7ZM-04

自然言語処理モデルを用いたウミガメのスープの問題作成支援手法とその実装「UmigaMaker's」

角谷康太† 大橋錬† 松永康太† 高木亜蘭‡ 濱川礼†
中京大学工学部情報工学科 † 中京大学大学院工学研究科情報工学専攻 ‡

1 背景・目的

本論文では，ユーザの用意した文章からウミガメのスープの問題を作成支援するシステム「UmigaMaker's」について述べる．ウミガメのスープは，出題者が提示する問題に対して解答者が質問をもとに状況を絞り込んで答えを導く対話形式クイズの一種である．

ウミガメのスープの問題形式には解答に必要な最小限の情報とミスリードが含まれる．しかしこれらの特徴を考慮して問題を作成するには，解答者にどの程度情報を残すべきか調節しなければならないため作問のハードルは高い．

そこで我々は, 頭の中で浮かんだ情景などをもとに問題が生成可能になれば作問をする際の参考に出来ると考え，「UmigaMaker's」を開発した．

2 手法

作問するにあたってユーザのストーリーを問題作成の土台にすることで, 核心となる部分を「答え」，それ以外の部分を「問題文」に分解する．この手順を踏まえ，図 1 に示す工程で問題文生成を行う．初めに，問題文生成部でユーザの入力したストーリーから問題文を GPT-2[1] を用いて複数生成する．次に，問題文選択部では生成された問題文とストーリーの関係性を算出して閾値を元に問題文の選択を行う．類似度算出には BERTscore[2]，含意関係算出には BERT[3] を利用した．最後に，ミスリード部では形態素解析で品詞と単語の頻出度を特定して単語の挿入場所の選択を行い，RoBERTa[4] を利用して前後の単語から間に入る単語を予測して挿入する．

A Method for Supporting the Creation of Sea Turtle Soup Problems Using Natural Language Processing Models and Its Implementation "UmigaMaker's"
Kota Sumiya, Ren Ohashi, Kota Matunaga, Aran Takagi, Rei Hamakawa
† School of Engineering Chukyo University
‡ Department of Information Engineering,Graduate School of Engineering Chukyo University

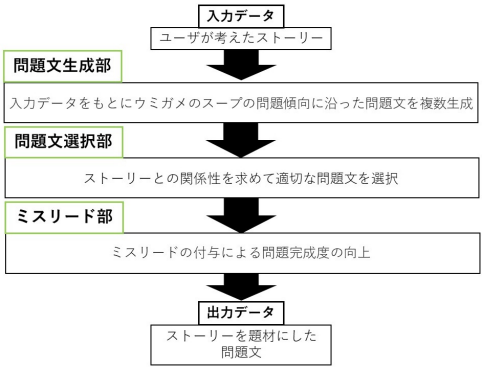

図 1 提案手法

3 システム概要

3.1 問題文生成部

GPT-2 は rinna 社の事前学習済モデル [5] をウミガメのスープ問題集のストーリー・問題文のペア 532 組でファインチューニングを行った．データセットの統計情報を表 1 に示す．この統計をもとに GPT-2 の出力パラメーターを設定した．

表 1 データセットの統計

項目	値
平均ストーリー文字数	129
平均問題文字数	78
平均文字数比	0.62
最大文字数比率	1.78

3.2 問題文選択部

問題文選択部では, 2 つの手法を組み合わせて問題文を選択する．

3.2.1 BERTscore による類似度計算

「ストーリーとの関連性がない」・「答えが混ざっている」問題文の除外を行う．事前に用意したストーリー 20 個の評価値と作問経験者の意見を元に表 2 の F 値の閾値とその条件を定義した．また 97 点以上の問題文は答えが含まれる傾向があるためその文章は必ず除外する．

表 2　F 値の閾値とその条件

F 値閾値	閾値条件
83~89 点	平均 F 値: 83~89 点 ストーリー文字数：40~60 字
90~92 点	平均 F 値: 90~92 点 ストーリー文字数：60~80 字
93.5~96.3 点	条件該当なしの場合の基本値
96~97 点	平均 F 値: 96~97 F 値で 97 点以上の問題文なし

3.2.2　含意関係認識

除外されなかった問題文とストーリーの含意関係を確認する．BERT を京都大学黒橋・褚・村脇研究室が公開している日本語 SNLI データセット [6] を用いてファインチューニングを行い, BERTscore で判定が難しい否定文と類義語の評価を行う．また，含意関係のある問題文が複数ある場合は BERTscore の点数を元に 1 つに絞り込む．

3.3　ミスリード部

ミスリード部では，形態素解析によって頻出度の高い名詞を選出した後に RoBERTa によって補完を行った文章を複数生成する．最後に，Word2Vec[7] によって原文との類似度がより高い文章を選択して出力する．

3.4　実行例

「UmigaMaker's」の実行例を表 3 に示す．下線部はミスリード部の入力を示している．入力されたストーリーに対して問題文生成部は 4 つの問題文を生成した．生成された問題文に問題文選択部は予め定義した閾値を元に問題文を 1 つ選択した．ミスリード部は挿入場所に「出勤時間」「男」の前を選択, 挿入単語を「私の」「その」に決定した．

表 3　UmigaMaker's の実行例

	問題文
入力	ある朝に起こった話。ふと目が覚めると時間は普段の出勤時間を過ぎていた。しかし、男が働いている会社はフレックスタイム制を導入している。男はその日はいつもより遅い時間で働く予定だったので慌てる様子もなく出勤した。
出力	ふと目が覚めると、時間は普段の私の出勤時間を過ぎていた。しかしその男は全く慌てる様子もなく出勤した。一体、なぜ?

4　評価

ウミガメのスープのルールを理解しているが作問をしたことがない 16 人に「UmigaMaker's」を利用してもらい，生成した問題文の妥当性, 被験者が自作した場合との作問時間の比較について評価を行った．結果は表 4 と図 2 で示す．作問時間については被験者より短縮することが可能になったが，生成された問題文の完成度は低くユーザが参考にし難い評価となった．考えられる原因として，学習に使用したデータセットのバリエーションが乏しく，入力したストーリーに合致する文体がなかった場合に突飛な文章が生成されてしまう点が挙げられる．

表 4　システムとユーザの作問時間 (秒)

	UmigaMaker's	ユーザ
平均時間	41.3	305.5

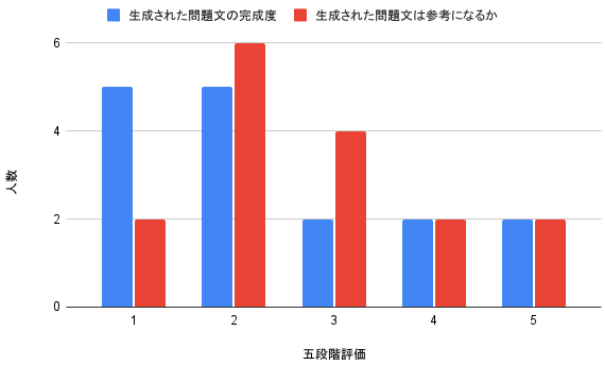

図 2　生成された問題文の評価

5　展望

「UmigaMaker's」の利用による問題作成の価値を向上するために，今後はデータセットの追加や生成過程の見直しを行って精度向上を目指したい．

参考文献

[1] Radford, A., et al. : Improving language understanding by generative pre-training. Technical report, OpenAI
[2] Zhang, T., et al. : Bertscore: Evaluating text generation with Bert
[3] Devlin, J., et al. : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
[4] Liu, Y., et al. : RoBERTa: A Robustly Optimized BERT Pretraining Approach
[5] rinna 株式会社, "日本語 GPT-2/BERT の事前学習モデルを開発しオープンソース化", 2021/8/25, https://rinna.co.jp/ニュース/f/日本語 gpt-2bert の事前学習モデルを開発しオープンソース化, (参照 2022/1/7)
[6] 吉越卓見, 他著：機械翻訳を用いた自然言語推論データセットの多言語化
[7] Mikolov, T., et al. :Efficient Estimation of Word Representations in Vector Space