変分質問回答ペア生成による質問応答モデルの汎化性能と頑健性の向上

篠田一聡 †‡　菅原朔 ‡　相澤彰子 †‡
† 東京大学大学院情報理工学系研究科　‡ 国立情報学研究所
shinoda@is.s.u-tokyo.ac.jp　{saku,aizawa}@nii.ac.jp

1 はじめに

文章を読解して質問に答えるシステムを作るために機械読解の研究が盛んに行われている [1, 2, 3]. そのためにはニューラルネットを訓練するための大量の質問回答ペアが必要であるが，その作成には膨大な人手コストを要する．そこでコスト削減のため，自動で質問回答ペアを生成してデータ拡張を行う研究が注目を集めている [4, 5, 6].

一方で質問応答モデルが訓練データと違う分布の分布外 (out-of-distribution; OOD) データセットには汎化しないこと [7] や質問応答モデルが頑健ではない困難なテストデータが存在すること [8, 9, 10] も実用上問題である．データ拡張を適切に行えばモデルの過学習を回避してこれらの問題を解決できる可能性がある [10, 8] が，質問回答ペア生成の既存研究ではこのような特性はあまり注目されて来なかった．

本研究では，データ拡張によって訓練データの質問回答ペアの多様性を向上することで，質問応答モデルの汎化性能と頑健性の向上を目指す．訓練データの多様性を向上するために，多様な質問回答ペアを生成する「変分質問回答ペア生成モデル」を提案する．評価用データセットには分布内テストデータに加えて，分布外テストデータと困難なテストデータを用いて汎化性能と頑健性を評価する．本研究の貢献は以下の 3 点である．

- 変分質問回答ペア生成モデルを提案し，既存のモデルに比べて質問回答ペアの多様性を大幅に向上させた．このモデルを用いて合成データセットを作成した．
- 作成した合成データセットでデータ拡張を行った結果，分布内テストデータにおいて既存の質問回答ペア生成手法に比肩する結果が得られた．
- ターゲットの分布がモデルにとって未知であるにも関わらず，提案手法は分布外テストデータへの汎化性能と困難なテストデータへの頑健性の向上に寄与することがわかった．

2 変分質問回答ペア生成モデル

条件付き変分自己符号化器を拡張することで，多様な質問回答ペアを単一の文章から生成可能な深層生成モデルを提案する．SQuAD [1] などの人が作成した質問応答データセットでは，一つの文章 c に複数の回答 a が存在し，文章と回答のペアに対して複数の質問 q が作成されうることから，回答と質問をそれぞれ異なる潜在空間に埋め込むことが妥当であると考えられる．従って，2 つの互いに独立な連続潜在変数 z と y を導入し，回答と質問を多様化する役割をそれぞれに持たせる．以下の章ではモデルの詳細について説明する．

2.1 目的関数

提案モデルは確率的潜在変数を含むため，対数周辺尤度を直接計算することができない．その代わりにその変分下限を最大化することで訓練ができる．しかし，変分自己符号化器の変分下限を最大化すると KL 項の値が 0 に収束して復号化器の出力が潜在変数によらず一定になってしまう posterior collapse 問題が起こることが知られている [11]．提案モデルでもその現象が確認されたため，この問題を回避するために Burgess ら [12] によって提案された KL 項の値を明示的に制御する手法を用いる．よって，提案モデルの目的関数 $\mathcal{L}$ は以下のようになる．

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{q_\phi(z,y|q,a,c)}[\log p_\theta(q|y,a,c) + \log p_\theta(a|z,c)] \\ - |D_{\mathrm{KL}}(q_\phi(z|a,c)||p_\theta(z|c)) - C_a| \\ - |D_{\mathrm{KL}}(q_\phi(y|q,c)||p_\theta(y|c)) - C_q| \end{aligned} \quad (1)$$

θ は復号化器の，ϕ は符号化器のパラメータであり，D_{KL} は KL ダイバージェンス，$C_a, C_q \geq 0$ は KL 項

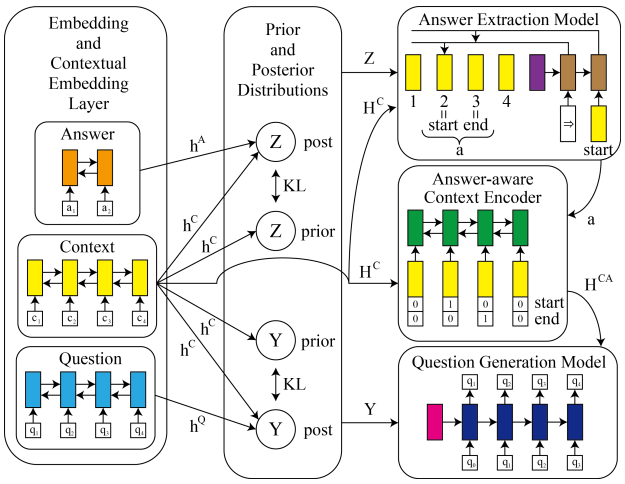

図 1　提案モデルの全体像．各構成要素の入力と出力が矢印とともに示されている．

の値を制御するためのハイパーパラメータである．

2.2　モデル構造

モデルの全体像を図 1 に示す．以下各構成要素について説明する．

2.2.1　Embedding and Contextual Embedding Layer

文章/質問/回答中の各単語の埋め込みベクトルと文字レベル埋め込みベクトルを結合し，異なる BiLSTM に入力する．双方向の LSTM の最終時刻の隠れ状態ベクトルを連結したものを $h \in \mathbb{R}^{2d}$，各時刻の隠れ状態ベクトルを連結したものを $H \in \mathbb{R}^{L \times 2d}$ とする．図 1 ではこれらがどの入力から得られたものかを示すために右肩に添字を付した．L は系列の長さで d は隠れ状態ベクトルの次元数である．

2.2.2　Prior and Posterior Distributions

潜在変数の事前確率と事後確率は分散共分散行列は対角行列の多変量正規分布に従うと仮定する．潜在変数 z と y の事前確率と事後確率の平均と分散の対数は h^C, h^Q, h^A を線形層に入力して得られる．その後に，サンプリングされた z と y はそれぞれ回答抽出モデルと質問生成モデルに入力する．

2.2.3　Answer Extraction Model

ここでは回答抽出を 2 ステップの自己回帰的な復号とみなして，回答の確率を $p(a|c) = p(c_{start}|c)p(c_{end}|c_{start}, c)$ のようにして求める．回答抽出モデルには，潜在変数 z を線形層に入力して得られる出力を初期状態の隠れ状態ベクトルとするように pointer network を拡張したものを用いる．この拡張によって潜在変数 z から回答 a への写像を学習することが可能になる．H^C を attention スコアの計算に用いる．

2.2.4　Answer-aware Context Encoder

回答の位置を考慮した文章についての情報を得るために，さらに BiLSTM を用いる．H^C に回答の始点と終点の one-hot ベクトルを連結したものを BiLSTM に入力し，各時刻の双方向の隠れ状態ベクトルを連結して $H^{CA} \in \mathbb{R}^{L \times 2d}$ を出力として得る．

2.2.5　Question Generation Model

潜在変数 y と H^{CA} を元に質問を生成する．質問生成モデルには LSTM 復号化器と注意機構，コピー機構 [13] を用いて質問中の単語を先頭から順に予測する．ここでも y を線形層に入力して得られる出力を復号化器の初期状態の隠れ状態ベクトルとする拡張を行う．H^{CA} は注意機構とコピー機構で用いる．

3　実験・結果

3.1　データセット

データセットには Wikipedia の文章からクラウドワーカーが作成した約 10 万の質問回答ペアからなる SQuAD v1.1 [1] を用いる．SQuAD はテストデータが公開されていないため，質問生成の先行研究でよく用いられる SQuAD-Du [14] を用いる．データのサイズは訓練データが 70,484，検証データが 10,570，テストデータが 11,877 である．

3.2　回答抽出

提案モデルが抽出した回答と SQuAD-Du の回答の一致度に加えて，抽出された回答の多様性を検証するために，提案モデルで各文章から 50 の回答をランダムに抽出した．C_a は様々な値を検証したが，紙幅の関係で限られた値の結果のみ報告する．

評価指標　回答同士の一致度の評価には Proportional Overlap (Prop.) と Exact Match (Exact) を用いた [5]．それぞれの指標について Precision と Recall を計算して多対多の評価に用いる．多様性の評価には抽出された回答から重複を省いた後の総数を Dist スコアとして定義して用いた．

ベースライン　固有表現抽出（NER）と，BiLSTM-CRF で回答抽出を学習している HarestingQG (HarQG) [5] をベースラインに用いた．

	Relevance				Diversity
	Precision		Recall		Dist
	Prop.	Exact	Prop.	Exact	
NER	34.44	19.61	64.60	45.39	30.0k
HarQG	45.96	33.90	41.05	28.37	-
Ours					
$C_a = 0$	**58.39**	**47.15**	21.82	16.38	3.1k
$C_a = 5$	30.16	13.41	**83.13**	**60.88**	71.2k
$C_a = 20$	21.95	5.75	72.26	42.15	**103.3k**

表 1 テストデータにおける回答抽出の結果.

		Relevance				Diversity		
	N	B1-R	ME-R	RL-R	Token	D1	E4	SB4
SemQG	50	**62.32**	**36.77**	**62.87**	7.0M	15.8k	18.28	91.44
Ours								
$C_q = 0$	50	35.57	18.31	33.92	7.6M	14.4k	17.33	97.61
$C_q = 5$	50	44.19	25.84	45.18	11.5M	19.0k	19.71	82.59
$C_q = 20$	50	48.19	25.29	48.26	4.9M	**22.4k**	**19.72**	**44.41**

表 2 テストデータにおける質問生成の結果.

結果 表 1 に結果を示す．C_a を適切な値に設定することで，先述の posterior collapse 問題を回避して多様性を向上しつつより高い再現度を達成した．SQuAD-Du よりも多様性が向上しているため，Precision の低さは必ずしも抽出された回答の質が低いことを意味しないと考える．

3.3 質問生成

参照質問の再現度と生成された質問の多様性を評価するために，文章と回答を入力としてランダムに 50 の質問を生成させた．

評価指標 再現度の評価には BLEU-1 (B1) / Meteor (ME) / ROUGE-L (RL) の Recall (-R)，多様性の評価には Dist-1 (D1) / Ent-4 (E4) / Self-BLEU-4 (SB4) を用いた．

ベースライン SemanticQG (SemQG) [6] をベースラインとして用いる．提案手法と公平な比較をするために，Diverse Beam search で復号を行い，スコアの高い 50 の質問文を各入力について生成した．

結果 表 2 に結果を示す．提案モデルは既存手法に比べて多様な質問を生成できている．また，提案手法の B1/ME/RL の Recall が SemQG のビーム幅 10 の時のスコア 48.59/24.86/46.66 と同等であることから，各入力につき少なくとも 1 つはベースラインに匹敵するスコアの質問を提案モデルが生成できていることが分かる．C_q が 0 の時に SB4 が 100 に近いということは各入力に対してほとんど同じ質問ばかり生成しており，posterior collapse 問題が起きているが，C_q を適切な値に設定することでそれを回避できている．

3.4 合成データセットの構築

以上の結果から，複数の設定で構築したモデルで生成すればより多様なデータが得られると考えられる．そこで $(C_a, C_q) = (5,5), (5,20), (20,20)$ の 3 つの設定でモデルを構築し，訓練データ中の各文章からそれぞれランダムに 50 の質問回答ペアを生成した．得られた合成データセットをそれぞれ $\mathcal{D}_{5,5}$, $\mathcal{D}_{5,20}$, $\mathcal{D}_{20,20}$ とおく．人手評価（付録 B）によるとこれらは質の低いデータを含むため，回答が長すぎるもの，質問が短すぎるか長すぎるもの，疑問詞を含まないものを省き，質問文中の n-gram の繰り返しを削除した．以下では基本的にこれら 3 つのデータセットを混合したものをデータ拡張に用いる．

3.5 質問応答

質問回答ペア生成の既存研究に従って BERT-base [15] を質問応答モデルとして使う．ハイパーパラメータは元論文に準拠する．データ拡張を行う際は合成データセットで 1 エポック学習した後にオリジナルの訓練データで 2 エポック学習する．§3.2, 3.3 と同様 HarQG と SemQG をベースラインに用いる．

3.5.1 半教師あり質問応答

質問回答ペア生成によるデータ拡張（半教師あり質問応答）の評価を行う．表 4 に半教師あり質問応答の結果を示す．提案手法は SemQG に匹敵する精度向上を達成した．BERT-base よりパラメータ数の多い BERT-large についても同様の実験を行ったが，いずれの手法でも精度向上は見られなかった．

提案した各合成データセットが精度にどれくらい寄与しているかについて分析するために検証データでアブレーションを行った．その結果，3 つのうちどの合成データセットを除いても EM / F1 スコアが 0.3 / 0.1~0.3 ポイント 減少した．詳しい結果は付録 A に載せる．よっていずれの合成データセットも精度向上に寄与していることが分かる．

3.5.2 分布外データセットへの汎化

既存研究において質問応答モデルは分布外データセットに対してうまく汎化できないことがわかって

	Generalization to OOD QA Datasets			Robustness to Challenge Test Sets				
Data	NewsQA	TriviaQA	NQ	non-Adv	Adv	Easy	Hard	Implications
SQuAD-Du	32.81/49.21	37.40/47.57	55.35/67.70	78.15/85.73	42.86/50.16	82.49/90.13	67.43/75.59	49.43/64.72
+HarQG	32.85/48.46	36.42/45.84	54.97/66.20	76.65/85.15	**51.79**/56.52	82.00/89.67	66.04/73.01	49.24/63.47
+SemQG	**33.86/50.51**	37.56/47.50	**58.19**/69.81	78.91/86.21	46.43/51.82	83.50/**91.05**	67.73/75.02	49.72/65.08
+Ours	32.81/49.25	**38.19/47.72**	58.02/**70.06**	**79.00/86.73**	**51.79/59.00**	**83.87**/90.94	**68.75/76.10**	**50.63/66.26**
Target	42.61/62.90	55.80/61.66	74.19/83.03	-	-	-	-	-

表 3　質問応答モデルの汎化性能と頑健性の評価

	Dev		Test	
Data	EM	F1	EM	F1
SQuAD-Du	80.12	87.85	72.69	84.08
+HarQG	79.49	87.05	72.32	83.31
+SemQG	81.02	88.53	**73.59**	**84.72**
+Ours	**81.49**	**88.61**	73.11	84.53

表 4　半教師あり学習

いる [7, 16]．ここでは提案手法が分布外データセットへの汎化性能を向上するかを評価する．評価用データセットには NewsQA [2]，TriviaQA [3]，Natural Questions (NQ) [17] を用いる．NQ については Sen ら [16] に倣って長い回答から短い回答を抽出するタスクに再定式化した．参考としてそれぞれのテストデータと同じ分布の訓練データ（Target）で質問応答モデルを訓練した時の精度も報告する．

表 3 の左に結果を示す．提案手法は TriviaQA では最も高い精度を達成し，NQ では SemQG に匹敵する精度となった．NewsQA ではほとんど向上は見られなかった．Wikipedia 以外のニュース記事等の文章を活用することが必要だと考えられる．

3.5.3　困難なテストデータへの頑健性

3 種類の困難なテストデータを用いて質問応答モデルの頑健性を評価する．

- 質問のパラフレーズ [8]: SQuAD の質問のパラフレーズ（**non-Adv**）と本来の回答と似ている回答の近くの単語を使って敵対的に作成したパラフレーズ（**Adv**）の 2 つのテストデータ
- 難しい質問 [9]: SQuAD の検証データを単純なヒューリスティクスで分割して得られる簡単な質問（**Easy**）と難しい質問（**Hard**）のサブセット
- 一貫性を要する質問 [10]: SQuAD の質問回答ペアを含意する質問回答ペアからなるテストセット（**Implications**）

結果　表 3 の右に結果を示す．non-Adv, Adv, Hard, Implications のすべての困難なテストデータにおいて提案手法がもっとも頑健性の向上に寄与することがわかった．また表 4 で SemQG と提案手法は同等の精度であったが，表 3 によると SemQG は Easy のみで精度を向上させている一方で，提案手法は Easy と Hard 双方で精度を向上させる効果があることが分かる．SemQG はより SQuAD に近い質問を生成することを目的としている [6] ことから，合成データセットの質の向上よりも多様性の向上の方がより頑健性の向上に寄与する可能性が示唆される．

4　おわりに

本研究では多様な質問回答ペアを生成可能なモデルを提案し，合成データセットを構築した．これを用いた半教師あり学習において既存手法と匹敵する精度を達成した．そして提案手法は分布外データセットへの汎化性能だけでなく困難なテストデータへの頑健性の向上に寄与することがわかった．質問応答のための質問回答ペア生成の研究 [5, 6, 18, 19, 20] の中ではこのような結果を報告している研究は本研究が初であり，質問回答ペアの多様性の向上がもたらす効果について新しい知見が得られた．さらに提案したデータ拡張手法は多様性の向上のみを目的としており，既存手法 [10, 8] とは異なりターゲット分布を意識したデータ拡張手法ではないことも注目すべきである．実応用においてモデルが答えることが困難な質問を訓練時に事前に知っておくことは困難なため，訓練時にターゲット分布が未知なままそれらへの汎化性能を向上させることは重要である．BERT-large のようなパラメータ数の多いモデルへのデータ拡張手法の考案は今後の課題としたい．

参考文献

[1]Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics, 2016.

[2]Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200. Association for Computational Linguistics, 2017.

[3]Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551, 2017.

[4]Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. Semi-supervised qa with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1040–1050. Association for Computational Linguistics, 2017.

[5]Xinya Du and Claire Cardie. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917. Association for Computational Linguistics, 2018.

[6]Shiyue Zhang and Mohit Bansal. Addressing semantic drift in question generation for semi-supervised question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2495–2509, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7]Alon Talmor and Jonathan Berant. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence, Italy, July 2019. Association for Computational Linguistics.

[8]Wee Chung Gan and Hwee Tou Ng. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019.

[9]Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10]Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy, July 2019. Association for Computational Linguistics.

[11]Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics.

[12]Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv e-prints*, page arXiv:1804.03599, Apr 2018.

[13]Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics, 2017.

[14]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)*, pages 1342–1352. Association for Computational Linguistics, 2017.

[15]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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[16]Priyanka Sen and Amir Saffari. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online, November 2020. Association for Computational Linguistics.

[17]Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

[18]Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA Corpora Generation with Roundtrip Consistency. *arXiv e-prints*, page arXiv:1906.05416, Jun 2019.

[19]Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. Asking questions the human way: Scalable question-answer generation from text corpus. In *Proceedings of The Web Conference 2020*, WWW ' 20, page 2032–2043, New York, NY, USA, 2020. Association for Computing Machinery.

[20]Dong Bok Lee, Seanie Lee, Woo Tae Jeong, Donghwan Kim, and Sung Ju Hwang. Generating diverse and consistent QA pairs from contexts with information-maximizing hierarchical conditional VAEs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 208–224, Online, July 2020. Association for Computational Linguistics.

A　アブレーション

表 5 にアブレーション結果を示す．

Data	EM	F1
SQuAD-Du+Ours	81.49	88.61
$-\mathcal{D}_{20,20}$	81.14	88.52
$-\mathcal{D}_{5,5}$	81.04	88.39
$-\mathcal{D}_{5,20}$	81.00	88.48

表 5　検証データでのアブレーション

B　人手評価

表 6 に人手評価の結果を示す．ベースラインに SemQG を用いた．Amazon Mechanical Turk (AMT) を用いて SQuAD は 100，その他は 200 ずつ質問回答ペアを評価した．項目は質問が文法的におかしくなくて意味がわかるかどうか，質問が文章と関連しているかどうか，回答が質問に対する回答として正しいかどうか [19]，回答が文章のメイントピックに関連しているかどうかの 4 つである．

Experiments		SemQG	$\mathcal{D}_{5,5}$	$\mathcal{D}_{20,20}$	SQuAD
Question is well-formed	No	2.9%	23.1%	27.8%	2.3%
	Understandable	34.5%	16.0%	17.0%	10.5%
	Yes	62.6%	60.9%	55.1%	87.2%
Question is relevant	No	2.5%	9.5%	11.5%	4.0%
	Yes	97.5%	90.5%	88.5%	96.0%
Answer is correct	No	2.8%	28.8%	30.5%	7.5%
	Partially	21.8%	28.1%	26.6%	11.8%
	Yes	75.4%	43.2%	42.9%	80.6%
Answer is important	No	1.5%	10.0%	5.0%	6.0%
	Yes	98.5%	90.0%	95.0%	94.0%

表 6　質問回答ペアの人手評価