複数の質問形式を利用した分類型の質問応答による薬物タンパク質間関係抽出

山田 晃士　三輪 誠　佐々木 裕
豊田工業大学
{sd22439,makoto-miwa,yutaka.sasaki}@toyota-ti.ac.jp

概要

近年，質問応答による関係抽出手法が提案され，高い性能を達成している．しかし，質問文は人手で作った固定のものが使われており，質問文の表現が抽出結果に与える影響は明らかでない．また，一般分野以外のタスクへの有効性は未知数である．加えて，質問応答モデルは関係抽出を考慮して設計されていない．本研究では，質問応答を用いた関係抽出手法の薬物タンパク質間関係抽出への適用を目指し，有効な質問形式の調査と，関係抽出に合わせた質問応答モデルの提案を行う．DrugProt データセットを用いて評価を行い，質問文の形式の関係抽出の性能への影響と薬物タンパク質間関係抽出における質問応答の有効性を確認した．

1　はじめに

近年，関係抽出に対して質問応答を用いる手法 [1, 2, 3] が高い性能を示している．質問応答を用いた関係抽出手法では，まず，候補となるすべての関係ラベルに対して質問文テンプレートを作成し，そこに関係抽出の対象とする用語ペアの片方を当てはめることで質問文を作成する．作成した質問文と用語ペアを含む元の文で質問応答を行い，回答がテンプレートに当てはめていない用語であった場合にテンプレートが持つ関係が存在するとして関係抽出を行う．質問応答を利用した関係抽出手法は，質問文の表現を調整することで 1 つの関係を複数の視点から抽出することができるという利点がある．

一方で，従来手法は人手で作成した固定したテンプレートを利用しており，質問文の表現の違いが関係抽出の結果にどのような影響を与えるのかは明らかにされていない．また，質問応答を用いた関係抽出手法は一般ドメインのデータセットを対象としており，薬学分野などの専門性の高いドメインに対する有効性は検証されていない．加えて，質問応答モデルが関係抽出のために設計されていないという問題点もある．

そこで，本研究では，質問応答による関係抽出手法の薬物タンパク質への適用を目的として，薬物タンパク質間関係抽出のデータセットである DrugProt データセット [4] に対して 2 種類の質問文テンプレートのセットを作成し，質問文の形式の違いが関係抽出の結果に与える影響について調査を行う．また，質問応答を用いた関係抽出において，関係の有無を決定づけるテンプレートに含まれていない用語が回答かどうかを直接二値分類により判定する，関係抽出に特化した分類型の質問応答モデルを提案する．本研究の貢献は以下の通りである．

- 薬物タンパク質間関係抽出に対して質問応答を用いた手法を適用し，有効性を確認した．
- 質問文の形式が関係抽出の性能へ影響を与えることを確認した．
- 関係抽出への利用に特化した分類型の質問応答モデルを提案した．

2　関連研究

2.1　双方向の質問応答による関係抽出

Cohen らは，関係の始点と終点の用語をそれぞれ含む 2 つの質問文を用いて，双方向の質問応答を行うことで関係抽出を行う手法 [1] を提案している．まず，候補となるそれぞれの関係ラベルに対して，関係の始点の用語を当てはめて終点を回答する質問文のテンプレートと，終点の用語を当てはめて始点を回答するテンプレートの 2 種類を作成する．テンプレートから作成した 2 つの質問文を用いてそれぞれ質問応答を行い，片方でも回答が質問文に当てはめていない方の用語であった場合にテンプレートが持つ関係があると判定する．これを全ての関係ラベ

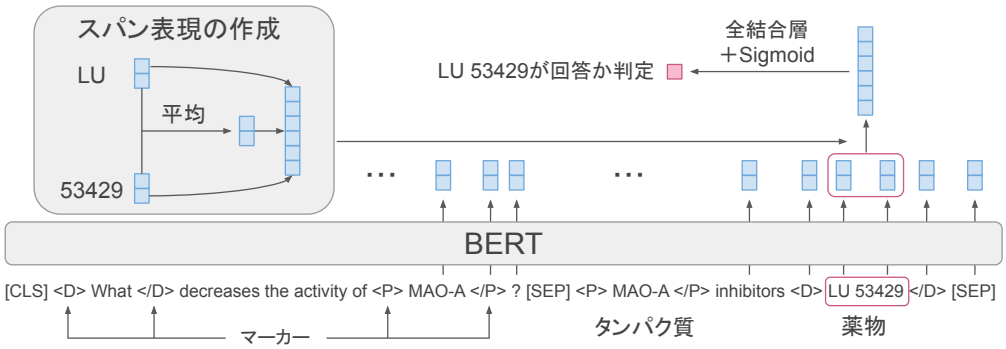

図 1　2 値分類による質問応答モデル

ルに対して繰り返すことで関係抽出を行っている．

2.2　関係抽出に適した質問応答

著者らは，質問応答を用いた関係抽出を行う際に，関係抽出タスクで与えられる情報を利用して質問応答を行う手法 [5] を提案した．入力文と質問文に含まれる用語の前後と，質問文の回答に対応する疑問詞の前後にデータセットで与えられる用語の情報をマーカーとして挿入することで，関係抽出タスクで与えられる情報を利用した質問応答を行った．マーカーを用いて挿入する情報として，データセットで与えられる関係の項のタイプを利用することで関係抽出性能が向上することを報告した．

3　提案手法

本研究では，質問応答を用いた関係抽出を薬物タンパク質間関係抽出に適用する際の有効な質問形式の調査と，関係抽出に適した質問応答モデルの提案を目的とする．3.1 節では質問応答を用いた関係抽出を薬物タンパク質間に適用し，質問文の表現による関係抽出の性能への影響について調査する際に必要な質問テンプレートの作成について，3.2 節では関係抽出に適した質問応答モデルについてそれぞれ説明する．

3.1　質問文テンプレートの作成

質問応答を用いた関係抽出手法を薬物タンパク質間関係抽出に適用するため，本研究で学習・評価に用いる DrugProt データセット [4] に含まれるすべての関係ラベルに対して質問文テンプレートを作成した．2.1 節で説明した双方向の質問応答による関係抽出を行うため，それぞれの関係ラベルに対して薬物を当てはめてタンパク質を答える質問とタンパク質を当てはめて薬物を答える質問の 2 種類を作成した．また，質問文の形式が関係抽出の性能に影響を与えるかどうかを調査するため，2 つ目の質問文テンプレートセットとして，質問文の形を統一したものも作成した．13 種類ある関係ラベルに対して双方向の質問応答をするため，1 つの質問文テンプレートセットは 26 個の質問文テンプレートを持つ．質問文テンプレートはいずれもデータセットのアノテーションガイドライン [6] を参考に作成した．作成した 2 つの質問文テンプレートセットは付録 A に示した．

3.2　分類型の質問応答モデル

従来の質問応答モデルをそのまま用いる関係抽出では質問の回答と用語の一致を確認するが，関係抽出では質問の回答が関係抽出の対象の用語であるかどうかを判別できればよい．そこで，本研究では，図 1 に示すような質問文に含まれない用語のスパン表現からその用語が質問の回答であるかどうかを判別する 2 値分類モデルを利用した，分類型の質問応答モデルを提案する．

まず，BERT (Bidirectional Encoder Representations from Transformers) [7] エンコーダを利用して質問文と入力文の各トークンの表現ベクトルを得る．ここで，質問文と入力文を入力する際は，2.2 節で説明した手法を用いて，関係抽出を行う対象の薬物タンパク質の用語ペアの前後と回答に対応する疑問詞の前後に用語のタイプをマーカーとして挿入する．BERT の SEP トークンを利用して連結した質問文と入力文 $S = \{w_1, w_2, \cdots\}$ を BERT エンコーダに入力し，各トークンの表現 $\mathbf{h}_i$ を得る．

$$\{\mathbf{h}_1, \mathbf{h}_2 \cdots\} = \mathrm{BERT}(w_1, w_2, \cdots) \tag{1}$$

得られた表現ベクトルのうち，回答候補の用語のス

表 1　2 種類の質問文セットと比較手法の F 値 (%)

関係	TS1	TS2	マーカー	文分類	学習事例数	開発事例数
INDIRECT-DOWNREGULATOR	**76.4**	73.4	74.1	75.6	1,330	332
INDIRECT-UPREGULATOR	75.5	75.1	**79.2**	74.7	1,379	302
DIRECT-REGULATOR	**69.2**	66.5	67.8	67.5	2,250	458
ACTIVATOR	73.2	**74.5**	73.3	72.4	1,429	246
INHIBITOR	**86.0**	85.3	84.7	84.5	5,329	1,152
AGONIST	77.1	77.0	**77.3**	75.3	659	131
AGONIST-ACTIVATOR	33.3	**53.3**	28.6	0	29	10
AGONIST-INHIBITOR	**100**	66.7	57.1	0	13	2
ANTAGONIST	**91.8**	91.7	88.5	88.4	972	218
PRODUCT-OF	61.5	62.2	**67.1**	64.4	921	158
SUBSTRATE	65.5	**68.8**	65.9	65.9	2,003	495
SUBSTRATE_PRODUCT-OF	0	0	0	0	25	3
PART-OF	73.5	72.2	71.5	**74.1**	886	258
マイクロ平均	**76.9**	76.5	76.2	75.9	—	—

パンに含まれるトークンの表現から，スパンの表現ベクトルを作成する．ここで，s と e はそれぞれ用語のスパンの開始位置と終了位置を，$\mathbf{h}_{span}$ は用語のスパンの表現ベクトルを表す．

$$\mathbf{h}_m = \text{mean}(\mathbf{h}_s, \cdots, \mathbf{h}_e) \tag{2}$$

$$\mathbf{h}_{span} = \text{Concat}(\mathbf{h}_s, \mathbf{h}_m, \mathbf{h}_e) \tag{3}$$

作成したスパンの表現ベクトルに全結合層と Sigmoid 関数を適用することで，その用語が質問の回答である確率を得る．

$$p_{span} = \text{Sigmoid}(W_{span}\mathbf{h}_{span} + \mathbf{b}_{span}) \tag{4}$$

得られた確率が閾値より大きい時，その用語が回答であるとし，双方向の質問応答に対して片方でも回答が対象の用語であれば，テンプレートが持つ関係があると判定する．これをすべての関係に対して繰り返すことで関係抽出を行う．また，複数の関係に対して回答が用語である確率が閾値より大きい時，その用語ペアに複数の関係ラベルがあるとして，マルチラベルな予測を行う．

4　実験設定

4.1　データセット

薬物タンパク質間関係抽出のデータセットとして，DrugProt データセット [4] を用いて学習および開発データでの評価を行った．このデータセットは薬物及びタンパク質を含む薬学文献のアブストラクトで構成され，薬物・タンパク質間には 13 種類の関係が設定されている．また，関係を持たない薬物とタンパク質のペアも存在する．評価指標にはマイクロ F 値を用いる．

4.2　実験環境

実装にはプログラミング言語 Python 3.7.11 を用いた．また，深層学習ライブラリとして PyTorch 1.10.0 [8] を，事前学習モデル利用のため Transformers 4.18.0 [9] を使用した．計算機には CPU に Intel(R) Xeon(R) W-3225 及び Intel(R) Xeon(R) CPU E5-2698 v4 を，GPU に NVIDIA RTX A6000 及び NVIDIA Tesla V100-DGXS-32GB を用いた．

4.3　比較手法

作成した 2 種類のテンプレートセット（**テンプレートセット 1**（**TS1**），**テンプレートセット 2**（**TS2**））を用いて，それぞれモデルの学習を行い F 値の比較を行った．また，質問文の代わりに関係固有のマーカーと用語を利用した場合（**マーカー**）との比較を行った．これは，図 1 を例とすると，“What decreases the activity of MAO-A?” という質問文の代わりに，調べたい関係ラベル INHIBITOR 固有のマーカーを用いた “<INHIBTOR> MAO-A?” を使用して学習・評価を行ったものである．加えて，質問応答を用いたモデルの有効性を調査するため，BERT の CLS トークンを用いて文分類を行ったモデル（**文分類**）をベースラインとして比

表 2　テンプレートセットによって関係抽出の予測結果が異なる事例（文中の太字が薬物，下線がタンパク質）

正解ラベル	PART-OF
テンプレート セット 1	What has a structural relationship to DRUG? What is structurally related to PROTEIN?
テンプレート セット 2	What includes DRUG? What is included in PROTEIN?
TS1 のみが正解	To identify key **amino acids** involved in factor IX activation,
TS2 のみが正解	about 60 **amino acids** forms a discrete domain, which is unique among the LeuRSs

較を行う．いずれの手法も，事前学習モデルとして大規模な生物医学文献を用いて事前学習された PubMedBERT-base-uncased-abstract-fulltext [10] を用いた．また，最適化手法として Adam [11] を用いており，学習率は 3e-6 に設定した．また，3.2 節で説明した質問応答モデルにおける閾値は 0.7 とした．

5　結果と考察

5.1　関係抽出性能の比較

作成した質問文テンプレートを用いた場合と比較手法それぞれについて，F 値を比較した結果を表 1 に示す．TS1 と 2 は作成した 2 種類のテンプレートセット 1 と 2 を用いた際の結果，マーカーは質問文の代わりに関係固有のマーカーを用いた結果，文分類は BERT の CLS トークンを用いて文分類として関係抽出を行った結果である．

ベースラインである文分類による関係抽出に対して，質問応答形式で関係抽出を行ったテンプレートセット 1，2，マーカーの 3 つと比較すると，訓練事例数の少ない関係クラスのうち，AGONIST-ACTIVATOR と AGONSIT-INHIBITOR において，予測性能の向上が見られた．また，マイクロ平均についても最大で 1.0%ポイントの向上が見られた．このことから，質問応答を利用した関係抽出手法は，薬物タンパク質間関係抽出においても有効であることがわかった．

次に，テンプレートセット 1，2 とマーカーの質問応答を用いた 3 つの手法で比較すると，各関係クラスで F 値が変化していることがわかるため，質問文の表現が関係抽出性能に影響していることがわかる．また，INDIRECT-UPREGULATOR と PRODUCT-OF の 2 つのクラスでは質問文の代わりにマーカーを用いた手法が，テンプレートセットを用いた他 2 つの手法と比較して，高い F 値が見られた．

5.2　解析

質問文の表現の違いが関係抽出の結果に影響を与えることを抽出結果が変化した事例で確認する．表 2 は正解ラベルを予測するための 2 種類の質問文テンプレートセットに対して，片方のテンプレートセットを使用した場合にのみが正しく関係ラベルを予測できている事例を示したものである．テンプレートセットはそれぞれ上が薬物を当てはめる質問，下がタンパク質を当てはめる質問になっており，DRUG と PROTEIN を文中の薬物とタンパク質の用語に置き換えて質問文を作成する．

片方のテンプレートセットのみが正解した事例はどちらも，正解したテンプレートセットでは 2 つの質問の両方で正しく答えられている一方で，間違っている方のテンプレートセットの 2 つ質問の両方に対して間違った予測をしている．このことから，同じ関係クラスを予測するための質問文であっても表現を変えることで大きく抽出結果が変わってしまうことあることがわかる．

6　おわりに

本研究では，質問応答を用いた関係抽出手法の薬物タンパク質間関係抽出への適用を目的に，DrugProt データセットに対して 2 種類の質問文テンプレートのセットの作成による有効な質問形式の調査と，質問の回答が用語であるかを判定する 2 値分類モデルによる関係抽出に特化した分類型の質問応答の提案を行った．DrugProt データセットにおける評価から，質問文の形式が関係抽出の性能に影響すること，薬物タンパク質間関係抽出においても質問応答を用いた手法が有効であることを示した．

今後は質問応答を用いた関係抽出手法に対して有効な質問形式や組み合わせの調査を行い，1 つの関係クラスを複数の視点から抽出可能な手法の実現を目指す．

謝辞

本研究は JSPS 科研費 JP20K11962 の助成を受けたものです.

参考文献

[1] Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. Relation classification as two-way span-prediction. **arXiv preprint arXiv:2010.04829v2**, 2021.

[2] Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. Entity-relation extraction as multi-turn question answering. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1340–1350, Florence, Italy, July 2019. Association for Computational Linguistics.

[3] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. In **Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)**, pp. 333–342, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[4] Martin Krallinger, Obdulia Rabal, Antonio Miranda-Escalada, and Alfonso Valencia. DrugProt corpus: Biocreative VII Track 1 - Text mining drug and chemical-protein interactions, June 2021.

[5] 山田晃士, 三輪誠, 佐々木裕. 項の表現に着目した質問応答による関係分類. 言語処理学会第 28 回年次大会, 2022.

[6] Obdulia Rabal, Jose Antonio López, Astrid Lagreid, and Martin Krallinger. DrugProt corpus relation annotation guidelines [ChemProt - Biocreative VI], June 2021.

[7] 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, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, Vol. 32, pp. 8026–8037, 2019.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. **arXiv preprint arXiv:1910.03771**, 2019.

[10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. **ACM Transactions on Computing for Healthcare**, Vol. 3, No. 1, p. 1–23, Jan 2022.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

A　作成した質問文テンプレート

本研究で作成した 2 つの質問文テンプレートセットを表 3 と 4 にそれぞれ示す.

表 3　作成した質問文テンプレート 1（上が薬物を当てはめる質問，下がタンパク質を当てはめる質問）

関係	質問文
INDIRECT-DOWNREGULATOR	What does DRUG downregulate via other targets? What downregulates PROTEIN via other targets?
INDIRECT-UPREGULATOR	What does DRUG upregulate via other targets? What upregulates PROTEIN via other targets?
DIRECT-REGULATOR	What does DRUG directly regulate? What directly regulates PROTEIN?
ACTIVATOR	What does DRUG increase the activity of? What increases the activity of PROTEIN?
INHIBITOR	What does DRUG decrease the activity of? What decreases the activity of PROTEIN?
AGONIST	What does DRUG act as an agonist to? What acts as an agonist to PROTEIN?
AGONIST-ACTIVATOR	What does DRUG increase activity by acting as an agonist to? What acts as an agonist on PROTEIN to increase its activity?
AGONIST-INHIBITOR	What does DRUG decrease activity by acting as an agonist to? What acts as an agonist on PROTEIN to decrease its activity?
ANTAGONIST	What does DRUG act as an antagonist to? What acts as an antagonist to PROTEIN?
PRODUCT-OF	What produces DRUG by the enzymatic reaction? What is the product of the enzymatic reaction of PROTEIN?
SUBSTRATE	What does DRUG act as a substrate for? What acts as a substrate for PROTEIN?
SUBSTRATE_PRODUCT-OF	What causes the enzymatic reaction with DRUG as substrate and product? What is the substrate and product of the enzymatic reaction of PROTEIN?
PART-OF	What has a structural relationship to DRUG? What is structurally related to PROTEIN?

表 4　作成した質問文テンプレート 2（上が薬物を当てはめる質問，下がタンパク質を当てはめる質問）

関係	質問文
INDIRECT-DOWNREGULATOR	What is indirectly downregulated by DRUG? What indirectly downregulates PROTEIN?
INDIRECT-UPREGULATOR	What is indirectly upregulated by DRUG? What indirectly upregulates PROTEIN?
DIRECT-REGULATOR	What is directly regulated by DRUG? What directly regulates PROTEIN?
ACTIVATOR	What is increased the activity by DRUG? What increases the activity of PROTEIN?
INHIBITOR	What is decreased the activity by DRUG? What decreases the activity of PROTEIN?
AGONIST	What is acted by DRUG as the agonist? What acts as the agonist of PROTEIN?
AGONIST-ACTIVATOR	What is acted and increased the activity by DRUG as the agonist? What acts as the agonist and increases the activity of PROTEIN?
AGONIST-INHIBITOR	What is acted and decreased the activity by DRUG as the agonist? What acts as agonist and decreases the activity of PROTEIN?
ANTAGONIST	What is acted by DRUG as the antagonist? What acts as the antagonist of PROTEIN?
PRODUCT-OF	What produces DRUG in the enzymatic reaction? What is produced by the enzymatic reaction of PROTEIN?
SUBSTRATE	What is acted by DRUG as the substrate? What acts as the substrate of PROTEIN?
SUBSTRATE_PRODUCT-OF	What is acted by DRUG as the substrate and produces it in the enzymatic reaction? What acts as the substrate and is produced in the enzymatic reaction of PROTEIN?
PART-OF	What includes DRUG? What is included in PROTEIN?