

機械読解精度向上のための英文読解問題データの分析

藤田 和成^{1,a)} 浅沼 爽汰¹ 田村 亮介¹ 坂井 優介¹ 町田 翔² 延澤 志保^{1,b)}

概要: 機械が人間と同じようにテキストを読んで質問に答えられるようにする処理は機械読解と呼ばれ、質問に対する正解率で理解度を測り評価とすることが一般的である。本研究では、中高生向けの英語の試験問題を元に作られた RACE タスクを機械読解タスクの評価対象として採用する。RACE タスクの精度向上のためデータの分析と実験を行った結果、類義語など同様の意味内容を持つ語句の認識の必要性を確認した。

キーワード: 機械読解, RACE, 英文読解, 類義語.

Analysis of RACE Data for the Improvement of Machine Reading Comprehension

KAZUMASA FUJITA^{1,a)} SOTA ASANUMA¹ RYOSUKE TAMURA¹ YUSUKE SAKAI¹ SHO MACHIDA²
SHIHO HOSHI NOBESAWA^{1,b)}

Keywords: Machine Reading Comprehension, RACE, English Reading Comprehension, Synonym.

1. はじめに

機械が人間と同じようにテキストを読んで質問に答えられるようにする処理は機械読解と呼ばれ、質問に対する正解率で理解度を測り評価とすることが一般的である。機械読解により、機械が膨大な量の書籍や書類のデータを分析し、欲しい情報を理解しやすい形でまとめることができるようになると考えられる。

本稿では、英文読解問題を対象とした機械読解タスク、特に、与えられた本文全体の理解が必要な設問を対象として、精度向上のための手法を検討する。

2. SQuAD データセット

現在主流の機械読解タスクは、質問文に対する解答が本文中に明示されている設問を対象として、本文中の 1 文から解

答を生成するものが多い。この形式の処理を想定した英文読解問題データセットのひとつに SQuAD (Stanford Question Answering Dataset) 1.1[1] がある。SQuAD1.1 データセットは Wikipedia の記事を元にして作られた 107,785 個の本文と 536 個の質問文から成る。SQuAD1.1 データセットの問題の 1 例を図 1 に示す。図 1 で、P は本文 (passage), Q は質問文 (question), A は解答 (answer) を示す。例え

P:	In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers."
Q1:	What causes precipitation to fall?
A1:	gravity
Q2:	Where do water droplets collide with ice crystals to form precipitation?
A2:	within a cloud

図 1 SQuAD1.1 の設問例

¹ 東京都市大学知識工学部情報科学科
Faculty of Knowledge Engineering, Tokyo City University

² 東京都市大学大学院工学研究科情報工学専攻
Graduate School of Engineering, Tokyo City University

a) fujita17@ipl.cs.tcu.ac.jp

b) shiho@tcu.ac.jp

ば図1では、Q1に対して、質問文を構成する precipitation や fall といった自立語を含む文を本文中で検索し、当てはまる文の構成要素の中で解答として適切なものを推定して出力している。機械読解タスクは質問応答タスクの一種と考えられるが、この例は質問応答の中でも文章理解を必要とせず語句の検索を主軸とする比較的単純なタスクと位置付けることができる。

3. RACE データセット

本文の内容の理解が必要な質問応答タスクの実現には、まず、機械読解タスクの中でも複数の文を組み合わせる解答を行う形式の設問への対応が必要と考えられる。

この形式の設問を含むデータセットとして、本稿では、RACE データセット [2] に着目する。RACE は中国の12歳から18歳の中高生向けの英語の試験問題を元に作成されたデータセットであり、27,933個の本文とこれに対応する97,687個の質問文からできている。それぞれの本文に対して複数の設問が設定されており、それぞれの設問は4つの選択肢の中から正解を1つ選ぶ形式である。設問の個数は本文ごとに異なる。RACE データセットでは、SQuAD と異なり、本文中の1文のみから解答を推定することは難しい。RACE データセットの1例を図2に示す。図2の

P:	Many best-selling milk brands such as mengniu, yili and bright were discovered to contain melamine, which is usually used to make plastic. The bad milk products have sickened more than 5,300 babies and killed four. (後略)
Q:	Why do some students stop drinking milk?
C1:	because their parents have enough time to prepare other foods for them.
C2:	because they get tired of drinking the same type of milk.
C3:	because some harmful materials are found in some kinds of milk.
C4:	because milk becomes more and more expensive .

図2 RACE の設問例

うち、Pは本文、Qは質問文、Cは選択肢 (choice) であり、赤で表示した選択肢 (C3) が正解である。この例では、質問文にある some students stop drinking milk という句は本文中には出現せず、本文全体から milk に melamine が含まれているとわかったこと、melamine を含む bad milk を原因とした被害が出ていることを読み取った上で、選択肢それぞれとの関連を推定する必要がある。さらにこの例では、正解選択肢 C3 に含まれる harmful materials との句は本文中に出現しておらず、この句が本文中の melamine のことを指すと推定する必要がある。

RACE データセットのサイズを表1に示す。RACE データセットは中学生向けの設問 (RACE-M) と高校生向けの設問 (RACE-H) とに分かれているが、ここではこの2種

表1 RACE データセットのサイズ [2]

	平均語数			語彙数
	本文	質問文	選択肢	
RACE-M	231.1	9.0	3.9	32,811
RACE-H	353.1	10.4	5.8	125,120
RACE	321.9	10.0	5.3	126,629

類を合わせたものを RACE データセットとして用いる (表1[2])。表1に示すとおり、本文の語数の平均は321.9語と長く、英文読解問題としては比較的内容が多い。

このように RACE データセットは機械読解タスクの中でも難易度が高く、Lai[2]らが Gated-Attention Reader を用いた手法で実現した44.1%が正解率としては現時点では最高であり、人間の正解率73.3%にはほど遠い結果である。

4. RACE データセットの特徴

RACE データセットのような複雑な機械読解タスクの正解率を上げるには、正解率に影響する問題点を洗い出す必要がある。

本稿では、Laiらの手法でテストデータとして用いられた設問4,934問 [2] を対象として考察を行った。Laiらはテストデータ4,934問すべてについて本文、質問文、選択肢、正解解答を開示している。

本研究では、Laiらと同様の手法で同じテストデータに対して追実験を行った結果 (表2) を基に、RACE データセットの特徴について考察を行う。表2に示すとおり、追

表2 RACE テストデータ

質問文数	
正解	2,716 問 (44.1%)
不正解	2,758 問 (55.9%)
合計	4,934 問 (100.0%)

実験の結果 Laiらの実験結果 [2] と同じ44.1%の正解率を得ることができた。

4.1 質問文の種類

質問文の形式ごとの設問数と正解率を図3に示す。図3

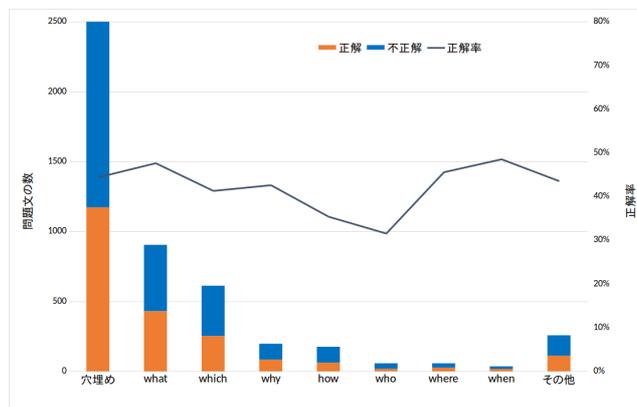


図3 質問文の形式による正解率と設問数の比較

では、質問文の種類として、質問文の先頭に WH 疑問詞等が出現するもの 7 種類 (what, which, why, who, where, when, how), 穴埋め形式の質問文, その他の形式の質問文の合計 9 種類に分けてそれぞれの設問数と正解率を比較している. 折れ線グラフが, 各質問形式の正解率を示す. 棒グラフは各質問形式の設問数であり, オレンジ色が正解の設問数, 青色が不正解の設問数である.

正解率は, 質問形式による大きな差は見られない (図 3). 正解率の平均は 42.3% で, 最も低い who 型で 31.6%, 最も高い when 型で 48.6% である. 図 3 では who 型が特に低いように見えるが, who 型は設問数がテストデータ中の 1.2% と少なく, 他と比べて正解率が低いとは断定できない. このことから, RACE テストデータでは質問形式による難易度の差は大きくないものと考えられる.

図 3 に示したとおり, このテストデータには穴埋め問題が多く含まれる. 穴埋め問題の割合は, テストデータ全体の 53.5% に上る. したがって, 穴埋め問題の正解率の向上が, RACE データセットでの正解率の向上に寄与するものと考えられる. 穴埋め問題は, 質問文が本文の内容と合うように質問文内の括弧に入る語句を選ぶ形式である. 穴埋め問題の例を図 4 に示す. 本文中に質問文とほぼ同様の

Q: according to the passage, the following are wrong except that ().

図 4 RACE データセットの穴埋め問題の例

内容を持つ 1 文が含まれる場合には, 穴埋め問題は文中の構成要素から構文的意味的に最適な要素を選択する処理となる. しかし, 図 2 の例のように本文全体を対象として最適な要素を推定する場合には, 候補の絞り込みが複雑になり, 難易度が上がる. 図 4 の例は本文の内容と合わない選択肢を求める内容であり, これは本文中に類似しているが内容が異なる記述がある場合と, そもそも本文中に該当する記述がない場合とが考えられ, さらに難易度が高いものと考えられる.

4.2 選択肢の語数

RACE データセットは質問文に対して 4 つの選択肢が与えられる形式である. 選択肢を消去法で選ぶにせよ, 最も可能性の高いものを推定するにせよ, 推定処理の対象となるのは選択肢の構成要素に他ならない. したがって, 選択肢の構成要素が少ないほど, 推定に用いる情報が少なくなり, 正解率が下がる可能性が考えられる. テストデータを対象に, 語数ごとに正解率の比較を行った結果を図 5 に示す. ただし, それぞれの設問について選択肢が 4 個あり, 各選択肢の語数は一定ではないため, ここでは正解選択肢の語数で各設問をグループ化した. すなわち, 図 5 で語数 1 の設問は, 正解選択肢の語数が 1 の設問を示し, この設問の他の選択肢の語数については考慮しない. 基本的に,

1 つの質問文に対する選択肢は語句のみあるいは文の形等, 形態が揃っており, 同じ質問文に対する選択肢の語数が極端に異なることはない. 正解選択肢の語数が 20 以上のものはまばらなため, 図 5 では語数 19 までを表示した折れ線グラフは, 各語数ごとの正解率を示す. 棒グラフは各語数ごとの正解設問数と不正解設問数である. 図 5 の結果

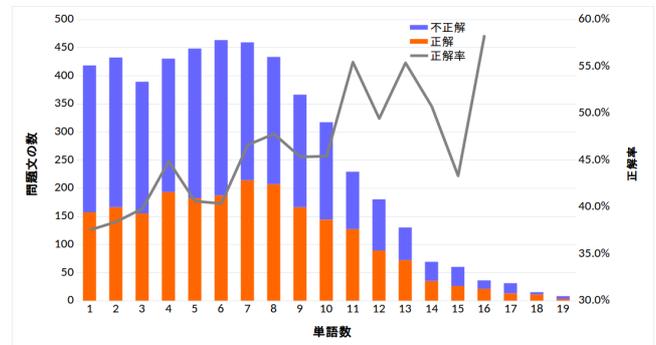


図 5 語数による正解率と設問数の比較

より, 正解選択肢内の語数が少ないほど正解率が低いことが分かった. 選択肢内の情報が少ないために正解率が落ちたと考えられる.

図 5 を見ると, 語数ごとの設問数が語数 3 で一度減少しており, RACE データセットの設問は語数 3 までの短い選択肢と, 語数 4 以上の長い選択肢とに分けられることがわかる. 表 1 にあるとおり 選択肢の平均語数は 5.3 語である. 正解選択肢の語数が少ない設問の例を図 6 に示す. 正解

Q: it can be inferred that riva grill is ()
C1: a supermarket
C2: a water sports center
C3: a restaurant
C4: a mall

図 6 正解選択肢の語数が少ない設問の例

選択肢の語数の少ない設問は, 図 6 の穴埋め問題のように文の一部を埋める形式の他, WH 疑問詞に対して文の形を成さず語句のみで答える形式のものがある. いずれにしても, そのほとんどが, 文の形を成していない. これに対して, 選択肢の語数が多いものは, 図 7 に示すように, 穴埋め問題の場合であっても, 選択肢が文の形態であるものが多い. 選択肢が文の形態を成していない場合には構成要

Q: we can infer from the passage that ()
C1: it is a very complex process for bugs to produce oil
C2: it is not worthwhile for mr. pal to do the experiment
C3: it is safe to use the excretion produced by bugs
C4: it is necessary to use bugs to produce crude oil

図 7 正解選択肢の語数が多い設問の例

素の間の関係が推定し辛く, このことが正解率を下げる一因となっている可能性が考えられる.

先に述べたとおり 1 つの質問文に対する選択肢は原則として形態が統一されている. 図 5 に示すように, 選択肢の

種類は、文の形態を成す(図7)か否(図6)かで分類できるものと考えられる。

4.3 選択肢の構成要素の本文中での出現状況

図5では、正解選択肢を構成する語の数によって正解率に差が出ることを示された。この理由として、もう1点、選択肢の構成要素の本文中での出現の有無が関わる可能性が考えられる。語数の少ない選択肢の場合、構成要素となっている語が本文中に出現しない場合には本文との関連を推定する材料が少なく、推定に失敗する可能性がある。

選択肢の語の本文中での出現状況を調べるため、正解不正解に関わらずすべての選択肢について、本文中に選択肢内の語がどのくらい含まれるかの調査を行った結果を図8に示す。グラフの横軸は語数ごとに選択肢をグループ化し

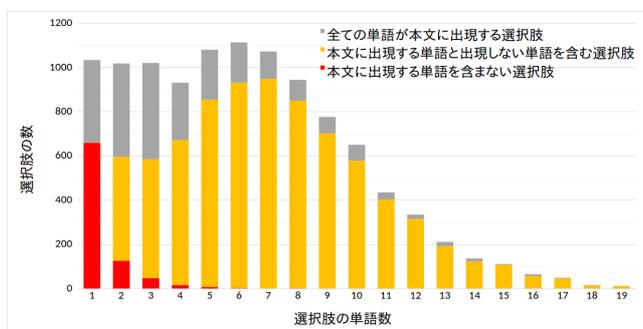


図8 選択肢に含まれる語の本文中での出現状況

たものであり、縦軸はそれぞれのグループに属する選択肢について、本文中での語の出現状況ごとに選択肢数を示したものである。赤色は選択肢内の語が本文にひとつも含まれない選択肢の数、灰色は、選択肢内の語が本文にすべて含まれる選択肢の数、黄色はそれ以外の選択肢の数を示す。例えば選択肢に含まれる語数が1の場合には、その語が本文に含まれる場合(図8灰色)と含まれない場合(図8赤色)のどちらかに分かれる。図8では、前置詞や冠詞のような一般的な語も含めて数えているため、語数2以上の選択肢のグループでは、語数が多くなるにつれて黄色(選択肢に含まれる語の一部が本文中に出現)が増加している。

選択肢内の語がひとつも本文中に出現しない選択肢(図8赤色)は、語数2以上の選択肢では多くない。その反面、選択肢内の語が本文にすべて含まれている選択肢も3割と比較的少ない。このことから、RACEデータセットでは本文中に出現しない語を選択肢に頻繁に用いることがわかる。図8を見ると語数が増えるにつれて選択肢内のすべての語が本文中に出現する選択肢の割合が明らかに小さくなる。これは、語数の多い選択肢では語の言い換えを問う設問が増える可能性を示す。

ここで、正解選択肢のみに着目し、正解選択肢の構成要素の本文中での出現状況が正解率に与える影響を考察する。図9は、テストデータを正解した設問と不正解の設問

とに分け、それぞれ、正解選択肢に含まれる語の本文中での出現状況ごとに設問の割合を比較したものである。図

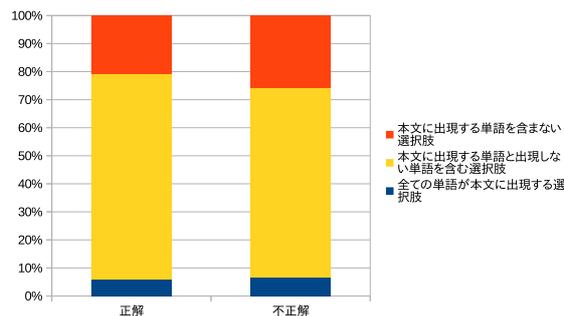


図9 選択肢構成要素の本文内出現状況と正解率

9が示すとおり、正解選択肢の構成要素が本文中にひとつも出現しない場合には、不正解の可能性が高くなる。しかし、構成要素が本文中に出現しない正解選択肢でも正解している設問もあり、正解選択肢の構成要素の本文中での出現の有無によって正解率に大きな差が出るわけではない。RACEデータセットは選択肢が与えられる形式のため、正解選択肢が十分な情報を有していない場合でも、他の選択肢が正解の可能性が低いと推定されれば、消去法で正解選択肢が最尤と判定される可能性がある。図5と図8から、少なくとも語数の少ない選択肢については本文中での出現の有無が正解率に影響を与えている可能性があり、今後この点についてさらに検討する余地がある。

正解の選択肢と本文で違う語を使うことで機械が正解することができなかったRACE設問の例を図10に示す。図

P:	you can decorate , preview and print your stamps all at your own home. you are able to use any of your own photos as part of your postage stamp as well.
Q:	with the system of the united states postal service , people can ().
C1:	buy photo stamps at a very low price
C2:	send their international mails for free
C3:	design their own photo stamps
C4:	have their letters delivered faster than before

図10 語の言い換えが不正解の原因と考えられる例

10では本文中の *decorate* が正解選択肢では *design* と言い換えられている。図9では語の出現状況による正解率の差は明確ではないが、RACEテストデータ中には図10のように選択肢の中の重要な語句が言い換えられている例が多数見られた。英文読解問題では語彙知識も採点対象となることを考えると、語の言い換えが正解率に影響する可能性はあると考えられる。

5. 類義語に着目した機械読解タスクの正解率向上

ここまでの考察から、RACEデータセットの特徴は以下の3点にまとめられる。

- 穴埋め問題が多い.
- 選択肢の形態 (語句のみか, 文を成しているか) によって正解率に差があり, 文の形態を成していない場合の正解率が低い.
- 本文中の語句を質問文や選択肢で他の語句に言い換える設問が多い.

これらの特徴のうち, 3 番目の語句の言い換えは, 1 番の穴埋め問題にも, 2 番の語数の少ない選択肢の問題にも影響を与える項目である. そこで, 正解率向上の1 要素として, 言い換えへの対応を考える. 本文中の語句が質問文や選択肢で他の語句に言い換えられている場合, これらの語句同士の関連を推定し類義語句をまとめることができれば, 正解率の向上が期待できる.

5.1 単語埋め込みによる類義語対応

単語埋め込みとは, 語と語の類似度を測るために語をベクトル表現に置き換えて, 語同士の距離を測る手法である.

本稿では, 単語埋め込みの次元数と語数とについて, それぞれ比較実験を行った.

5.2 単語埋め込み次元数の影響

単語埋め込みには Glove[3] の Glove6B を利用する. Glove6B は Wikipedia の記事から語ベクトルを生成したもので, 40 万語を含んでいる. 次元数と類義語の認識率の関係を調べるため Glove6B の次元数を 50, 100, 200, 300 と変えて語情報を増やして実験を行った結果を図 11 に示す. 横軸は Glove6B の次元数, 縦軸に各次元数での正解率を示す. 語の次元数を増して語情報を増やすと類義語の認識

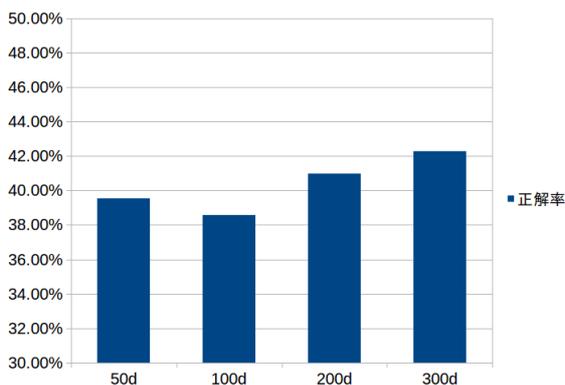


図 11 単語埋め込み次元数ごとの機械読解タスク正解率

率が上がり, 機械読解タスクの正解率が向上する (図 11). Glove6B では 300 次元以上のデータがないため次元数がさらに増えた場合については確認できていないが, 次元数を絞り過ぎると正解率が下がる恐れがあり, 適切な次元数についてさらに考察が必要である.

5.3 単語埋め込みでの語数の影響

単語埋め込みの語数の影響を調べるため, 語数 40 万語の

Glove6B に加えて 220 万語を含む Glove860B と 200 万語を含む Lexvec[4] の 3 種類で機械読解タスクの実験を行った. Glove840B は CommonCrawl というニュースサイトを元にして語ベクトルを生成した単語埋め込みである [3]. Lexvec はニュースサイトを元にして語ベクトルを生成した単語埋め込みで, いくつかの語類似度タスクにおいて Glove より良い成果を出している. 実験の結果を図 12 に示す. グラフの横軸は使用した単語埋め込みを, 縦軸は各単語埋め込みの機械読解タスクでの正解率を示す. この 3

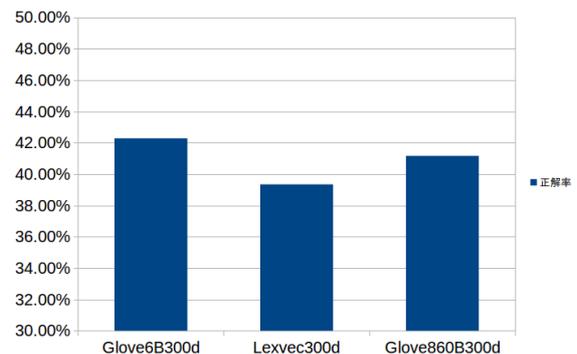


図 12 単語埋め込み語数ごとの機械読解タスク正解率

種類の比較では, 単語数 40 万語の Glove6B が最も正解率が高い. しかし, 単語数 200 万語の Lexvec と 220 万語の Glove840B では Glove840B の方が正解率は高く, 正解率の差は単語数の違いよりもむしろ含まれる語彙によるものと考えられる. これは, RACE データセットが中高生向けの問題であり, 語彙が約 13 万語 (表 1) と比較的小さいためと考えられる.

6. まとめ

機械読解タスクは, 複雑な問題に対応し得る質問応答システムの実現に絡む, 興味深いタスクである. 機械読解タスクは未だ十分な正解率を達成しているとは言えず, 本稿ではその改善を目的として, 比較的複雑な設問から成る英文読解問題データセットについて考察を行った. 本稿では本文と設問との間の語句の言い換えが正解率の向上に関与する可能性を指摘した. また本稿では, これを類義語の問題と捉えて単語埋め込みによる正解率向上の可能性を検討した結果について報告した.

参考文献

- [1] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P.: SQuAD: 100, 000+ Questions for Machine Comprehension of Text, *CoRR*, Vol. abs/1606.05250, (2016).
- [2] Lai, G., Xie, Q., Liu, H., Yang, Y. and Hovy, E. H.: RACE: Large-scale Reading Comprehension Dataset From Examinations, *CoRR*, Vol. abs/1704.04683, (2017).
- [3] Pennington, J., Socher, R. and Manning, C. D.: GloVe: Global Vectors for Word Representation, pp. 1532–1543 (2014).
- [4] Salle, A., Idiart, M. and Villavicencio, A.: Enhancing the LexVec Distributed Word Representation Model Using Positional Contexts and External Memory, *CoRR*, Vol. abs/1606.01283, (2016).