答えのないことを答えるMachine Reading Comprehension

中西 真央　　小林 哲則　　林 良彦
早稲田大学理工学術院
nakanishi@.pcl.cs.waseda.ac.jp

1 まえがき

Machine Reading Comprehension (MRC) はマシンの言語理解能力を開発・評価するためのタスクであり，一般に文章に対する質問応答によって評価される．

現在の MRC の問題点の一つに，質問の答えが与えられる対象の文章中に存在することを前提としていることが挙げられる．マシンの言語理解能力を人間に近づけるためには，答えのない質問についてもそれを識別できることが必要であるが，この機能・能力の研究開発には，答えのないことが簡単には判別できないような質問を含むデータセットが必要になる．しかし，このような要件を満たすデータセットの作成には莫大な時間やコストを要する．

そこで本研究では，既存のデータセットから自動的に答えのない質問データを作成する簡潔な手法を提案する．さらに，答えのない質問に対して，その質問が文章中に答えを持たないことを識別することの難易度を自動的に付与する方法を提案する．これにより，作成するデータセットが一定の難易度を持つように制御できることを示す．

2 背景

2.1 言語理解の測定ツールとしての MRC

マシンに言語理解能力を与えることは，自然言語処理や人工知能における中心的な課題である．MRC はマシンの言語理解能力を開発・評価するためのタスクであり，定められた文章に関して設定された質問に正しく解答できるかにより，マシンによる理解度を測る．

2.2 答えのない質問を含む MRC の意義

現在 MRC のデータセットが数多く公開されている．それらのほとんどは，文章に対する質問が文章を読んで答えられることを前提としている．このため多くの MRC システムは，学習された手掛かりに基づいて最善と考えられる答えを解答する．しかしながら本来的な言語理解能力を考えるならば，マシンは文章内に記述のない内容に関する質問に対し「解答できない」と答えるべきである．以上の問題意識から，本研究では答えのない質問を含む MRC タスク，および，そのためのデータセット作成方法を提案する．

3 MRC のタスク分類

3.1 タスクとデータセット

MRC において，どのような言語理解能力に着目し，それをどのように計測するかによって，様々なタイプのタスク (解答選択，空欄補充，文章からの抜き出し，数の数え上げなど) が考えられる．また，タスクのタイプに即して，データのソースや文章の形式・長さが異なるデータセットが作成されてきた．

例えば，代表的な空欄補充のタスクとして，CNN/Daily Mail [1] や Children Book Test (CBT) [2] がある．CNN/Daily Mail における空欄はすべて固有名詞であるのに対し，CBT では解答に要求される能力・機能に応じて，固有名詞，一般名詞，動詞，前置詞といった多様な品詞の単語が空欄となる．空欄を補充する単語は 10 の選択肢のなかから最も適切なものを選択するが，与えられる文章は 20 文という一定の長さを持つ．

一方，本研究で当面の対象としている SQuAD [3] は，ウィキペディアの 1 パラグラフからなる文章を読み，質問に対する解答を文章中から抜き出して解答するタスクである．

言語理解能力だけでなく，より高度な推論を必要とするような質問を含むデータセットとして bAbI [4] がある．これは 20 の異なるタイプの質問からなるデータセットであり，yes/no で解答する問題から，物の数を数える問題，空間位置関係を推論する問題などを含む．

3.2 対象とするタスクの種類

本研究では，Stanford Question Answering Dataset (SQuAD) と呼ばれるデータセットを当面の対象とする．このデータセットは文章[1]，質問，答えを要素として持ち，解答は与えられた文章中のある連続区間に制約されている．SQuAD における一つの文章は英語の Wikipedia の一パラグラフであり，パラグラフにつき最大 5 つの QA がアメリカ，カナダのクラウドワーカーによって作成された．ここで，97%のワーカーの同意が得られない QA は低品質な QA として排除されている．

[1] Passage

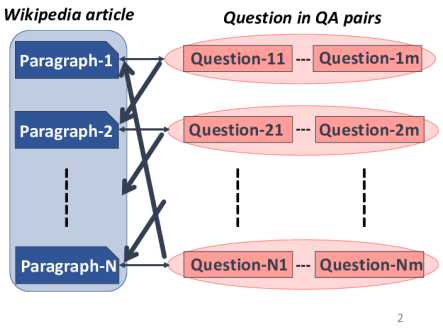

図 1: 作成方法

SQuAD を用いた MRC における評価指標には Exact Mach (EM) と F1 の 2 つがある．EM は予測解答が答えと寸分違わず一致することを要件とする評価指標であるのに対し，F1 は適合率と再現率から求めるものであり，予測した解答と正解の答えが部分的に一致している場合も考慮される．

4 答えのない質問を含むデータセットの作成

4.1 答えのない質問の作成方針

対象の文章とは全く関連のない質問を組み合わせることにより，答えのない質問は容易に作成することはできるが，これでは言語理解能力の開発・測定という目的にそぐわない．すなわち，答えのない質問を含むデータセットを作成する際には，識別の難易度を高く保つことが最も重要である．

本研究では，答えがない質問とは文章中に解答の手がかりが存在しない質問であるとする．このとき，ある文章に対する質問を適度に関連する他の文章に対する質問と入れ替えることで答えのない質問を作成する手法を提案する．すなわち，文章間の類似度を算出し，類似度が高い文章間の質問を入れ替えることで答えのない質問を作成する．

4.2 SQuAD を用いたデータセットの作成

上記の方針に基づき，図 1 に示す方法により，SQuAD データセットに対して文章と文章からでは答えられない質問及び答えの組を作成した．SQuAD における文章は Wikipedia の一パラグラフであり，ある主題に関する内容を述べているものと想定できる．そこで，あるパラグラフに対する質問を同じ記事の次のパラグラフに対する質問とした．これにより，適度の関連性を保った文章と質問の組が作成できる．ただし，作成された質問のうち答えが文章中に出現しているものは取り除いた．図 2 に提案する手法で作成した文章と答えのない質問例を示す．

Passage:
Like other digital music players iPods can serve as external data storage devices. Storage capacity varies by model ranging from 2 GB for the iPod Shuffle to 128 GB for the iPod Touch (previously 160 GB for the iPod Classic which is now discontinued).

Question:
Which company produces the iPod?

図 2: 作成した文章と答えのない質問例

4.3 答えのない質問に対する難易度の付与

このような簡潔な手法により答えのない質問を作成するが，作成されたそれぞれの質問に答えがないことを識別する難易度は一定ではない．答えのない質問の難易度を推定することができれば，データセット中の質問を一定の難易度に保つような制御が可能となり，データセットの利用価値が高まる．

ここで問題となるのは，答えのない質問の難易度を定義する基準であるが，本研究では自動的に実行できる機械的手段による識別の正解率をこの基準とする．これにより，客観的な基準に基づき，さらに，人手による作業コストが不要な難易度付与が可能となる．

具体的には，以下の 9 種類の特徴量を利用する識別器を構成し，各特徴量の答えの有無の識別率に対する寄与度を調べることにより，難易度付与において有効な特徴を調査した．

- IF 文章と質問それぞれ単体の特徴量
  - IF-1. 文章全単語 単語ベクトル平均ベクトル g_d
  - IF-2. 質問全単語 単語ベクトル平均ベクトル g_q
  - IF-3. 文章全単語 重み付き単語ベクトル平均ベクトル gt_d
  - IF-4. 質問全単語 重み付き単語ベクトル平均ベクトル gt_q
- SF 文章と質問の組の類似度に関する特徴量
  - SF-1. 単語ベクトルコサイン類似度最大値 $max(w_{cos})$
  - SF-2. 単語ベクトルコサイン類似度平均値 $E(w_{cos})$
  - SF-3. BLEU スコア
  - SF-4. TF-IDF ベクトルコサイン類似度 t_{cos}
  - SF-5. 重み付き単語ベクトルコサイン類似度 gt_{cos}

ここで，単語ベクトルは GloVe[5] から取得し，100 次元とする．SF-1，SF-2 における単語ベクトルコサイン類似度は，質問の各単語に対して与えられる．ある文章 d に対する質問 q の m 番目の単語の単語ベクトルを $\boldsymbol{v}_{\boldsymbol{m}}^{\boldsymbol{q}}$ とするとき，$\boldsymbol{v}_{\boldsymbol{m}}^{\boldsymbol{q}}$ に対する単語ベクトルコ

表 1: 難易度分類基準決定のための識別実験結果

特徴量		RF	LR	SVM	AdaBoost
特徴量	IF	0.607	0.563	0.565	0.554
	SF	0.847	0.852	0.851	0.852

表 2: SF 特徴量 寄与率の調査結果

抜いた特徴量	$max(w_{cos})$	$E(w_{cos})$	BLEU	t_{cos}	gt_{cos}
正解率	0.851	0.812	0.830	0.852	0.851

サイン類似度 $w_{cos,m}$ を文章の単語数を N とし以下のように定義する．

$$w_{cos,m} = max\{cos(\boldsymbol{v_m^q}, \boldsymbol{v_n^d}) | \ n \in \mathbb{N}, 0 \leq n \leq N\}$$

Random Forest (RF)，Logistic Regression (LR)，SVM，AdaBoost により識別器を構成し，識別の正解率 (accuracy) を調査した結果を表 1 に示す．実験においては，作成したデータセットのうち答えのある質問 45,000，答えのない質問 45,000 を使用し，10 分割交差検定を行った．

この結果から，文章と質問の組の類似度に関する特徴量 (SF) が識別に有効であることがわかったので，SF の各特徴量の寄与度を ablation test により調べた．ここで識別器には AdaBoost を使用した．その結果を表 2 に示す．SF-2. $E(w_{cos})$ を抜いた場合に最も識別率が悪くなることから，これが最も識別に寄与していると考え，この特徴量を難易度分類の基準とする．

4.4 作成したデータセット

上記の基準により，答えのない質問を $0.5 \leq E(w_{cos}) \leq 1.0$ の範囲に限定し，さらに $E(w_{cos})$ の値により，以下のような難易度分類を付与した．

- LEVEL1（易）: $0.5 \leq E(w_{cos}) < 0.6$
- LEVEL2　　: $0.6 \leq E(w_{cos}) < 0.8$
- LEVEL3（難）: $0.8 \leq E(w_{cos} \leq 1.0$

作成した答えのない質問・答え及び文章の組と，元来の SQuAD の文章及び質問・答えの組を答えのある質問として新しいデータセットとした．表 3 に作成されたデータセットの諸元を示す．

表 3: SQuAD 作成した答えのない質問データ数

		範囲	train	dev
答えあり質問			87599	10570
答えなし質問	ALL	$0.5 \leq E(w_{cos}) \leq 1.0$	79522	9302
	LEVEL1	$0.5 \leq E(w_{cos}) < 0.6$	7914	862
	LEVEL2	$0.6 \leq E(w_{cos}) < 0.8$	31526	3641
	LEVEL3	$0.8 \leq E(w_{cos}) \leq 1.0$	40082	4799

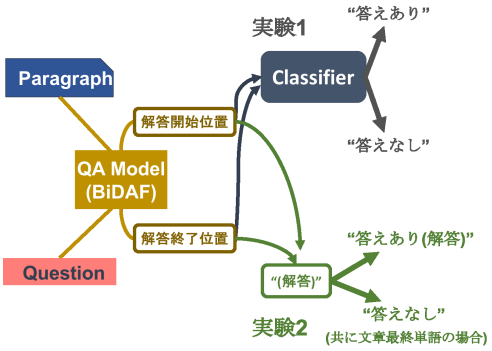

図 3: 識別実験 実験 1，2

5 答えの有無の識別実験

5.1 識別実験の設定

作成したデータセットに対し，Bi-Directional Attention Flow (BiDAF) [6] を既存解答予測モデルとして用い，2 つの識別実験 (識別実験 1，2) を行った．これらの実験の構成を図 3 に示す．

実験 1: 既存モデルを用いて解答を予測する際に得られる答えの候補及び各候補の確信度を利用する．BiDAF を含む多くの既存解答予測モデルは，抜き出す解答の開始・終了単語を各々確信度と共に予測する．そこで，この 2 つの確信度を特徴量として識別器へ入力する．開始単語の確信度を yp_{start}，終了単語の確信度を yp_{end} とする．識別器には Random Forest，Logistic Regression，SVM，AdaBoost を用いた．作成したデータセットのうち答えのある質問 39,120，答えのない質問 39,120 を使用し，10 分割交差検定を行った．

実験 2: 陽に答えのない質問クラスを追加する．このためには，答えがないことを開始・終了位置を用いて表現する必要があるが，予備実験の結果から，開始・終了単語の双方が文章の最終単語であると表現することとした．作成したデータセットのうち，学習に答えのある質問とない質問をそれぞれ 79,522，テストに答えのある質問とない質問をそれぞれ 9,302 使用した．

5.2 実験結果

実験 1 (表 4): 全識別器において 2 つの特徴量 yp_{start}, yp_{end} を使用した場合の識別率が高く，特に Random Forest による識別結果が 81.4%で最高正解率であった．

実験 2 (表 5): この場合の解答の有無の識別結果は 89.4%であった．また答えのある質問に注目した場合，予測した答えの EM スコアは 62.1%であった．SQuAD について答えのないクラスを追加せずに実験した場合の EM スコアは 68.0%であるので，大きく減少した．

表 4: 識別実験 1 実験結果

識別器		RF	LR	SVM	AdaBoost
特徴量	yp_{start}	0.808	0.768	0.761	0.809
	yp_{end}	0.788	0.757	0.752	0.791
	yp_{start}, yp_{end}	0.814	0.778	0.772	0.810

答えがあるのに関わらず答えがないと識別したのは，答えのある質問全体の 5.93%であった．一方，答えがあると識別しつつもその答えを誤ったのは，答えのある質問全体の 32.0%であった．答えのないクラス追加前に答えを誤った質問が全体の 32.0%であることから，増加は確認できなかった．従って，答えのある質問に対して答えがないと誤識別したことが EM スコアが大きく減少した原因である．

以上の実験結果から，解答候補の始点・終点を適切に表現することにより，既存の解答モデル (BiDAF) によっても答えのないことを一定の精度で識別できるが，答えのある質問に対する精度に盛況を与えることから，既存モデルを用いた文章に対する質問の答えの有無識別には改善の余地があると言える．

表 5: 識別実験 2 実験結果

	全体	答えあり質問	答えなし質問
2 値分類 正解率	0.894	0.941	0.849
答えを含めた正解率	-	0.621	-

5.3 難易度分類の妥当性検証実験

4.3 節の手法により付与した難易度が妥当であるかの検証を行った．各難易度の答えのない質問に対し，識別実験 1,2 における学習済みモデルを使用した識別実験を行い，どの程度答えのない質問を正しく同定できたかという再現率を求めた．ここで，実験 1 では 4 つの識別器のうち AdaBoost の学習済みモデルを使用し，yp_{start}, yp_{end} 両特徴量を使用した．

この結果を表 6 に示す．表 6 から難易度が向上するに従って答えのない質問の再現率がすべて低下していることがわかる．このことから，提案する難易度付与の手法が一定の妥当性を持つものと判断できる．

表 6: 難易度分類の検証実験結果

	実験 1	実験 2
LEVEL1	0.865	0.960
LEVEL2	0.855	0.946
LEVEL3	0.762	0.720

6 おわりに

マシンによる言語理解能力の開発・測定のための重要な手段である Machine Reading Comprehension に，答えのない質問を含めるべきであることを主張し，非常に簡潔な手法により答えのない質問を含むデータセットを作成する手法を提案した．さらに，自動的に難易度を付与する手法を提案した．この手法では人手を全く用いずに答えのない質問を作成できるためコストや時間の削減に繋がる．

提案手法により作成されたデータセットに対し，既存の解答予測モデルを使用した識別実験を行った．その結果，難易度が上昇するに従い答えの有無の識別率が低下することを確認した．これは，作成したデータセットの妥当性を示す．また，難易度の最も高いデータセット (LEVEL3) に対する答えの有無の識別の正解率は，72.0%と低く，これは，本論文で作成したデータセットの解答有無識別について，既存手法には改善の余地があることを意味している．よって，今回作成した難易度の高いデータは，MRC のベンチマークとしての有用であると考えられる．

今後の課題としては，より識別の難しい答えのない質問データの作成やその識別が挙げられる．識別の難しい答えのない質問の作成方法として，質問中の固有名詞を文中に登場する他の固有名詞と入れ替えたり，質問中の形容詞や副詞を対義語と置き換えたりすることで答えのない質問を作成する方法などを検討している．

参考文献

[1] K. M. Hermann, T. Koisk, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, “Teaching machines to Read and Comprehend,” In Advances in Neural Information Processing Systems (NIPS 2015). F.

[2] F. Hill, A. Bordes, S. Chopra, and J. Weston, “The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations,” In International Conference on Learning Representations (ICLR 2016), pp.1–13, 2016.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” In Empirical Methods in Natural Language Processing (EMNLP 2016), no. ii, pp. 2383–2392, 2016.

[4] J. Weston, A. Bordes, S.Chopra, and T. Mikolov, “Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks,” In The computing Research Repository (CoRR), abs/1502.05698, 2015.

[5] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation,” In Empirical Methods in Natural Language Processing (EMNLP 2014), pages 1532–1543, 2014.

[6] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “Bi-Directional Attention Flow for Machine Comprehension,”In International Conference on Learning Representations (ICLR 2017), pp. 1–12, 2017.