを構築する場合、概念辞書を用いる手法は人手で単語間の関係を記述する必要があるが、Web 検索を用いる手法ではテンプレートを用意することで生成可能である。

5 おわりに

本稿では、単語間の関係性を用いた話題誘導応答文の生成について提案し、生成された応答文の自然性について主観評価を行った。その結果、概念辞書に記述のある関係を用いた応答文や、Web 検索により計算した相互情報量が高い関係性を用いた応答文において自然性の高い話題誘導応答文が生成されることが確認できた。今後の課題として、応答文テンプレートの自動生成や辞書に登録されていない話題からの話題誘導応答文生成手法の検討がある。

参考文献

- [1] 河原 他, “音声対話システム.” オーム社, 2006.
- [2] F. Mairesse et al., Proc. ACL 2010, pp. 1552–1561, 2010.
- [3] 濱, 信学技報. HCS 95-3, Vol. 95, No. 88, pp. 9–14, 1995.
- [4] G. A. Miller et al., International Journal of Lexicography, Vol. 3, No. 4, pp. 235–244, 1990.
- [5] NAIST 情報科学研究科, <http://isw3.naist.jp/Contents/Research-ja/Lablist-ja.html>
- [6] P. Pantel et al., Proc. ACL 2006, pp. 113–120, 2006.
- [7] M. P. Kato et al., Proc. CIKM 2009, pp. 27–36, 2009.
- [8] CiNii, <http://ci.nii.ac.jp>