BERT を用いた TOEIC 問題自動生成システムにおける意味問題の生成

Generating Semantic Questions in an Automatic TOEIC Question Generation System Using BERT

玉田周太郎 [1]　原田恵雨 [1]　佐藤奈々恵 [1]

Shutaro Tamada[1], Keiu Harada[1], and Nanae Satou[1]

[1] 苫小牧工業高等専門学校
[1]National Institute of Technology Tomakomai College

Abstract: In this study, we developed TOEIC Listening & Reading Test Question Generation System. This system generates semantic questions in Part 5 (single sentence fill-in-the-blank questions). The system allows the learner to solve a large number of problems. The implementation method is as follows: First, a sentence is generated by GPT-2. Next, we generate word candidates with Word2Vec and BERT. Word2Vec considers word meaning. BERT considers context. Finally, the system removes words from the candidate choices that are inappropriate as choices. As a result, the system is able to output valid questions for the most part. However, there are some problems that are not valid.

1. 緒言

急速なグローバル化により日本国内の大学や企業などにおいて英語力が要求されることが増えている。

英語力の指標として、TOEIC Listening & Reading テストのスコアを活用することが大学入試で約 4 割[1]、新卒採用で約 5 割[2]となっていることから TOEIC スコアを向上させる必要性が高まっていると考えられる。TOEIC の類似問題を多数解くことに対するニーズが存在すると考えられることから、TOEIC の問題を自動生成することができるシステムを作成することにする。

先行研究[3]では、TOEIC の Part5（短文穴埋め問題）における派生語問題（約 27%）及び動詞の活用形問題(約 6%)についての実装を行っている。

本研究においては Part5 において最も出題割合が高い単語の意味を問う問題（約 53%）について自動生成することを目的とする。なお、本研究において”TOEIC”と表記するものは、TOEIC Listening & Reading テストを指す。

2. 提案方法

2.1.システムの概要

システムの概要(1 題生成する場合)を図 1 に示す。

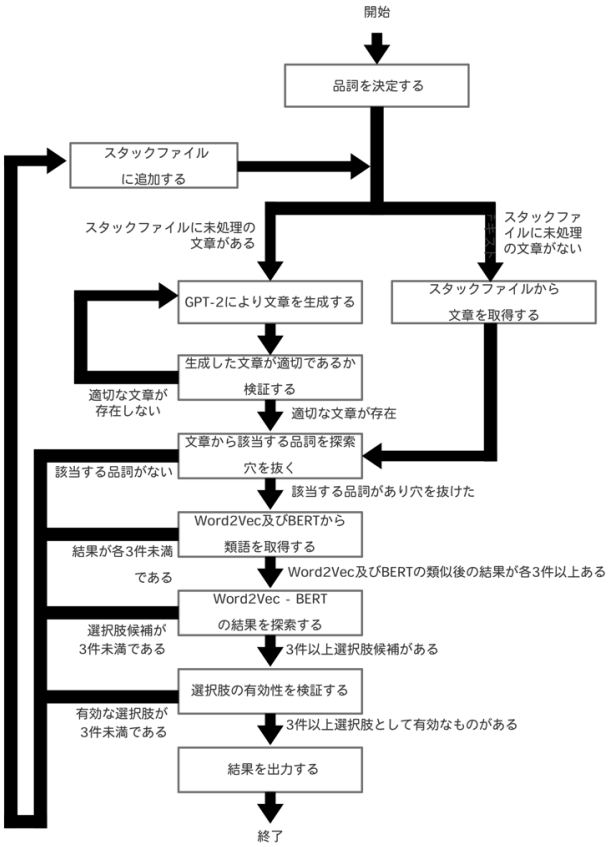

図 1: 自動生成システムの概要

品詞の種類は、指定された問題数を生成した時点で発生確率（動詞：15%、名詞：17%、副詞：23%、形容詞：17%、前置詞または従属接続詞：18%、調整接続詞：10%）が公式問題集の調査に基づいた値に近くなるよう、その時点でのすでに生成された品詞の種類を保持して決定時点での品詞の出現確率を計算し、ランダムで決定する。

文章の生成には、GPT-2[4]を使用する。GPT-2 は、OpenAI によって開発された 800 万の Web ページのデータセットでトレーニングされた 15 億のパラメータをもつ大規模な Transformer をベースとした言語モデルであり、テキスト内の単語が与えられた時、次の単語を予測することができる。GPT-2 で文章を生成するための入力は、TOEIC 公式問題集とインターネット上の英語学習サイトの英文から取得した 501 文の中からランダムに 1 文を選択した文章としている。GPT-2 の文章生成には、Hugging Face Transformers ライブラリ[5]を使用している。生成した文章にゆれを出すために、パラメータの temperature を 0.6 としている。一般的には temperature は 1 から 0.7 にするが、文章のゆれを大きくするため 0.6 とした。また、Beam 探索を行い、次の単語の最も高い確率のもの選択する Greedy 探索と比較して隠された高確率の単語列を見落とすリスクを低減している。

生成された文章には問題として使用することが不適切である文章が含まれている。生成するために GPT-2 に入力した文章と生成された文章が同一である場合、単語数が 8 文字未満又は 30 文字以上である場合、However, And, But が先頭に含む文章である場合、記号（[]()_*:/!?”）を含む、ダブルクォーテーションが 1 箇所又は 3 箇所以上、クォーテーションが 5 箇所以上含まれて居る文章である場合、先頭文字が大文字ではない文章である場合、最後の文字がピリオドではない文章である場合、人名や企業名、商品名や暴力的・差別的な単語を収録したブラックリストに該当する単語が含まれている場合は不適切な文章として除外する。

適切であるとされた文章の穴埋めの生成は、Python ライブラリである NLTK[6]を使用して単語ごとに分割した上で品詞を取得する。システム開始時に決定した品詞に該当する単語を探索し、該当する各単語からランダムで穴を開ける場所を決定する。

文脈を考慮しない単語ベクトルによる距離が近い単語を出力する Word2Vec による結果（上位 400 件）から、穴が抜かれた文章全体を入力し前後の文脈を元に穴埋め箇所に当てはまる単語の候補を出力する BERT の結果（上位 8000 件）の重複している単語を削除したものを選択肢候補とする。Word2Vec[7]はニューラルネットワークを用いて単語を分散表現に変換するモデル群であり、お互いに共通の文脈を持つ単語ベクトルが近くなり、そうでない単語同士はベクトル空間上で遠く配置されるようになっている。単語ベクトルは Google News のデータセットの一部(約 1000 億語)で訓練された”GoogleNews-vectors-negative300.bin” を使用している。BERT (Bidirectional Encoder Representations from Transformers)[8]は、すべての層で左右両方の文脈を条件付けることにより、ラベルのついていない文章から深い双方向表現を事前学習するように設計されており、出力層をファインチューニングするだけで、タスク固有のアーキテクチャを大幅に変更することなく、事前訓練することが可能である。本研究で使用している学習済みパラメータは、BooksCorpus (800M words)と English Wikipedia (2,500M words)を事前学習用コーパスとした”bert-large-uncased”である。これにより、意味は近いが文脈上当てはまらない選択肢を生成できると考えられる。

接続詞に限り、接続詞の候補が少ないため、CSV ファイルに接続詞の分類を保存し、穴埋めで正答となった接続詞の分類以外の接続詞からランダムで表示することで処理時間の短縮を図っている。

選択肢候補の中には、選択肢として不適切である単語が含まれている。穴埋め箇所が文章の先頭である場合以外に単語に大文字が存在する場合、正答と誤答候補で品詞が一致しない場合、正答と同義語又は対義語が一致する場合、語幹が正答と同一である場合、正答と誤答間の GloVe による COS 類似度が 0.05 以下又は 0.8 以上の場合、ブラックリストに該当する場合、選択肢候補に記号また数字を含む場合、正答に ing / ed を含むが、誤答に ing / ed を含まない場合（意味問題ではなく文法的な問題となってしまうため）、他の選択肢と同一又は類似している場合、生成された単語がスペルミスであると考えられる場合、副詞の場合において選択肢の単語が 4 文字未満である場合は除外する。また、GPT-2 による文書の Perplexity スコアが正答を誤答が上回る場合(誤答の方が自然な文章だとされた場合)も除外する。

除外した選択肢候補が 3 以上になった場合は、テキストファイルに出力する。3 に満たない場合、文章自体は検証済みであり他の問題に使用可能であるためスタックファイルに穴あけ前の生成した文章を保存する。これは、GPT-2 による文章生成には時間を要してしまい、システム全体の処理時間短縮のためである。

2.2. システムの検証方法

本研究における目的は、TOEIC と類似した問題を

生成することにあるため、「意味の近さ」、「文章間の文脈の近さ」、「難易度・妥当性」の 3 点から生成した問題の質を検証する。

「意味の近さ」は WordNet[9]、Word2Vec 及び GloVe を用いた COS 類似度の傾向から検証する。WordNet は、英語の大規模データベースであり、本検証項目においては NLTK を通して使用する。類似度は分類学における 2 つの語義の深さと、Least Common Subsumer (最も具体的な祖先ノード) の深さに基づき 2 つの語義がどの程度似ているかを示すスコアを返すウー・パルマー類似度 (wup_similarity)[10]を用いて検証する。

Word2Vec を用いた COS 類似度は、Word2Vec によって求められた正答と誤答の単語ベクトル間の COS 類似度を示す。学習済みモデルには”GoogleNews-vectors-negative300.bin”を使用する。

GloVe[11]を用いた COS 類似度は、GloVe（Global Vectors for Word Representation）による単語ベクトルでの COS 類似度を示している。GloVe は Python ライブラリである spaCy において訓練済みパイプライン”en_core_web_lg”を通して使用する。

「文章間の文脈の近さ」は GPT-2 を用いた文章の Perplexity スコアの傾向から検証する。Perplexity スコアは、その文のトークンの並びが発生する確率であり、その文の自然差の評価として解釈することもできる。なお、発生する確率が高い方がスコアは小さくなる。なお、学習済みパラメータは”gpt2-large”を使用している。

「難易度・妥当性」は 18 歳～19 歳の 61 名に対してシステムで生成した文章を解いてもらう。同日に実施した TOEIC 模擬問題の得点の傾向（Part5 に該当する箇所）と分析する。得点を IRT 分析し、問題ごとの傾向を分析することで、問題としての妥当性を調べる。

3. 結果と考察

3.1. システムによって生成される問題

システムによって生成した問題の一例を以下に示す。なお、正答を太字で示す。（実際には別項目に出力される）

Directions: A word or phrase is missing in each of the sentences below. Four answer choices are given below each sentence. Select the best answer to complete the sentence. Then mark the letter (A), (B), (C), or (D) on your answer sheet.

No.1 This will remove the _____ and prevent it from spreading to other areas of the body .
(A) gloss (B) bleach
(C) stain (D) dab

No.2 We want to make sure that our _____ are treated with respect and dignity , said the company in a statement .
(A) retirees (B) laidoff
(C) employees (D) staffers

No.3 He will be replaced by former CEO _____ co-founder of the Canadian-based company , John G. Brown .
(A) since (B) till
(C) and (D) without

No.4 _____ you have any questions , please contact the manufacturer .
(A) If (B) Though
(C) Until (D) While

No.5 The company said it expects to generate $ 1.4 billion in revenue in the fourth quarter , up from $ 2.7 billion a year _____ .
(A) recently (B)hastily
(C) subsequently **(D) earlier**

No.6 The news agency said it had been _____ to find any other sources of revenue for the year .
(A) unable (B) unresponsive
(C) unconvinced (D) unprepared

No.7 The company is building a robot that will be able to perform tasks such as picking up objects and moving them _____ .
(A) aimlessly **(B) around**
(C) everywhere (D) erratically

No.8 The EIA estimates that the cost of crude oil rose by $ 1.6 trillion in the first quarter of 2016 , up _____ $ 2.1 trillion the previous year .
(A) onto (B) via
(C) upon **(D) from**

No.9 The project is expected to be completed by the end of the year , _____ to a statement from the company .
(A) summarizing **(B) according**
(C) outlining (D) estimating

No.10 The new system will be available to all employees , including those who are _____ on the payroll , for the first time in more than a decade .

(A) already (B) lately
(C) rightly (D) sparsely

No.2 の問題の選択肢である”retirees”（定年退職者[複数]）、 ”employees”（従業員[複数]）、”staffers”（新聞社の従業員・新聞記者[複数]）は意味的には区別できずに、回答者が正答を導くことは困難である。また、No.8 のように数字が多いなど実際の TOEIC にはないような問題が一部に出現することがある。

3.2. 問題生成にかかる時間について

システムにより問題を作成する際にかかる所要時間の分布について図 2 に示す。

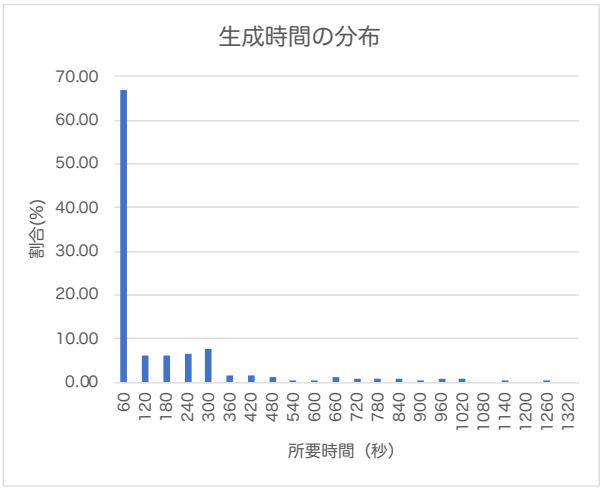

図 2 :システムによる問題生成の所要時間の分布

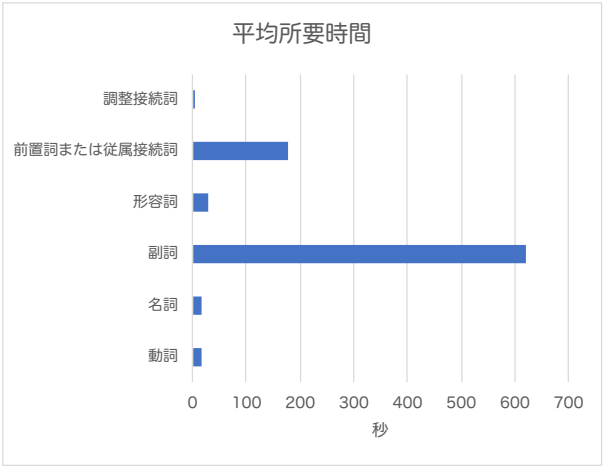

図 3:システムによる問題生成の品詞ごとの所要時間の平均

図 2 より、大半の問題は 60 秒以内で生成完了するが、一部の問題において 10 分を越す時間を要している。また、図 3 より所要時間を増大させている原因が副詞であることが確認できる。これは、GPT-2 において選択肢として使用できる副詞が発生する確率が他の品詞に比べて低く、何度も生成を繰り返してしまうためであり、この問題を解決するためには単文穴埋め問題の元になる文章の生成過程を検討する必要がある。

3.3. 意味的な類似度についての結果と検証

WordNet による ウ ー ・ パ ル マ ー 類 似 度 (wup_similarity)の分布を図 4 に、Word2Vec による COS 類似度を図 5 に、GloVe による COS 類似度を図 6 に示す。なお、COS 類似度は 1.0 に近いほど類似している。

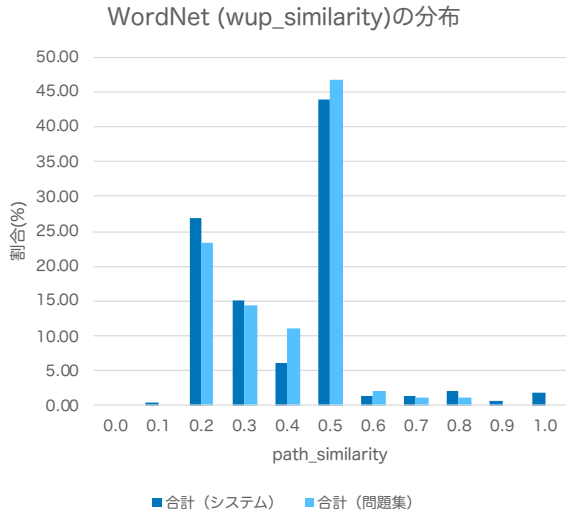

図 4: WordNet(wup_similarity)の分布

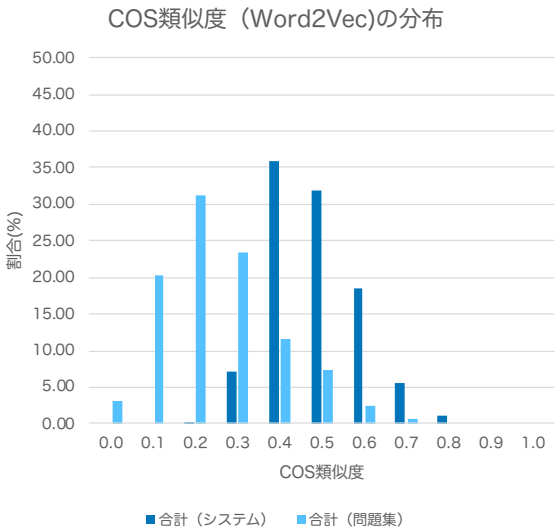

図 5 : Word2Vec による COS 類似度の分布

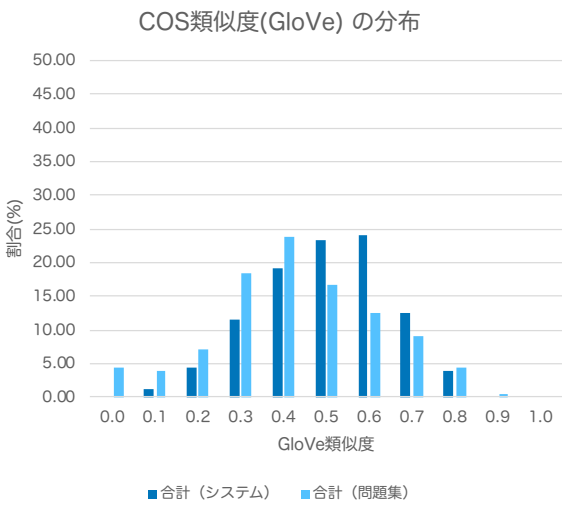

図 6 : GloVe による COS 類似度の分布

WordNet の結果の分布の傾向は図 2 より TOEIC 問題集と同程度である。また、COS 類似度の結果の分布の傾向は、図 5 及び図 6、システムで生成された問題の方が問題集の問題に比べて Word2Vec による類似度で 2 割、GloVe による類似度で 0.5 割程度高い傾向である。このことから、「単語間の意味の近さ」は、TOEIC と同程度～類似度が高い傾向で、TOEIC に比べてやや正答と誤答の見分けがつきにくい問題であると考えられる。

3.4. 文脈的な類似度についての結果と検証

文章の自然さを示す GPT-2 による Perplexity スコアの正当と誤答の差を図 7 に示す。なお、正答と誤答の Perplexity スコアの差（文章の自然さの違い）が大きいほど、問題は正解しやすいと考えられる。

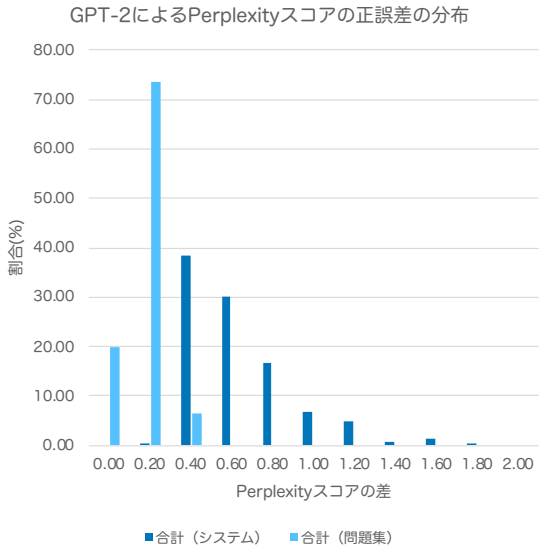

図 7: GPT-2 による Perplexity スコアの差の分布

「文章間の文脈の近さ」については、図 7 より、文脈を考慮すると全体的に正答を選択しやすい問題が多いが、TOEIC の問題と比べて問題間の難易度のばらつきが多いと考えられる。

「文章間の文脈の近さ」及び「単語間の意味の近さ」の結果より、単語だけで考えると正答を選択しにくく、文章全体の情報を元に推測すると正答を選択しやすい問題が多いことから、単語の知識のみだけではなく、文法の知識を勉強した成果が現れやすいような問題であると考えられる。

3.5. 難易度・妥当性ついての結果と検証

「難易度・妥当性」について、高専 4 年生に解いていただいた TOEIC 模擬問題の Part5 該当箇所の得点（10 問）とシステムによる問題の各問題の正答率の関係を図 8 に示す。なお、点は TOEIC 模擬問題の正答数該当者における各問題の正答率を、線は回帰直線を示している。検証に使用した問題は、3.1 の項目で示したシステムで生成した問題である。

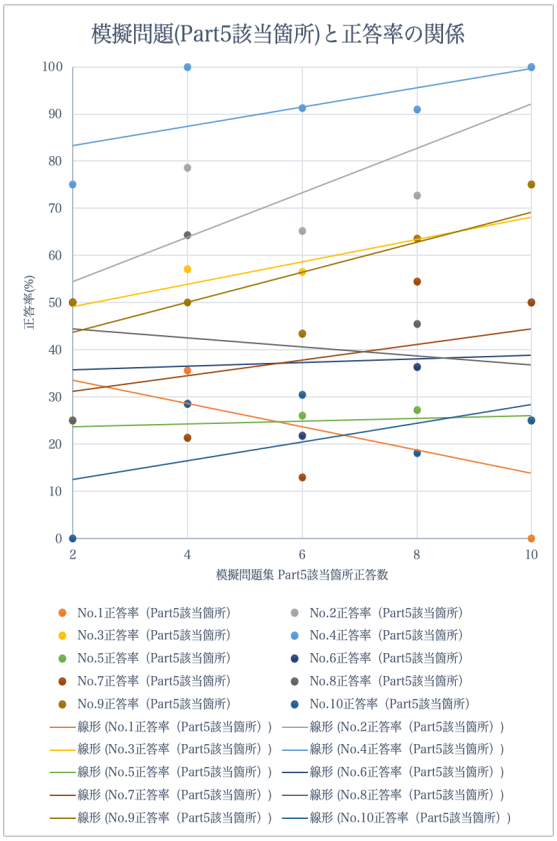

図 8:模擬問題(Part5)と正答率の関係

図 8 より、大半の問題は TOEIC の Part5 該当箇所の得点が上がるにつれて正答率も上がる傾向になっていることが確認できる一方、10 問中 2 問は得点が上がるにつれて正答率も下がる傾向にありこの問題は問題として成立しておらず不適切な問題であると

いえることから、次の IRT 分析の過程においてはこの 2 問を除いた 8 問で検証することにする。
図 9 に IRT 分析における項目特性曲線（ICC）を、図 10 にテスト情報曲線（TIC）を示す。ICC は縦軸が能力値（値が大きいほど能力が高い）を横軸が正答率を、TIC は縦軸が能力値を横軸がその能力値の学習者に対して能力をどの程度測定できているかを示している。

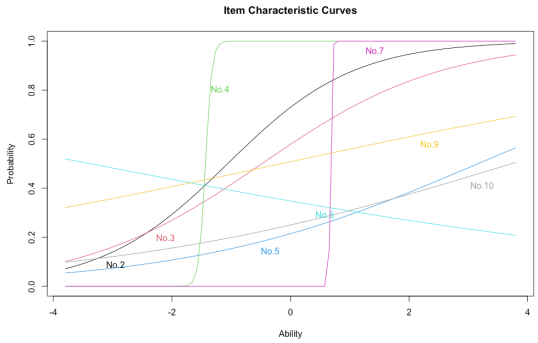

図 9: IRT 分析（項目特性曲線・ICC）

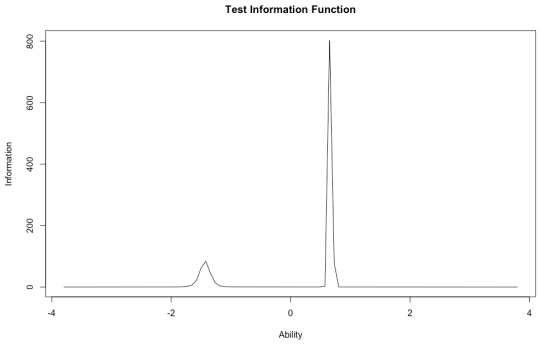

図 10 : IRT 分析（テスト情報曲線・TIC）

図 9 の IRT 分析（ICC）より、No.4 や No.7 などの特定の能力が測れる問題（ある得点の能力で正答率が大きく変化する）や No.2、No.3 などの問題としては成立している問題（能力が低い人は解けずに能力が高い人は解ける）が存在している一方、No.5 や No.10 のような能力があるのにも関わらず解けないまたは難しすぎる問題や、No.6 のような能力があがるにつれ正答率が下がる問題が存在することが確認できる。また、図 10 より特定の能力しか正確に測れていないことが確認できる。
このことより、システムによって生成された問題には、能力が上がるにつれて解けなくなる問題として成立していない問題が一定数含まれているが、問題間の問題の難易度のばらつきがやや大きいものの、大半の問題は妥当であると考えられる。

4. 結言

このシステムは GPT-2 により生成した文章から穴埋め箇所を決定し、Word2Vecd による正答の類似単語から BERT による穴埋め箇所の候補の単語の差を使用して選択肢を生成する。
生成した問題には、一部不適切な問題が混入するが TOEIC の意味問題とある程度類似する問題を作成することができる。このシステムを利用することで、TOEIC に類似した問題を多数解くことができる。

参考文献

[1] 一般財団法人 国際ビジネスコミュニケーション協会 :”TOEIC® Program 大学 の入学試験における活用状況,” https://www.iibc-global.org/toeic/official_data/univ_research.html (2020)
[2] 一般財団法人 国際ビジネスコミュニケーション協会 :”英語活用実態調査 企 業・団体 ビジネスパーソン 2019,” https://www.iibc-global.org/library/default/toeic/official_data/lr/katsuyo_2019/pdf/katsuyo_2019_corpo.pdf (2019)
[3] 鳥木瑛司,原田恵雨, 佐藤奈々恵: “IRT 分析を用いた TOEIC 問題自動生成システム”, 北第 20 回複雑系マイクロシンポジウム (2021)
[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever: Language Models are Unsupervised Multitask Learners, OpenAI Blog (2019)
[5] Patrick von Platen , How to generate text: using different decoding methods for language generation with Transformers, https://huggingface.co/blog/how-to-generate (2020)
[6] NLTK Project, NLTK Natural Language Toolkit Documentation, https://www.nltk.org (2022)
[7] Tomas Mikolov, Greg Corrado, Kai Chen, Jeffrey Dean : Efficient Estimation of Word Representations in Vector Space ,ICLR Work-shop Papers (2013)
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova : BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Google AI Language (2018)
[9] George A. Miller : WordNet: A Lexical Database for English, Communications of the ACM Vol. 38, No. 11: 39-41. (1995)
[10] Zhibiao Wu, Martha Palmer: Verb Semantics and Lexical Selection, ACL '94: Proceedings of the 32nd annual meeting on Association for Computational LinguisticsJune, p.133–138 (1994)
[11] Jeffrey Pennington, Richard Socher, Christopher D. Manning : GloVe: Global Vectors for Word Representation , Association for Computational Linguistics : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) ,p.1532-1543(2014)