価値判断を伴うファクトイド質問応答

関 直哉 †‡ 水野 淳太 ‡ 呉 鍾勲 ‡ Julien Kloetzer‡ 飯田 龍 †‡ 鳥澤 健太郎 †‡
† 奈良先端科学技術大学院大学 先端科学技術研究科
‡ 国立研究開発法人 情報通信研究機構（NICT)
seki.naoya.si8@is.naist.jp
{junta-m, rovellia, julien, ryu.iida, torisawa}@nict.go.jp

1 はじめに

ユーザはしばしば「糖尿病患者は何を食べるといい？」のように、何かをうまくやる，もしくはまずいことを避けるための質問を発する．このような質問のうち，名詞句を回答とするものを，本研究では「価値判断を伴うファクトイド質問」（Opinion Factoid Question Answering，以下 OFQA）と呼ぶ．従来の質問応答システムでは，価値判断を伴うファクトイド質問には対応できない．その結果，例えば質問「糖尿病患者は何を食べるといい？」に対し，「ケーキ」といった不適切な回答が出力されてしまう．

本研究では，日本語の OFQA データセットを作成し，これと，価値判断を考慮していないファクトイド QA のデータセットや一見質問応答とは関連の薄い因果関係認識認識に関するデータセットを組み合わせて fine-tuning した，BERT ベースの質問応答手法を開発した．この精度はセンチメント分析等，複数の異なるタスク用の分類器/抽出器を組み合わせたパイプライン方式よりも精度が高かった．これらのパイプライン方式の開発では，価値判断の根拠（例えば，「糖尿病患者がケーキをたべてはまずい」の「まずい」）を表す入力中の表現を特定するため追加的なアノテーション等も行ったが，それらの追加アノテーションや各種モジュールを使用しなくても，BERT がいわゆる end-to-end でより高い精度を達成できたことは興味深い．

2 データセットの構築

本研究で構築した OFQA データセットは，質問・回答候補・回答候補が抽出された文（以下，元文）の 3 つ組と，質問に対し回答候補が適切か否かの二値ラベルで構成される．本研究では，関らのファクトイド QA（以降，FQA と呼ぶ．価値判断は考慮されていない）データセット [1] の訓練事例から抽出した 7,887 事例（抽出条件の詳細は付録を参照）と，大規模情報分析システム WISDOM X[2][3] で新規に収集したファクトイド質問応答の出力からデータセットを作成する．WISDOM X は，ユーザから与えられた質問からバイナリパタンもしくはユナリパタンと呼ばれるパタンを抽出し，それらパタンを用いて Web 文書 40 億件から回答候補を検索する．表 1 に検索結果の例を示す．

表 1 バイナリパタンに合致して得られた検索結果の例

質問	糖尿病患者は何を食べる？
回答候補	ケーキ
元文	糖尿病患者がケーキを食べてましたが，まずいと思います．
パタン	A が B を食べる A=糖尿病患者

バイナリパタンとは，係り受けで繋がった 2 つの変数化された名詞をもつ「A は B を引き起こす」のようなパタンを指す [4]．ユナリパタンは 1 つの変数化された名詞をもつ「A を引き起こす」のような形式で表されるパタンを指す．WISDOM X では，Web40 億文書に形態素解析など処理を行い，バイナリ（ユナリ）パタンを抽出し，検索用インデックスを構築する．検索の際は，与えられた質問からパタンとパタンの変数に当てはまる名詞（例：A が B を食べる, A=糖尿病患者）を抽出し，検索用インデックスに照合して回答候補を検索する．また，柔軟な照合を行うために，抽出されたパタンと含意関係にあるパタンも利用して回答候補を検索する [4]．こうしたパタンを用いることで単純に名詞等でキーワード検索するよりも，より少数の，正解の可能性が高い回答候補を絞り込むことができる．

アノテーション作業では，表 2 に示したような質問，回答候補，元文の 3 つ組を対象に，質問に対して回答候補が適切か否かを判定する．ただし，質問

は「といい？」といったような表現で表される価値判断を伴う質問となっているため，その価値判断も含めて判定を行う．例えば，この表の質問に対しては，表 1 にある元文から抽出された回答候補「ケーキ」は「糖尿病患者が食べるといいもの」ではないので，正解とはみなされない．

表 2　質問・回答候補．元文の 3 つ組の例

質問	糖尿病患者は何を食べるといい？
回答候補	たまねぎ
元文	糖尿病患者が**たまねぎを**を食べるのは，血糖値を下げるのに効果的です．

また，上記のアノテーション作業時，元文中に回答候補の価値判断の根拠（以降，根拠表現と呼ぶ）があれば，その根拠表現を特定する作業も同時に行なった．表 2 の例では，「血糖値を下げるのに効果的です」が根拠表現となる．抽出した根拠表現は後述する比較手法で用いる．

質問文については，Web40 億文書から質問文を疑問代名詞の有無など簡単なルールで抽出した後，質問一般性判定によるフィルタリングを行なった．この判定では，例えば「IoT で何が変わるのか？」のように，質問文が単独で知識，意見を求めている質問として成立していれば質問一般性が成り立つと判断する．一方，「この英単語の意味わかりますか？」のように，文脈がないと理解できない質問については質問一般性が成り立たないと判断する．本研究では，NICT が開発した BERT ベースの質問一般性判定モデルを用いて質問一般性を判定する．判定の結果，約 161 万件のファクトイド質問が獲得できた．

このようにして得られたファクトイド質問について，UniLM[5] を用いた質問変換器で WISDOM X で検索しやすい形式の質問に言い換えた質問を作成する．例えば，Web から抽出した質問が「Q1. 糖尿病患者って何を食べるんですかねえ？」だとすると，「糖尿病患者は何を食べる？」のように言い換えた質問を作成する．ついで，言い換え後の質問に「といい？」または「とまずい？」のいずれか，価値判断を表す表現をランダムに選んで付加し，価値判断を伴う質問を作成する．なお，「といい？」といった表現を付加すると不自然な質問となる場合があるため，最終的に人手で質問が意味的に自然か否かを判定した．約 161 万件のファクトイド質問からサンプリングした 10,000 件の質問に対して上記判定を行い，7,643 件の質問を得た．

ついで，WISDOM X は価値判断を伴う質問に対応していないため，「といい？」あるいは「とまずい？」を付加する前段階の質問について，WISDOM X で回答候補と元文を検索する．質問あたりの回答候補の検索上限数は 10 件とし，回答が重複しないようにした．この結果，7,643 質問に対し 29,659 件の回答候補と元文の組が収集できた．

これらの回答候補には，価値判断を考慮しなくても正解とは言えない回答候補も含まれる．そうした回答候補を大量にアノテーションするのは効率が悪いため，関らの FQA 分類器 [1] を用い，価値判断を考慮しない場合に正解とみなされる事例のみをアノテーションすることにした．さらに，元文に根拠表現が含まれる可能性の高い事例を抽出するため，元文の末尾の表現に着目してフィルタリングを行った（詳細は付録参照）．最終的に，5,273 件の質問，回答候補，元文の組が得られた．一方，評価時にこのフィルタリングの影響を軽減するため，本研究では，1,774 件の質問・回答・元文の組をフィルタリング抜きで作成し，テストデータに加えた．

これらのデータに加えて，関らの FQA データセット [1] の訓練事例の正例から抽出した 7,887 事例を加えて，価値判断も考慮した正例／負例のアノテーション，さらには前述した価値判断の根拠の特定のアノテーションを行った．回答候補の適切さ及び根拠表現の特定のアノテーションは一つのインスタンスに対して，３名のアノテータが行い，正例／負例のラベルは多数決で決定した．異なるアノテータによるラベルの一致度を示すカッパ値はテストデータに追加したフィルタリングしていないデータに関してが 0.53，それ以外が 0.57 であり，良好なアノテーションであると考える．ついで，質問が重複しないようにこれらのデータを以下の表 3 にあるように訓練，開発，テストの３種に分割した．

表 3　OFQA データセットの統計値

	データ件数	正例件数（割合）
訓練	8,222	1,068(13.0%)
開発	1,648	221(13.4%)
テスト	5,064	558(11.0%)

3　手法

本研究では，提案手法，自明なベースライン手法の他に比較手法として複数のモジュールを組み合わせたパイプライン方式も二つテストした．以下では，まず，それらのパイプライン方式を説明する．なお，提案手法も含めて，以下で述べるベース

ライン，比較手法で使われている各モジュールは，すべて全て 22 億文からなる因果関係文を含む Web コーパス（サイズは約 353GB）[6] で事前学習した $\text{BERT}_{\text{LARGE}}$ をベースとして用いている．

3.1 比較手法

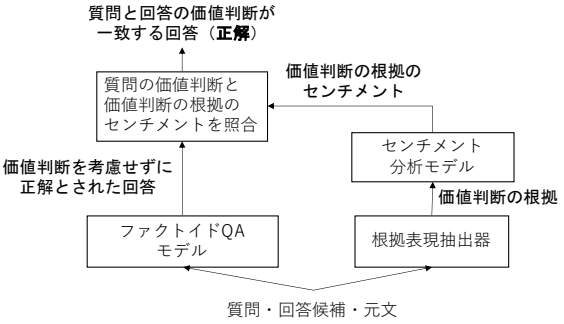

図 1 根拠表現を用いるパイプライン方式

パイプライン方式の一つ目は，ファクトイド QA モデル [1]，根拠表現抽出器，センチメント分析を組み合わせたパイプライン方式である（図 1）．この手法では，ファクトイド QA モデルで正解とされた回答に関して，根拠表現抽出器を用いて価値判断の根拠を特定し，その根拠のセンチメントが質問の価値判断と一致している場合（つまり，「といい？」の質問に対しては positive なセンチメント，「とまずい？」の質問に対しては negative なセンチメント）のみ，回答を正解とする．

ファクトイド QA モデルは BERT を関らの FQA のデータセットと OFQA データセットの二つでマルチタスクで fine-tuning したものを用いる．根拠表現抽出器には BERT のトークン分類モデル（つまり，各トークンの出力に二値の softmax を接続し，そのトークンが根拠表現にふくまれるかどうか判定するモデル）を用い，前述したように，データセットに付与された価値判断の根拠表現で fine-tuning した．センチメント分析器も同様に BERT を用いた二値分類器である．（詳細は付録）．

もう一つのパイプライン手法は，ファクトイド QA，WISDOM X のどうなる型 QA[7][8][6] とセンチメント分析器のパイプライン方式である（図 2）．WISDOM X のどうなる型 QA は「A とどうなる？」というタイプの質問に A の因果的帰結で回答するサービスである．例えば，質問「糖尿病患者は何を食べるといい？」と回答候補「ケーキ」から「糖尿病患者はケーキを食べるとどうなる？」というどうなる型質問を作成し検索すると，「救急車で運ばれる」といった因果的な帰結句が得られるが，これらの帰結句は回答候補に対する価値判断の根拠とも考えられる．この手法では，こうした見方に基づき，帰結句にセンチメント分析を適用した結果と質問の価値判断が一致した回答で，ファクトイド QA が正例とみなした回答だけを正解とする．ファクトイド QA とセンチメント分析器は前述したものを再利用した．

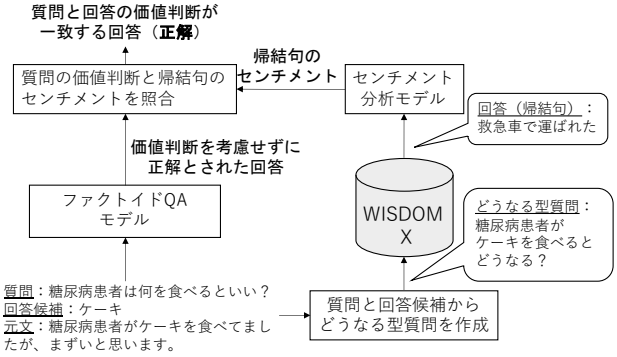

図 2 どうなる型 QA を用いるパイプライン方式

3.2 提案手法

本節では，提案手法として，BERT の一段階目のファインチューニングを OFQA と異なるタスク，二段階目を OFQA で学習する二段階 fine-tuning を提案する．より具体的には，関らの FQA データセット [1] と，Hashimoto らの因果関係認識データセット [7] を合わせたデータで一段階目の fine-tuning を行い，FQA と因果関係認識に対する性能の平均が最も性能の高いモデルを二段階目の fine-tuning に用いた．価値判断を伴わない FQA と OFQA では回答が正解となる基準は異なるが，どちらもファクトイド質問を対象とした質問応答であり，必要とする知識も共通する部分があると考えられる．また，前述したように，元文中の回答候補にまつわる表現の因果的帰結 (「糖尿病患者がケーキを食べる」に対して「救急車で運ばれる」) は価値判断と密接に関係があると考えることもできる．提案手法では，これらの仮説に基づき FQA と因果関係認識のデータセットで一段階目の fine-tuning を行い，それらのデータセットが含む知識を OFQA で使うことを狙う．

4 実験

提案手法及びベースライン，パイプライン方式の比較手法で用いる FQA モデルの fine-tuning 時には表 6（付録を参照）のパラメータを探索した．前

述したパイプライン方式の比較手法の他に，ベースライン手法として前述した $\text{BERT}_{\text{LARGE}}$ を (1) 関らの FQA データセット，(2) OFQA データセット，(3)FQA データセットと OFQA をマージしたデータ（FQA+OFQA）の 3 通りのデータセットで fine-tuning をしたモデル，また，(4) 提案手法で用いた因果関係，FQA，OFQA を段階を踏んで fine-tuning をするのではなく，一気にマルチタスクで学習したモデルの 4 種類を評価した．

表 4 に OFQA のテストデータに対する実験結果を示す．なお，実験時間が足りなかったため，パイプライン（どうなる型 QA）は OFQA のテストデータの一部（1,774 件）に対する実験結果である．提案手法 (テストデータ一部）はそれとの比較のため，提案手法を同じテストデータで評価したものとなる．また，提案手法 (-因果関係) は，提案手法の 1 段回目の fine-tuning 時に因果関係のデータセットを使わなかったモデルである．

表 4　OFQA のテストデータで評価した結果

手法	Recall	Precision	F1
パイプライン（根拠表現）	58.1	68.6	62.9
パイプライン（どうなる型 QA）	45.2	74.0	56.2
(1) ベースライン (FQA)	60.2	23.0	33.3
(2) ベースライン (OFQA)	60.5	63.9	62.2
(3) ベースライン (FQA+OFQA)	63.4	65.8	64.6
(4) マルチタスク	65.2	61.5	63.3
提案手法	68.8	66.4	**67.6**
提案手法 (テストデータ一部)	61.9	66.7	64.2
提案手法 (-因果関係)	64.5	66.1	65.3

提案手法の精度はベースラインも含め比較手法のいずれをも上回った．特に，提案手法が (4) マルチタスク及び提案手法 (-因果関係）の性能を上回ったことはそれぞれ 2 段階 fine-tuning，因果関係データセットの有効性を示している．特に，因果関係という一見 OFQA，ファクトイド QA とかけ離れたデータセットがポジティブな影響を持っていることは興味深い．また，パイプライン手法は二つとも提案手法よりも高い精度を達成することはできなかった．これはこれらのパイプライン手法が根拠表現やセンチメント，さらには因果関係に関係する別のデータで学習した各種モジュールを使っていることを考えると興味深い．つまり，この結果は，OFQA データセットだけから BERT が学習した価値判断の方が，少なくとも我々の設定では，それらの追加の学習データで別途学習した内容よりも高品質であることを意味しており，BERT の end-to-end 学習の強力さを示すものであると考えられる．

また，FQA のみで学習したベースライン (1) は精度が極めて低く，OFQA+FQA の組み合わせ (3) がベースラインの中では一番精度が高い．なお，OFQA で高い精度を達成するためには，そのベースとなっている価値判断の判定と，そもそも価値判断を考慮せずに FQA の正解を正しく認識することの両者が必要であると考えることができる．ここで，OFQA のデータセットはあくまで価値判断の判定を適切に学習することに貢献しているのであって，価値判断を考慮しない FQA の精度向上によって間接的に OFQA タスクの精度向上を果たしているわけではないことを確認するため，追加で以下の実験を行った．この実験では，（価値判断を考慮しない）FQA のテストデータでベースラインの精度を評価した．（表 5）．

表 5　FQA のテストデータで評価した結果

手法	Recall	Precision	F1
(1) ベースライン (FQA)	79.4	78.9	**79.1**
(2) ベースライン (OFQA)	6.7	37.4	11.4
(3) ベースライン (FQA+OFQA)	78.6	78.4	78.5

これによると，OFQA だけで学習したベースライン (2) も，FQA と OFQA の両者で学習したベースライン (3) も，FQA だけで学習したベースライン (1) の性能を下回った．これは，OFQA のデータセットは価値判断を考慮しない FQA の性能向上には貢献しないことを示している．このことから，FQA を学習データに追加することで学習されているのは確かに価値判断であることが分かる．

5　おわりに

本研究では「価値判断を伴うファクトイド質問」を新たなタスクとして提案し，データセットを作成の上，そのタスクのための質問応答モデルを開発した．因果関係認識や FQA といった別タスクの学習データを利用することで性能が向上することを示した．モデルは BERT をいわば end-to-end で学習したものであり，別途アノテーションした価値判断の根拠やセンチメント分析を利用したパイプライン方式よりも精度が高かった．これはいわば，このタスクにおいてはパイプライン方式を考えた人間の開発者を BERT が上回ってしまったと考えることもできるかもしれない．

参考文献

[1] 関直哉, 水野淳太, 門脇一真, 飯田龍, 鳥澤健太郎. ファクトイド質問応答における BERT の pre-trained モデルの影響の分析. 言語処理学会 第 26 回年次大会, 2020.

[2] Masahiro Tanaka, Stijn De Saeger, Kiyonori Ohtake, Chikara Hashimoto, Makoto Hijiya, Hideaki Fujii, and Kentaro Torisawa. WISDOM2013: A large-scale web information analysis system. In *The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations*, pp. 45–48, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[3] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP systems for analyzing textual big data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 263–267, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[4] Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Chikara Hashimoto, Jong-Hoon Oh, Motoki Sano, and Kiyonori Ohtake. Two-stage method for large-scale acquisition of contradiction pattern pairs using entailment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 693–703, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[5] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, December 2019.

[6] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. 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)*, pp. 5816–5822, Hong Kong, China, November 2019. Association for Computational Linguistics.

[7] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 987–997, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[8] Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks, 2017.

[9] 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*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[10] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

付録

A　データ収集

Web から抽出した質問には不要な記号（「★」,「↑」,「＊」,「Q1.」等）や質問の内容に大きく影響しない単語列（「ところで」,「そもそも」,「技術における企業の最責任者, ＣＴＯの仕事とは？」の下線部等）が含まれ，そのまま WISDOM X に入力すると十分な数の回答候補が得られない可能性がある．そこで，Web から抽出した質問を対象に，不要な部分を取り除いて簡単な形式に言い換えた質問を生成する．言い換えた質問の生成には NICT が開発した UniLM[5] ベースの質問変換器を用いた．

末尾表現は，WISDOM X のファクトイド質問応答で用いられるバイナリ（ユナリ）パタン [4] が合致した箇所以降の単語列を指す．表 1 の例では「が，まずいと思います.」が末尾表現となる．回答候補に対する価値判断の根拠は，この末尾表現に出現しやすいと考えられるため，2 節のデータ収集では，価値判断の根拠が含まれる可能性の低い末尾表現が 4 文字未満の事例は取り除いた．

本研究では，関らの FQA データセットの訓練事例 174,765 事例から，(1) 回答候補がユナリパタン [4] で抽出されている，(2) 回答候補が正解である，(3) 元文の末尾表現が 4 文字以上存在する，(4) 質問一般性が成り立つ，の条件を全て満たす事例を抽出する．(4) について，FQA データセットの質問の一部は WISDOM X によりルールベースで生成されており，文法的，意味的に不自然な質問が含まれる場合があるため，質問一般性判定によりフィルタリングを行う．(1) から (4) の条件で抽出した結果，7,887 件の質問，回答候補，元文の組が得られた．

次に，質問に「といい？」もしくは「とまずい？」を付加して価値判断を伴う質問を作成する．「といい？」「とまずい？」のどちらを付加するかは，組となる元文の末尾表現のセンチメントで決定し，センチメントがポジティブであれば「といい？」を，ネガティブあれば「とまずい？」を質問に付加する．センチメントの特定には BERT ベースのセンチメント二値分類器を用いた（C を参照）．最後に，価値判断を伴う質問，回答候補，元文を組としアノテーション作業を行なった．

B　fine-tuning のパラメータ探索

本研究では，表 6 のパラメータの組み合わせに対し個別に fine-tuning し，開発データで F1 スコアが最も高いモデルを選択した．

表 6　fine-tuning のパラメータ探索

パラメータ	探索範囲
学習率	{1,2,3}e-5，{1,3,5,7,9}e-6
エポック数	1,2,3

C　センチメント分析

本研究では，センチメント分析を事象を表す文が与えられた際，その事象がポジティブなのかネガティブなのかを分類する二値分類のタスクとする．センチメント分析モデルは，NICT が開発した $\text{BERT}_{\text{LARGE}}$ を用いたセンチメント二値分類器を用いる．データセットは NICT がクロールした Web40 億文書から抽出された事象と事象に対する感情表現のペアから作成されたセンチメント分析データセットを用いる（表 7）．上記データセットで学習したモデルをテストデータで評価すると非常に高い精度でセンチメントを分類できていることが分かる（表 8）．

表 7　センチメント分析データセットの統計値

	データ件数	正例件数（割合）
訓練	34,330,398	19,172,347(55.8%)
開発	25,000	13,998(56.0%)
テスト	25,000	13,935(55.7%)

表 8　センチメント分析モデルの性能

Recall	Precision	F1	Ave.P
85.1	85.7	85.4	93.4

D　根拠表現抽出

本研究では，(1) 最長一致 (longest) と完全一致 (exact)，(2) minor，(3) ofqa-neg の条件を組み合わせて根拠箇所を選択し，根拠表現抽出データセットを作成した．

最長一致は，アノテータが抽出した根拠箇所が一部重なる場合，重複しない部分も含めた最長の単語列を根拠表現とする．例えば，根拠表現として「ビタミン C が豊富に含まれ，美容にも良い」,「美容にも良いし，風邪予防にもなる.」という単語列が抽出された場合，「ビタミン C が豊富に含まれ，美容にも良いし，風邪予防にもなる.」を根拠表現とする．完全一致は，アノテータが抽出した根拠箇所のうち，重複する単語列（「美容にも良い」）を選択する．minor は，(1) で根拠箇所が決まらなかった場合に，(1) で最終的に洗濯されなかった根拠表現，アノテータ 3 人のうち 1 人のみが抽出している根拠表現からランダムに 1 つを選択する．ofqa-neg は，回答候補が適切な事例に対し抽出された根拠表現に加え，回答候補が不適切な事例に対し抽出された根拠表現も利用する．

本研究では，予備実験において最も性能が高かった exact+minor+ofqa-neg の組み合わせを用いる．表 9 に根拠表現抽出データセットの統計値を示す．

根拠表現抽出に用いるトークン分類モデルの入力は質問・回答候補・元文の 3 つ組とし，「⟨ 質問の単語列 ⟩[SEP]⟨ 回答候補の単語列 ⟩[SEP]⟨ 元文の単語列 ⟩」の形式で入力する．評価指標は SQuAD[9][10] の評価に用いられる Exact Match(EM)，F1 スコアを用いる．パイプライン手法で用いた根拠表現抽出器を根拠表現抽出データのテストデータで評価すると，EM で 76.6，F1 スコアで 80.8 の性能が得られた．

表 9　根拠表現データセットの統計値

	データ件数	根拠表現が存在する事例数（割合）
訓練	8,222	2,402(29.2%)
開発	1,648	491(29.8%)
テスト	5,064	1,366(27.0%)