2L5-OS-07b-2

語の意外度に基づき話題展開する非タスク指向型対話システム

Topic Extension Method for Non-task-oriented Dialogue System Based on Serendipity Score

伊藤 直之 Naoyuki Ito　西川 侑吾 Yugo Nishikawa　大野 和久 Kazuhisa Ohno　松本 征二 Seiji Matsumoto　中川 修 Osamu Nakagawa

大日本印刷株式会社 C&I 事業部 ICT 開発本部 フロンティアテック・ラボ
Frontier Technology Lab, ICT Development Division, Communication & Information Operations, Dai Nippon Printing Co., Ltd.

1. はじめに

近年、雑談システムのような非タスク指向型対話システムの研究が盛んである。非タスク指向型対話システムは、対話によって解決すべき明確な課題を持たない対話システムのことであり、チケット予約や店舗での接客で利用されるタスク指向型対話システムと併用することで、ユーザの満足度向上や、ユーザに対して安心感や親近感を与えるといった効果が期待できる。

非タスク指向型対話システムが発話を生成するために、ユーザとの対話内容から話題を推定する必要がある。直前の対話に出現した重要語を利用して、重要語の集合から話題を自動判定し、その話題に関連するフレーズを対話システムが発話するのが一般的である。

話題を判定して適合する内容をシステムに発話させる場合、過去の話題やユーザの嗜好に過度に最適化されてしまい、新たな話題提供の無い会話になるという問題がある。実際の人間同士の雑談では、現在の話題と関連性はあるが当たり前すぎない話題への遷移がよく行われ、対話の活性化につながると指摘されている[藤本 04]。また、推薦システムにおいても、セレンディピティ(思いがけない発見)の重要性が指摘されており、目新しさや意外性の要素が求められている[奥 13]。

本稿では、ウェブサイトのテキストデータとカテゴリ情報を利用して、対話中の話題に対して意外性を持ったサイト記事を判定し、意外性を持った応答文を生成する手法を提案する。生活総合情報サイト All About(オールアバウト)のデータをコーパスとして利用し、意外性のある記事を判別する評価実験を行ったので結果を報告する。

2. 従来手法

対話システムにおけるユーザの発話内容から対話行為タイプを推定する手法として、bag of words 手法が多く用いられる。この手法では、発話内容から単語の頻度ベクトルを作成し、学習済みの分類器を用いて分類を行う[Chu-Carroll 99] [Alshawi 03]。この手法は、大量の学習データがあれば、高い精度で対話行為タイプを推定することができるため有用である。

しかし、非タスク指向型対話システムにおいて、従来型の bag of words 手法によりユーザの発話内容と関連度の大きい話題を判定して発話する場合、システム側の発話内容がユーザにとっても想起しやすい、ありきたりな内容になる可能性が高いと考えられる。また、対話を盛り上げるためには、常識的な内容のみでなく、相手に意外だと思わせるような発話が必要であるといわれている[太田 09]。

連絡先: 伊藤 直之, 大日本印刷株式会社 C&I 事業部 ICT 開発本部フロンティアテック・ラボ 〒141-8001 東京都品川区西五反田 3-5-20 E-mail: Itou-N12@mail.dnp.co.jp

3. 提案手法

本研究では、入出力のモダリティがテキストのみのテキスト対話システムを対象とする。対話システムが発話する文の生成を下記の 5 段階で行う。

1. ユーザ発話からの特徴語抽出
2. ユーザ発話と記事との関連度算出
3. ユーザ発話と記事カテゴリとの関連度算出
4. 記事の意外性スコア算出
5. 記事からの発話文生成

本研究では、発話文生成のためにウェブサイトのデータをコーパスとして用いる。今回、具体的なウェブサイトとして、All About (http://allabout.co.jp/) を利用した。All About は、ガイドと呼ばれる専門家が仕事や趣味などのテーマを紹介、解説するサイトである。All About には生活全般に関する多様な分野の記事が含まれるため、幅広い話題を扱う非タスク指向型対話システムの発話文生成への利用に有効と考えた。

All About データについて

All About サイトには、サイト利用者が所望の記事を探す際に参照するための約 1300 のカテゴリ(テーマ)が設定されており、各カテゴリの下に複数の記事が属する構造となっている。例えば、“掃除”というカテゴリの下には、『お風呂を制する者、カビを制す!』『玄関掃除のコツ』などの記事が属しており、“留学”というカテゴリには、『海外留学・海外進学までの流れ』『世界の舞台で活躍する!スポーツ留学』などの記事が属している。表 1 に、All About データの概要を示す。なお、本研究では 2014 年 2 月時点のデータを用いた。

表 1 All About データ概要 (※2014 年 2 月時点)

カテゴリ数	1,316
記事数	145,083 件
カテゴリごとの平均記事数	1066.8 件

以下、本章では各処理の詳細を述べる。

3.1 ユーザ発話からの特徴語抽出

対話システムとユーザとの間で行われる対話の話題を推定するため、ユーザの発話したテキストに対して形態素解析処理を行い、テキストに出現する形態素のうち、名詞、動詞、形容詞、未知語を特徴語として抽出する。特徴語として抽出する品詞は、発話における自立語のうち重要であると考えられるもの[樋口 08]を選定した。抽出した各形態素の出現頻度による語ベクトル

を生成する。なお、形態素解析には MeCab [Kudo 04]を用いた。ユーザ発話と、生成される語ベクトルの例を表 2 に示す。

表 2 ユーザ発話から生成される語ベクトル

好きな選手はやっぱりイチローだね。 イチローの派手なプレーが好き！	
語	出現頻度(回)
イチロー	2
好き	2
選手	1
派手	1
プレー	1

3.2 ユーザ発話と記事との関連度算出

ユーザ発話と各記事との関連度を算出するため、All About 記事の本文部分のテキストに対して形態素解析を行い、3.1 と同様に、名詞、動詞、形容詞、未知語の出現頻度を算出し、TF-IDF による特徴語ベクトルを生成する。

次に、ユーザ発話と、各記事の特徴語ベクトルから関連度を算出する。記事の長さによらず適切な類似度を求めるため、関連度の評価関数として SMART 尺度 [Singhal 96] を用いる。ユーザ発話と最も類似する記事の関連度が 1.000 になるように正規化する。ユーザ発話から生成した語ベクトル(表 2)との関連度が高いと判定された記事を表 3 に示す。

表 3 ユーザ発話と記事との関連度

順位	カテゴリ	記事タイトル	関連度
1	野球	イチローを超えた「イチロー」	1.000
2	メジャーリーグ	イチロー、旅の途中。目指すは世界の頂点	0.979
3	野球	パイオニアと天才の共通点とは	0.928
4	メジャーリーグ	勝ちにこだわるイチローが見せた驚異的スライディング	0.919
5	メジャーリーグ	念願のメジャーベンチ入りを果たしたマリナーズ・川崎	0.894

3.3 ユーザ発話とカテゴリとの関連度算出

3.2 で算出したユーザ発話との関連度の高い記事に含まれる語やフレーズは、ユーザの発話内容と似通った分野のものである可能性が大きいといえる。対話システムの発話生成にそのまま利用した場合に、常識的な内容ばかりになり、ユーザに意外だと思わせるような発話ができないと考えられる。そこで、記事が属するカテゴリとユーザ発話との関連度を用いて、ユーザ発話に対して意外性のある記事を判定する。

各カテゴリごとに、属する記事をすべて結合し、カテゴリごとに1つの文書が存在すると見なし、3.1、3.2 と同様の方法で、ユーザ発話とカテゴリとの関連度を算出する。ユーザ発話から作成した語ベクトル(表 2)とカテゴリとの関連度の例を表 4 に示す。

表 4 ユーザ発話とカテゴリとの関連度

順位	カテゴリ	関連度
1	メジャーリーグ	1.000
2	野球	0.893
	≀	
19	高校野球	0.391
20	ニューヨーク	0.377

3.4 記事の意外性スコア算出

3.2 で算出したユーザ発話と記事との関連度と、3.3 で算出したユーザ発話とカテゴリとの関連度を用いて、各記事に対して意外性を考慮したスコアを算出する。本研究では、下記の 2 条件を満たす記事をユーザ発話に対する意外性が高いと見なす。

① ユーザ発話との関連度が小さいカテゴリに属する記事
② ユーザ発話との関連度が大きい記事

例えば、“イチロー”という話題に関連するユーザ発話に対して、“メジャーリーグ”や“野球”のような関連度の高いカテゴリに属する記事はユーザ発話との関連度が高く、意外性が低いと考えられる。一方、ユーザ発話との関連度が低い、“ニューヨーク”のようなカテゴリに属する記事の中に、イチローに言及している記事があれば、ユーザ発話と関連していながら、かつ、“ニューヨーク”という異なる観点で語られた記事ということになり、意外性が高まると考えられる。

ユーザ発話 U に対して、カテゴリ C に属する記事 Ac のスコア $Score(U,Ac)$ を式(1)で算出する。

$$Score(U, A_C) = sim(U, A_C) - \alpha \cdot sim(U, C) \quad \cdots \ (1)$$

$sim(U,Ac)$ はユーザ発話 U と記事 Ac との関連度を、$sim(U,C)$ はユーザ発話とカテゴリ C との関連度を表す。α はユーザ発話とカテゴリとの関連度をどの程度考慮するかのパラメータである。この式により、ユーザ発話と、記事が属するカテゴリとの関連度が高い場合にスコアにペナルティを与える。

スコア算出の例を表 5 に示す。記事『イチローを超えた「イチロー」』は、ユーザ発話との関連度が 1.000 と高い記事であるが、ユーザ発話とカテゴリ“野球”との関連度が 0.893 と高いため、スコアは、0.107 と小さくなる。一方、記事『野球の殿堂にイチローの最多安打特別展示も追加！ 野球の町クーパーズタウン』は、ユーザ発話との関連度が 0.624 と低い記事であるが、ユーザ発話とカテゴリ“ニューヨーク”との関連度が 0.377 と低く、スコアは 0.247 と前述の記事に比べて大きくなる。

表 5 スコアの算出例(α=1.0)

カテゴリ	記事タイトル	*sim(U,Ac)*	*sim(U,C)*	*Score(U,Ac)*
野球	イチローを超えた「イチロー」	1.000	0.893	0.107
ニューヨーク	野球の殿堂にイチローの最多安打特別展示も追加！ 野球の町クーパーズタウン	0.624	0.377	0.247

3.5 記事からの発話文選択

3.4 で算出したスコアが大きく、話題に対して意外性が高いと判定された記事から、対話システムの発話文を選択する。

コーパスからの発話文生成には、命題テンプレートを用いてテキストから命題部分を抽出しモダリティ表現を付与する手法[樋口 08]や、文に含まれる語の TF-IDF 値、語共起、文の長さ

を指標として対話を盛り上げる文をスコア付けする手法[太田 09]などがある。

本研究の手法で、ユーザにとって意外性のある発話をするためには、意外性が高い記事の内容を適切に表現した文を選択する必要がある。そのため、既存手法により、意外性が高いと判定された記事から発話文を生成しても、ユーザにとって意外な内容になるとは限らない。

そこで、本研究では、All About の記事のタイトルもしくは記事の冒頭の部分を、対話システムの発話文として利用することとした。タイトルや記事の冒頭部は、記事の内容を簡潔に表現した要約であることが多いため、対話システムが発話する文として適していると考える。タイトルもしくは記事の冒頭の部分に対して「～ですよ」「～かな？」といった表現を追加し、発話文を生成する。

4. 評価実験

本研究で算出するスコアにより、ユーザ発話に対して意外性の高い記事を判別できることを確認する。

4.1 実験方法

著者が 5 つの話題(t1～t5)を設定し、各話題について実験用のユーザ発話サンプルを作成した(表 6)。ユーザ発話サンプルを入力とし、1)既存手法:ユーザ発話と記事との関連度のみを利用する方法、2)提案手法:カテゴリとの関連度も利用して意外性を考慮する手法、の 2 通りで提示される記事を評価した。

評価は被験者 7 名による主観評価とし、各手法により算出されるスコア上位 10 件の記事について評価した。評価観点は、①ユーザ発話の内容に対して記事の内容が適合しているか、②ユーザ発話の内容に対して意外性があると思うか、の 2 点とした。式(1)のパラメータ α は 1.0 に設定した。

表 6 実験に用いたユーザ発話サンプル

話題	ユーザ発話
t1	スポーツは好きだよ。メジャーリーグを良く観るけど。 好きな選手はやっぱりイチローだね。 イチローの派手なプレーが好き！
t2	田舎で暮らしたいかも。 別荘とか憧れだなぁ。
t3	東京の空気もきれいになったね。 でも都会に出てきて花粉症になったよ。 田舎の方が花粉は多いはずなのになぁ。
t4	最近、健康が気になるから週末に何かスポーツしたいと思ってるんだ。スポーツジムとかどうかなあ。五反田だとどの辺にあるんだろう。ヨガもちょっと気になるかも。ホットヨガってなんだろう。
t5	ホワイトデーのお返しでセンスあるなと思われるものは何かある？プレゼント選ぶのって苦手なんだよなあ。 とりあえずデパートに行ってみよう。

4.2 実験結果及び考察

既存手法と、提案手法により提示された記事について被験者が評価した結果を表 7 に示す。

既存手法では、話題に適合していると判断された記事の割合は 69.7%、意外性があると評価された割合は 27.7%であった。一方、意外性を考慮した提案手法では、話題に適合していると判断された割合が 61.4%と、既存手法と比較して、-8.3 ポイントと若干劣るものの、記事の意外性については 42.0%と、既存手法に比べて 15 ポイント以上の向上が確認できた。

表 7 実験結果

	既存手法	提案手法
①話題に適合している	69.7%	61.4%
②意外性がある	25.2%	42.0%

表 8 に、既存手法により話題との関連度が高い記事を示す。また、表 9 に、提案手法が提示した記事のうち、被験者により意外性が高いと評価された記事の例を示す。表 8 と表 9 を比較すると、提案したスコア付けの手法によりユーザ発話に対して観点の異なるカテゴリの中から、関連性を持ちそうな記事を抽出できていることがわかる。

表 8 既存手法:話題との関連度が高い記事

話題	カテゴリ	記事タイトル
t1	メジャーリーグ	勝ちにこだわるイチローが見せた驚異的スライディング
t2	リゾートマンション	一度は失敗、二度目で実現した田舎暮らし
t3	掃除	花粉症に負けない住まい
t4	疲労回復法	マグマヨガってどんなもの？
t5	暮らしの歳時記	【ホワイトデー】の由来

表 9 提案手法:意外性があると評価された記事

話題	カテゴリ	記事タイトル
t1	社会ニュース	選手だけじゃない！日本人メジャーリーガー
t2	お笑い・バラエティ	近ごろ「田舎番組」が流行るワケ
t3	アミューズメントスポット	癒しのテーマパークで花粉対策セラピー
t4	家計簿・家計管理	運動する、しない？こんなに違う将来の家計
t5	子育て	ママに感謝！のホワイトチョコフォンデュ

スコアが上位の記事から、3.5 で述べた手法により対話システムの発話文を生成した上でのユーザとの対話例を表 10 に示す。

表 10 ユーザとの対話例

ユーザ: スポーツは好きだよ。メジャーリーグを良く観るけど。好きな選手はやっぱりイチローだね。イチローの派手なプレーが好き！

システム:日本人として、選手だけではなく、審判、球団関係者、指導者などですでにメジャーで活躍したり、目指している人は少なくないですよ。

ユーザ: 東京の空気もきれいになったね。でも都会に出てきて花粉症になったよ。田舎の方が花粉は多いはずなのになぁ。

システム: 癒しのテーマパークで花粉対策セラピーかな？

5. まとめ

本稿では、非タスク型対話システムにおいて、システムの発話に意外性を持たせる目的で、ウェブサイトの記事に意外性を考慮したスコア付けをする手法を提案した。発話文生成については、意外性のある記事のタイトルや記事冒頭を使用し対話システムの発話に利用した。All About データを用いた実験により、ユーザ発話との適合性を大きく低下させることなく、意外性がある記事を提示できることを確認した。

今後の課題として、発話文生成の高度化がある。意外性が高いと判定された記事のタイトルや記事冒頭部だけでなく、対話システムが有効な文を自動的に判定し、発話文生成する処理を組み込めば、より対話を盛り上げることが可能と考えられる。また、今回は直近のユーザ発話のみを入力として発話文を選定したが、対話の文脈を利用することが挙げられる。一連の対話の流れから、システムが話題展開すべきタイミングを自動判定し、適切なタイミングで、意外性のある発話をする仕組みが必要である。さらに、ユーザごとに属性や趣味・嗜好などのプロファイルを用いて、ユーザに合わせた意外性の判定をすれば、雑談のような対話では効果が高いと考える。これらの手法を実装し、より精度を高めていきたい。

謝辞

本研究の実施にあたって、株式会社オールアバウト (All About,Inc.)より、All About サイトのデータを提供いただきました。

参考文献

[藤本 04] 藤本英輝,高梨克也,河野恭之,木戸出正継:概念的関連性に基づく雑談の話題転換点分析, 第 18 回人工知能学会全国大会, pp.2G3-01. (2004)

[奥 13] 奥 健太 セレンディピティ指向情報推薦の研究動向, 知能と情報(日本知能情報ファジィ学会誌)Vol.25, No.1, pp.2-10.(2013)

[Chu-Carroll 99] J. Chu-Carroll and B.Carpenter: Vector-based Natural Language Call Routing, Computational Linguistics, 25(3),pp.361-388 .(2009)

[Alshawi 03] H. Alshawi, Effective Utterance Classification with Unsupervised Phonotactic Models. In the Proceedings of HLTNAACL, Edmonton, Canada, (2003)

[太田 09] 太田 知宏. 鳥海 不二夫. 石井 健一郎. 発話生成を目的とした Wikipedia からの文抽出, 第 23 回人工知能学会全国大会 論文集 , pp.1-4. (2009)

[Kudo 04] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237. (2004)

[Singhal 96] A. Singhal, C. Bukley, and M. Mitra "Pivoted document length normalization," in Proeedings of the 1996 ACM SIGIR Conferene on Researh and Development in Information Retrieval, pp. 21-29. (1996)

[樋口 08] 樋口真介,ジェプカラファウ,荒木健治:Web を利用した連想単語及びモダリティ表現による雑談システム, 言語処理学会第 15 回年次大会, PA1-7, pp.175-178. (2008)