＜ 修 士 論 文 ＞

深層学習を使った雑談対話システム
（要　旨）

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：２０２１年度
学籍番号：6020105
氏　　名：長田　帆貴
指導教員：市川　治
提出年月日：２０２２年１月 12 日

背景と目的

対話システムの研究においてタスク指向型と非タスク指向型の 2 種類の研究がされてきた。タスク指向型の対話システムは、ユーザ側に明確な達成目標があり、天気や道案内等の閉じたドメインで対話が行われることを仮定できる場合が多い。一方、非タスク指向型の雑談対話システムはユーザ側に明確な達成目標がなく対話自体が目的であり、対話システムは多様な話題を扱う必要がある。これに対して近年、非タスク指向型対話システムの研究において Twitter や Wikipedia のような大規模コーパスを用いたシステムの発展がある。

既存の大規模コーパスを雑談対話モデルへの学習データとして用いる場合の問題点として、その言語資源が文語でということがある。ユーザの対話欲求を満たす雑談対話システムにはユーザが対話システムに親しみを持ち対話を続けたくなる言葉遣いが求められる。しかし、十分な質と量のある口語調の対話データを作成することは一般に高コストである。そこで本研究ではテレビ字幕データから対話ペアを抽出し、対話モデルの学習データとすることで口語調の発話が可能な対話モデルの作成を目指す。

提案手法と使用データ

Dialogue Breakdown Detection Challenge(DBDC)とは、雑談対話システムとユーザの対話ログに対して、対話の破綻を検出することを目的としたモデルを開発し、その精度を競う学術的なコンペティションである。DBDC においてはユーザと対話システムの対話ログに対して複数のアノテータが破綻ラベルを付与したデータが提供されている。本研究ではテレビ字幕から対話ペアを抽出する対話破綻検出器の学習に DBDC で提供されているデータを用いる。

テレビ字幕データは場面単位で付与されている字幕を時系列順に対話ペアとして加工し対話破綻検出器で推論することで対話として成立しているペアを抽出する。DBDC で提供されているようなユーザと対話システムの対話で起こる対話破綻のパターン加え、字幕中にはナレーションが入ったり、モノローグが入ったりする対話破綻のパターンが考えられる。捉えることのできる対話破綻検器の作成をするため、本研究では LightGBM（提案法 1）・、回帰ベースの BERT モデル（提案法 2）・、ソフトラベルを用いた BERT モデル（提案法 3）の 3 種類の対話破綻検出器の作成し評価を行った。対話破綻検出器の評価には DBDC のデータ・筆者がアノテーションした字幕データの 2 つを用いて評価を行った。

実験の結果、BERT による回帰モデルの対話破綻検出器が評価用字幕データにおいて最も precision が高かった。また、字幕データにおける非破綻クラスの precision は DBDC3 評価用データにおける非破綻クラスの precision よりも低下している傾向が見られた。BERT による対話破綻検出器では、LightGBM モデルに比べその低下が少なく、BERT の

ほうが LightGBM よりも高い対話破綻検出の汎化性能を得ていることが示唆された。

実験結果を踏まえ、BERT による回帰モデルの対話破綻検出器を用いて字幕データから対話抽出を行った。抽出した対話ペアを対話モデルの学習データとして用いて対話ペアの対話コーパスとしての有用性を検証する。評価用字幕データの非破綻クラスの precision がもっとも高かったモデルを用いてテレビ字幕約 1700 万行から対話ペア約 140 万行の抽出を行った。字幕データから抽出した対話ペアを用いて対話モデルの学習を行い、口語調の発話が可能な対話モデルの作成を行った。対話モデルは embeding 層と 2 層 LSTM、1 層の Attention 機構からなるモデルとして作成した。

結果と今後の課題

本論文では DBDC の対話データを用いて対話破綻検出器を作成し、テレビ字幕中に含まれる口語の対話ペアを抽出し、口語調の対話モデルの作成を目指した。作成した対話モデルが口語調の発話を生成できることを確認できた。しかし、対話モデルが一定割合で非文または文脈に合わない発話を生成することも確認した。対話モデルの学習に使用したデータをテレビ字幕から抽出するために使用した対話破綻検出器の precision が低く学習データに対話ペアとして不適切なデータが混在してしまったためだと考えられる。

DBDC の学習データ・評価データは対話モデルと人の対話ログである。一方、字幕データは、放送された字幕を時系列でつなぎ対話ペアにしたものである。本論文の DBDC 評価データ・テレビ字幕データを用いた対話破綻検出器の実験では、テレビ字幕データの精度は DBDC 評価データよりも大幅に低下することが確認できた。このことから、字幕データと DBDC では破綻のパターンが異なることが想定される。よって、提案法 2 と提案法 3 においては、テレビ字幕データの一部を学習データとしてアノテーションし、DBDC のデータ学習した対話破綻検出器のファインチューニングを実施することが有効であると考えられる。