登場人物を考慮した日本語文章読解テストセットの作成 *

渡会 拓人 † 横井 康孝 † 土屋 雅稔 †‡
† 豊橋技術科学大学 工学部 情報・知能工学課程 / ‡ 大学院 情報・知能工学専攻

1 はじめに

計算機によって自然言語で記述された文章を理解することは，自然言語処理分野における重要な目標の一つである．この目標を達成するには，**文章理解とは何か**を定義する必要があるが，直接的に定義することは非常に困難である．そのため，文章理解を直接的に定義する代わりに，対象となる文章が理解できている場合に限って達成できるタスクを定義することによって，文章理解の度合いをベンチマークするという方針が広く採用されている．

この方針のタスク定義には，4 択問題 [9, 11] や質問応答 [10]，空所穴埋め問題 [4, 8] などがある．4 択問題は，複数文からなるパッセージ C，パッセージ C に記述されている内容に関する質問文 q，解答候補集合 A が与えられた時，正解 $a \in A$ を選ぶタスクである．質問文 q とパッセージ C が理解できている場合，かつ，質問文 q とパッセージ C が理解できている場合に限って，正解 $a \in A$ が選択できるように設計されていれば，4 択問題は文章読解のベンチマークとして利用可能と考えられる．質問応答は，解答候補集合 A が与えられない条件で正解 a を答えなければならない点のみが 4 択問題と異なるため，4 択問題と同様に文章読解のベンチマークとして利用可能と考えられる．

空所穴埋め問題 [4, 8] を，以下のように定義する．対象となるテキストに連続して出現する $n+1$ 個の文 s_i^{i+n+1} を，先行する n 個の文 s_i^{i+n} と，直後の文 s_{i+n+1} に分割する．ここで，先行する n 個の文 s_i^{i+n} を文脈文集合 C，文 s_{i+n+1} の一部を空所に置き換えた文を質問文 q と呼ぶ．空所穴埋め問題は，文脈文集合 C，空所を含む質問文 q，空所に挿入し得る候補単語集合 A が与えられた時，正しい単語 $a \in A$ を選ぶタスクである．文脈文集合 C と質問文 q が理解できている場合，かつ，文脈文集合 C と質問文 q が理解できている場合に限って，正しい単語 $a \in A$ が選択できるように設計されていれば，空所穴埋め問題は文章読解のベンチマークとして利用可能と考えられる．

空所穴埋め問題は，4 択問題や質問応答と比較すると，テストセットを機械的に作ることが可能であるという利点がある．先に述べた通り，文脈文集合 C と質問文 q は，対象テキストの一部から機械的に生成できる．また，候補単語集合 A は，実際に空所に入る正解単語 a と，正解単語と類似した機能を持つ単語を文脈文集合 C から抽出するという方針により，機械的に生成可能である．それに対して，4 択問題や質問応答では，パッセージ C とは別に，質問文 q を用意する必要がある．近年では，クラウドソーシングによって質問文 q を作成する手法が広く用いられているが，作業者のバイアスによる問題が指摘されている [12].

以上から，本稿では，文章読解のタスク定義として，空所穴埋め問題に注目する．英語で記述された文章に対する空所穴埋め問題テストセットとしては，Children Books' Test (CBT)[4] が広く用いられているが，日本語で記述された文章に対する空所穴埋め問題テストセットは，筆者らの知る限りでは存在しない．そのため本稿では，日本語で記述された児童向け小説・物語を対象とする空所穴埋め問題テストセットを作成する手順と，作成結果について述べる．

2 関連研究

CBT[4] は，Project Gutenberg[*1] に収録された英語で記述された児童向け書籍から作成された空所穴埋め問題テストセットである．CBT の文脈文集合 C の長さ n は 20，候補単語集合 A の大きさ $|A|$ は 10 である．文

* Construction of Japanese Reading Comprehension Testset Considering Characters

† Watarai Takuto, School of Engineering, Department of Computer Science and Engineering Toyohashi University Technology
‡ Yasutaka Yokoi, Masatoshi Tsuchiya, Graduate School of Computer Science and Engineering, Toyohashi University Technology

[*1] https://www.gutenberg.org/

脈文集合 C に対して Stanford Core NLP Toolkit[13] を用いて固有表現抽出し，抽出された固有表現から無作為に選択した単語を用いて，候補単語集合 A を作成している．固有表現には，人名，地名，組織名，時間表現，割合表現などがあるが，これらは全て区別せずに候補単語集合 A が作成されている．そのため，文脈文集合 C 全体を理解せずとも，直前の文 s_{i+n} と質問文 q のみから正解単語 $a \in A$ を推定可能であるという偏りが指摘されている [6]．このように，空所穴埋め問題テストセットを，文章読解のベンチマークとして用いるためには，候補単語集合 A の作成方法は非常に重要である．

3 日本語文章読解テストセットの作成

3.1 対象書籍の選定

本稿では，日本語文章読解テストセットを作成する対象テキストとして，青空文庫[*2]に注目する．青空文庫は，著作権が期限切れし public domain となった書籍を，誰もがアクセスできるようインターネット上で公開している活動である．文章読解テストセットを作成するには，明確な文章構造を持ち，解釈のぶれが少ない物語文が適していると考えられる．そのため，CBT と同様に，児童向け小説を対象としてテストセットの作成を行うことにする．

青空文庫では，各書籍に対して，表題，著者，出版年，仮名遣いなどのメタデータに加えて，日本十進分類法 (NDC) に基づく分類コードが付与されている．さらに，児童向け書籍の場合は，NDC 分類コードに「K」を前置した分類コードが付与されている．そこで本稿では，青空文庫において，「K913」(児童向け小説 · 物語) という分類コードが付与されている書籍を対象とする．また，青空文庫には，青空文庫には旧字 · 旧仮名の書籍も多く収録されている．しかし，既存の形態素解析器や係り受け解析器の多くは旧字 · 旧仮名を未知語として処理することになり，解析精度が低下することが予想される．そのため，本稿では，新字 · 新仮名の書籍のみを対象とする．

以上の方針に従い，青空文庫から対象書籍を取り出した結果を表 1 に示す．対象となる書籍は，2018 年 12 月 1 日現在の青空文庫の Git リポジトリ[*3]に収録されている書籍の 5.4% (810 冊) である．対象書籍の著者を，表 2 に示す．対象書籍の著者は 54 名いるため，表 2 には，対象書籍数が多い上位 10 名を示した．

表 1　青空文庫に収録されている書籍数

全書籍数	14,999 冊
児童向け小説 (K913 に分類される書籍) 数	1,085 冊
対象書籍 (新字 · 新仮名の児童向け小説) 数	810 冊

表 2　対象書籍の著者 (対象書籍数が多い順に 10 名を抜粋)

著者	対象書籍数
小川未明	388 冊
宮沢賢治	64 冊
夢野久作	50 冊
新美南吉	43 冊
楠山正雄	42 冊
海野十三	37 冊
江戸川乱歩	33 冊
豊島与志雄	30 冊
竹久夢二	18 冊
小酒井不木	14 冊
計 54 名	計 810 冊

3.2 作成手順

空所穴埋め問題の作成手順は，おおよそ以下のステップからなる．

1. 対象書籍を章単位に分割し，各章を文に分割する．
2. 文から候補単語を抽出する．
3. 閾値以上の候補単語が得られる箇所から，空所穴埋め問題を作成する．

以下，各ステップの詳細を述べる．

3.2.1 対象書籍の整形

先に述べた通り，空所穴埋め問題とは，質問文 q に設けられた空所の単語を，文脈文集合 C に基づいて推定するタスクである．文脈文集合 C が第 i 番目の章，質問文 q が第 $i+1$ 番目の章に属している場合を仮定する．この場合，文脈文集合 C と質問文 q は，同一書籍内ではあるものの，非常に異なるコンテキスト下で出現しているため，文脈文集合 C から質問文 q の空所を推定できるとは考えにくい．そこで，本稿では，適切な空所穴埋め問題を作成するには，文脈文集合 C と質問文 q が同一の章に連続して出現していることを条件とする．

[*2] https://www.aozora.gr.jp/
[*3] https://github.com/aozorabunko/aozorabunko.git

小田さんはじめ、私たちは呆気にとられて俊夫君の顔を見つめました。

消えた証拠

しばらくして小田さんは、「それでは君は白痴を尋問するのか？」と尋ねました。

図 1　見出し

この条件を満たすため，最初に，対象書籍を章単位に分割する．しかし，青空文庫に収録されている書籍には，明示的かつ統一的な章区切り情報は付与されていないため，見出しを手がかりとして章単位に分割する．見出しの例を，図 1 に示す．図 1 では「消えた証拠」が見出しである．青空文庫の書籍テキストファイルでは，見出しの前後には 2 つ以上の改行文字が連続して出現することが大半であるので，これを手がかりとして見出しを抽出し，書籍を章単位に分割する．

次に，各章を文単位に分割する．本稿では，句点に基づいて文を区切るという方針を採る．ただし，カギ括弧で囲まれた台詞中に出現する句点については，区切らないこととした．

3.2.2　候補単語の抽出

本稿では，候補単語として，以下の 5 種類を考える．

1. 人名
2. 地名
3. 組織名
4. 登場人物名
5. 固有物名

2 節で述べた通り，全ての固有表現を候補単語の対象とすると，不適切なテストセットになる可能性がある．この問題を回避するために，文章読解の対象として，もっとも重要と考えられる種類の固有表現に限定して考える．

図 2 に，JUMAN++ 2.0.0RC2 と KNP 4.19 の組み合わせを利用して固有表現抽出を行なった結果を示す．図 2 より明らかに，候補単語として抽出したい重要な登場人物名などの各種の固有表現が抽出できていないことが分かる．そのため，各書籍に登場する人名，地名，組織名，登場人物名のデータベースを人手作成し，抽出結果を補う．人手により作成したデータベースの諸元を表 3 に示す．

種類	抽出できなかった固有表現
人物名	怪人二十面相，丁七唱 (ちょうしちとなう)，アア，ノロちゃん，ギネ，フウフィーヴオ，ペンネンネンネンネン・ネネム，エキモス
組織名	月世界探検隊，少年探偵団，イーハトーヴ火山局，読売新聞
土地名	カマジン国，六天山塞，地獄の一丁目，ハンムンムンムンムン・ムムネ市，火星

図 2　JUMAN++/KNP を用いて抽出できなかった固有表現

表 3　人手作成した固有表現および登場人物データベース

種類	エントリ数
人名	1250
地名	746
組織名	168
登場人物名	449
固有物名	228

3.3　作成結果

テストセットに採用された書籍の統計情報を，表 4 に示す．

作成した問題データの具体例を図 3 に示す．CBT と同様の形式で構成されており，1 単語が隠されて空所が設けられた質問文 q，空所に挿入できる 5 個の候補単語からなる解答候補集合 A，質問文 q の直前の 20 個のパッセージ C，正解 $a \in A$ が確認できる．

4　おわりに

本稿では，日本語の空所穴埋め問題において，CBT のように広く標準的に使用されるテストセットは存在していないことから，青空文庫に収録されている児童向け小説・物語の内，新字・新仮名遣いの書籍を対象として空所穴埋め問題テストセットを作成した．今後は，作成したテストセットを用いて，日本語の省略を考慮

表 4　対象テキストの諸元

筆者数	28
書籍数	582
章数	113
文数	21,015
単語数	2,044,518

1 いつかの、おとうさんの童話のような、ふとった鶏が、この小舎に来るのかとおもうと僕はたのしみです。
2 金井君も時時みに来ます。
3 おかあさんは鶏を飼ってもたべさせるものがないので、生物は困るといっています。
4 僕は生物は何でも好きです。
5 鶏は、吉田さんのおじさんが、宇都宮から持ってきて下さるのだそうです。
6 吉田さんのおじさんは、お仕事のことで、たびたび東京へいらっしゃいます。
7 早く鶏のおうちが出来て、宇都宮の鶏が来るといいと思います。
8 今日は日曜日なので、僕は金井君と二人で雑司ヶ谷の坂井君のおうちへ約束しておいた竹をもらいに行きました。
9 金網のかわりに、竹の細いので格子をつくってやるのです。
10 目白へ出て、学習院の通りを歩いていると、僕たちぐらいの男の子が、
11 「八王子へ行くのはこの道を行ったらいいの」とききます。
12 破れたシャツと、あしの出たつぎはぎだらけのズボンで、小さい風呂敷包を持っています。
13 髪の毛が随分のびていて大人のようにつかれた顔をしています。
14 僕たちは八王子を知りません。
15 「君はどこから来たの」
16 金井君がたずねました。
17 「遠いところから来たの‥‥」
18 「遠いところってどこなの」
19 「深谷というところから歩いて来たの」
20 「へえ、深谷ってどこだい、健ちゃん知ってる‥‥」
21 深谷というのは、どこだか知らないけれども、おかあさんは、ねぎの話が出ると、すぐ、XXXXXのねぎはおいしかったというから、ねぎの出来るところから来たのかも知れないと思いました。

深谷
包|宇都宮|雑司ヶ谷|深谷|八王子

図 3　作成した問題データ

した文章読解モデルの研究を行う予定である.

参考文献

[1] “Attention-over-Attention Neural Networks for Reading Comprehension”,Yiming Cui,Zhipeng Chen,Si Wei,Shijin Wang,Ting Liu,Guoping Hu(2017)

[2] “The Stanford CoreNLP Natural Language Processing Toolkit“,Christopher D. Manning,Mihai Surdeanu,John Bauer,Jenny Finkel,Steven J. Bethard,David McClosky

[3] “Text Understanding with the Attention Sum Reader Network”,Rudolf Kadlec,Martin Schmid,Ondrej Bajgar &Jan Kleindienst(2016)

[4] “THE GOLDILOCKS PRINCIPLE: READING CHILDREN ’ SBOOKS WITH EXPLICIT MEMORY REPRESENTATIONS”,Antoine Bordes,Sumit Chopra&Jason Weston(2016)

[5] “Extended Named Entity Hierarchy“,Satoshi Sekine, Kiyoshi Sudo, Chikashi Noba(2002)

[6] “How Much Reading Does Reading Comprehension Require?A Critical Investigation of Popular Benchmarks“, Divyansh Kaushik, Zachary C. Lipton(2018)

[7] “Natural Language Comprehension with the EpiReader“Adam Trischler,,Zheng Ye,Xingdi Yuan,Kaheer Suleman(2016)

[8] “Teaching Machines to Read and Comprehend“, Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom(2015)

[9] “MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text“, Matthew Richardson, Christopher J.C. Burges, Erin Renshaw(2013)

[10] “SQuAD: 100,000+ Questions for Machine Comprehension of Text“, Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, Percy Liang(2016)

[11] “「大学入試問題を解く」ことから見える言語, 知識, 世界理解に関する研究課題 “, 宮尾 祐介 and 川添 愛, 人工知能学会誌, Vol. 27, No. 5 (2012)

[12] “Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment”, Masatoshi Tsuchiya, Proc. of LREC2018, pp. 1506–1511 (2018)

[13] “The Stanford CoreNLP Natural Language Processing Toolkit”, Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. Proc. of ACL2014 System Demonstrations, pp. 55–60 (2014)