

Universal Transformerを使用した対話破綻検出

Dialogue Breakdown Detection using Universal Transformer

桑原 健太^{*1} 大村 英史^{*1} 桂田 浩一^{*1}
Kenta Kuwahara Hidefumi Ohmura Kouichi Katsurada

^{*1} 東京理科大学理工学研究科情報科学専攻

Department of Information Sciences, Graduate School of Science and Technology, Tokyo University of Science

In recent years, a lot of dialogue systems have been developed at many institutions. However, most of these systems sometimes cause the problem of dialogue breakdown because of the systems' inappropriate utterances, which makes it impossible to continue the conversation between a user and the system. In this paper, we propose a dialogue breakdown detection system using a universal transformer. Since our system can model the detailed structure of input sentence, it is expected to predict the dialogue breakdown accurately than the previous systems. To show its performance, we conducted an experiment in which our system is compared with a previous study using an RNN. The experimental result shows that our system can obtain a higher F-measure metrics than the previous method in the task of dialogue breakdown challenge.

1. はじめに

近年、非タスク指向の対話システムが注目を集めている。非タスク指向の対話システムは日常的に利用できる利便性の高いシステムであるが、現状の技術水準ではしばしばシステムの不適切な発話によってユーザとの対話が続行不可能になる状態（対話破綻）をしばしば引き起こすという問題がある。そこでこの問題に対処するために対話破綻を検出・回避する技術やリカバリする技術が検討されてきた。

このうち対話破綻の検出に関しては、対話破綻を引き起こすかどうかを予測させる「対話破綻検出チャレンジ」が近年開催されている[1]。このチャレンジにおいて小林らはRecurrent Neural Network(RNN)で構成されたNeural Conversational Model(NCM)を用いて文脈をモデル化し、対話破綻を検出する方法を提案した[2]。さらに久保らは小林らのモデルの問題点であった「対話破綻検出チャレンジ」の配布データ以外の対話データが必要であるという点を、学習時のデータに対話破綻状態が否かの情報を加えることによって改善し、高い性能を示した[3]。しかし自身の検出器の更なる性能向上には、文構造に関する特徴量を入力に与えることが必要であると示唆していた。

我々は久保らのモデルをベースに、Universal Transformer [5]を用いた対話破綻検出器を提案する。Universal Transformer は、自動翻訳の分野で広く用いられ始めたTransformer [4]というモデルを翻訳以外の自然言語タスクに適応可能にしたモデルである。これらのモデルでは注意機構(Attention)という、2つの系列データの要素間の関係性を学習する機能を持つ。これによって、入力文の文構造を学習することができるため、久保らが述べた問題を解決できると考えられる。

本稿では2章で対話破綻検出チャレンジの概要と小林らの手法、久保らの手法について述べ、3章で提案手法を説明する。その後、4章で評価実験について解説し、5章で本研究のまとめを述べる。

2. 対話破綻検出チャレンジの概要と関連研究

2.1 対話破綻検出チャレンジ

対話破綻検出チャレンジは、ユーザとシステムの対話ログ中の各システム発話において対話破綻を引き起こしているか否かを判定するタスクを行うコンペティションである。データとして対話システムとユーザの行った対話が提供されており、システム発話には複数人のアノテータによって「O」、「T」、「X」のラベルがつけられている。それぞれ、「破綻を引き起こしていない」、「破綻を引き起こしていないが違和感がある」、「破綻を引き起こしている」という意味を持つ。チャレンジの参加者は、このラベルの確率分布を予測することである。以下に破綻の例とアノテーションの例を示す。

ユーザ：猫に付いて行ったらなにか発見できますか。
システム：猫は元気です O:0 T:9 X:21

各ラベルの隣の数字は、そのラベルをつけた人数を表している。この例ではシステムがユーザの質問を無視した発話をしているため、多くのアノテータはシステムの発話が破綻を引き起こしたと判断している。

2.2 対話モデルを使用した対話破綻検出

対話破綻を検出するには、まず各発話を破綻検出に必要な特徴量を含むベクトル表現に変換し、このベクトル表現を用いて各ラベルの確率分布を予測するモデルを構築すればよい。小林らはVinyalsらが提案したNeural Conversational Model (NCM)という対話モデルを用いる方法を提案した[2]。NCMはRNNを使用したEncoder-Decoder型の対話モデルである。Encoderはユーザ発話を入力として受け取り、発話の意味を表す潜在表現へと圧縮する。DecoderはEncoderで得られた潜在表現を素性としてシステム発話を生成させる。図1上に小林らが用いたモデルを示す。小林らは事前に学習したNCMにユーザ発話とシステム発話のペアを与え、EncoderとDecoderの最終時刻の内部状態を特徴量として検出器を構築した。しかし、対話破綻

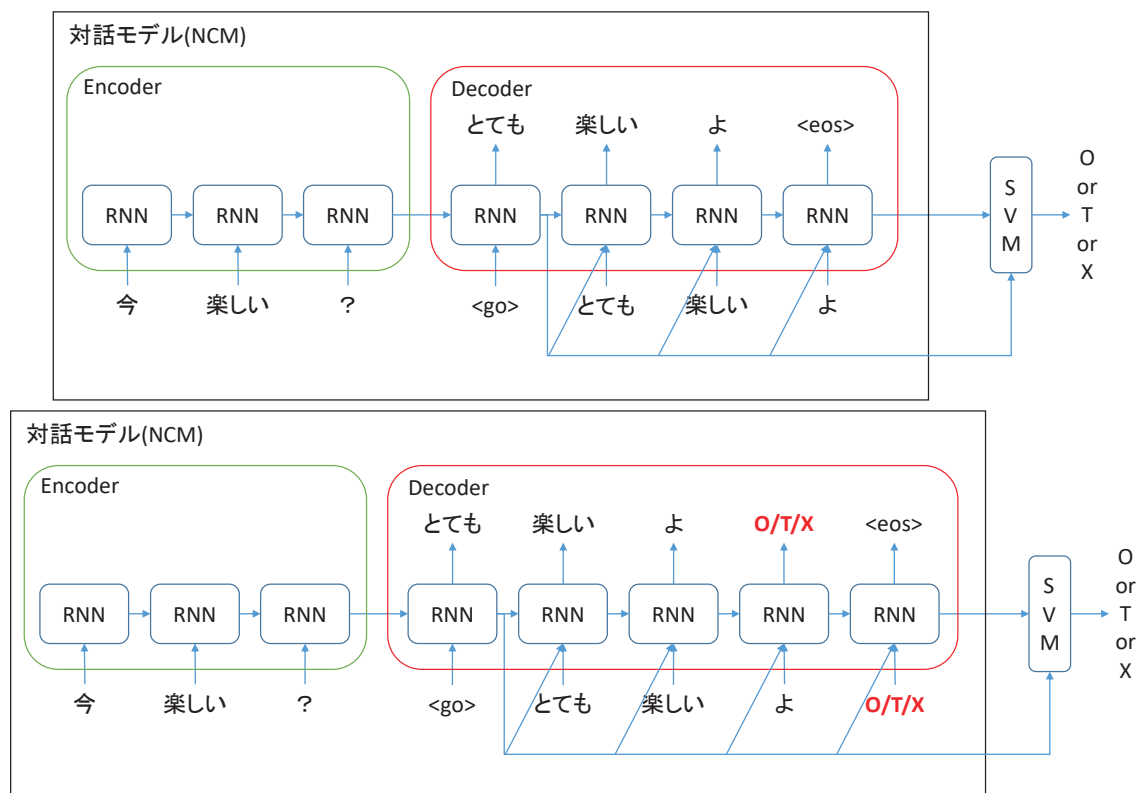


図 1:小林らの手法(上)と久保らの手法(下)

検出チャレンジで配布されているデータには破綻しているデータが含まれている。このデータを用いて NCM を学習させると、破綻している対話も学習することになる。すると NCM の内部状態は、破綻しているか否かを区別しなくなるため、得られる特徴量を使った破綻の検出は困難となる。そこで小林らは別途正常な対話データを用いて学習を行っていた。

これに対し久保らは NCM の学習を行う際にシステムの発話文の末尾に「O」、「T」、「X」をアノテーション分布に従って確率的に挿入するという手法を提案した[3]。図 1 下に久保らが用いたモデルを示す。破綻してない対話と破綻している対話が末尾のラベルによって区別できるため、この手法によって破綻している対話かどうかを区別しながら学習することが可能になった。その結果、第二回対話破綻検出チャレンジにおいて配布データのみを使用しながらも最も高い F 値を示した。しかしながら、自身のモデルに対し、ユーザ発話とシステム発話に同じ単語が出現した場合には、破綻であっても破綻していないと判断してしまう場合があるとの問題点を挙げていた。この問題を解決するためには文構造や体裁をチェックする特徴量が必要であると述べている。

3. 提案モデル

本稿で提案するモデルは、久保らのモデルをベースに、NCM を Universal Transformer で学習したものである。以下ではまず Universal Transformer の概要を説明した後に、本手法の詳細について述べる。

3.1 Universal Transformer

Universal Transformer のベースとなる Transformer [4]は注意機構(Attention)の導入によって、系列データの各要素の

処理において別の系列の特定の要素に着目した処理を行えるようにしたニューラルネットワークの一種である。注意機構によって構文や意味構造を学習できることから、Transformer は機械翻訳タスクにおいて RNN モデルより優れた性能を示している。しかし、Transformer は再帰的構造を持たないため RNN より帰納バイアスが弱く、学習データに含まれない長さのデータには対応できなかった。この問題に対し、Transformer に再帰的構造を付与したモデルが Universal Transformer [5]である。

Universal Transformer は Attention とフィードフォワードネットワーク(FFN)のみで構成された Encoder-Decoder モデルのニューラルネットワークである。系列 X の処理における系列 Y の各要素の重要度を表現する Attention は、学習される重み Q, K, V を用いて以下の式によって定式化される。

$$QX = Q * X, KY = K * Y, VY = V * Y$$

$$\text{Attention}(X, Y) = \text{softmax}(QX * KY^T / \sqrt{d}) * VY$$

ここで d はスケーリング因子のパラメータである。softmax 関数によって X の処理に対する、Y の各要素の影響を求めている。つまり、Attention によって X のある単語と Y のある単語との意味的つながりや文法的つながりを学習可能になる。X = Y とする Attention は Self-Attention と呼ばれ、X を文とした場合、単一の文内の構文や意味構造を表現する。Universal Transformer の Encoder 部は、Self-Attention と単語ごとに処理される FFN の 2 つのネットワークを 1 つの層としている。層への入力是最初に self-attention で処理され、次に単語ごとに FFN によって処理される。この層を用いて多層にしたネットワークが Encoder となる。再帰的構

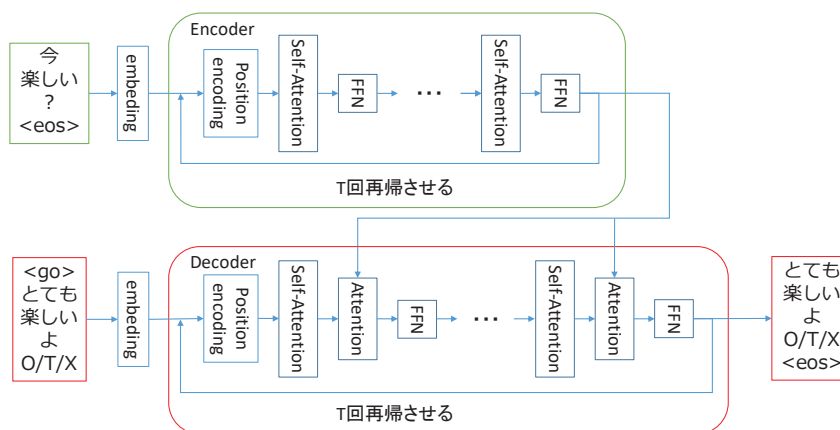


図 2: Universal Transformer を用いた対話モデル

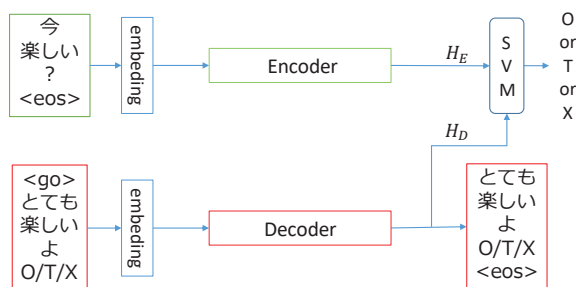


図 3: 提案手法 (Universal Transformer を用いた破綻検出器)

造を付与するため、Universal Transformer では Encoder の出力を再び Encoder の入力として与えている。この処理は複数回行われる。最終的な Encoder の出力は最終時刻の最終層の出力となる。Decoder は Encoder とほとんど同じ構成であるが、Self-Attention と単語ごとの FFN の間に、Encoder の出力との Attention の処理が加わる。この Attention によって Encoder の入力との関係をモデル化することができる。

3.2 Universal Transformer を用いた破綻検出法

提案手法では久保らが提案した RNN を用いた手法と同様に最初に学習データのシステム発話の末尾に破綻ラベルをアノテーション分布に従って挿入する。次にユーザ発話とシステム発話のペアを用いて Universal Transformer を学習させる。破綻を検出する際には、まずユーザ発話とシステム発話のペアを学習した Universal Transformer に与え、Encoder の最終時刻の出力 H_E と、Decoder の最終時刻の出力 H_D を得る。この 2 つの特徴量には self-attention によって、発話の構文や意味構造などの情報が含まれていると見なすことができる。また、 H_D にはユーザ発話とシステム発話の関係性もモデル化されていると考えられる。従って H_E, H_D は発話の文法的な表現や意味的表現、ユーザ発話とシステム発話の関係を表していると期待される。Universal Transformer によって得られた H_E, H_D は、破綻検出のための特徴量として判別器の Support Vector Machine (SVM) に与えられる。以上に述べた提案手法の対話モデルを図 2 に、破綻検出システム全体の構成を図 3 に示す。

4. 評価実験

4.1 実験概要

RNN を使用したモデルと提案手法の比較を行った。モデル構築に使用したデータセットは第一回対話破綻検出チャ

レンジの学習用データ (rest1046) と開発用データ (DBDC1dev)、および第二回対話破綻検出チャレンジの学習用データ (DBDC2dev) である。rest1046, DBDC1dev は同じ対話システム DCM を用いて収集されており、それぞれ 1046 対話、20 対話が収録されている。DBDC2dev に関しては DCM に加え DIT, IRS という対話システムを加えた 3 つの対話システムを用いて得られた対話から構成されており、それぞれ 50 対話ずつが収録されている。対話モデルの学習にはこれらすべてを用いた。破綻検出器である SVM の学習には対話モデルの学習に用いたデータから rest1046 を除いたデータを用いた。rest1046 は破綻ラベルの偏りが大きいので破綻検出器の学習データには適さないと判断したためである。

評価データには第二回対話破綻検出チャレンジの評価用データ (DBDC2test) を用いた。DBDC2test も DBDC2dev と同様に DCM, DIT, IRS を用いて 50 対話ずつ収集されている。第二回対話破綻検出チャレンジでは各対話システムのデータごとに評価しているため、本実験においてもデータごとに評価する。以下ではそれぞれの評価データを DCMtest, DITtest, IRStest と表す。評価には対話破綻検出チャレンジで用いられた以下の 5 つの評価尺度を用いる。

- Accuracy : 全ラベルの一致率
- F-measure(X) : 破綻ラベル X の検出の精度を示す F 値。X の適合率と再現率の調和平均によって表される。
- F-measure(T+X) : ラベル T を X と同一のラベルとして扱った際の F 値
- JS Divergence(O,T,X) : Jensen-Shannon Divergence によるアノテーション分布との分布間距離
- Mean Squared Error(O,T,X) : アノテーション分布との分布間の平均二乗誤差

表 1: 破綻検出の結果

データ	モデル	Accuracy	F-measure (X)	F-measure (T+X)	JSD (O,T,X)	MSE (O,T,X)
DCMtest	RNN	0.480	0.511	0.682	0.134	0.070
	Universal Transformer	0.475	0.549	0.811	0.131	0.067
DITtest	RNN	0.596	0.705	0.882	0.087	0.045
	Universal Transformer	0.605	0.707	0.907	0.081	0.041
IRStest	RNN	0.525	0.609	0.763	0.139	0.074
	Universal Transformer	0.500	0.616	0.813	0.143	0.076

表 2: 各モデルの Precision と Recall

データ	モデル	Precision (X)	Recall (X)	Precision (T+X)	Recall (T+X)
DCMtest	RNN	0.418	0.657	0.779	0.607
	Universal Transformer	0.395	0.904	0.762	0.866
DITtest	RNN	0.562	0.947	0.849	0.917
	Universal Transformer	0.555	0.973	0.857	0.964
IRStest	RNN	0.491	0.801	0.743	0.784
	Universal Transformer	0.457	0.944	0.711	0.950

対話モデルの学習にはミニバッチ学習を行い、バッチサイズは 10 とした。中間層は 256 次元とし、中間層は 1 層とした。最適化手法は学習率 0.1 の確率的勾配降下法を用い、目的関数には予測単語ごとの交差エントロピー誤差を用いた。前処理として Mecab による形態素解析を行った。辞書には mecab-ipadic を用いた。Universal Transformer に関しては、再帰回数は 6 回に設定した。

4.2 実験結果と考察

破綻検出の結果を表 1 に示す。Universal Transformer を用いた手法が RNN を用いた手法を概ね上回っている事が分かる。以下に RNN で破綻検出に失敗して Universal Transformer では成功したシステム発話とその直前のユーザ発話の例を示す。

ユーザ：明日は猛暑らしいですから
システム：猛暑は欲しいですね

この例はどちらにも「猛暑」という同じ単語が含まれているため、2.2 節で述べた通り RNN では破綻と判定することができなかったと考えられる。しかし、このような例に対しても Universal Transformer では、破綻を検出できるようになった。

Universal Transformer では Accuracy に関して RNN を上回ることができていない。この原因として、O のラベルが付与された発話の多くを X と誤識別してしまったことが挙げられる。両者のモデルの特徴を調べるため、破綻のラベル X が付与されたデータの Precision (適合率) と Recall (再現率)、および T を X と同一のラベルとした際の Precision と Recall の値を表 2 に示す。DITtest の Precision(T+X) を除いて Precision は RNN の方が高く、Recall は Universal Transformer の方が高いことが分かる。この結果から Universal Transformer は O や T の発話を X と出力したケースが多いことを示しており、Accuracy が低いのもこの誤認識の多さが原因であると考えられる。今後は Precision を向上するための方法を検討する必要がある。

5. おわりに

本研究では Universal Transformer で学習させた対話モデルを用いた対話破綻検出法を提案した。RNN を用いた破綻検出器と比較した結果、破綻検出精度の F 値に関して Universal Transformer が上回るという結果が得られた。これは Self-Attention によって文構造や意味構造が学習され、RNN を用いたモデルにおける問題点を解決できたためと考えられる。しかしながら対話破綻検出の適合率は RNN を用いたモデルを下回る結果となったため、今後は適合率を向上するための特徴量について検討したい。

参考文献

- [1] 東中竜一郎, 船越孝太郎, 稲葉通将, 荒瀬由紀, 角森唯子, "対話破綻検出チャレンジ 2", 第 78 回言語・音声理解と対話処理研究会(第 7 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-B505-19, pp.64-69, 2016.
- [2] 小林颯介, 海野裕也, 福田昌昭, "再帰型ニューラルネットワークを用いた対話破綻検出と言語モデルのマルチタスク学習", 第 75 回言語・音声理解と対話処理研究会(第 6 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-075-09, pp.41-46, 2015
- [3] 久保隆宏, 中山光樹, "Neural Conversational Model を用いた対話と破綻の同時学習", 第 78 回言語・音声理解と対話処理研究会(第 7 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-B505-26, pp.94-97, 2016.
- [4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is all you need", arXiv:1706.03762, 2017
- [5] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Lukasz Kaiser, "Universal Transformers", arxiv:1807.03819, 2018