ニューラルネットに基づく質問文生成モデルのクロスドメイン評価*

王 怡青† 龍 梓 † 土井 俊弥 † 韓 炳材 † 宇津呂 武仁‡
筑波大学大学院 システム情報工学研究科† 筑波大学システム情報系‡

1 はじめに

近年，質問応答 (QA) システムに関する研究が注目を集めている．その中でも，特に，ニューラルネットワークを用いて，QA システムを構築する研究が多く行われている．ニューラルネットワークを用いて QA システムを訓練するためには，大規模な質問応答訓練文が必要である．そのため，文献 [6] では，大規模知識ベース Freebase [1] を利用して質問応答訓練文が作られた．文献 [6] では，Freebase の意味表現から質問文を作るためのモデルを訓練し，そのモデルを用いて，自動的に大規模な質問文集合を作成した．しかし，Freebase に収録されている語彙知識の範囲は，主に欧米を中心とする地域に関する事実であり，ドメインが限定されている．このため，文献 [6] において作成された質問応答訓練文を用いて質問応答システムを作成する場合においても，同様に，限定されたドメインの質問応答システムしか作成できない点が問題となる．そこで，本論文では，ニューラルネットワークに基づく質問文生成モデルに対して，欧米地域中心の知識に関するドメインから，日本に関する知識のドメインへのクロスドメイン評価を行う．本論文では，文献 [6] で用いられた SimpleQuestion [2] 中の三項関係事例，および，自動生成された 3,000 万文の質問文から抽出した 190 万組の訓練事例に対して，RNN(Recurrent Neural Network) の一種である LSTM(Long Short Term Memory) [7] を適用することにより，質問文生成モデルを訓練する．ここで，各訓練事例は，⟨ 主語，述語，目的語 ⟩ の三項関係事例およびそれに対応する質問文で構成される．本論文では，190 万組の訓練事例を用いて訓練した質問文生成モデルのインドメイン評価およびクロスドメイン評価を行なった．インドメイン評価においては，訓練事例中の主語・目的語を抽象化した placeholder を用いることにより，質問文生成性能が改善された．次に，クロスドメイン評価においては，placeholder で表現された訓練事例を用いた質問文生成モデルの訓練によって，クロスドメインへの適用が可能であるという結果が得られた．

*Cross-Domain Evaluation of Question Generation based on Neural Network

†Yiqing Wang, Zi Long, Syunya Doi, Bingcai Han, Graduate School of Systems and Information Engineering, University of Tsukuba
‡Takehito Utsuro, Faculty of Engineering, Information and Systems, University of Tsukuba

2 ニューラルネットに基づく質問文生成

本論文では，LSTM(Long Short-Term Memory) モデル [7,8] を用いて，三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ から質問文を生成するモデルを訓練する．訓練する際には，主語，述語，目的語中の各単語をこの順に並べた単語の系列を入力系列 $\boldsymbol{x} = (x_1, \ldots, x_N)$ とみなして，質問文生成モデルを訓練する．本論文では，LSTM モデルの実装においては，Google が公開したオープンソースの深層学習ツールである TensorFlow[1]を用いた．また，LSTM モデルの階層数，ベクトルの次元数等の詳細設定およびパラメータ設定においては，文献 [4] の設定を用いた．

3 同一ドメインの質問文生成

本節では，まず，同一ドメインでの質問文生成モデルの訓練および評価手順，評価結果について述べる．

3.1 Freebase から作成された質問文データセット

文献 [2,6] においては，Freebase [1] 中の語彙知識として三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ を用いる．Freebase はエンティティ間の関係知識を集めた語彙知識データベースの一つである．Freebase 中には，3 億個以上の三項関係の事例が収録されている．例えば，Freebase におけるエンティティ “James Cameron” および “film director” が，関係 “profession” によって関係付けられている．この関係において，“James Cameron” が主語，“film director” が目的語，関係 “profession”

[1] https://www.tensorflow.org/

表 1: 本研究の訓練・評価事例

(a) 同一ドメインの質問文生成

名称	事例数	説明
訓練用 190 万事例	1,909,292	30M Factoid Question-Answer Corpus [6] 中の事例のうち，主語，目的語とも SimpleQuestion [2] に含まれる事例を選定．さらに，⟨ 主語，述語，目的語 ⟩ の組が評価用 2 万事例と重複する事例を除外．
評価用 2 万事例	21,300	SimpleQuestion の評価用事例

(b) ドメイン適応可能性の評価

名称	事例数	説明
日本ドメイン 32 事例	32	SimpleQuestion 中の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ を参考にして，日本ドメインの三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ を人手で作成
SimpleQuestion32 事例	32	SimpleQuestion の評価用 2 万事例のうち，日本ドメイン 32 事例の述語 (およびカテゴリ) を持つ事例をまず選定．次に，同一述語 (およびカテゴリ) をもつ事例集合に対して，質問文自動生成モデル適用時の BLEU 値の平均値を求めて，BLEU 平均値に最も近い BLEU 値を持つ事例を選定した．

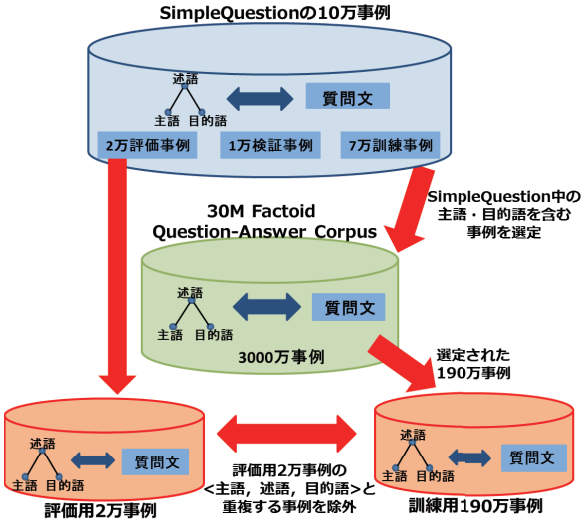

図 1: 本論文における訓練・評価用事例集合作成の流れ

が述語である．

文献 [2,6] において三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ から質問文を作成する際には，

- 主語と述語の情報を質問文に含める．
- 目的語が質問文の回答となる．

の規則性を用いる．例えば，三項関係事例 ⟨James Cameron, profession, film director⟩ からは，質問文

What is James Cameron's profession ?

が生成される．

SimpleQuestion データセット [2] は，Freebase [1] 中の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ に対して，人手で質問文を作成することによって作られた．訓練用事例約 7 万，評価用事例約 2 万，検証用事例約 1 万の合計約 10 万個の事例から構成される．

次に，文献 [6] においては，TransE [3] モデルによって，Freebase 中の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ に対する分散表現[2]を学習した後，前節で述べた SimpleQuestion 中の 7 万訓練事例によって質問文自動生成モデルを訓練し，この質問文自動生成モデルを Freebase 中の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ に適用することによって 3,000 万例の質問文を自動生成した．この手順により構成したデータセットを 30M Factoid Question-Answer Corpus と呼ぶ．

3.2 本研究で用いる訓練・評価事例

本研究で用いる質問文自動生成モデルの訓練・評価用事例を選定する詳細な手順を図 1 に示す．また，選定された訓練・評価事例の詳細を表 1(a) に示す．

図 1 の手順においては，前節の 30M Factoid Question-Answer Corpus 中の 3,000 万例の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ と質問文の組の中から，質問文自動生成モデル訓練用の事例 190 万例をまず選定する．この際には，SimpleQuestion 中の 10 万事例中の主語，目的語のみを含む三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ に限定して選定を行うことにより，できる限りスパース性を避けた事例集合を作成した．ただし，選定された事例集合においては，SimpleQuestion の 2 万評価事例中の三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ と重複する事例は除外して選定を行う．

3.3 placeholder

文献 [6] においては，主語，目的語のスパース性の問題を回避するために，主語，目的語を抽象化した placeholder 方式を用いる．placeholder としては，1 カテゴリのみのもの，および 82 カテゴリのものの二種類を用いる．

[2]この分散表現を学習する際には，SimpleQuestion データセット中の主語・目的語を含む三項関係事例 ⟨ 主語, 述語, 目的語 ⟩ を中心に分散表現訓練事例 3,000 万例を選定することによって，できる限りスパース性を避けた分散表現を学習している．

表 2: 日本ドメイン 5 事例および SimpleQuestion 5 事例とその評価

(a) 日本ドメイン 5 事例および SimpleQuestion 5 事例

ID	SimpleQuestion ドメイン		日本ドメイン	
	〈 主語，述語，目的語 〉	質問文	〈 主語，述語，目的語 〉	質問文
1	richard t.sullivan -notable types- author	who was richard t.sullivan	hayao miyazaki -notable types- film director	what is hayao miyazaki notable for?
2	jeremy c.owens -profession- actor	what is jeremy c.owens a professional at ?	hayao miyazaki -profession- film director	what is hayao miyazaki's profession?
3	dmitriy mamin-sibiryak -gender- male	what gender is dmitriy mamin-sibiryak	hayao miyazaki -gender- male	what is the gender of hayao miyazaki?
4	maimi yajima -place of birth- saitama prefecture	where in japan was maimi yajima born ?	ryoma sakamoto -place of birth- kouchi prefecture	where is ryoma sakamoto's place of birth ?
5	bruce wasserstein -organizations founded- wasserstein perella & co.	which organization is founded by bruce wasserstein	masayoshi son -organizations founded- softbank group	what company did masayoshi son found ?

(b) 評価結果

ID	SimpleQuestion ドメイン			日本ドメイン		
	生成された質問文 (placeholder を単語に変換後)	BLEU	主観評価 (正解：1, 不正解：0)	生成された質問文 (placeholder を単語に変換後)	BLEU	主観評価 (正解：1, 不正解：0)
1	what is richard t . sullivan ?	41.1	0	what is hayao miyazaki ?	57.9	0
2	what is jeremy c . owens ' s profession ?	30.2	1	what is hayao miyazaki's profession ?	100.0	1
3	what is the gender of dmitriy mamin-sibiryak ?	30.2	1	what is the gender of hayao miyazaki ?	100.0	1
4	where was maimi yajima born ?	57.0	1	where was ryoma sakamoto born ?	18.8(2-bleu)	1
5	what organization was founded by bruce wasserstein ?	36.6	1	what organization was founded by masayoshi son ?	23.2(2-bleu)	1

3.4 質問文生成の精度評価

SimpleQuestion の 2 万評価事例を対象として，自動評価尺度である BLEU [5] を用いて同一ドメインの質問文を生成するモデルの評価を行なった結果を表 3 に示す．この結果から分かるように，主語，目的語とも単語のままのモデルと比較して，主語，目的語ともいずれかの placeholder としたモデルの方が BLEU 値が 10 ポイント以上改善した．また，全モデルの中では，主語に対して 82 カテゴリの placeholder を用い，目的語に対して 1 カテゴリの placeholder を用いた場合に最も高い性能となった．

これらのモデルの間の優劣を分析するために，

- モデル 1: 「主語・目的語両方とも単語」，
- モデル 2: 「主語・目的語両方とも 1 カテゴリの placeholder」，
- モデル 3: 「主語は 82 カテゴリの placeholder，目的語は 1 カテゴリの placeholder」

の三種類のモデルによる質問文生成例の比較結果を以下に述べる．

まず，三項関係事例 〈 主語, 述語, 目的語 〉 として ⟨alex golfis, place of birth, athens⟩ に対する質問文をモデル 1 で生成した結果の質問文は，

where was alex _UNK born ?

となり，主語の “golfis” の部分が_UNK(未知語) となったのに対して，モデル 2 およびモデル 3 においては正しく質問文が生成できた．このことから，placeholder を用いたモデルの方が，単語のみを用いたモデルよりも質問文生成性能が高いことが分かる．

次に，三項関係事例 〈 主語, 述語, 目的語 〉 として ⟨avocados, genre, short film⟩ に対する質問文を，主語・目的語両方とも 1 カテゴリの placeholder であるモデル 2 で生成した結果の質問文は，

what genre of music does avocados play ?

となり，映画ジャンルが音楽ジャンルに置換されてしまった．それに対して，主語は 82 カテゴリの placeholder，目的語は 1 カテゴリの placeholder であるモデル 3 で生成した結果の質問文は，

what type of film is avocados ?

となり，映画ジャンルであるという情報が忠実に質問文中に表現され，モデル 2 よりも性能が高いことが確認できた．

表 3: 各モデルの評価結果 (訓練事例：「訓練用 190 万事例」，評価事例：「評価用 2 万事例」)

各モデルにおける主語・目的語の表現（単語または placeholder）		BLEU
主語	目的語	
1 カテゴリ	1 カテゴリ	36.2
82 カテゴリ	1 カテゴリ	37.8
1 カテゴリ	82 カテゴリ	36.7
82 カテゴリ	82 カテゴリ	37.2
単語	単語	26.0

4 クロスドメインの評価

前節で述べたように，訓練事例と評価事例の間でドメインが同一の場合には，主語，目的語を placeholder として訓練した質問文生成モデルを用いることによって，質問文生成の性能が大幅に改善した．そこで本節では，前節の評価結果において最も高い性能を達成した「主語は 82 カテゴリの placeholder，目的語は 1 カテゴリの placeholder」のモデルを用いて，訓練文とは異なるドメインの評価事例を対象として，クロスドメイン評価を行なった．

4.1 評価手順

本節では，表 1(b) の手順によって，質問文生成モデルの訓練事例とはドメインが異なるクロスドメインの評価事例集合「日本ドメイン 32 事例」を作成し，さらに，これを用いて，質問文生成モデルの訓練事例と同一ドメインのインドメインの評価事例集合「SimpleQuestion32 事例」もあわせて作成し，これらの評価事例を用いてドメイン適応可能性の評価を行なった．作成された評価事例集合のうちの 5 事例を表 2(a) に示す．また，これらの評価事例に対して生成された質問文を各文ごとに客観評価した結果，および，主観評価した結果を表 2(b) に示す．ただし，主観評価においては，参照用質問文と同義の場合に「正解: 1」と判定し，完全には同義ではない場合に「不正解: 0」と判定する．また，評価事例全体に対する評価結果を表 4 に示す．

4.2 評価結果

この評価結果においては，参照用質問文と完全に一致する生成結果は，インドメインの SimpleQuestion ドメインでは 0 文，クロスドメインの日本ドメインでは 4 文であったが，BLEU 値は 40 ポイント台となり高い値となった．また，主観評価の結果においては，インドメイン，クロスドメインとも，評価事例全体の 75%において正解となるという高い評価結果が得られた．さらに，「主観評価において正解と判定された質問文を参照用質問文に置き換えて算出した BLEU 値」が 80 ポイント台となるという高い評価結果が得られた．このように，placeholder を用いた質問文生成モデルによって，インドメイン，クロスドメインを問わず高い性能を達成できることから，本論文のタスクであるニューラルネットによる質問文生成のタスクにおいては，クロスドメインへの適用が可能であることが分かった．

表 4: 質問文 32 文全体の評価結果

	Simple Question ドメイン	日本 ドメイン
32 文中の完全一致文数	0 (0.0%)	4 (12.5%)
主観評価において正解となる文数	24 (75.0%)	24 (75.0%)
BLEU 値	40.9	48.0
主観評価において正解と判定された質問文を参照用質問文に置き換えて算出した BLEU 値	84.9	84.2

5 おわりに

本論文では，文献 [6] で用いられた SimpleQuestion [2] 中の三項関係事例，および，自動生成された 3,000 万文の質問文から抽出した 190 万組の訓練事例に対して，RNN の一種である LSTM [7] を適用することにより，質問文生成モデルを訓練した．そして，この質問文生成モデルのクロスドメイン評価を行った．その結果，placeholder で表現された訓練事例を用いた質問文生成モデルの訓練によって，クロスドメインへの適用が可能であるという結果が得られた．

参考文献

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. SIGMOD*, pp. 1247–1250, 2008.

[2] A. Bordes, N. Usunier, S. Chopra, and J. Weston. Large-scale simple question answering with memory networks. *ArXiv e-prints*, 2015.

[3] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proc. 26th NIPS*, pp. 2787–2795, 2013.

[4] Z. Long, T. Utsuro, T. Mitsuhashi, and M. Yamamoto. Translation of patent sentences with a large vocabulary of technical terms using neural machine translation. In *Proc. 3rd WAT*, pp. 47–57, 2016.

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pp. 311–318, 2002.

[6] I. V. Serban, A. García-Durán, Ç. Gülçehre, S. Ahn, S. Chandar, A. C. Courville, and Y. Bengio. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proc. ACL*, pp. 588–598, 2016.

[7] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proc. NIPS*, pp. 3104–3112, 2014.

[8] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. E. Hinton. Grammar as a foreign language. In *Proc. ICLR*, pp. 2773–2781, 2015.