

# 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

中野 哲寛

荒木 雅弘

京都工芸繊維大学 大学院工芸科学研究科 情報工学専攻

nakano@ii.kit.ac.jp araki@kit.ac.jp

## 1 はじめに

現在まで対話システムはタスク指向型と非タスク指向型の2種類が研究されてきた。タスク指向型の対話システムにおいては、QAシステム等のある目的に対する対話システムが研究されてきた。一方、人間同士での話には目的のない対話が行われる事が多いため、非タスク指向型の対話システムが開発されていくようになった。近年ではTwitterやWikipediaのような大規模データを用いて、雑談機能を拡張させるような研究が行われてきた。具体的な例としてはTwitterの文章を返答として用いる手法 [1][2] や、検索エンジンから得られた情報を返答として用いる手法 [3] 等が報告されている。

しかし対話に必要な要素である話題展開についての研究 [2][3] は少なく、どのように話題展開を行うべきなのかについて研究されてこなかった。そのため本研究では、表1のような話題展開を行うシステムを提案する。

我々は長く続く雑談の実現を目的とした発話制御システムを開発しており図1にシステムの構成を示す。このような発話制御システムの構築には、ユーザが発話しやすいと感じるシステム応答をどのように生成するのが重要となってくる。これはシステムの構成にあたり、ユーザ・システム間で得られる対話から対話破綻指数と対話継続指数を学習することで、適切なシステム応答を選択できると考えて行っている。ここで述べている対話破綻指数は、ユーザ・システム間で得られた発話が適切かどうかを判定する指標であり、対話継続指数はシステム応答を受けてユーザが次の発話をしやすいかどうかを評価した指標である。

これらの指標に基づいて学習されたパラメータを用いて対話制御を行うためには、各発話に対してシステム応答を行うようなモジュールが必要となる。システム応答を生成するためのモジュールの候補としてルールベースや統計ベース等の手法があげられるが、それぞれのモジュールには一長一短があるためユー

表 1: システムの対話例

発話者	発話
User	今年は阪神の藤浪は調子よかったね
System	今年は阪神の鳥谷も調子よかったね
User	今年はガンバ大阪もよかったよね
System	今年は浦和レッズもよかったよね

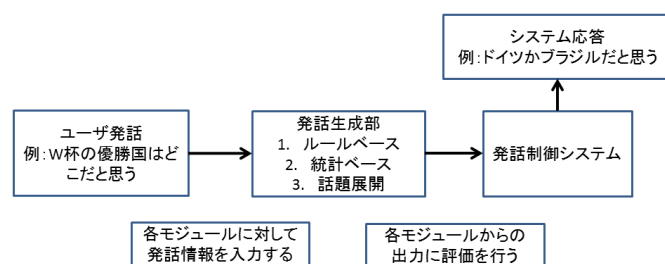


図 1: 対話制御システムの構造

ザ・システム間の発話を進めるようなモジュールの一つとして、従来ではあまり研究されていなかった話題展開手法のみを行うモジュールの作成を行った。

本手法では単語分散表現の学習においてword2vec<sup>1</sup> [4] を用いている。これは深層学習を行うことで、単語の関係を効率的に学習することができると報告されている。さらに大規模データに対して学習を行っても他の手法より短時間で効率的に学習を完了できると報告されていること [5] より、単語分散表現を用いた対話システムの提案をしている。

## 2 関連研究

話題展開を行った先行研究として Higashinaka らの手法 [2] や前田らの手法 [3] が挙げられる。Higashinaka らの手法 [2] では、Twitter データに対して制限をか

<sup>1</sup><https://bitbucket.org/yoavgo/word2vecf>

けることでより適切なシステム応答を学習することができるという報告されている。前田らの手法 [3] では、話題語として抽出された固有名詞をスタック上に積み上げることで、ユーザが興味を持っている話題を掘り下げるといった手法を用いて話題展開を行っている。この際に形態素解析から出現した名詞に対して興味ベクトルを計算することで、ユーザが興味を持っているであろう単語を検索エンジンから獲得し、その単語から発話生成を行うことで話題展開を達成している。

単語分散表現では代表的な手法として Micolov らの手法 [6] が挙げられる。Micolov らの手法 [6] では、単語の関係を示すための手法として深層学習を用いた単語分散表現を提案している。単語分散表現を用いることで対象の単語の類義語を検出できるため、さまざまな研究 [7][8] が行われている。また単語分散表現の学習を行う際に Skip-gram 法や CBOW (continuous bag-of-words) 法を用いて学習を行うことができるが、学習精度の向上に向けてさまざまな手法が提案されている。その中でも Levy らの手法 [4] では、単語分散表現を学習する際に単語の係り受け情報を用いることでより効率的な学習ができることが報告されている。Levy らの手法 [4] の特徴として対象となる単語の属するカテゴリが、Micolov らの手法 [6] に比べ明確に検出できることも報告されている。そのため従来法よりもカテゴリ上で関連した単語が検出されることが日本語のデータに関しても期待される。実際に Sugiyama らの手法 [9] では係り受け情報を用いた対話システムが効果を発揮したことが報告されている。これによって日本語の係り受け情報も英語と同様に、単語分散表現から取得される単語の関係性を従来法 [6] よりもカテゴリ上で保持したものであることが考えられる。

本手法では係り受け情報を考慮した単語分散表現の学習を行うために、日本語の wikipedia の全データに対して CaboCha<sup>2</sup> を用いた係り受け解析を行うことで係り受け情報を含めた単語分散表現の学習を行っている。

### 3 話題展開を行うための発話生成

単語分散表現を用いて話題展開を行う際に、固有名詞に注目して自然な返答を生成することが有効である [3] と報告されている。その中でも Sugiyama らの手法 [9] では、名詞や述語構造に注目することでユーザの意図を理解した対話システムの提案を行っている。本

手法では以下のような順番で話題語の情報を取得することで、システム応答の作成を提案している。

1. 固有名詞・一般名詞
2. 上記以外の名詞

Sugiyama らの手法 [9] においても名詞に注目した処理が必要であることが述べられている。これは実際に話題語が名詞であるため、名詞情報の考慮が重要とされていることに基づく。全ての名詞に着目することで話題語を取得していない場合での話題展開の効果の検証を行い、どの程度効果が表れるのかを調べることで話題語が単語分散表現においても、情報を保持したまま話題展開ができるのかを確認している。

単語分散表現を用いて判定された単語をどのようにシステム応答として採用するのかということが、対話システムにおいては重要な要素である。実際に単語分散表現によって適切とされる単語がシステム応答としてふさわしい場合においても、日本語特有の言い回しによって不適切な文章となってしまうため対話システム自体の評価を下げる一因となってしまう可能性がある。そのため本手法では、ユーザ発話から検出された単語を単語分散表現から得られた結果と入れ替えることでシステム応答の生成を行っている。また名詞が存在しない発話に対してはシステム応答を行うことができないため、入力をそのままシステム応答として採用している。システム応答としての文章の正確さはユーザの発話に依存することにはなるが、単語分散表現を用いて求められた単語が正しい場合には文章として違和感がないものと想定している。

本手法では図 2 のようなシステムを用いている。またユーザ発話に対して MeCab<sup>3</sup> を用いた形態素解析を行うことで、どのような単語が入力されているのか解析を行い、その結果に応じて学習された単語分散表現から得られた情報を用いてシステム応答を生成している。またユーザ発話中の固有名詞からシステム応答を生成する場合は、名詞にかかっている機能語を「も」に変えることで文章の自然さを維持している。

### 4 評価実験

評価を行う際にどのような話題において本手法が有効となる話題展開が行われているのかを比較するために、事前にある話題についてシステム・ユーザ間で会話されたデータ [10] の 200 対話を用いている。この時 1 対話につきユーザが 10 回システムに発話をおこな

<sup>2</sup><https://code.google.com/p/cabochoa/>

<sup>3</sup><https://code.google.com/p/mecab/>

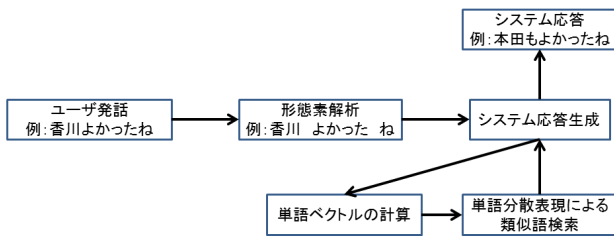


図 2: 話題展開システムの構造

表 2: 話題展開の評価

	発話ターン数
展開および文法間違いなし	226
どちらかのみ成功	243
展開および文法間違い	1531
全ターン数	2000

い、システムが 10 回ユーザに対して発話を行っており、各対話データ中のそれぞれシステム応答を提案手法のものに変えることで主観評価をおこなっている。主観評価を行う際に評価した点としては、以下の 2 点を考慮している。

1. ユーザ発話を受けてシステム応答が適切であるか
2. システム応答に文法的な間違いがないか

また本手法では、word2vecf で単語分散表現の学習を行う際に wikipedia 上の全データ<sup>4</sup>を用いて学習を行っている。また係り受け解析を行う際に未知語の解消を行うために、wikipedia のタイトルページ<sup>5</sup>を用いて辞書データの拡張を行っている。この際に辞書データは ipadic 辞書<sup>6</sup>を用いており、以下の条件を設けて辞書データの拡張を行っている。

- 「一覧」・「登場人物」・「日付」・「括弧」が対象単語の語尾に表れない
- 形態素解析による結果で一つの名詞に解析されない
- 形態素解析による結果で必ずその単語が名詞を含むこと

上記の 3 点を満たすものを一般名詞として登録している。これらの処理を施すことで未知語に対する対応を行っている。また単語分散表現を用いた場合、関係が近いと判定された単語を  $N$  個返答するようになっていく。今回は  $N = 10$  としてこの  $N$  個からランダムにシステム応答を選択する設定を行った。

<sup>4</sup>2014 年 8 月 25 日での wikipedia 日本語データ

<sup>5</sup><http://dumps.wikimedia.org/jawiki/20140825/>

<sup>6</sup><https://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz>

## 4.1 話題展開に成功した例

主観的な評価を行い話題展開に成功したと評価した例の一部を紹介する。表 3 に成功例を載せている。発話生成に関してはあまり処理を行っていないにもかかわらず、話題に広がりを持つていくことが分かる。このことからユーザは返答による文章の正確さよりも、単語のつながりから文章の意味を理解していることが考えられる。実際に本研究ではユーザの発話履歴等を考慮することなく、ユーザの発話中に存在している名詞情報から似ている単語を選択しているため、ユーザはユーザ発話と同じ文章構造でも意味が良く似ている単語であれば話題がある程度続いていると評価できることがわかる。

## 4.2 話題展開に失敗した例

話題展開の失敗例では表 4 に一部の失敗例を載せており、ユーザ発話に処理が行われた発話のみを取り上げて紹介している。結果より発話生成の処理を行っていない影響や、単語分散表現を用いて決定する際の単語の選び方に問題があることが分かる。実際に単語分散表現を用いることで同様の意味を持った違う言い回しを選択したものをシステム応答として採用しているため、発話の自然さが保持されていないことが大きな問題である。このためシステム応答として単語がうまく検出されている場合でも、文章構成による問題で不自然である応答文が見受けられた。また、関連性が高いと判定された単語の 10 個のなかから選択をして入れ替えを行っているため、同音異義語や全く関係のない単語が検出されていることも見て取れる。このように関係を検出できるからこそ、どのように単語を選択していくべきなのかを考慮する必要がある。

## 4.3 結果の分析

得られた表 2 から単語分散表現を用いた話題展開による効果として、名詞を入れ替えることによって約 10% は話題展開に成功していることが見て取れた。また文法の間違いを解消することで残り約 10% も話題展開できる可能性があることが分かる。また名詞以外の形容詞等に注目することで間違いがあるものからも、話題展開を行うことができることも想定される。このような結果になった要因としては、ユーザがシステム側に質問しているターンが多いため話題を展開することができなかったことが、大きな要因としてあげられる。

表 3: 話題展開の成功例

話題	ユーザ発話	システム発話
旅行	私は福岡に行きました。	私は名古屋も行きました。
音楽	メンデルスゾーンとか	ショパンとか
朝食	朝はパンが食べたいです	朝はチーズも食べたいです
スポーツ	私はサッカー観戦によく行きます。	私はバスケットボール観戦によく行きます。

表 4: 話題展開の失敗例

話題	ユーザ発話	システム発話
天気	そうだね。今日は天気が悪いね。涼しいけど。	そうだね。今日はとことん悪いね。涼しいけど。
ラーメン	醤油とんこつのパンチが効いたやつにしよう	醤油オリーブオイルのパンチが効いたやつにしよう
場所	じゃあお化け屋敷なんかはどうですか。	じゃあメイド喫茶もなんかはどうですか。
レストラン	今日のメニューは何でしょうね	今日の譜面も何でしょうね

前節までの例を見ていると文章の自然さが保てないため話題展開が失敗していることがよく分かる。また失敗例としてあげていたように正しい名詞が検出できているような結果が見受けられた。そのため発話生成を工夫して行う必要があり、その結果によっては提案手法がより効果的になる可能性が高い。

また形態素解析に応じて単語分散表現の学習における精度が依存しているため、辞書データの拡張を効果的に行うことでより効果的な学習を行えるように辞書データを改良する必要がある。

## 5 おわりに

本手法では、単語分散表現を用いた話題展開システムの提案を行っている。話題語として名詞のみに着目した話題展開をおこなっており、発話生成の際に名詞に対して単語分散表現を用いて関連した単語に入れ替えることによる効果が見受けられた。これにより名詞が文章中においてどれだけ重要な要素であるのかが判明した。

その一方で実際にどれだけ似ている単語を選択すべきなのかまた、以前に話された話題語の情報を考慮する必要があるどのように話題後を保持していくべきなのかが今後重要になってくる。そのため今後は、ユーザとシステム間で交わされた全ての話題語の情報を考慮することまた、他品詞である形容詞や接続詞の情報をを用いてシステム応答を生成することで、名詞のみの場合に対してどのような変化があり、どういった効果を与えることができるのかを検証する必要がある。

## 参考文献

- [1] 稲葉通将, 神園彩香, 高橋健一. Twitter を用いた非タスク指向型対話システムのための発話候補文獲得. 人工知能学会論文誌, Vol. 29, No. 1, pp. 21–31, 2014.
- [2] Ryuichiro Higashinaka, Nozomi Kobayashi, Toru Hirano, Chiaki Miyazaki, Toyomi Meguro, Toshiro Makino, and Yoshihiro Matsuo. Syntactic filtering and content-based retrieval of twitter sentences for the generation of system utterances in dialogue systems. *Proc. IWSDS*, pp. 113–123, 2014.
- [3] 前田和希, 宋きん, 國政裕友樹, 豊田博之, 韓東力. 雑談システムにおける話題展開の性能向上. 言語処理学会第 16 回全国大会論文集, 2010.
- [4] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 2, , 2014.
- [5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *In Proceedings of NIPS*, 2013.
- [7] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. A neural network for factoid question answering over paragraphs. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 633–644, 2014.
- [8] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.’s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.
- [9] Hiroaki Sugiyama, Toyomi Meguro, Ryuichiro Higashinaka, and Yasuhiro Minami. Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. *Proc. SIGDIAL*, pp. 334–338, 2013.
- [10] 東中竜一郎, 船越孝太郎. Project next nlp 対話タスクにおける雑談対話データの収集と対話破綻アノテーション (第 5 回対話システムシンポジウムおよび一般). 言語・音声理解と対話処理研究会, Vol. 72, pp. 45–50, 2014.