

クラウドベースの対話システムにおける時節情報を用いた発話分類

Classification of Utterances Using Temporal Relation in Cloud-based Dialogue System

平田 和暉 *1
Kazuki Hirata下川原 英理 *2
Eri Shimokawara高谷 智哉 *3
Tomoya Takatani山口 亨 *4
Toru Yamaguchi*1*2*4 首都大学東京
Tokyo Metropolitan University*3 トヨタ自動車株式会社
Toyota Motor Corporation

In recent years, research on dialogue system is prosperous. we developed spoken dialogue system as the previous research. One of the challenges of the system that we developed in previous research is that temporal relation cannot be considered. In order to solve this problem, we consider a system which automatically judges whether the temporal relation is contained in robot response sentences. We used two methods as the classification method. The first is simple classification by word pattern matching. The second is classification using MLP (Multi-Layer Perceptron.) As a result, the classification method using MLP was more useful than the word pattern matching. However, it was not practical with current precision. We found that more training data is necessary.

1. はじめに

近年、対話ロボットの研究が盛んに行われている。対話ロボットには大きく分けてタスク指向型対話を目的としたロボットと非タスク指向型対話を目的としたロボットの二つの種類に分類される。タスク指向型対話とは受付の案内や電話でのカスタマー対応等、特定の目的を達成するための対話である。非タスク指向型対話とは雑談のように、特定の目的がない対話の事である。我々は先行研究として非タスク指向型対話を目的とした音声対話システムを開発した。[1] 先行研究で開発したシステムの概要を図1に示す。対話にはロボットの応答文を格納したデータベースを用いる。話者の発話に対する返答として適切な応答をデータベースから選択してロボットが発話する事で、話者と会話を行う。ロボット応答文を格納したデータベースは端末内のホームサーバとクラウドサーバによって管理される。会話を行う際には、ホームサーバ内にあるデータベースを用いてロボットは会話を行い、その会話のログデータは逐次ホームサーバに蓄積される。そして、その日の終わりに全ての端末に蓄積した会話のログをクラウドサーバにアップロードし、会話データから得られる新しい情報をデータベースに追加した後、各端末のホームサーバ内のデータベースを更新するシステムを構築した。しかし、現状の対話システムでは話者との雑談が成立しているとは言い難い。我々はより人間らしい雑談を行う音声対話ロボットの実現を目指す。

2. 雑談時における時節の影響

先行研究で構築した対話システムを用いて高齢者を対象として、6月、7月に会話実験を行った。その実験において、被験者からロボットの応答に違和感を感じるという評価が得られた。具体的に被験者が違和感を感じた会話の例を表1に示す。表1より、ロボットが「今日は寒くて水が張ってました」と応答を返している。しかし、この会話実験を行った時期は6、7月であり、暖かい時期に行っている。実際の会話の環境にそぐわない応答をロボットが返している事が分かった。光田らの研究[2]では、ロボットが人と会話を行う上で考慮すべき言外の情報は多様であり、また、言外の情報が会話に与える影響は

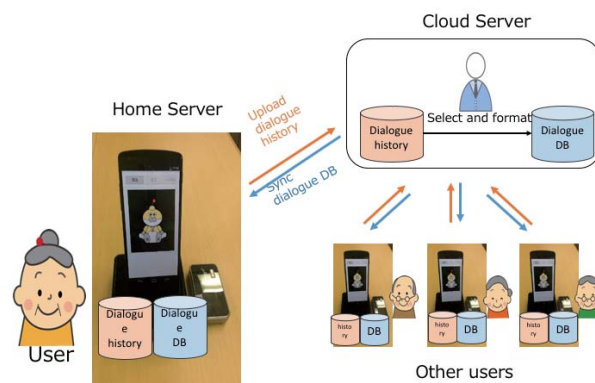


図1: 先行研究として開発した音声対話ロボット

大きいと述べている。我々は会話に影響を与える言外の情報の一つとして時節情報に着目した。単語、または文に適切な時節情報(四季、日中の時間帯、日付など)を推定する研究は行われている。[3][4]しかし、これらの研究ではテキストに対し時節情報を推定しているが、そもそもそのテキストが時節情報を付与する必要があるのかは考慮していない。光田らの時期を判定するシステム[5]では、新聞記事からあらかじめ時節情報が含まれる表現、文を手で抽出し、抽出したデータに対しどれだけ適切な時節情報を付与できるかを評価している。ロボットの応答文に対し時節情報を含む文であるかどうかを自動で判定する事ができれば、ロボットが時節にそぐわない発話することを回避できる。時節情報を含む文を本論文では「時節文」と定義し、本論文では以降、時節文と呼ぶ。また、ロボットの応答文が時節文と判定される要因となる単語を時節語と定義する。本論文ではロボットの応答文を時節文かどうかを判定する分類問題を検討することを目的とする。

表 1: 時節にそぐわない発話をした場合の雑談履歴.

U: ユーザ, R: ロボット	発話文
U:	こんにちは
R:	こんにちは
U:	今日は暖かいね
R:	今日は寒くて氷が張っていました (時期にそぐわない応答)
U:	そうねー 氷が張ると寒いねー
R:	春が好きです。桜が咲いてきれいですよね。
U:	早く春がくるといいね 私も春は大好きです
R:	一人暮らしですか？
U:	父と二人です
R:	今は小学校からパソコン習っているんですね

3. 提案手法

時節文かどうかを判定する手法として、二つの手法を本論文で提案する。一つ目の手法として、時節語を用いた単語マッチングによる分類を提案する。また、二つ目の手法として多層パーセプトロン (MLP: Multi Layer Perceptron) を用いた分類を提案する。

3.1 単語マッチングによる分類

一つ目の手法として時節語を用いた単語マッチングによる分類を行う。時節語の例を表 2 に示す。時節語は 1~12 月のいずれかの範囲、またはある特定の日を連想させる単語である事がわかる。

表 2: 時節語の例

時節語
盆踊り
花火
梅雨
コタツ
熱中症
冬至
スイカ

単語マッチングの処理を図 2 に示す。入力となる発話を MeCab [6] によって単語毎に分割する。そして、分割した単語リストの中に時節語が 1 単語以上含まれる場合、その文を時節文と分類する。また、分割した単語リストの中に時節語を含まない場合、その文を非時節文と分類する。上記のように、単語マッチングはシンプルな条件判定処理によって構成する。

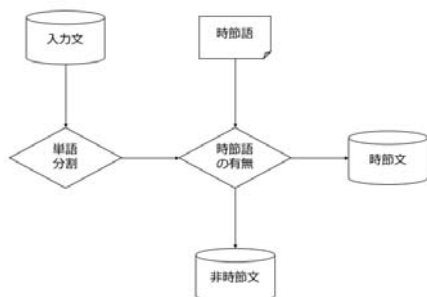


図 2: 単語マッチングの概要

3.2 MLP による分類

3.2.1 時節文の特徴抽出・学習

特徴抽出と学習の処理を図 3 に示す。単語マッチングと同様に、入力となる文を MeCab によって単語毎に分割する。そして、単語リストの先頭と末尾に空文字を二つずつ追加し、各単語を中心として 5 単語ずつ抽出する。そして抽出した 5 単語のリストに対し、各単語を Word2Vec[7] を用いて 100 次元のベクトルに変換した後、100 次元ベクトルを連結する事で 500 次元のベクトルを生成する。この 500 次元の特徴ベクトルを入力として学習を行う。また、その特徴ベクトルを生成した 5 単語の中に時節語を含む場合、その特徴ベクトルを正例として学習する。時節語を含まない場合は、その特徴ベクトルを負例として学習する。

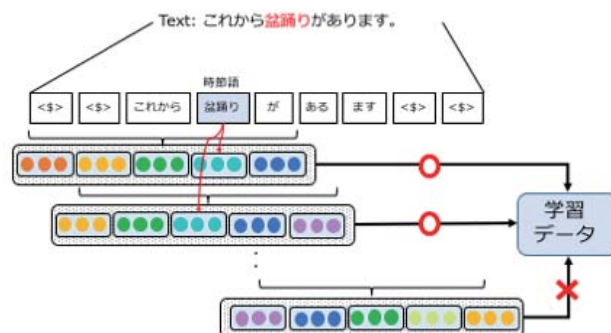


図 3: 特徴ベクトルの生成

学習器として用いた MLP の構成を図 4 に示す。入力層のノード数は 500、中間層はノード数が 355、隠れ層は 5 層、出力層のノード数は 500 である階層型ニューラルネットワークを構築した。活性化関数には ReLU (正規化線形関数)、最適化手法には SGD (確率的勾配降下法) を用いた。出力層では入力の文が、どの程度時節文であるかを示す尤度と、どの程度時節文でないかを示す尤度を、それぞれ 0 から 1 の範囲で出力する。

3.2.2 時節文の判定処理

時節文の判定処理を図 5 に示す。未知の文に対し、前節で述べた特徴抽出方法と同様に、一文から複数の 500 次元ベクトル生成する。この 500 次元のリストを MLP の入力とする

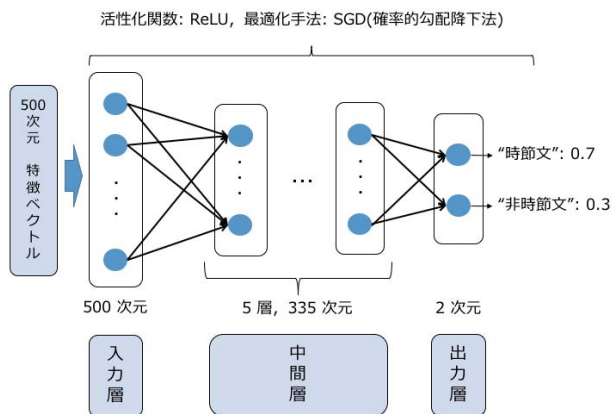


図 4: MLP の構成

ことで、各ベクトルに対する時節文である事を示す尤度を出力する。その中で時節文である尤度が最も高い値が閾値をより大きい場合、その文を時節文と判定する。今回は閾値を 0.1 と設定した。

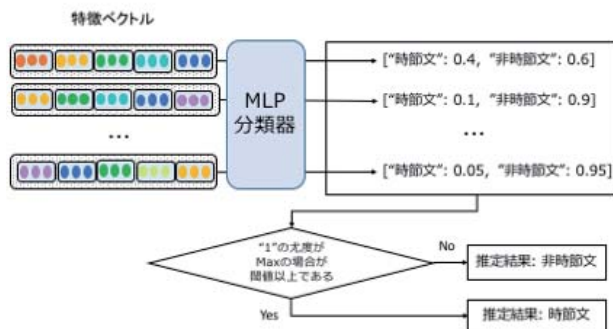


図 5: MLP による時節文判定の処理

3.3 提案手法に用いるデータ

提案手法における入出力の概要を図 6 に示す。評価データにはロボットの応答文が格納されている。学習データには時節文かどうかのアノテーションがされた文が格納されている。また、学習データにある時節文とアノテーションされた文から時節語を抽出する。単語マッチングには、抽出した時節語と評価データを入力として分類を行う。MLP の学習には、時節語と時節文かどうかのアノテーションを入力として生成した正例データ・負例データを使用する。MLP を用いた判定処理では、評価データと MLP の学習モデルを入力として分類を行う。

また、各提案手法において使用するデータ、処理の相違を表 3 に示す。単語マッチングでは時節語を必要とするのに対し、MLP では時節語、時節文、学習モデルを必要とする。また、単語マッチングでは MeCab を用いて単語分割したリストを入力とするのに対し、MLP では単語分割した後に Word2Vec によって変換した特徴ベクトルを入力とする。このような相違点があげられる二つの提案手法を比較し、評価する。

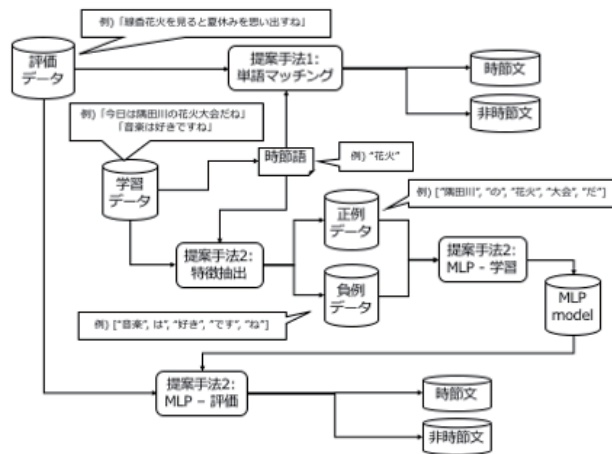


図 6: 提案手法における入出力データ

表 3: 各提案手法における使用するデータの相違点

	単語マッチング	MLP
時節語	必要	必要
時節文	不要	必要
学習モデル	不要	必要
MeCab による単語分割	必要	必要
Word2Vec による特徴ベクトル変換	不要	必要

4. 提案手法の評価実験

4.1 学習データと評価データの設計

ロボットの応答文として使用している 2,478 件の文を評価データとして用いる。この 2,478 件の各応答文に 7 名の被験者が時節文かどうかのアノテーションを行った。7 名の被験者の内、過半数の被験者が時節文と判定した文を時節文と定義し、提案手法の評価を行った。また、時節文のアノテーションを正確に行う信頼性が高い特定の被験者 3 名を用意し、同様に 2478 件の評価データに時節文のアノテーションの作業を行った。この 3 名がアノテーションを行った 2,478 件のデータを学習データとして用いる。その際、信頼性の高い 3 名の被験者の中で、誰も時節文と判定しなかった文を重要度 0、1 名が時節文と判定した文を重要度 1、2 名が時節文と判定した文を重要度 2、3 名全員が時節文と判定した文を重要度 3 と定義する。それによって、三つの学習セット A, B, C を設計した。学習セット A は重要度 1 以上の時節文を正例として使用した場合、学習セット B は重要度 2 以上の時節文を正例として使用した場合、学習セット C は重要度 3 の時節文を正例として使用した場合の学習データとする。また、三つのデータセット全てに共通して、負例には重要度 0 の文を用いた。それぞれの学習セットのデータ件数を表 4 に示す。

表 4: 各学習セットのデータ件数の内訳

	学習セット A	学習セット B	学習セット C
正例の文数	176	74	24
負例の文数	2,302	2,302	2,302
時節語	106	62	28

4.2 時節文判定の結果

評価指標として、再現率 (Recall) と適合率 (Precision) の二つの指標を用いた。二つの提案手法における各学習データを用いた場合の時節文判定結果を表 5, 表 6 に示す。そして、各分類の再現率と適合率をまとめたグラフを図 7 を示す。

表 5: 単語マッチングによる分類結果

		予測結果					
		学習セット A		学習セット B		学習セット C	
		pos	neg	pos	neg	pos	neg
真の結果	true	156	2	128	30	103	55
	false	610	1,710	279	2,041	165	2,155

表 6: MLP による分類結果

		予測結果					
		学習セット A		学習セット B		学習セット C	
		pos	neg	pos	neg	pos	neg
真の結果	true	143	15	115	43	72	86
	false	133	2,187	69	2,251	46	2,274

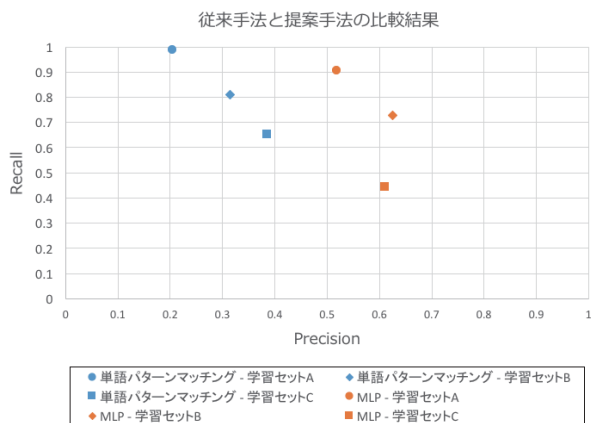


図 7: 各分類結果の比較

表 5 より、単語マッチングでは学習セット A を用いた場合の再現率の値が最も高い事がわかる。これは学習セット A に含まれる時節語の種類数が最も多い事が起因している。そのため、学習セット A を用いた場合の適合率は最も低い。表 6 より、MLP でも学習セット A を用いた場合の再現率が最も高い。適合率に比べ再現率の値の方が高い理由として、時節文と判定する閾値が低い事が考えられる。本論文では再現率を目標値として使用しているため、閾値を低くした結果として適合率が下がってしまった。図 7 より、単語マッチングでは全ての学習セットにおいて適合率の値が MLP より大きく下回っている。また、MLP において学習セット A を用いた場合の再現率は 90% を超えている。これらのことから、MLP は単語マッチングに比べ再現率を大きく損なわずに、より高い適合率で時節文判定を行うことができる。

また、MLP の分類では全ての学習セットの場合において再現率の値が単語マッチングより下回っている。その原因として、学習データの不足が考えられる。表 4 より、特に重要度 3 の時節文は 28 件と圧倒的にデータ数が少ない事がわかる。そのため、特に学習セット C では十分な学習を行っていないと

考えられる。それぞれ学習セット A を用いた場合の単語マッチングによる分類と MLP による分類において目標値を達成した。また、適合率においては学習セット A を用いた単語マッチングより、学習セット A を用いた MLP による分類の方が高い精度を出している。よって、現状では学習セット A を用いた MLP による分類が最も有効である。

5. おわりに

本論文で使用した学習データが少ないことが課題として挙げられた。しかし、普段の会話で使用する発話のほとんどが時節文でないため、時節文の件数を大量に取得するにはさらに多くのロボットの応答文を生成する必要がある。また、本論文の提案手法では入力となるロボットの応答文が時節文かどうかの 2 値分類を行った。しかし、時節文の重要度毎に分けて学習する事で 2 値分類ではなく他クラス分類を行う事ができる。それによって、時節文判定の中間値を表現する事ができると考えられる。今後は対話データの収集と多クラス分類による時節文判定システムを検討する。

参考文献

- [1] E. Sato-Shimokawara, S. Nomura, Y. Shinoda, H. Lee, T. Takatani, K. Wada, and T. Yamaguchi. A cloud based chat robot using dialogue histories for elderly people. Vol. 2015-November, pp. 206–210, 2015.
- [2] 光田航, 東中竜一郎, 松尾義博. 複数の作業グループを用いた対話における言外の情報の類型化. Technical Report 3, dec 2016.
- [3] 吉川義紀, 芋野美紗子, 土屋誠司, 渡部広一. E-015 文の意味を考慮した常識的時間判断システムの構築 (e 分野: 自然言語・音声・音楽, 一般論文). 情報科学技術フォーラム講演論文集, Vol. 13, No. 2, pp. 235–236, aug 2014.
- [4] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, Vol. 2003, p. 40. Lancaster, UK., 2003.
- [5] 小畑陽一, 渡部広一, 河岡司. 単文の名詞と動詞から時間/季節を判断するメカニズム. Technical Report 1(2000-ICS-123), jan 2001.
- [6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pp. 230–237, 2004.
- [7] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.