E-010

Freebase を用いた対話システムにおける話題の選択
Selecting Topics in Dialogue Systems using Freebase

北村　悠起†　　高瀬　裕‡　　中野　有紀子‡
Yuki Kitamura　Yutaka Takase　Yukiko Nakano

1. はじめに

近年のロボット技術の発展にともない，OHaNAS[1] や Pepper[2] のような対話ロボットの実用化が進んでおり，対話機能の高度化も求められている．我々は，その 1 つとして，ロボットから幅広い話題を提供することは重要な対話機能であると考える．話題の種類や幅が限定されていると，ユーザはロボットとの会話にすぐに飽きてしまい，会話は長く継続しないだろう．しかし，話題の種類を増やすためにルールやコンテンツを開発するのには，大きなコストを要する．

そこで本研究では，大規模な知識ベースを用いることにより，現在の話題と関連した話題を展開する対話システムを開発する，これにより，会話が長続きする対話ロボットを目指す．

2. 関連研究

Han et al. [1] は，文章中の人名を抽出し，さらにその人名に関連するキーワードを Freebase [2] (後述)を用いて検索し，これを対話システムの応答文に取り込んでいる．中野ら [3] は，発話内にある名詞に対し，Word2Vec [4] で関連する名詞を選び，これを用いてシステムの応答文を生成している．

これらの研究を参考にし，本研究では，Freebase を利用して，固有名詞だけでなく一般名詞に対しても関連語を見つける手法を提案する．さらに，Freebase から得られるカテゴリ情報を用いて，Word2Vec から得られる関連語を絞り込むことにより，不適切な関連語の選択を防ぐための処理も提案する．

3. 大規模知識ベース Freebase

本研究では，大規模知識ベースとして，Freebase を用いる．Freebase はデータが互いにリンクされている大規模なデータベースである．データは全て主語，述語，目的語の 3 つの要素を持ったトリプル構造となっている．このトリプル構造を利用することにより，必要な情報を効率的に得ることが可能となる．本研究では，検索速度の向上のため，Freebase 中の必要なデータのみを格納したデータベースを作成した．このデータベースには，Freebase の述語の一つである Notable type が含まれている．Notable type とはその名詞の種類を表わしており，例えば，「東京都」の Notable type は「日本の都道府県」となる．この Notable type を用いることで，同じカテゴリから関連語が選択でき，この関連語を用いたより適切な応答文生成が可能になると考えた．

4. 提案システム

提案システムの主な構成を図 1 に示す．まず入力理解部では，会話文から名詞を抽出し，関連語検索部では，抜き出した名詞の関連語を大規模知識ベース Freebase と Word2Vec を用いて選択する．出力生成部では，関連語検索部が決定した関連語を新たな話題とし，この名詞を使った応答文を生成する．

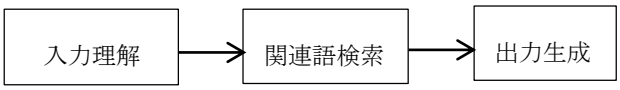

図 1　提案システム

4.1 入力理解部

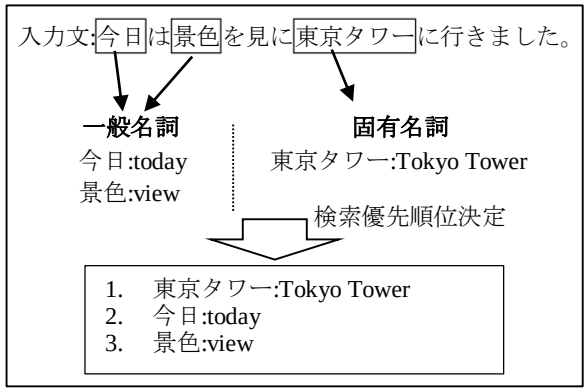

図 2　入力理解部

図 2 に入力理解部の処理を示す．まず形態素解析器 juman [5] を用いてユーザの入力文から名詞を抽出する．例えば，「東京タワー」が「東京」と「タワー」に分かれてしまうことを避けるため，連続した名詞は 1 つの複合名詞として扱う．また，Freebase は英語で書かれているため，これを使った関連語検索を行うために，Microsoft Translator API [6] を用いて抽出した名詞を英語に翻訳した．

次に，複数の名詞が得られた場合，関連語検索の優先順位を決定する．本研究では，固有名詞と一般名詞に分類し，固有名詞を一般名詞に優先させることとした．検索時に記事が見つからないなどの問題が生じたときには，次の順位の名詞を検索する．

4.2 関連語検索部

図 3 に関連語検索部の処理を示す．

記事特定：まず，Freebase を用いて，抽出された名詞を記事名とする記事を検索する．しかし，検索結果で同じ記事名を持つ記事が複数見つかる場合がある．例えば，「東京タワー」という名詞を検索すると，59 個の記事が見つかる．その一部を表 1 に示す．

† 成蹊大学大学院理工学研究科
‡ 成蹊大学理工学部

[1] http://www.takaratomy.co.jp/products/omnibot/ohanas/
[2] http://www.softbank.jp/robot/consumer/products/

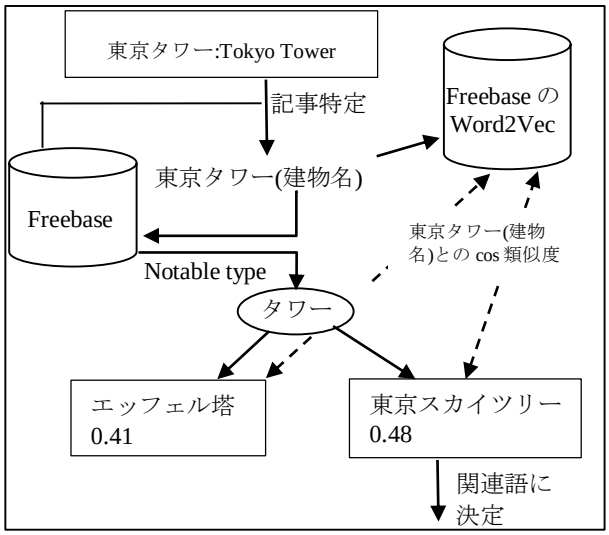

図 3　関連語検索部

表 1　「東京タワー」の記事(一部)

記事の内容	トリプル数
「東京タワー」という音楽アルバム名	26
「東京タワー」という建物	570
「東京タワー」という曲名	25

そこで，記事名を主語とするトリプル数を調べ，トリプル数の一番多い記事をその名詞における代表的な記事であると仮定した．「東京タワー」の例であれば，表 1 から，トリプル数が 570 個と最も多い「東京タワー」という建物の記事を名詞「東京タワー」の記事とする．次にこの記事の Notable type を調べる．例えば，「東京タワー」という建物の記事から得られる Notable type は「タワー」となる．

関連語の決定：まず，Word2Vec を用いて関連語の候補を見つける．Word2Vec を使うと単語間の関連度(cos 類似度)を算出することが容易であり，これに基づき関連が強い語を抽出できる．本研究では Freebase を学習データとして作成された Word2Vec のモデルを用いた．

表 2　Word2Vec と Notable type の結果

記事名	Notable type	cos 類似度
東京スカイツリー	タワー	0.48
通天閣	場所	0.45
エッフェル塔	タワー	0.41
お台場	地理的特徴	0.40

表 2 の例では，「東京タワー」との類似度の高い名詞，上位 4 つが列挙されている．次に，これらの記事の Notable type を調べると，「東京スカイツリー」と「エッフェル塔」は「タワー」である．以上から，「東京タワー」と同じ Notable type を持つ候補の中で，cos 類似度が最も高い「東京スカイツリー」が関連語として決定される．

4.3 出力生成部

応答文はあらかじめ用意したテンプレートと関連語を用いて作成する．テンプレートは，「(名詞)へ行ったことがありますか」や「(名詞)を知っていますか」といったものを用意する．例えば，「 (名詞)を知っていますか」という文の「(名詞)」に，関連語検索部で得られた「東京スカイツリー」を挿入することにより，「東京スカイツリーを知っていますか」という応答文を生成し，話題を展開する．

5. 実行例

実装システムは，テキスト入力での対話を行う．その対話の一例を図 4 示す．太字の部分は現在の話題を示す．この例では，「東京タワー」から「東京スカイツリー」，そして「広州塔」へと話題が展開している．

ユーザ:今日，**東京タワー**に行きました
システム:そうなのですか．ところで，**東京スカイツリー**を知っていますか
ユーザ:はい，知っていますよ
システム:ぜひ，**東京スカイツリー**について教えてください
ユーザ:**東京スカイツリー**は高さ 634m のタワーです
システム:そうなのですか．ところで，東京スカイツリーと同じ種類で**広州塔**を知っていますか

図 4　実行例

6. おわりに

本研究では，ユーザ発話に関連した話題を展開する対話システムを目指し，大規模知識ベース Freebase および Word2Vec を用いて現在の話題に関連した語を選定する方法を提案した．そして，この関連語を用いて現在の話題から別の話題へ誘導する応答文を生成した．

今後の課題として，現在は出力生成のテンプレートが不十分であるため，これを追加し，様々なバリエーションの応答文を生成できるようにする必要がある．また，実際にユーザと対話をさせることにより，本手法の有効性を検証する予定である．

参考文献

[1] Sangdo Han, Jeesoo Bang, Seonghan Ryu, Gary Geunbae Lee. Exploiting knowledge base to generate responses for natural language dialog listening agents. The 16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)
[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247–1250. ACM.
[3] 中野 哲寛，荒木 雅弘，雑談対話システムにおける単語分散表現を用いた話題展開手法，言語処理学会第 21 回全国大会論文集，2015
[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
[5] 日本語形態素解析シ ス テ ム JUMAN ：http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN.
[6] Microsoft Translator api: http://www.microsofttranslator.com/dev/