マルチタスク学習を用いた系列変換タスクの品質推定

喜友名 朝視顕　吉村 綾馬　金子 正弘　小町 守
東京都立大学
{kiyuna-tomoshige,yoshimura-ryoma,kaneko-masahiro}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1 はじめに

品質推定（Quality Estimation: QE）の目的は，人手による参照文を利用せず，タスクに応じて設計された評価項目に対して，システムの出力文を自動評価することである．QE の性能は，人手評価との相関係数で評価するのが一般的である．QE の手法として，人手評価値付きのデータセットを用いて人手評価値に最適化する教師あり手法があり，高い性能を示している [1, 2]．しかし一般に，教師あり手法で用いるデータセットの作成には多大なコストが必要であるため，表 1 に示すように，学習に利用できるデータセットのサイズが小さいという問題がある．そのため，モデルの性能は，ドメインやデータセットに対して一貫性がない [10][1)]．

本研究では，この問題に取り組むために，複数のタスクの QE のデータセットを用いるマルチタスク学習手法として，3 種類の方法を提案した：(1) 1 つのタスクのすべての評価項目のデータを用いる手法，(2) すべてのタスクのすべての評価項目のデータを用いる手法，(3) 複数のタスクの同じ評価項目のデータのみを用いる手法．

複数の系列変換タスク（文法誤り訂正，言い換え生成，テキスト平易化，スタイル変換）の QE で実験を行い，学習時のドメイン（in-domain）に加え，学習時と異なるドメイン（out-of-domain）でメタ評価を行った結果，多くのタスクと評価項目の組について，マルチタスク学習により性能が向上した．特に，訓練データのサイズが小さいテキスト平易化の文法性および意味保存性に関する QE において，in-domain と out-of-domain の両方で大幅に性能が向上した．一方，性能が最も良い手法は，評価項目により異なることがわかった．

1） ニュースやフォーラムといったトピックの違いだけでなく，出力したシステムの違いも含めてドメインと呼ぶ．

2 関連研究

2.1 系列変換タスクの QE

QE の教師あり手法で用いるデータセット（表 1）にはそれぞれ，いくつかの評価項目に関する人手評価値が付いている．その評価項目は，タスクに固有のものと，タスク間で共通しているものがある．Yamshchikov と Shibaev [6] は，言い換え生成とスタイル変換強度の 2 つのタスクに共通している意味保存性に関して，13 種類の自動評価尺度を調査した．

文法誤り訂正における人手評価値付きデータセットを用いた文法性の QE 手法は，Napoles ら [11] が提案している．Yoshimura ら [12] は，文法性だけでなく流暢性や意味保存性の 3 項目の人手評価付きデータセットを作成した．大量のラベルなしデータを使って事前学習された BERT [13] を用いて，各評価項目について，人手評価に最適化する教師あり学習を行った．

これまでの研究では単一タスクの単一の評価項目についての人手評価値のみを用いているのに対し，本研究では，複数の評価項目あるいは複数のタスクのデータセットを用いて，人手評価に最適化する教師あり手法を提案する．また，本研究でも事前学習モデルされた BERT を用いる．

機械翻訳タスクも系列変換タスクの一つである．機械翻訳タスクの QE の場合，原文とシステムの出力文を入力するが，これらは言語が異なる．本研究では，入力を単一言語（英語）に揃えるため，機械翻訳タスクの人手評価値付きのデータセットは除外した．

2.2 マルチタスク学習

Liu ら [14] は，事前学習モデルとマルチタスク学習を用いる，Multi-Task Deep Neural Network (MT-DNN) を提案し，自然言語理解タスクで実験を行い性能が

表 1　本実験で使用したデータセットの評価項目と文または文対の数．**太字**はタスク間で共通している評価項目を示す．

タスク	評価項目	in-domain				out-of-domain
		#Train	#Dev	#Test (in)	文/文対	#Test (out)
文法誤り訂正 [3, 4]	**文法性**	1,518	747	754	文	1,312 文 × 13 システム
言い換え生成 [5, 6]	**意味保存性**	5,749	1,500	1,379	文対	13,223 文対
テキスト平易化 [7, 8]	**文法性**，**意味保存性**，平易性	404	101	126	文対	272 文対
スタイル変換 [9]	**文法性**，**意味保存性**，スタイル変換強度	1,758	219	219	文対	—

向上することを示した．また，ドメイン適応の実験を行い，学習済みの MT-DNN は学習済みの BERT よりも少量のデータで高い性能に達することを示した．

自然言語理解タスクの GLUE ベンチマーク [15] は 9 つのタスクから成る．そのうち，3 つは文対が意味的に等価であるかを判定するタスクであり，4 つは自然言語推論タスクである．Liu らは GLUE ベンチマークについて，タスクの情報を考慮せずすべてのデータを用いて共通のモデルを作成した．一方，本研究では複数の系列変換タスクの QE について，すべてのデータを用いる手法の他に，タスクまたは評価項目が同じデータのみを用いる手法を提案する．共通点のあるデータを明示的に選択することで，タスク間または評価項目間に共通する特徴に着目しやすくなり精度が向上することが期待できる．

3　QE のためのマルチタスク学習

本研究では，QE のための，複数の評価項目あるいは複数のタスクのデータセットを用いるマルチタスク学習手法として，次の 3 つを提案する．

- 単一のタスクの，すべての評価項目のラベルを用いてマルチタスク学習を行う手法（single-task, multi-aspect: **mlt-aspect**）
- すべてのタスクの，単一の評価項目のラベルのみを用いてマルチタスク学習を行う手法（multi-task, single-aspect: **mlt-task**）
- すべてのタスクの，すべての評価項目のラベルを用いてマルチタスク学習を行う手法（multi-task, multi-aspect: **all**）

複数のデータセットを用いて学習を行うため，モデルには，マルチタスク学習を行う MT-DNN [14] を用いる．マルチタスク学習にはいくつかの手法があるが，MT-DNN は一般的に用いられている [16] ハードパラメータ共有 [17] による手法を採用している．ハードパラメータ共有の場合，モデルは図 1 に示すように，タスク間で共有される層とタスク固有の出力層に分けられる．一部の層を共有し，複数のタスクに対して同じパラメータを使うことで，タスク間で共通する特徴に着目することができるようになり，特定のタスクに過学習しにくくなる．

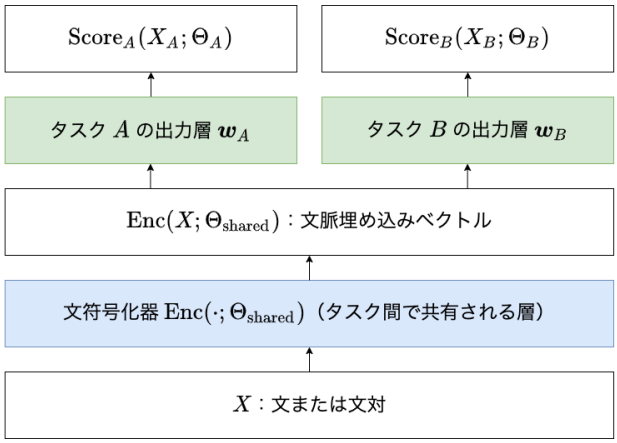

図 1　MT-DNN の概略図．

モデルの入力 X は，先頭に入力の先頭を表す特殊トークン，各文の末尾に文の区切りを表す特殊トークンを追加して，各文を連結した系列である．入力が 1 文の場合も同様である．

Θ を MT-DNN の学習パラメータ，$\Theta_{\mathrm{shared}} \subset \Theta$ をタスク間で共有される層の学習パラメータとする．タスク間で共有される層は文符号化器 $\mathrm{Enc}(\cdot;\Theta_{\mathrm{shared}})$ である．文符号化器 $\mathrm{Enc}(\cdot;\Theta_{\mathrm{shared}})$ は，入力 X に対し，文脈埋め込みベクトル $\mathrm{Enc}(X;\Theta_{\mathrm{shared}})$ として，入力の先頭を表す特殊トークンの文脈埋め込みベクトルを返す．タスク t の出力層は，タスク t の入力 X_t の文脈埋め込みベクトル $\mathrm{Enc}(X_t;\Theta_{\mathrm{shared}})$ に対し，評価値 $\mathrm{Score}_t(X_t;\Theta_t)$[2] として，タスク固有のパラメータベクトル $\boldsymbol{w}_t$ との内積を返す：

$$\mathrm{Score}_t(X_t;\Theta_t) = \boldsymbol{w}_t^{\mathrm{T}} \cdot \mathrm{Enc}(X_t;\Theta_{\mathrm{shared}}). \qquad (1)$$

ただし，Θ_t は $\Theta_{\mathrm{shared}} \cup \boldsymbol{w}_t$ である．

ミニバッチ確率的勾配法を用いて，単一タスク t のミニバッチ b_t ごとにモデルのパラメータ Θ_t を更

2）　簡単のため，$\mathrm{Score}_t(\mathrm{Enc}(X_t;\Theta_{\mathrm{shared}});\Theta_t)$ を $\mathrm{Score}_t(X_t;\Theta_t)$ と表記する．

新する．これにより，モデルは各タスクの目的関数の合計におおよそ最適化される．タスク t の目的関数 $L_t(\Theta_t)$ は，データセットに付いている人手評価値 y_t と，モデルの推定値 $\mathrm{Score}_t(X_t;\Theta_t)$ との二乗誤差である：

$$L_t(\Theta_t) = \frac{1}{|b_t|} \sum_{(X_t,y_t)\in b_t} (y_t - \mathrm{Score}_t(X_t;\Theta_t))^2. \quad (2)$$

あるタスクと評価項目の組を，MT-DNN における 1 つのタスクとみなし，QE のためのマルチタスク学習を行う．以下の 2 段階で学習を行い，文または文対から対応する評価項目の評価値を推定する回帰モデルを作成する．

1. 文符号化器 $\mathrm{Enc}(\cdot;\Theta_{\mathrm{shared}})$ を BERT [13] などの事前学習済みモデルで初期化し，目的関数 $\sum_{t\in T}\sum_{b_t\in B_t} L_t(\Theta_t)$ に従い，学習を行う．
2. 1 で作成したモデルに対して，目的関数 $\sum_{b_{t'}\in B_{t'}} L_{t'}(\Theta_{t'})$ に従い，再学習を行う．

ただし，T は学習に用いるタスクの集合，B_t はタスク t のミニバッチの集合，t' は目的のタスク（主タスク）である．

4 評価実験

4.1 実験設定

本実験では，4 種類の系列変換タスクの，人手評価値が付いているデータセットを用いて，マルチタスク学習による QE の有効性を検証する．各データセットの評価項目と，文または文対の数は表 1 のとおりである．また，評価項目が同じでも，タスクにより評価値の範囲が異なる．各データセットの詳細は付録 A を参照．モデルの入力は，文法誤り訂正のときは文，その他のタスクのときは文対とする．

MT-DNN は，著者らによって公開されている実装[3)]を用いた．モデルの詳細は付録 B を参照．

各手法の性能の評価（メタ評価）には，人手評価値とのピアソンの積率相関係数を用いる．文法誤り訂正の out-of-domain はシステム単位，その他は文単位で評価を行う．

4.2 比較手法

初期化済みの MT-DNN に対し，単一のタスク，単一の評価項目のラベルのみを用いて fine-tuning を行うベースライン手法（single-task, single-aspect: **sgl**）と提案手法を比較する．

3) https://github.com/namisan/mt-dnn

4.3 実験結果

表 2 に各評価項目ごとの実験結果を示す．言い換え生成の意味保存性とテキスト平易化の平易性を除くすべてのタスクと評価項目の組において，mlt-task または all が最も高い性能を示した．特に，訓練データのサイズが最も小さいテキスト平易化の文法性または意味保存性に関して，大幅な改善が見られた．このことは，複数の評価項目あるいは複数のタスクのデータセットを用いるマルチタスク学習が QE に有用であることを示唆しており，マルチタスク学習が低リソース問題の解決に役立つことがわかる．

大幅な改善が見られたテキスト平易化の out-of-domain における，人手評価値（human）と各手法による評価値を表 3 に示す．文法性と意味保存性については，人手評価との相関係数が高い手法ほど，モデルによる評価値と人手評価値との差が小さいことがわかる．文法性では，mlt-task が human と最も近く，意味保存性では，mlt-task や all が human と近い．平易性については，システムの出力をあまり正しく評価できていないことがわかる．

5 考察

マルチタスク学習の有効性が確認できたタスクと評価項目の組． テキスト平易化の意味保存性と，スタイル変換の意味保存性およびスタイル変換強度では，all が最も良く，各タスクの文法性は mlt-task が最も高い性能を示した．これらは，複数のドメインやデータセットを用いたことによる，学習に利用可能な言語表現の増加が直接，性能の改善に結びついた結果であると考えられる．

文法性に関しては，mlt-aspect は sgl より性能が高いことから，文法性以外の評価項目のデータを用いた場合のマルチタスク学習の有効性は示されている．しかし，all は mlt-task より性能が低い．これは，言い換え生成のデータセットの訓練データの量が最も多いことを考慮すると，all により意味保存性が支配的になったことが要因の 1 つであると考えられる．その場合，補助タスクのデータセットの量を調整したり，補助タスクの損失を小さくしたりする必要があるだろう．スタイル変換強度で all の性能が最も高いという実験結果は，スタイル変換強度も文法ではなく意味を扱う評価項目であることから，all は意味保存性が支配的であるという仮説と整合

表 2 各評価手法の各評価項目の評価値と人手評価との相関係数．開発データにおけるピアソンの積率相関係数の上位 10 個の平均値．

	文法誤り訂正–文法性			言い換え生成–意味保存性					
	Dev	Test (in)	Test (out)	Dev	Test (in)	Test (out)			
sgl	75.07	72.97	**96.74**	**90.27**	**85.87**	64.75			
mlt-task	**76.14**	**73.21**	95.88	90.08	85.54	64.19			
all	75.89	72.46	95.55	89.89	85.64	**64.76**			
テキスト平易化	文法性			意味保存性			平易性		
	Dev	Test (in)	Test (out)	Dev	Test (in)	Test (out)	Dev	Test (in)	Test (out)
sgl	65.28	41.25	74.63	50.66	44.59	70.26	**64.57**	**40.20**	65.94
mlt-aspect	70.98	44.59	77.71	55.69	46.28	71.90	63.07	36.81	68.17
mlt-task	**74.89**	**52.12**	**80.18**	60.18	46.68	76.57	—	—	—
all	72.99	48.08	78.23	**65.66**	**48.73**	**76.81**	61.10	35.61	**70.68**
スタイル変換	文法性			意味保存性			スタイル変換強度		
	Dev	Test (in)	Test (out)	Dev	Test (in)	Test (out)	Dev	Test (in)	Test (out)
sgl	78.18	71.76	—	53.06	54.11	—	69.18	69.33	—
mlt-aspect	78.21	**72.68**	—	52.15	50.69	—	68.25	67.47	—
mlt-task	**79.08**	**72.68**	—	54.22	**55.40**	—	—	—	—
all	77.59	72.40	—	**55.04**	51.32	—	**70.61**	**69.54**	—

表 3 テキスト平易化の out-of-domain における各評価項目の人手評価および各手法による評価値．赤字は原文とシステム出力の差分を示す．評価値は human / sgl / mlt-aspect / mlt-task / all の順である．

原文	The two Chinese surveillance vessels appeared on the scene on Tuesday, and blocked the Philippine warship from approaching the fishing boats.
システムの出力	The two Chinese surveillance vessels appeared on the scene on Tuesday. The two Chinese surveillance vessels blocked the Philippine warship.
文法性 3 / 3.16 / 3.09 / 3.08 / 2.85　意味保存性 2 / 2.73 / 2.40 / 1.98 / 2.60　平易性 2 / 3.07 / 2.83 / — / 2.89	
原文	Spain's government tried to plug a gaping hole in the country's banking system on Friday, but the fourth such attempt to tackle the fallout of a property crash fell short of expectations.
システムの出力	Spain 's government tried to plug. The fourth such attempt fell short.
文法性 2 / 1.93 / 1.84 / 1.95 / 1.76　意味保存性 1 / 1.90 / 1.52 / 1.42 / 1.38　平易性 1 / 3.10 / 2.84 / — / 3.14	

する．

マルチタスク学習の有効性が確認できなかったタスクと評価項目の組．　言い換え生成の意味保存性，テキスト平易化の平易性では，マルチタスク学習の有効性が示せなかった．言い換え生成の意味保存性については，主タスクの訓練データのサイズが大きく，本実験で用いた補助タスクが悪影響を及ぼした可能性がある．テキスト平易化の平易性については，平易性以外の評価項目の量を増やすと性能が悪化していることから，平易性以外の評価項目のデータを活用することは難しいと言える．

6　おわりに

本研究では，系列変換タスクの QE に取り組み，マルチタスク学習手法を 3 つ提案した．4 つのタスクのデータセットを用いた実験の結果，多くのタスクと評価項目の組について，マルチタスク学習により性能が向上した．特に，訓練データのサイズが小さいテキスト平易化の文法性および意味保存性に関する QE において，大幅に性能が向上した．一方で，3 つの手法はいずれも，言い換え生成の意味保存性とテキスト平易化の平易性では改善が見られなかった．また，性能が最も良い手法は，評価項目により異なることがわかった．

今後は，提案手法が全体的な改善に至らない要因を分析し，また，内積が負となる勾配を射影する手法 [18] や，入力の先頭にタスクと評価項目を区別する特殊トークンを追加し出力層を一つにする手法 [19] などを行いたいと考えている．

参考文献

[1] 嶋中宏希, 梶原智之, 小町守. 事前学習された文の分散表現を用いた機械翻訳の自動評価. 自然言語処理, Vol. 26, No. 3, pp. 613–634, 2019.

[2] Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of WMT*, pp. 101–105, 2019.

[3] Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. Predicting grammaticality on an ordinal scale. In *Proceedings of ACL*, pp. 174–180, 2014.

[4] Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. Human evaluation of grammatical error correction systems. In *Proceedings of EMNLP*, pp. 461–470, 2015.

[5] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of SemEval*, pp. 1–14, 2017.

[6] Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *arXiv:2004.05001*, 2020.

[7] Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. Shared task on quality assessment for text simplification. In *Proceeding of QATS*, pp. 22–31, 2016.

[8] Goran Glavaš and Sanja Štajner. Event-centered simplification of news stories. In *Proceedings of RANLP*, pp. 71–78, 2013.

[9] Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. Evaluating style transfer for text. In *Proceedings of NAACL*, pp. 495–504, 2019.

[10] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *TACL*, Vol. 8, pp. 539–555, 2020.

[11] Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *TACL*, Vol. 4, pp. 169–182, 2016.

[12] Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. SOME: Reference-less submetrics optimized for manual evaluations of grammatical error correction. In *Proceedings of COLING*, pp. 6516–6522, 2020.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pp. 4171–4186, 2019.

[14] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In *Proceedings of ACL*, pp. 4487–4496, 2019.

[15] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of BlackboxNLP*, pp. 353–355, 2018.

[16] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017.

[17] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of ICML*, p. 41–48, 1993.

[18] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv:2001.06782*, 2020.

[19] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *TACL*, Vol. 5, pp. 339–351, 2017.

[20] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of CoNLL*, pp. 1–14, 2014.

[21] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill™: A Bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pp. 569–576. 2007.

A　データセット

文法誤り訂正　in-domain のデータセットは，GUG (“Grammatical” versus “Un-Grammatical”) データセット [3] である．このデータセットには，学習者が書いた文に対して，文法性の人手評価値（1 から 4 の離散値）が付いている．

out-of-domain のデータセットは，CoNLL 2014 Shared Task [20] のデータセットおよび，それに対して Grundkiewicz et al. (2015) [4] の人手評価を用いた．このデータセットは，入力文 1,312 文と，それに対する 12 システムの訂正結果を含む．Grundkiewicz らは，人手で文ごとに評価した少量のデータを用いてレーティングアルゴリズムである TrueSkill [21] を用いて訂正システム単位の人手評価スコアを算出した．相関係数は，入力を含む 13 システムごとの評価値の平均値と，Grundkiewicz et al. (2015) の Table 3c の人手評価値を用いて計算した．

言い換え生成　in-domain のデータセットは，STS (Semantic Textual Similarity) ベンチマーク [5] である．このデータセットには，意味保存性の人手評価値（0 から 5 の連続値）が付いている．

out-of-domain のデータセットは，Yamshchikov and Shibaev (2020) [6] である．このデータセットには，意味保存性の人手評価値（1 から 5 の連続値）が付いている．

テキスト平易化　in-domain のデータセットは，QATS (Quality Assessment for Text Simplification) データセット [7] である．このデータセットには，複数のシステムの出力文に対して，文法性，意味保存性，平易性の人手評価値（bad，ok，good の 3 値）が付いている．本実験では，bad，ok，good をそれぞれ 1，2，3 として扱った．また，学習用の 505 文対のうち，8 割を学習用，2 割を開発用に利用した．本実験で学習に用いたデータセットのうち，テキスト平易化のデータセットが最も小さい．

out-of-domain のデータセットは，Glavaš and Štajner (2013) [8] である．このデータセットには，文法性，意味保存性，平易性の人手評価値（1 から 3 の離散値）が付いている．

スタイル変換　使用したデータセットは，Mir et al. (2019) [9] である．このデータセットには，複数のシステムの出力文に対して，文法性，意味保存性，スタイル変換強度の人手評価値（1 から 5 の連続値）が付いている．スタイル変換強度は，あるスタイルに対する入力文と出力文のスタイルの違いの度合いを示す指標であり，同じスタイルの場合は 1，完全に異なるスタイルの場合は 5 となる．Mir et al. (2019) で考慮するスタイルは感情である．本実験では，8 割を学習用，1 割を開発用，1 割を評価用に利用した．

B　モデル

MT-DNN は，著者らによって公開されている実装[4)]を用いた．MT-DNN の文符号化器の初期化に使用する事前学習済みモデルは，transformers 2.3.0 [5)] の `bert-base-cased` を用いた．MT-DNN の各ハイパーパラメータは，以下の組合せに対してグリッドサーチを行い，開発データにおけるピアソンの積率相関係数が最も大きいモデルを選択した．その他のハイパーパラメータは既定値を用いた．

- 最大トークン数 $\in \{128, 256\}$
- バッチサイズ $\in \{8, 16\}$
- 学習率 $\epsilon \in \{2e{-}5, 3e{-}5, 5e{-}5\}$
- エポック数 $\in \{1, \ldots, 20\}$

4）https://github.com/namisan/mt-dnn
5）https://github.com/huggingface/transformers