Multi-Perspective Context Aggregation Networkの拡張による英文空所補充問題の一解法

玉城　悠仁† 　新妻弘崇 †† 　太田　学 ††

†, †† 岡山大学大学院自然科学研究科　〒 700–8530 岡山県岡山市北区津島中三丁目 1 番 1 号
E-mail: †tamaki@de.cs.okayama-u.ac.jp, ††{niitsuma, ohta}@cs.okayama-u.ac.jp

あらまし　英文空所補充問題は空所を含む英文に対して，空所にあてはまる語句を解答する問題であり，機械学習や深層学習のモデルの文脈理解度の評価に用いられる．近年，英文空所補充問題のための大規模データセットが公開されたことで，様々なニューラルネットワークモデルが提案されている．Wang らは選択肢つき英文空所補充問題データセットである CLOTH データセットを用いて，空所を含む英文と選択肢を同時に学習する Multi-Perspective Context Aggregation Network (MPNet) を提案した．本研究ではこのモデルを拡張し，英文空所補充問題の解答に取り組む．CLOTH データセットで学習し，3 種類の英文空所補充問題テストデータに対して正答率などで評価した．Wang らの MPNet を拡張した提案モデルでは，テスト用 CLOTH データセットでの高校入試問題に対する正答率が 58.5%であり，ベースラインである KenLM の正答率 58.8%と同程度だったが，もとの Wang らの MPNet の正答率 47.2%を上回った．

キーワード　自然言語処理，ニューラルネットワーク，英文空所補充

1 はじめに

空所補充問題は一部が空所となっている文章に対して，空所に当てはまる語句や文を予測するタスクであり，機械学習や深層学習のモデルの評価に用いられる代表的なタスクの一つである．特に近年では，英文空所補充問題のための大規模データセットが公開されたことで，ニューラルネットワークを用いた様々なモデルが提案されている．

英文空所補充問題のための代表的な大規模データセットには，CNN/Daily Mail データセット [1]，Story Cloze Test データセット [2]，CLOTH データセット [3] などが挙げられ，それぞれ問題の形式が異なる．CNN/Daily Mail データセットはニュース記事と，その要約であるクエリのペアからなる．すなわち，クエリの一部が空所となっており，当てはまる語句を予測する．Story Cloze Test データセットは 4 文からなる文章と，そのあとに続く文の選択肢が 2 つ与えられ，正しい選択肢を予測する問題である．Story Cloze Test データセットは LSDSem 2017 Shared Task [4] における英文空所補充問題のデータセットとしても採用された．CLOTH データセットは複数の空所を含む小説であり，空所ごとに 4 つの選択肢が与えられ，空所内に当てはまる語句を予測する問題である．CLOTH データセットは中国の高校入試や大学入試における英語科目の長文問題をもとに作成されている．

本稿で取り組む英文空所補充問題は CLOTH データセットのような，文章中の空所に当てはまる語句を選択肢から予測する形式の問題である．空所を 1 箇所含む英文に対して，空所に当てはまる語句を選択肢から予測する．CLOTH データセットの英文空所補充問題の例を表 1 に示す．表 1 中の “___” は空所を表し，正答の選択肢は太字で表す．表 1 に示すように CLOTH データセットの英文空所補充問題には内容理解を問う問題や，語彙を問う問題，時制や人称といった英文法を問う問題などが含まれている．

Wang らは CLOTH データセットを用いて，空所を含む英文と選択肢の両方を学習に利用する Multi-Perspective Context Aggregation Network を提案した [5]．本稿では Wang らのモデルを [5] の表現にならい，“MPNet” と呼ぶ．MPNet は CLOTH テストデータの高校入試問題において 53.2%，大学入試レベルの問題において 49.0%の正答率を示した．本稿では Wang らの MPNet を拡張し，英文空所補充問題の解答に取り組み，その性能を評価する．

本稿は次のような構成となっている. 2 節で，関連研究について述べ，3 節では，提案する MPNet の拡張について説明する．4 節では，評価実験の内容と結果を示し，5 節では，その結果について考察する．6 節では，本稿のまとめと今後の課題について述べる．

2 関連研究

2.1 統計的言語モデルを用いた英文空所補充

2011 年から始まったプロジェクト “ロボットは東大に入れるか”[1] は，大学入試センター試験 [2] および東京大学入試において高得点をとることを目的としている．このプロジェクトの英語科目では問題を一文問題，複数文問題（問題や選択肢が複数文からなるもの），長文問題の 3 つに大きく分類している．一文問題はさらに文法・語法・語彙問題，語句整序完成問題，発話文生成問題の 3 つに分類され，この文法・語法・語彙問題が我々の取り組む英文空所補充問題と一致する．

1 : https://21robot.org/
2 : https://www.dnc.ac.jp/

表 1　CLOTH データセットの英文空所補充問題の例

空所つき英文	選択肢			
She pushed the door open and found nobody there. “I am the ____ to arrive.” She thought and came to her desk.	last	second	third	**first**
As I expected, Lois didn’t pass over the ____ when she saw me the next day.	**issue**	outcome	criterion	message
He then bought a new pair of glasses ____ the man.	to	**for**	after	with
He liked to draw pictures and he thought these pictures ____ good.	is	are	was	**were**

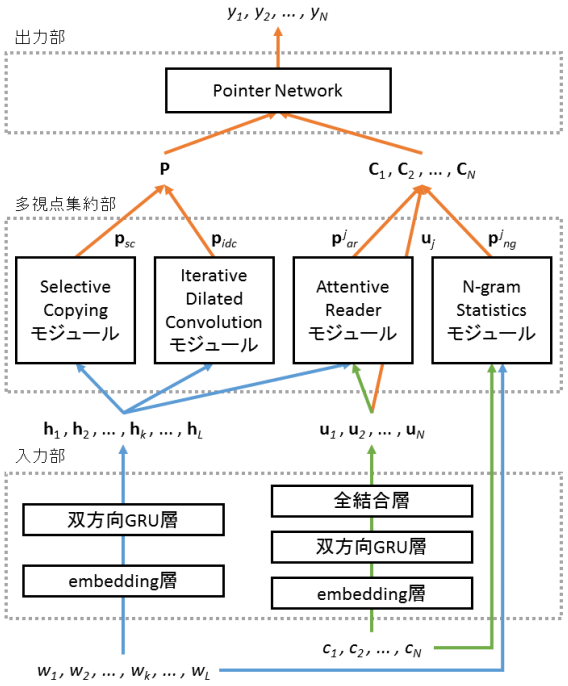

図 1　Multi-Perspective Context Aggregation Network (MPNet) の概略図

東中ら [6] は統計的言語モデルである KenLM [7] を用いて，文法・語法・語彙問題の英文空所補充問題に取り組んだ．空所に選択肢を補充した英文に対して，KenLM で対数尤度を算出することで単語の並びとして尤もらしい選択肢を予測する．東中らは約 178GB の Common Crawl コーパス [3] を学習データとして 7-gram 言語モデルを構築した．また，大学入試センター試験の本試験および追試験の過去問，代ゼミセンター模試，ベネッセ模試，独自に収集したその他の問題を合わせた，合計 552 問の英文空所補充問題をテストデータとした．東中らの手法では，テストデータにおける正答率は 82.2%であった．さらに，文長による対数尤度の正規化や，数詞を桁数のみ分かるように変換するなどの正規化を利用すると正答率は 85.7%まで向上した．

2.2　Multi-Perspective Context Aggregation Network (MPNet)

Wang らの提案した Multi-Perspective Context Aggregation Network (MPNet) は，空所を含む英文と各選択肢の語句を入力として，選択肢に対する確率分布を出力するモデルである．MPNet の概略図を図 1 に示す．MPNet は入力部，多視点集約部，出力部の 3 つで構成される．

入力部は空所を含む英文および各選択肢の語句のエンコーダである．embedding 層と双方向 GRU 層を通して，空所を含む英文は単語ごとにベクトル化される．選択肢の語句は embedding 層を通した後，別の双方向 GRU 層を通して語数に関わらず固定長のベクトルへ変換される．なお，MPNet では空所を表す記号 “____” も 1 単語として扱う．図 1 において w_i, $i \in [1, L]$ は空所を含む英文の i 番目の単語を表し，L は文長を表す．また，c_j, $j \in [1, N]$ は j 番目の選択肢を表し，N は選択肢数を表す．入力部を通して $\{w_i\}_{i=1}^{L}$ は $\{\mathbf{h}_i\}_{i=1}^{L}$ へ，$\{c_j\}_{j=1}^{N}$ は $\{\mathbf{u}_j\}_{j=1}^{N}$ へと変換される．

多視点集約部は入力部のエンコード結果をもとに，複数の視点からそれぞれベクトルを生成し，それらを連結したベクトルを出力する．Wang らは多視点集約部を以下の 4 つのモジュールで構成した．

Selective Copying モジュール　空所が英文中の k 単語目，すなわち w_k であるとすると，このモジュールは入力部の出力 $\mathbf{h}_k$ のみを抽出し，出力する．$\mathbf{h}_k$ は双方向 GRU 層の出力であるため空所の前後の情報が反映されている．このモジュールの出力を $\mathbf{p}_{sc}$ と表す．

Iterative Dilated Convolution モジュール　畳み込みニューラルネットワーク (convolutional neural network: CNN) は画像処理や自然言語処理において一般的に用いられており，一定間隔離れた要素を畳み込む dilated convolution を多層化することは自然言語処理において有効である [8] [9]．このモジュールは dilation rate の異なる 4 層の CNN で構成されている．このモジュールの出力を $\mathbf{p}_{idc}$ と表す．

Attentive Reader モジュール　英文空所補充問題においては空所から離れた語句と選択肢を関連づけて考えることも必要である．このモジュールは Chen らの提案した Attentive Reader [10] に基づくものであり，$\{\mathbf{h}_i\}_{i=1}^{L}$ と $\{\mathbf{u}_j\}_{j=1}^{N}$ による attention を利用することで空所から離れた語句の情報も選択肢の予測に反映させる．このモジュールの $\mathbf{u}_j$ に対する出力を $\mathbf{p}^j_{ar}$ と表す．

N-gram Statistics モジュール　空所に入る語の予測にコロケーション情報を反映させるため，このモジュールは空所に選択肢を補充した英文についての n-gram($n \in [1, 5]$) の出現回数を対数正規化して利用する．このモジュールの c_j に対する出力を $\mathbf{p}^j_{ng}$ と表す．

これら 4 つのモジュールの出力および，$\{\mathbf{u}_j\}_{j=1}^{N}$ を連結し，多視点集約部は $\mathbf{P} = [\mathbf{p}_{sc}; \mathbf{p}_{idc}]$ と $\mathbf{C}_j = [\mathbf{u}_j; \mathbf{p}^j_{ar}; \mathbf{p}^j_{ng}]$, $j \in [1, N]$ を出力する．ここで “;” はベクトルの連結を表す．

出力部は Pointer Network 機構 [11] を用い，多視点集約部の出力をもとに各選択肢に対する確率分布を出力する．MPNet の損失関数はこの出力と教師データとの交差エントロピー誤差であり，この損失が最小となるように学習を行う．

3 : https://registry.opendata.aws/commoncrawl/

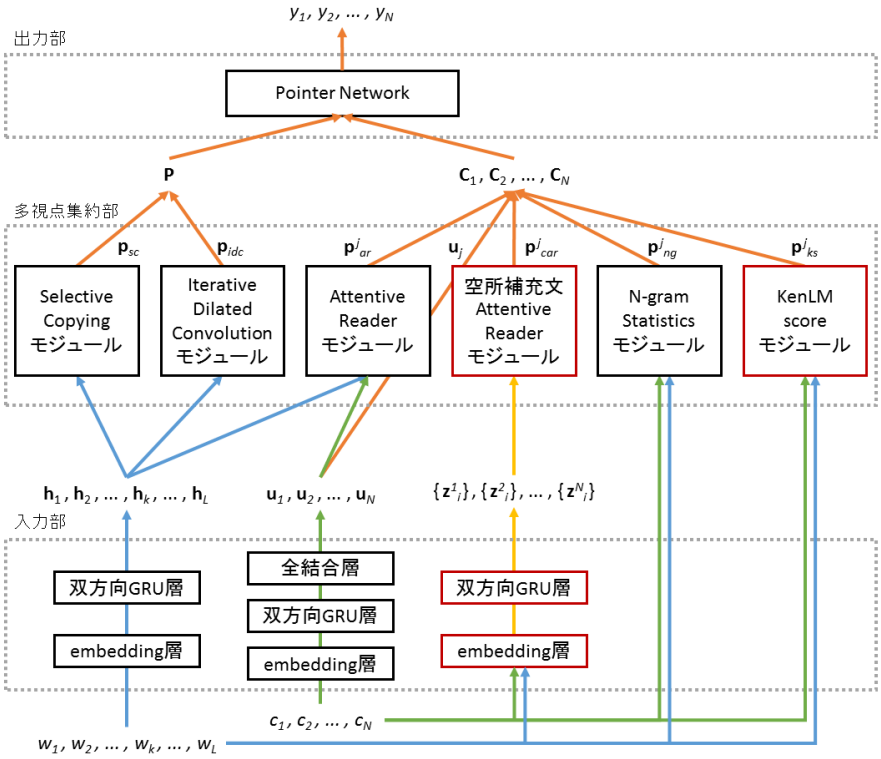

図 2　提案モジュールを追加した MPNet の概略図

3　Multi-Perspective Context Aggregation Network (MPNet) の拡張

本節では 2.2 節で述べた Wang らの MPNet の拡張について述べる．MPNet の多視点集約部において Wang らの設計した 4 つのモジュールとは別の 2 つのモジュールを提案する．本稿で提案するモジュールを “空所補充文 Attentive Reader モジュール” および “KenLM Score モジュール” と呼び，それぞれ詳細を述べる．提案モジュールを追加した MPNet の概略図を図 2 に示す．図 2 において赤色で示す箇所が，本稿で提案する拡張部分である．

3.1　空所補充文 Attentive Reader モジュール

図 1 に示した Wang らの Attentive Reader モジュールの出力は，式 (1) および式 (2) で算出される．ここで $\mathbf{W}_{ar}$ は重みを表し，$\mathbf{b}_{ar}$ はバイアス項を表す．

$$\alpha^j_i = \mathrm{softmax}_i(\mathbf{u}^T_j \mathbf{W}_{ar} \mathbf{h}_i + \mathbf{b}^T_{ar} \mathbf{h}_i),\ i \in [1, L],\ j \in [1, N] \quad (1)$$

$$\mathbf{p}^j_{ar} = \sum_{i=1}^{L} \alpha^j_i \mathbf{h}_i \quad (2)$$

Wang らは式 (1) のように attention の算出に空所を含む英文から得られる $\{\mathbf{h}_i\}_{i=1}^{L}$ を用いたが，本稿で提案する空所補充文 Attentive Reader モジュールでは，空所に各選択肢を補充した英文を用いて attention を算出する．Wang らの MPNet の入力部に新たに双方向 GRU 層を追加し，空所を含む英文と同様に空所に選択肢を補充した英文も単語ごとにベクトル化する．これにより得られる出力を $\{\mathbf{z}^j_i\}_{i=1}^{L_c}$ とする．ここで j は何番目の選択肢を補充するかを表し，L_c は空所に選択肢を補充した英文の文長を表す．選択肢の語句が 1 単語である場合にのみ L と L_c が一致する．空所補充文 Attentive Reader モジュールの $\mathbf{u}_j$ に対する出力を $\mathbf{p}^j_{car}$ と表し，式 (3) および式 (4) で定義する．ここで $\mathbf{W}_{car}$ は重みを表し，$\mathbf{b}_{car}$ はバイアス項を表す．

$$\alpha^j_i = \mathrm{softmax}_i(\mathbf{u}^T_j \mathbf{W}_{car} \mathbf{z}^j_i + \mathbf{b}^T_{car} \mathbf{z}^j_i),\ i \in [1, L_c],\ j \in [1, N] \quad (3)$$

$$\mathbf{p}^j_{car} = \sum_{i=1}^{L_c} \alpha^j_i \mathbf{z}^j_i \quad (4)$$

3.2　KenLM Score モジュール

Wang らは英文のコロケーション情報を反映させるために N-gram Statistics モジュールを利用した．しかし，n-gram の n が大きいと，未知語や出現頻度の少ない語が含まれるとき，その単語の影響により，その単語を含む n-gram の出現回数が下がりやすい．これにより未知語や出現頻度の少ない語だけでなく，その n-gram に含まれる他の単語の情報まで反映しにくくなる．

そこで本稿では n-gram に含まれる他の単語の情報を空所の予測に反映させるため，統計的言語モデルである KenLM を利用する．2.1 節で述べたように KenLM は文に対して対数尤度を算出するが，文中に未知語が含まれる際はその算出方法を変更し，未知語以外の残りの語を用いて対数尤度を算出する．j 番目の選択肢を補充した英文を s_j とするとき，KenLM Score モジュールの s_j に対する出力を $\mathbf{p}^j_{ks}$ と表し，式 (5) で定義する．ここで $f_n(s)$ は n-gram モデルの KenLM によって算出される英文 s の対数尤度を表す．

$$\mathbf{p}^j_{ks} = [f_1(s_j); f_2(s_j); \ldots; f_n(s_j)],\ (n = 5) \quad (5)$$

図 2 では提案モジュールを 2 つとも追加しているが，4 節の

モデルの拡張の評価実験では，Wang らの MPNet に上記のモジュールを追加，あるいは Wang らのモジュールと入れ替えて実験を行う．

4 評価実験

4.1 実験の設定

4.1.1 実験データ

学習データおよび検証データには CLOTH データセットを用いる．CLOTH データセットは中国の高校入試および大学入試の長文問題から作られた英文空所補充問題であり，学習データには 5,513 の長文からの 76,850 問，検証データには 805 の長文からの 11,067 問が含まれる．CLOTH データセットでは空所に入る語句の選択肢が 4 つ与えられる．

テストデータには CLOTH データセット，大学入試センター試験の英文空所補充問題，Microsoft Research Sentence Completion Challenge データセット [12] の 3 つを用いる．本稿では Microsoft Research Sentence Completion Challenge データセットを “MSR データセット” と呼ぶ．

CLOTH データセットのテストデータには，中国の高校入試問題である 335 の長文からの 3,198 問，中国の大学入試問題である 478 の長文からの 8,318 問が含まれる．無作為に選出した 3,000 問に対する Amazon Mechanical Turkers [4] による人手の正答率は，高校入試問題では 89.7%，大学入試問題では 84.5%であった [3]．

大学入試センター試験では大問 2A が英文空所補充問題であり，2000 年度から 2016 年度に行われた本試験および追試験のうち，空所が 1 箇所のみである 324 問を評価の際のテストデータとして使用する．大学入試センター試験では空所に入る語句の選択肢が 4 つ与えられる．

MSR データセットは 1,040 問の英文空所補充問題であり，英文はアーサー・コナン・ドイルの小説シャーロック・ホームズシリーズ 5 冊から抽出したものである．英文中の 1 箇所が空所となっており，選択肢が 5 つ与えられている．無作為に選出された 100 問に対する人手の正答率は 91%であった [12]．また，MSR データセットは選択肢が 5 つであるが，CLOTH データセットで学習する MPNet 拡張モデルは選択肢を 4 つにする必要があるため，MPNet 拡張モデルのテスト時には，正答でない選択肢を無作為に 1 つ除く．

4.1.2 実験データの前処理

CLOTH データセットおよび，CLOTH データセットで学習したモデルに対するテストデータについては以下の前処理を行う．ここで英文の分割，トークン化および品詞を表すタグの付与には NLTK [13] を用いる．

（1） 英文の分割およびトークン化

（2） 各トークンへ品詞を表すタグの付与

（3） 記号を表すタグが付与されたトークンの削除

（4） 数詞を表すタグが付与されたトークンの “NUM” への置換

（5） 大文字の小文字への変換

上記の処理をした後，英文に空所が含まれない場合には，空所を含む英文の前の英文と連結し，複数の英文をまとめて MPNet 拡張モデルへの入力とする．また，CLOTH データセットにおいては 1 文中に空所が複数含まれる英文も存在するが，MPNet 拡張モデルへの入力にそのような英文が含まれる場合には，空所が 1 つとなるように残りの空所を正答の語句で補充した複数の入力へと分割する．例えば “New-born babies can hold things ____ either of their hands, but ____ about two years they usually like to use their right hands. ” のような空所を含む英文があるとする．正答の選択肢はそれぞれ “with” と “for” である．この場合には “New-born babies can hold things ____ either of their hands, but **for** about two years they usually like to use their right hands. ” のように一部の空所を正答の語句で補充し，モデルへの入力とする．また，同様にして “New-born babies can hold things **with** either of their hands, but ____ about two years they usually like to use their right hands. ” もモデルへの入力とする．

4.1.3 モデルのハイパーパラメータ

MPNet 拡張モデルのハイパーパラメータは Wang らと同じ値を用い， [5] 中に記載されていないハイパーパラメータや提案した拡張部分のハイパーパラメータは以下のように設定する．

図 2 の MPNet 拡張モデルの入力部において，全ての embedding 層の出力は 300 次元とし，初期値には GloVe の学習済みベクトル [14] を用い，本モデルの学習中にさらに最適化されるようにする．また，全ての GRU 層のユニット数は 128 とし，入力系列 $\{w\}$ の最大長は 80 単語とする．双方向 GRU 層の出力に適用する dropout 率は 0.5 とする．

多視点集約部の Iterative Dilated Convolution モジュールは，2 ブロックで構成され，各ブロックは dilation rate が 1 および 3 の CNN 層からなる．CNN 層におけるフィルタ数は 128，カーネルサイズは 3 とする．2 ブロックへの入力前にバッチ正規化および ReLU 関数を適用し，最終出力において max pooling を適用する．

損失関数の損失は交差エントロピー誤差とし，最適化は Adam [15] によって行う．学習回数は 30 回とし，検証データにおいて最も正答率の高いモデルをテストで用いる．また，計算資源の都合により，本モデルに 30,000 語の語彙を与え，embedding 層でのベクトル化に用いる．この語彙は，2017 年 12 月 2 日時点での英語版 Wikipedia の記事データにおける出現頻度上位 30,000 語である．本モデルではこの語彙に含まれない単語は未知語を表す “UNK” として扱う．

4.2 性能評価

Wang らの MPNet に 3 節で提案したモジュールを追加，あるいは Wang らのモジュールと入れ替えて実験を行い，拡張したモデルを評価する．評価尺度には英文空所補充問題の正答率を用いる．MPnet は与えられた選択肢 4 つに対する確率分布を出力し，その確率が最も高い選択肢を空所に補充すべき選択肢と

4 : https://www.mturk.com/

表 2　4 択のテストデータ対する各モデルの正答率 (%)

モデル	テストデータ			
	CLOTH-M (3,198 問)	CLOTH-H (8,318 問)	センター試験 (324 問)	MSR (1,040 問)
Wang らの MPNet	47.2	38.8	37.0	29.2
+ 空所補充文 AR	47.2	37.5	32.1	25.8
+ KenLM Score	58.2	**47.2**	41.3	**32.5**
+ 空所補充文 AR + KenLM Score	57.3	46.0	40.1	30.9
− AR + 空所補充文 AR	46.5	38.6	32.4	26.8
− N-gram Statistics + KenLM Score	57.9	46.7	**42.2**	30.9
− AR − N-gram Statistics + 空所補充文 AR + KenLM Score	**58.5**	43.5	**43.5**	**31.7**
KenLM(7-gram)	**58.8**	**47.6**	40.7	31.4

表 3　テストデータの選択肢に現れる単語の語彙数と最も出現する単語

テストデータ	語彙数 (正規化した語彙数)	最も出現する単語	出現回数
CLOTH データセットの 高校入試問題 (3,198 問)	2,655 (0.207)	in	124
CLOTH データセットの 大学入試問題 (8,318 問)	5,992 (0.180)	when	109
大学入試センター試験 (324 問)	784 (0.604)	to	88
MSR データセット (1,040 問)	3,671 (0.705)	clozed	9

して予測する．また，ベースラインとして同じデータで学習した KenLM の 7-gram モデルを用いる．

実験結果を表 2 に示す．表 2 中の CLOTH-M，CLOTH-H はそれぞれ CLOTH データセットにおける高校入試問題，大学入試問題を表す．センター試験は大学入試センター試験の英文空所補充問題，MSR は MSR データセットを表す．また，– は Wang らの MPNet からのモジュールの除去を，+ はモジュールの追加を表す．空所補充文 AR，KenLM Score はそれぞれ 3 節で述べた空所補充文 Attentive Reader モジュール，KenLM Score モジュールを表す．AR，N-gram Statistics はそれぞれ Wang らの Attentive Reader モジュール，N-gram Statistics モジュールを表す．なお，[5] で示されている Wang らの MPNet による CLOTH データセットに対する正答率は表 2 中の値より高いものになっている．これは本実験で行ったデータの前処理やモデルのハイパーパラメータの差異，および学習時の乱数による影響が原因であると考えられる．

表 2 中では各テストデータに対して最も高い正答率と 2 番目に高い正答率を太字で表している．表 2 より，Wang らのモデルと比べ，空所補充文 Attentive Reader モジュールを追加したモデルでは正答率が低下した．KenLM Score モジュールを追加したモデルでは正答率が大きく向上した．特に大学入試センター試験に対しては，Wang らの MPNet の Attentive Reader モジュールおよび，N-gram Statistics モジュールを提案した空所補充文 Attentive Reader モジュールおよび KenLM Score モジュールに置き換えたモデルでの正答率は 43.5%であり，Wang らの MPNet の正答率 37.0% および KenLM の正答率 40.7%を上回った．他のテストデータに対しては，いずれの正答率もベースラインである KenLM の正答率とほぼ同じ値であった．

CLOTH データセットの大学入試問題は高校入試問題に比べ正答率が低い傾向にあり，モデルによる予測においても問題の難易度が影響するといえる．MSR データセットに対してはいずれのモデルでも他のテストデータに比べ正答率が低いが，これは問題の質の差であると考えられる．CLOTH データセットや大学入試センター試験の問題は 1 文以上の英文からなり，内容理解や語彙，文法を問う問題であるのに対し，MSR データセットは 1 文からなり，語彙を問う問題のみである．また，選択肢に現れる単語の種類についても MSR データセットと他のデータセットで異なる．テストデータの選択肢に現れる単語の語彙数，最も出現する単語とその出現回数を表 3 に示す．表 3 中の “語彙数” の項目にある ( ) 内の値は語彙数を問題数と 1 問あたりの選択肢数の積で割った値である．すなわち，総選択肢数で正規化した語彙数である．MSR データセットは他のデータセットに比べて，選択肢に現れる単語の種類が多い．また，前置詞や接続詞といった頻出単語が MSR データセットの選択肢には，ほとんど含まれておらず，単語の出現頻度に差が少ない．他のテストデータに比べて頻出単語が選択肢に現れにくいことから，統計情報を利用する KenLM では予測が難しく，正答率が低くなったといえる．

表 4　各テストデータに対して MPNet および MPNet+KS が正答した問題数と誤った問題数

テストデータ	両モデルで正答	MPNet のみ正答	MPNet+KS のみ正答	両モデルで誤答
CLOTH-M (3,198 問)	1,257 (39.3%)	254 (7.9%)	607 (18.9%)	1,080 (33.7%)
CLOTH-H (8,318 問)	2,305 (27.7%)	924 (11.1%)	1,628 (19.5%)	3,461 (41.6%)
センター試験 (324 問)	81 (25.0%)	39 (12.0%)	53 (16.3%)	151 (46.6%)
MSR (1,040 問)	150 (14.4%)	154 (14.8%)	188 (18.0%)	548 (52.6%)

5 考　　察

5.1 KenLM Score モジュールの有効性

4.2 節で示したように，Wang らの MPNet に KenLM Score モジュールを追加すると MPNet の正答率が大きく向上した．このモデルを本節では “MPNet+KS” と呼ぶ．MPNet+KS が正答した問題と Wang らの MPNet や KenLM の正答した問題の分布を調べ，それらの違いについて考察する．

5.1.1 Wang らの MPNet との比較

Wang らの MPNet が正答した問題と MPNet+KS が正答した問題の分布を調べる．各テストデータに対する MPNet および MPNet+KS の予測結果を表 4 に示す．表 4 中の ( ) 内の値は全問題数に対する割合を表す．表 4 より，どのテストデータに対しても Wang らの MPNet では正答できなかったが，MPNet+KS では正答できた問題が全体の 15%以上あり，KenLM Score モジュールが適切な予測に有効であったといえる．特に CLOTH データセットの高校入試問題に対しては，11 ポイントの正答率の向上につながった．一方で，Wang らの MPNet では正答できたが，MPNet+KS では正答できなかった問題も少なくない．

Wang らの MPNet のみが正答した問題，MPNet+KS のみが正答した問題について，その問題が問う内容の種類で分類し，Wang らの MPNet と MPNet+KS の予測結果の違いについて調べる．CLOTH データセットには内容理解を問う問題や，語彙を問う問題，時制や人称といった英文法を問う問題などが含まれている．MPNet のみが正答した問題，MPNet+KS のみが正答した問題について，その種類を分類する．CLOTH データセットの高校入試問題において，一方のモデルのみが正答した問題を，問う内容について分類した結果を表 5 に示す．表 5 中の “人称” は主語に対応する動詞の活用形の種類や，代名詞の格の種類に関する問題を表し，“名詞の単複” は名詞の単数形や複数形に関する問題や，名詞の冠詞に関する問題を表す．表 5 より，Wang らの MPNet に比べ MPNet+KS では語彙を問う問題や名詞の単複を問う問題に正答が多い．KenLM Score モジュールの追加により，コロケーション情報を予測に反映させたことで，これらの問題が適切に予測しやすくなったといえる．

5.1.2 KenLM との比較

5.1.1 項と同様に KenLM が正答した問題と MPNet+KS が正答した問題の分布も調べる．各テストデータに対する KenLM および MPNet+KS の予測結果を表 6 に示す．表 6 より，どのテストデータに対しても MPNet+KS と KenLM は正答した問題や誤答した問題がほとんど一致しており，MPNet+KS のみが正答した問題や KenLM のみが正答した問題はいずれも全体の 5%未満であった．すなわち，MPNet+KS と KenLM は予測結果がほとんど一致している．MPNet+KS は Wang らの MPNet の正答率を向上させたが，それは KenLM の予測によるところが大きいといえる．

表 5　CLOTH データセットの高校入試問題 (CLOTH-M) において MPNet または MPNet+KS の一方のモデルのみが正答した問題の種類

問題が問う内容	MPNet のみ正答した問題数	MPNet+KS のみ正答した問題数
内容理解	161	313
語彙	75	229
時制	0	4
人称	10	26
名詞の単複	8	35

KenLM のみが正答した問題，MPNet+KS のみが正答した問題について，その問題が問う内容の種類で分類し，KenLM と MPNet+KS の予測結果の違いについて調べる．CLOTH データセットの高校入試問題において，KenLM または MPNet+KS の一方のモデルのみが正答した問題を，問う内容について分類した結果を表 7 に示す．表 7 より，KenLM は MPNet+KS より語彙を問う問題での正答数が多いが，それ以外ではあまり差がみられない．

5.2 KenLM における学習データの大きさ

テストデータが異なるものの，本実験における KenLM の大学入試センター試験に対する正答率は，同じく KenLM を用いた東中ら [6] の実験の正答率 85.7%を大きく下回っている．前処理や正規化の方法も異なるが，東中らは約 178GB の学習データを用いたことから，学習データの大きさ正答率の関係を調べる．

4.2 節では KenLM の学習に CLOTH データセットを用いたが，ここでは英語版 Wikipedia の記事データを用いた実験を行う．英語版 Wikipedia の記事データを用いた代表的なデータセットの 1 つに text8 コーパス [5] がある．text8 コーパスは，Mahoney

5 : http://mattmahoney.net/dc/textdata

表 6　各テストデータに対して KenLM および MPNet+KS が正答した問題数と誤った問題数

テストデータ	両モデルで正答	KenLM のみ正答	MPNet+KS のみ正答	両モデルで誤答
CLOTH-M (3,198 問)	1,757 (54.9%)	124 (3.8%)	107 (3.3%)	1,210 (37.8%)
CLOTH-H (8,318 問)	3,557 (42.7%)	410 (4.9%)	376 (4.5%)	3,975 (47.7%)
センター試験 (324 問)	123 (37.9%)	9 (2.7%)	11 (3.4%)	181 (55.8%)
MSR (1,040 問)	290 (27.8%)	37 (3.5%)	48 (4.6%)	665 (63.9%)

表 7　CLOTH データセットの高校入試問題 (CLOTH-M) において KenLM または MPNet+KS の一方のモデルのみが正答した問題の種類

問題が問う内容	KenLM のみ正答した問題数	MPNet+KS のみ正答した問題数
内容理解	64	65
語彙	50	34
時制	5	0
人称	2	4
名詞の単複	3	4

表 8　各学習データにおける大学入試センター試験 324 問 (4 択) に対する KenLM の正答率 (%)

学習データ	正答率 (%)
CLOTH データセット (約 7MB)	40.7
Wikipedia の英文先頭 50MB	42.5
text8 コーパス (100MB)	45.0
Wikipedia の英文先頭 200MB	48.4
Wikipedia の英文先頭 1GB	54.0

によって 2006 年 3 月 3 日時点での英語版 Wikipedia [6] の記事データを整形して作られた 100MB の英文コーパスである．また，2017 年 12 月 2 日時点での英語版 Wikipedia の記事データに対して text8 コーパスと同様の前処理を行い，先頭から 50MB，200MB，1GB を取り出す．text8 コーパスおよび，これらの記事データ学習データとして実験する．このときの大学入試センター試験に対する正答率の変化を表 8 に示す．学習データが大きくなるにつれて正答率も上がることが確認でき，さらに学習データを増やせば，さらなる正答率の向上が予想される．

4.2 節で述べたように，KenLM を利用した KenLM score モジュールを追加することで MPNet の正答率を向上させたが，それは KenLM 自体の性能によるところが大きく，MPNet の学習による効果はあまり見られなかった．統計的言語モデルの KenLM は英文空所補充問題に対して高い効果があり，学習データを増やせば，さらなる正答率の向上が望める．MPNet においても学習データの拡張は有効であると考えられるが，CLOTH データセットのような選択肢つき英文空所補充問題は作成が非常に高コストであるため，半教師あり学習を適用し，MPNet においても学習データを拡充することが今後の課題の 1 つである．

6 : https://en.wikipedia.org/wiki/Main_Page

6 ま と め

本稿では Wang らの提案した英文空所補充問題を解くモデルである Multi-Perspective Context Aggregation Network において，その中間層である多視点集約部を拡張し，新たなモジュールを提案した．実験では，CLOTH データセットで学習し，CLOTH データセットのテストデータ，大学入試センター試験，Microsoft Research Sentence Completion Challenge データセットの 3 種類のデータに対してテストした．テスト用 CLOTH データセットでの高校入試問題に対する正答率は，Wang らの MPNet では 47.2%であったのに対し，統計的言語モデルである KenLM を用いたモジュールを MPNet に追加することで 58.2%に向上した．しかし，これは KenLM のみによる正答率 58.8%とほぼ同じであった．CLOTH データセットのような選択肢つき英文空所補充問題は作成が非常に高コストであるため，MPNet では半教師あり学習による学習データの拡充の効果を検証することなどが今後の課題として挙げられる．

文　献

[1] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1693–1701, 2015.

[2] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, 2016.

[3] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2344–2356, 2018.

[4] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 46–51, 2017.

[5] Liang Wang, Sujian Li, Wei Zhao, Kewei Shen, Meng Sun, Ruoyu Jia, and Jingming Liu. Multi-perspective context aggregation for semi-supervised cloze-style reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 857–867, 2018.

[6] 東中竜一郎, 杉山弘晃, 成松宏美, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩, 大和淳司. 「ロボットは東大に入れるか」プロジェクトにおける英語科目の到達点と今後の課題. 人工知能学

会全国大会論文集, Vol. JSAI2017, p. 2H21, 2017.

[7] Kenneth Heafield. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pp. 187–197, 2011.

[8] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.

[9] Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. Fast and accurate entity recognition with iterated dilated convolutions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2670–2680, 2017.

[10] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 2358–2367. Association for Computational Linguistics, 2016.

[11] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pp. 2692–2700, 2015.

[12] Geoffrey Zweig and Chris J. C. Burges. A challenge set for advancing language modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, WLM '12, pp. 29–36, 2012.

[13] Steven Bird and Edward Loper. Nltk: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.