外国語検定の面接試験において生成する質問の難易度選定

林鳴昊 [1]　伊藤滉一朗 [1]　松原茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科　[2] 名古屋大学情報連携推進本部
lin.minghao.b3@s.mail.nagoya-u.ac.jp　ito.koichiro.v1@s.mail.nagoya-u.ac.jp
matsubara.shigeki.z8@f.mail.nagoya-u.ac.jp

概要

外国語学習者の語学力を客観的かつ迅速に判定する需要が高まっている．読解力，聴解力，作文力の測定については，筆記試験やリスニング試験の作問あるいは採点の自動化が検討されている．一方，会話力の測定については十分に検討されていない．本稿では，外国語の会話力測定のための面接試験の自動化に向けて，生成する質問の難易度を選定する手法を提案する．本手法では，受験者の能力に応じて柔軟に質問の難易度を選定するために，受験者の返答の適切さを利用する．日本語学習者会話データベースの模擬試験データを用いて実験を行い，本手法の有効性を確認した．

1　はじめに

留学や就職で海外に渡る人が増加するにつれて，外国語学習者の語学力を評価する需要が高まっている．語学力の評価では，読解力，聴解力，作文力，会話力の測定が行われる [1]．このうち，読解力，聴解力，作文力の測定については，試験問題の作成や採点などにおいて，自動化が検討されている（例えば，[2, 3, 4]）．一方で，会話力の測定については，自動測定を目的とした研究があるものの [5]，その検討は十分ではない．

会話力を測定するための試験はその形態により，モノローグ試験と面接試験に分けられる．モノローグ試験では，事前に決められた質問に対する受験者の返答によって会話力を評価する．すなわち，モノローグ試験は，試験中に質問が変更されることがなく，静的な試験であるといえる．一方，面接試験は，面接官が受験者との会話を通して質問を柔軟に変更しながら，その返答に基づき会話力を評価するものであり，動的な試験であるといえる．

会話力の測定においては，実世界での会話の要素を試験に反映させることが期待されている [6]．面接試験は，モノローグ試験と比較して，高い信頼度で会話力を測定でき，実世界での人間同士の会話のシミュレートも容易である．そのため，近年のコンピュータを利用した言語アセスメントは，対話的または動的な方略を導入し，個々の学習者に焦点を当てた測定へと移行している [7]．

本稿では，外国語の会話力測定のための面接試験の自動化に向けて，生成する質問の難易度を選定する手法を提案する．本手法では，受験者の能力に応じて柔軟に質問の難易度を選定するために，受験者の返答の適切さを利用する．面接官と受験者の対話文脈に加えて，直前の返答の適切さを入力として，面接官の次の質問の難易度を選定する．難易度を選定するためのモデルとして，BERT [8] を採用した．模擬面接試験のデータセットを用いた実験を行い，本手法の有効性を確認した．

2　関連研究

会話力を測定するための試験は，モノローグ試験と面接試験に分けられる．これまで，モノローグ試験の自動化に関する研究が行われている [9, 10, 11]．採点の自動化に関しては，Siamese convolutional neural network [12] を用いて受験者の会話力を評価する研究 [9] や，線形回帰モデルを用いて受験者の返答を評価する研究 [10] が存在する．

試験の実施を含めた自動化を目指す研究も存在している．受験者に対して，文の読み上げや短文問題への返答などを機械的に要求する Pearson Test of English の結果から，会話力を評価する研究が行われている [13]．また，モノローグ試験を実施するための音声対話システムを構築する試みもある [11]．このシステムでは，読み上げタスクと繰り返しタスクが実施可能であり，流暢性，発音，繰り返しの精度に着目して，会話力を評価する．

モノローグ試験と面接試験の評価の間には，高い相関があることが示されているものの [6]，面接試

表 1　面接試験における会話の例

Q_1	どんなスポーツが好きですか．【易しい】
R_1	サッカーが好きです．　【適切】
Q_2	サッカーのルールを説明してくれますか．【難しい】
R_2	えっと，それは ... ゲームで ... すみません，説明できない．【不適切】
Q_3	あ，そうですか．じゃあ，サッカーって楽しいですか．【易しい】
R_3	はい，楽しいです．【適切】

験の方が高い信頼度で会話力を測定でき，実世界での人間同士の会話のシミュレートも容易である．また，会話力の測定においては，実世界での会話の要素を試験に反映させることが期待されている [6]．しかし，質問の難易度が動的に変化する面接試験に関しては，その自動化はあまり進んでいない．

3　面接試験の質問の難易度

外国語の会話力測定のための面接試験では，受験者の返答の適切さに応じて，面接官は質問の難易度を調整している [14]．

表 1 に，面接試験の会話例を示す．Q_i と R_i は，それぞれ，i 番目の面接官の質問と受験者の返答を表す．また，【】内は，質問の難易度または返答の適切さを表す．例えば，2 つ目の質問 Q_2 に対する返答 R_2 は適切ではない．受験者が Q_2 の質問に返答する会話力を備えておらず返答に困ったために，R_2 のような不適切な返答になり，これを受けて，面接官の次の質問 R_3 は，2 択で返答できる易しい質問となったと考えられる．このように面接官は，受験者の返答の適切さに応じて試験中に質問の難易度を動的に変えることで，受験者の言語能力を評価している [15]．そこで本稿では，外国語の会話力測定のための面接試験の自動化に向けて，生成する質問の難易度を選定する手法を提案する．

4　提案手法

4.1　手法の概要

外国語の会話力測定のための面接試験では，面接官は，受験者の返答の適切さに応じて，質問の難易度を調整している．そこで本手法では，面接官の質問方略に基づき，2 段階のアプローチをとる．図 1 に，本手法のアーキテクチャを示す．本手法は，返答の適切さ推定モデルと質問の難易度選定モデル

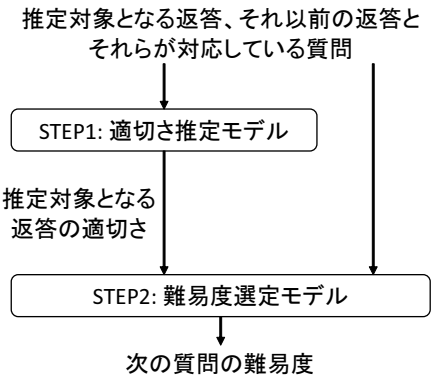

図 1　提案手法のアーキテクチャ

の 2 つから構成される．まず，面接試験の対話文脈から，推定対象となる返答の適切さを推定し，続いて，対話文脈と適切さの推定結果から，次の質問の難易度を選定する．推定および選定モデルには，BERT を用いる．BERT は，Transformer [16] アーキテクチャをベースに構築されており，対話システム関連のタスクでの適用例も多い [17, 18, 19]．

4.2　返答の適切さ推定

返答の適切さ推定モデルでは，面接試験の会話における，i 番目の受験者の返答 R_i の適切さを推定する．推定された適切さは，面接官が R_i の次に生成する質問の難易度の選定に利用される．返答の適切さの推定には，推定対象の返答を含む直前 2 組の質問と返答の組を用いる．すなわち，$[Q_{i-1}, R_{i-1}, Q_i, R_i]$ から，R_i の適切さを推定する．

返答の適切さ推定モデルは，BERT を fine-tuning することで構築する．[[CLS] Q_{i-1} [SEP] R_{i-1} [SEP] Q_i [SEP] R_i [SEP]] という系列をモデル全体の入力として与え，[CLS] トークンに対応する BERT の最終層の出力 $T_{[\mathrm{CLS}]}$ を，線形変換と softmax 関数からなる分類層に入力することで，R_i の適切さを出力する．なお，BERT は，2 種類の Segment embeddings を持つが，適切さ推定モデルでは，入力系列全体を 1 つのセグメントとみなし，系列全体で同種の Segment embeddings を適用している．参考として，モデル図を Appendix の A に示す．

4.3　質問の難易度選定

難易度選定モデルでは，面接試験の会話における，i 番目の受験者の返答 R_i の次に生成する質問の難易度を選定する．すなわち，Q_{i+1} の質問の難易度を選定する．難易度選定には，直前の 2 組の質問と

返答の組に加えて，直前の返答の適切さを用いる．直前の返答の適切さ（以降では，適切さラベルとも呼ぶ）を L とすると，$[Q_{i-1}, R_{i-1}, Q_i, R_i]$ と L を入力として，Q_{i+1} の難易度を選定する．

難易度選定モデルは，適切さ推定モデルと同様に，BERT を fine-tuning することで構築する．ただし，適切さラベル L を，特殊トークンとして入力系列に加える．[[CLS] L [SEP] Q_{i-1} [SEP] R_{i-1} [SEP] Q_i [SEP] R_i [SEP]] という系列を入力として与え，[CLS] トークンに対応する BERT の最終層の出力 $T_{[CLS]}$ を，線形変換と softmax 関数からなる分類層に入力することで，Q_{i+1} の難易度を出力する．

難易度選定モデルでは，Xiong ら [20] のラベルフュージングの手法を参考に，非自然言語であるラベルと自然言語である対話文脈を別のセグメントとみなし，それぞれに別の Segment embeddings を割り当てる．すなわち，[[CLS] L [SEP]] には 1 つ目のセグメントに対応する Segment embeddings を適用し，後続の系列には 2 つ目のセグメントに対応する Segment embeddings を適用する．参考として，モデル図を Appendix の A に示す．

5 実験

5.1 実験データ

本手法の有効性を検証するために，難易度選定実験を実施した．実験には，国立国語研究所が公開している日本語学習者会話データベース (JLCD)[1] の模擬面接試験データを用いた．模擬面接試験は，ACTFL Oral Proficiency Interview (OPI) の標準に従って行われ，実生活の中で効果的かつ適切に言語を使用する能力を評価するものである．約 20 分の試験を経て，受験者の会話力は，大きく分けて，初級，中級，上級，超級の 4 段階で評価される．

模擬面接データには，390 個の試験データが含まれている．本実験では，その一部分である 59 個の試験データを，実験データとして利用した．その書き起こしデータを発話単位に分割した結果，3,251 個の面接官の質問と受験者の返答に分割された．実験データを試験単位で約 8:1:1 に分割し，それぞれ，学習，開発，テストに用いた．

実験データ内の返答に対して，その適切さを人手で付与した．本実験では，適切さは，返答が適切か不適切かの 2 値とした．表 2 に，実験データにおける返答の適切さの分布を示す．実験データ内の質問に対しても，その難易度を人手で付与した．ACTFL-OPI 標準では，

1. 初級から中級程度の会話力では，2 択問題や単純な事実と結論を求める質問に適切に返答可能
2. 上級から超級程度の会話力では，物事に対する詳しい説明，描写，叙述，比較，個人的な意見を求める質問に返答可能

とされている．これらを踏まえて，本実験では，上記の 1 に対応する質問を易しい質問，2 に対応する質問を難しい質問とし，難易度を易しいか難しいかの 2 値で付与した．表 3 に，実験データにおける質問の難易度の分布を示す．

表 2 適切さの分布

適切さ	train	dev	test
適切	2,069	233	292
不適切	491	73	93

表 3 難易度の分布

難易度	train	dev	test
易しい	1,829	228	287
難しい	731	78	98

5.2 モデル設定

事前学習済みの BERT を fine-tuning することで，返答の適切さ推定モデルと質問の難易度選定モデルを実装した．本実験では，返答の適切さと質問の難易度は共に 2 値としているので，推定および選定モデルは共に 2 クラス分類を行うモデルである．いずれのモデルについても，事前学習済みの BERT として，東北大学が公開しているモデル[2] を用いた．

各モデルとも，バッチサイズを 16，最適化アルゴリズムを AdamW [21]，重み減衰を 0.01，ドロップアウト率を 0.1，入力トークンの最大長を 512 に設定した．学習率は，返答の適切さ推定モデルでは 1e-5, 質問の難易度選定モデルでは 2e-7 とした．また，各モデルとも，損失関数は Cross Entropy Loss とした．ただし，損失には各クラスの重みの逆数に比例する class-weight を適用した．各モデルとも，10 エポック学習し，1 エポックごとに開発データでの損失を計算し，その損失が最も小さかったモデルを用いて，テストデータに対する性能を評価した．

4 章で提案した難易度選定モデルの学習には，返答の適切さの正解ラベルを用いた．一方，評価では，返答の適切さ推定モデルで推定されたラベルを用いた．以降では，このように学習させた難易度選定モデルを [ours (noisy)] と表記する．[ours (noisy)] の評価時の性能は，返答の適切さ推定モデルの推定

1) https://mmsrv.ninjal.ac.jp/kaiwa/DB-summary.html

2) https://github.com/cl-tohoku/bert-japanese

誤りの影響を受ける．

5.3 評価指標と比較手法

本実験では，適合率，再現率，F1 値のマクロ平均を評価指標とした．また，本手法の性能を評価するために，比較手法を実装した．BERT を用いた返答の適切さ推定手法に対しては，返答によらずランダムに推定を行う手法を比較手法とした．返答が適切であると推定する確率は，学習データにおける適切な返答の割合と同じく 0.809 とした．以降では，BERT を用いた推定手法を [ours]，ランダムに推定する手法を [random] と表記する．

質問の難易度選定手法 [ours (noisy)] に対しては，以下の 4 つの手法を比較手法とした．

- [random]: 質問の難易度によらずランダムに難易度を選定する．易しい質問であると選定する確率は，学習データにおける易しい質問の割合と同じく 0.714 とした．
- [vanilla BERT]: 返答の適切さを利用せずに，直前の 2 組の質問と返答から難易度を選定する．
- [ours (noisy) w/o seg]: 返答の適切さを利用するが，入力系列全体で同種の Segment embeddings を適用している．
- [ours (correct)]: 評価時においても，返答の適切さの正解ラベルを用いて，難易度を選定する．本手法の性能は，返答の適切さの推定性能が 100%という理想的な環境下での参考値である．

5.4 実験結果

表 4 に，適切さ推定手法の性能を示す．[ours] は，比較手法である [random] を全ての評価指標で上回った．このことから，BERT を利用することで，返答の適切さの推定がある程度の性能で可能であることを確認した．また，[ours] の正解率は 0.758 であった．このことは，質問の難易度選定手法の性能評価において，[ours (noisy)] および [ours (noisy) w/o seg] が利用した返答の適切さラベルのうち，75.8%のラベルが正しいものであったことを意味する．

表 5 に，質問の難易度選定手法の性能を示す．提案手法 [ours (noisy)] は，返答の適切さを利用しない [random] と [vanilla BERT] を全ての評価指標で上回った．このことは，質問の難易度選定において，返答の適切さを利用することの有効性を示している．また，[ours (noisy)] は，[ours (noisy) w/o seg] に対しても，全ての評価指標で上回った．このことは，非自然言語である返答の適切さラベルと，自然言語である対話文脈を別のセグメントとみなし，それぞれに別の Segment embeddings を割り当てることの有効性を示している．なお，提案手法 [ours (noisy)] は，[ours (correct)] と比較して，全ての評価指標でわずかに性能が低下した．この結果は，返答の適切さ推定モデルの誤差によるものと考えられる．参考として，Appendix の B に，本実験における提案手法 [ours (noisy)] による難易度選定の例を示す．

表 4 返答の適切さ推定手法の性能

	適合率	再現率	F1
[random]	0.473	0.480	0.472
[ours]	**0.665**	**0.653**	**0.658**

表 5 質問の難易度選定手法性能評価

	適合率	再現率	F1
[random]	0.500	0.500	0.499
[vanilla BERT]	0.605	0.619	0.609
[ours (noisy) w/o seg]	0.626	0.638	0.631
[ours (noisy)]	**0.629**	**0.648**	**0.635**
[ours (correct)]	0.631	0.650	0.637

6 おわりに

本稿では，外国語の会話力測定のための面接試験において，生成する質問の難易度を選定する手法について述べた．本手法では，直前の返答の適切さを推定したのちに，その適切さを踏まえて，次の質問の難易度を選定する．また，返答の適切さを効果的に利用するために，BERT における Segment embeddings に着目したラベルフュージング手法を取り入れた．模擬面接試験のデータセットを用いた実験を行い，本手法の有効性を確認した．

本手法では，質問と返答については，直前 2 組を入力として利用したが，さらに前の質問と返答の利用も検討したい．また，面接官は，試験の初期段階ではプロービングに，後期段階ではレベルチェックに重点を置くというように，その質問方略は変化すると考えられる．このような変化を考慮するために，今後は，面接試験における経過時間などの利用も検討したい．

参考文献

[1] Donald E. Powers. The case for a comprehensive, four-skills assessment of English-language proficiency. **R & D Connections**, Vol. 14, pp. 1–12, 2010.

[2] Yi-Ting Huang, Ya-Min Tseng, Yeali S Sun, and Meng Chang Chen. Tedquiz: automatic quiz generation for ted talks video clips to assess listening comprehension. In **Proceedings of the 2014 IEEE 14th International Conference on Advanced Learning Technologies**, pp. 350–354, 2014.

[3] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1342–1352, 2017.

[4] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 180–189, 2011.

[5] Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier Van Dalen, and David Vandyke. Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English. In **Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 270–275, 2016.

[6] Jared Bernstein, Alistair Van Moere, and Jian Cheng. Validating automated speaking tests. **Language Testing**, Vol. 27, No. 3, pp. 355–377, 2010.

[7] Akbar Bahari. Computer-assisted language proficiency assessment tools and strategies. **Open Learning: The Journal of Open, Distance and e-Learning**, Vol. 36, No. 1, pp. 61–87, 2021.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[9] Su-Youn Yoon and Chong Min Lee. Content modeling for automated oral proficiency scoring system. In **Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 394–401, 2019.

[10] Klaus Zechner, Keelan Evanini, Su-Youn Yoon, Lawrence Davis, Xinhao Wang, Lei Chen, Chong Min Lee, and Chee Wee Leong. Automated scoring of speaking items in an assessment for teachers of English as a foreign language. In **Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 134–142, 2014.

[11] Febe De Wet, Christa Van Der Walt, and Thomas R. Niesler. Automatic assessment of oral language proficiency and listening comprehension. **Speech Communication**, Vol. 51, No. 10, pp. 864–874, 2009.

[12] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In **Proceedings of the 30th AAAI Conference on Artificial Intelligence**, pp. 2786–2792, 2016.

[13] Pearson Longman. **Official Guide to Pearson Test of English Academic**. Pearson Japan, 2nd edition, 2012.

[14] Gabriele Kasper. When once is not enough: Politeness of multiple requests in oral proficiency interviews. **Multilingua**, Vol. 25, No. 3, pp. 323–350, 2006.

[15] ACTFL. Oral proficiency interview familiarization manual, 2012. https://community.actfl.org.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in neural information processing systems**, Vol. 30, 2017.

[17] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 917–929, 2020.

[18] Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 167–177, 2021.

[19] Peiyao Wang, Joyce Fang, and Julia Reinspach. CS-BERT: a pretrained model for customer service dialogues. In **Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI**, pp. 130–142, November 2021.

[20] Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. Fusing label embedding into BERT: An efficient improvement for text classification. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 1743–1750, 2021.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **Proceedings of the 7th International Conference on Learning Representations**, 2019.

A　モデル図

返答の適切さ推定モデル，及び，質問の難易度選定モデルについて補足する．図 2 と図 3 に，それぞれのモデル図を示す．q_i^j および r_i^j は，それぞれ，i 番目の質問 Q_i と返答 R_i 中の j 番目のトークンを表す．質問の難易度選定モデルでは，R_i の適切さラベル L を入力に加えており，[[CLS] L [SEP]] には 1 つ目のセグメントに対応する Segment embeddings E_A を適用し，後続の系列には 2 つ目のセグメントに対応する E_B を適用する．

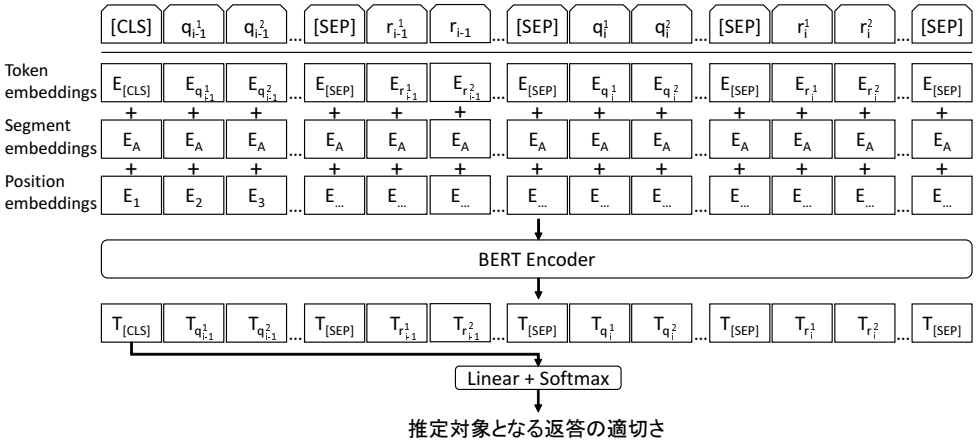

図 2　返答の適切さ推定モデル構造

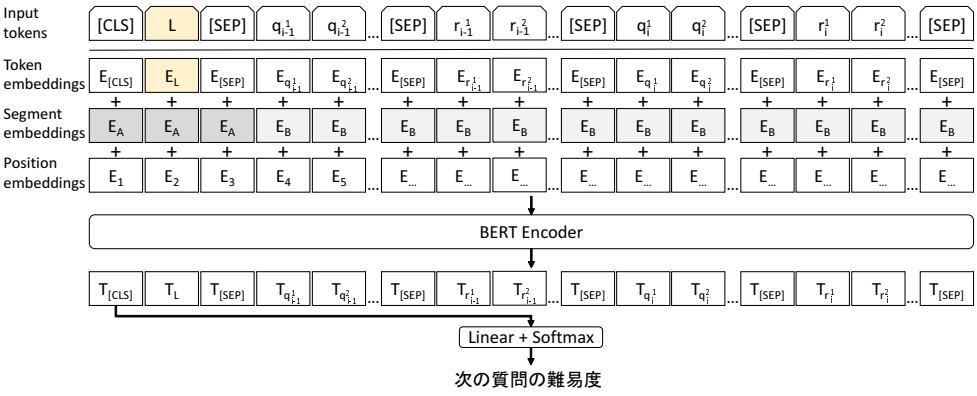

図 3　質問の難易度選定モデル

B　難易度選定の例

表 6 に，[ours (noisy)] と [vanilla BERT] による難易度選定の例を示す．適切さの列は，返答の適切さの推定結果を示している．なお，この推定結果は正しいものである．また，難易度の列は，データ中の正解ラベルを示している．この例では，[ours (noisy)] は正しく難易度を選定できたが，[vanilla BERT] は誤っている．[ours (noisy)] は，R_{55} が適切な返答であるという情報を利用して，R_{55} の次に生成される質問は難しい質問であると正しく選定できたものと考えられる．

表 6　[ours (noisy)] と [vanilla BERT] による難易度選定の例（Q_{56} の難易度が選定の対象）

	発話	適切さ	難易度	[vanilla BERT]	[ours (noisy)]
Q_{54}	そうするとローマ字みたいな音で送るんですか，今．	-	易しい	-	-
R_{54}	音で．	-	-	-	-
Q_{55}	うん今ローマ字みたいに，英語を使ってモンゴル語を送るっていうのは，ロシア語の文字をほんとは使うんだけども，英語の文字で，モンゴル語を．	-	易しい	-	-
R_{55}	あー，ロシア語の文字，と似ているんですけどもロシア語の文字じゃないよ．	適切	-	-	-
Q_{56}	えーえー，どこが違いますか．	-	難しい	易しい	難しい