複数文からなる文章読解タスクへのテキスト含意認識の適用可能性の検討

笠原要，平博順，永田昌明

NTT コミュニケーション科学基礎研究所
{kaname, taira}@cslab.kecl.ntt.co.jp, nagata.masaaki@lab.ntt.co.jp

1. はじめに

テキストの汎用的な意味処理方法の1つとして2005年に Recognizing Textual Entailment（テキスト含意認識、以下、RTE と略記）が提唱された[1]。与えられた文章（テキスト T）に対して別の文（仮説 H）の意味を含んでいる/いないかをシステムで判定させるものであり、質問応答、情報抽出、要約、機械翻訳など幅広い応用が期待されている。

本稿では、Reading Comprehension（文章読解、以下、RC と略記）タスクに RTE を応用可能か検討する。扱われるテキストの種類や対象とすべき文数が従来の RTE 研究のテストデータと異なるため、検討結果は RTE の汎用性を高めるための課題を明確化することに貢献できると考えられる。

2. 文章読解タスク

2.1 関連研究

RC タスクとは、国語や外国語学習者の文章読解力を計るためのテストをシステムで自動解答するものである。テストには文章（問題文章）があり、学習者はこれを読んで関わる質問に対して選択肢、単語、文章等の指定された形式で解答する。

これをシステムで行うためには、言語処理や知識処理技術を総合的に適用することが必要となる。そのため、それら要素技術の総合評価プラットフォームや成果を分かりやすくデモンストレーションする手段として有効である。また、読解問題を解く方法を工学的に表現・実装することができれば、学習者の答案の自動評価や問題作成の支援にも貢献できる。

Hirschman らは、問題文章に関わる 5W1H 型の質問に自動解答することを狙いとし、質問文と関連がある文を問題文章から抽出する Deep READ を提案している[2]。関根らは、小学 2 年生程度の国語テストを解答させるシステム「電脳優子」を提案している[3]。様々なタイプの問題解答の中で、問題文章に関連する文の空白部分に単語を埋める“穴埋め問題”の解答を試みている。このような問題文章に関する質問応答の RC 方式に対して RTE は、出力された解答文が問題文の内容を包含するか判定するチェック機能を提供できる。

また、小論文の自動評価において、解答文と問題文の“近さ”から内容の良否を評価する方法が提案されている[4]。このような検討においてテキスト間の内容の包含関係を判定する RTE を適用することで、解答文章が問題文章と関連しているかだけではなく、内容が矛盾していないか判定する手段を提供できる。

2.2 テキスト含意認識研究での文章読解タスク

第一回の RTE チャレンジ[1]では、情報検索や質問応答タスクとともに RC 応用を考慮したテストデータが作成されている。新聞記事のテキストに対して、高校生レベルの読解力で包含関係を判定できる文が作成され、仮説(H)とされている。

2 回目以降のチャレンジでは、テストデータ作成の応用先として RC 応用は想定されていない。一方、チャレンジで扱うテキストは1文程度と短いものであったが、RTE-3[3]からはテストデータの一部に段落レベルのテキストが含まれるようになっている。テキスト中の複数文を考慮することが必要となっており、文章読解へ応用も可能と考えられる。

2.3 文章読解タスクの設定

読解力に関する国語・語学テストには様々な形式があるため、本稿では、RTE をより直接評価できるようなタスクの設定を試みる。まず、読解力の評価において RTE に関係が深い側面を明確化し、それを直接的にテストできる問題、回答形式について論じる。

[読解力]

読解力の定義については、OECD による国際的な生徒の学習到達度調査（Programme for International Student Assessment, PISA）のための検討が参考になる。2003 年の PISA 調査のための検討[5]では、読解の仮定として 5 つの側面があると主張しているが、調査結果に基づく文部科学省での検討では、それを 4 点に集約している。

PISA 型「読解力」の特徴（文部科学省[6]）

情報の取り出し:	テキストの中の事実を切り取り、言語化・図式化
テキストの解釈:	書かれた情報から推論・比較して意味を理解
熟考・評価:	書かれた情報を自らの知識や経験に位置づけて理解・評価
論述:	内容、構造、形式、表現法を考慮して記述

この内、「テキストの解釈」が最も RTE と関連性が高いと考えられる。一方、「情報の取り出しは」、情報検索や情報抽出に関連が深い。そのため、複数段落からなる

長文のテキストを問題文章とした場合には、RTE 以外の検討が含まれてしまう恐れがあるので、内容としてはひとまとまりの段落程度の文章を扱うことで RTE をより直接的に評価できると考えられる。

一方、「熟考・評価」は学習者の意見があることを前提とするものであり、これを考慮することは検討のモデルを複雑化する恐れがある。さらに「論述」は内容を超えたテキストの表現なども扱うものであり、タスクで考慮する必要は当面ないと思われる。

以上の考察に基づき本稿では、RTE に関わる読解力として、

「段落レベルの分量の問題文章に対して、様々な解釈と比較できる能力」

とする。

[問題形式]

上記の読解力を計る問題形式として、複数の解釈が与えられたとき、それぞれの良否を解答する選択式の問題形式を採用する。最近のセンター試験の国語(現代)の問題でも半数の質問がこの形式であり、読解力を評価するために有効な問題形式であると考えられる。

選択肢の解釈の良否については、正答(選択肢の仮説の内容を問題文章が包含する)の個数が制限されている場合と制限されていない場合がある。RC タスクをシステム処理する際に、正答の個数がシステムに入力される場合は、その情報を有効活用することで解答精度を高めることができる。その場合 RTE の尺度は相対尺度、絶対尺度それぞれに扱われる。本稿では RTE の適応性をできるだけ幅広く検討することを目的としているので、正答個数については、固定/不定の両方を扱う。

このような設定に基づいて作成した RC のテスト問題について、次章で説明する

3. 問題作成

3.1 問題文

国語のテストでは、問題文として新聞記事とともに、小説や評論、物語も用いられている。これらは、新聞記事と比べて口語表現や省略表現が多い。また、小説や評論文では、小説独自の世界観や作者の主張や嗜好に基づいて書かれているため、事件等の事実の記述を中心とする新聞記事と内容が異なる場合がある。一方、規模の面から考えた場合、近年、ブログや掲示板などの個人発信型のインターネットコンテンツが急速に増加しており、その情報処理技術が望まれている。

このような状況を鑑み本稿では、京大と共同作成した KNB コーパス(Kyoto-University and NTT Blog コーパス)[7]の京都観光に関する 91 記事を用いた。先に述べた通り、段落レベルの記事を問題文章とする方針としたため、その中から 15 文以上が含まれる 43 記事を抽出し、その内の 10 記事を方式評価のための評価データとした。残り 33 記事について、適用方式検討のための分析データの問題文章とした。含まれる文数の平均は 21.15、単語数は 357.5(延べ)、164.4(異なり)であった。

3.2 問題作成

上記 33 件のブログ記事それぞれについて、問題を作成した。質問としては、次の2種類を設定した。

「次のブログ記事を読んで、1～4の説明文から適当なものを1つ選べ。」(正解数固定型)
「次のブログ記事を読んで、1～4の説明文が適当であるか、それぞれ答えよ。」(正解数不定型)

2つの問題で利用する解釈の選択肢とは1文で表現し、テキスト中の単語をできるだけ使うようにした。但し、主語として記事の「著者」が省略されていることが多いので、問題文中にない場合でも使用した。

4 つの正しい解釈を表す文を作成し、その内の3つについては、含まれる単語の一部を文章中で現れる意味が近い他の単語と置き換え、問題文章の内容を包含しないか、包含するか未知である解釈を作成した。作成されたデータの一例を下記に挙げる。

表1 問題の一例

[京都観光]大文字
先日、大文字のある山を登ってきました。自転車で銀閣寺よりやや上の辺りまで行って、そこから歩いて大文字のある場所。最初は流れる川沿いを気持ち良く歩いていたのですが、川を離れてからは階段、ひたすら階段と上り坂を繰り返してた気がします。どこまで歩けばいいのか、途中何度かそう思いながら、たまに引き返そうかともっていくうち、やっと、それらしき場所までたどり着きました。「あと少し」と思い歩いていくと、遂に、大文字に到着！「大」の字はいくつかのブロックから構成されていました。松明を掲げるのかな？そこからは、京都市の景色が一望できます。私のアパートのベランダからは大文字が見えるのですが、逆に大文字側から見て～とか(笑)しばらくそこに居ました。そろそろ空が暗くなり始めた頃。さらに上へ行けるみたいでしたが、さすがにそこは断念して、引き返しました。暗い中、たどってきた階段を下るのは少し怖かったです(汗)川が見えたとき、安心感がありました。あと少しで私の世界へ戻れる！また機会があれば、もう一度のぼってみたいです。

	選択肢(H)	正解(含意)
1	著者は銀閣寺に行った。	false
2	大文字のある場所までは自転車で行くことができる。	false
3	大文字のある場所まで著者は階段を登った。	true
4	著者は、大文字から著者のアパートを見ることができた。	false

得られた選択肢は平均 11.7 語、異なりの語彙数は平均 11.4 であった。

4. 適用可能性検討

4.1. RTE 方式

複数文からなるテキストに RTE を適用した場合の傾向を明らかにすることが本稿の主旨であるために、単純な bag-of-words モデルをベースとして問題文章や個々の文と仮説である説明文の RTE スコアを求め、それらの組み合わせの有効性を検証した。

RTE-1 チャレンジで提案されたモデルでベースラインとされている Pérez らの方法[8]を単純化して適用した。

(1) テキスト、仮説から標準型の単語列を抽出 (juman を使用)
(2) テキストを test、仮説を reference と見なして BLEU スコアを計算
(3) BLEU スコアをテキスト含意の尺度として用いる:
正解数固定型 RC: 値が大きな上位一定数を回答
回答数不定型 RC: 一定値以上の選択肢を回答

BLEU(Bilingual Evaluation Understudy)スコアとは、Papineni らによって提唱された機械翻訳結果の評価尺度である[9]。参考の翻訳(reference)との比較を複数の単語 N-gram (N=1,2,3,4)の一致度から計算する。完全に一致する場合は最大値の1となり、スコアが大きな程、品質が高いとされている。Pérez らはこれを TRUE/FALSE が半数ずつ含まれる RTE-1 のテストデータに適用し、約5割の精度となったことを報告している[8]。

RTE-1 でのテストデータと本稿で扱うテストデータでは、ソースとなるテキストの性質や TRUE/FALSE の比率が異なるため、N-gram の最大値を1から4まで変えてスコア計算をした。実際の計算には、NTCIR7 のプログラムを利用した(NTCIR7 Scoring tools for Patent Translation task http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/) 。

BLEU スコアを利用し、以下の3種類の RTE スコアを計算した。

- テキストの BLEU スコア [normal]
- 各文の BLEU スコア平均値 [average]
- 各文の BLEU スコアの最大値 [max]

normal は、テキストの長さが RTE スコアに依存しないと仮定して通常の方法を適用することに相当する。また、average と max は、テキストを構成する各文を単位としたものであり、average は、問題文章中の各文が含意判定に貢献すると仮定することに相当する。一方 max は、テキストからの「情報の取り出し」が必要と考えるものであり、問題文中の1文で含意判定が可能であると仮定することに相当する。

4.2. 実験

実験では、解答数固定型について3つの方法を比較し、最適な方法についてその RTE スコアが含意の評価値となるかについて確認した。

表1に方法の比較結果を示す。数値は、33 問題それぞれについて選択肢との RTE スコアを計算し、その最大値を取る選択肢を回答とした時の正答率の平均である。4 つの選択肢から 1 つを選ぶので、ランダムに回答した時の正答率 0.25 が評価値のベースラインとなる。

3つの方法の中では、複数文全体をテキストとみなす normal が最も高い数値を示している。一方、1文のみを選択する max では、正答率平均値がベースライン以下となっているため、テキストを構成する複数文それぞれから RTE スコアを計算する方法が有効であると示唆される。また、比較する N-gram の最大値としては、N=3 の時に正答率平均が最大となっている。選択肢を構成する単語としては問題文章の単語をできるだけ使うようにしたため、N=1 の時には包含しない仮説も誤って一致してしまう恐れがある。一方 N=4 の値が下がっているのは、仮説がテキストに比べて短いために、4-gram で一致する頻度が低い一方、一致した時の BLEU スコアの重み付けが高いために、長い単語列での偶然の一致が強く影響した可能性がある。

表2 RC(1択)での評価結果

method	averaged precision (n=33)			
	1-gram	1:2-gram	1:3-gram	1:4-gram
normal	0.333	0.333	0.364	0.273
average	0.303	0.303	0.212	0.182
max	0.182	0.212	0.182	0.182

次に、テストデータを正答数不定型の問題とみなしたときに、RTE スコアが含意の真偽を与える尺度となるか確認した(表3)。

TOTAL は、33 問の全選択肢 132 件、TRUE は、その内含意すると人手で判定した選択肢 33 件、FALSE は、含意しないと判定した選択肢 99 件(1問あたり3件)の normal の RTE スコアの平均値である。それぞれの最大 N-gram について、TRUE の RTE スコア平均値が FALSE より高くなっているが、有意の差ではなかった。そのため、真偽を判定するしきい値を変化させて正答率の変化を調べたが、全ての選択肢を FLASE と回答した場合の正答率 0.75 を超える結果は得られなかった(図1)。

本稿では、RTE の適用可能性を検討するために単純な方法を用いているためにこのような結果となった可能性がある。そのため、構文情報や文間の関連情報を利用するような高度な RTE 方式の適用により、改善される可能性がある。

表3 正答に対する RTE スコア比較

	number of samples	averaged RTE Score			
		1-gram	1:2-gram	1:3-gram	1:4-gram
TRUE	33	0.0328	0.0206	0.0117	0.0069
FALSE	99	0.0311	0.0192	0.0106	0.0049
TOTAL	132	0.0315	0.0196	0.0109	0.0054

4.3. 考察

正答数固定型の RC で最も正答率が高かった、問題文章全体を1つのテキストと考えて RTE スコアを計算する方法(normal、N=3)で正しく回答できなかった 21 事例についてその原因を分類した。原因として構文レベルの

比較ができていない事例が7件あった。スコア計算は単語列の比較のみに基づく単純な方法であったので、これらの事例は、構文解析結果を用いた含意認識方法の適用（例えば、文献[10]）で精度改善が期待される。

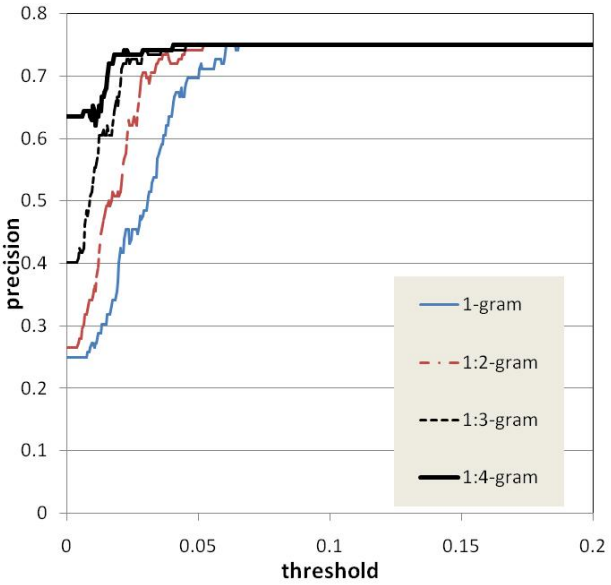

図 1 threshold に対する正答率の変化

一方、残りの21件はそれより高度な原因に起因していた。その一部は、「出かけてはどうだろう」→「勧めている」や「行こうとしたが断念した」→「行かなかった」という文内での言い換えに対応していないので RTE スコアが低かった。また、関連が低い文に遮られた複数文が仮説に対する含意形成に寄与しているためにスコアが低くなってしまった例も含まれていた。テキスト中で関連の高い文をまとめる、あるいは不足する情報を補うような手続きが必要であると予想される。

また、これらの誤解等の一部で、TRUE となる選択肢に対してある程度の RTE スコアを与えているが、FALSE となる選択肢により高い RTE スコアを与えているものが3割程度見かけられた。このような“ひっかけ”の選択肢に対しても、高度な RTE 方式を適用することでスコアが相対的に低下すると考えられるために、正答率の向上が期待される。

5. おわりに

本稿では、テキスト含意認識の適応領域を拡張することを目的として、複数文からなるブログに対して関連あ解釈文を選択肢を仮説として用意し、包含関係の真偽を問う読解タスクにRTEが有効であるかについて検証を行った

その結果、bag-of-words の単純な方式であっても正解数が決められているような問題については、チャンスレベル以上の正答を与えるので応用可能であることがわかった。数文のテキストについても段落レベル程度であれば、それをひとまとめに扱う方法がテキストの個別の文の RTE スコアを総合する方法よりも有効であることを確認し、正しく回答出来ない事例について簡単な分析を行った。一方、RTE スコアのみから含意の真偽を判定する問題については不十分であることがわかった。今後は、得られた知見を活用して複数文の RTE 方式の検討を進める予定である。

参考文献

[1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Qui˜nonero-Candela et al., editor, MLCW 2005, LNAI Vol. 3944, pp. 177-190, Springer-Verlag, 2006.

[2] Lynette Hirschman, Marc Light, Eric Breck and John D. Burger. Deep READ: a Reading Comprehension system. *In Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 325-332, 1999.

[3] 関根聡，齋藤真実，岡田美江，井佐原均．小学2年生の問題を解く－電脳優子2年生・概要．言語処理学会 第11回年次大会，pp.1068-1071, 2005.

[4] 石岡 恒憲．小論文自動採点．電子情報通信学会論文誌 Vol.92, No.12, pp.1036-1040, 2009.

[5] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The Third PASCAL Recognizing Textual Entailment Challenge, Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pp. 1-9, 2007.

[5] OECE. PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills, 2003.

[6] 文部科学省．読解力向上プログラム．http://www.mext.go.jp/a_menu/shotou/gakuryoku/siryo/05122201/014/005.htm. 2005.

[7] 橋本 力，黒橋 禎夫，河原 大輔，新里 圭司，永田 昌明． 構文・照応・評判情報つきブログコーパスの構築．言語処理学会 第15回年次大会，pp.614-617, 2009.

[8] Diana Pérez and Enrique Alfonseca. Using Bleu-like Algorithms for the Automatic Recognition of Entailmen. In Qui˜nonero-Candela et al., editor, MLCW 2005, LNAI Vol. 3944, pp. 191-204, Springer-Verlag, 2006.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Proc. of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311-318, 2001.

[10] 小谷通隆，柴田知秀，中田貴之，黒橋禎夫．日本語textual entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識．言語処理学会第14回年次大会，pp. 1140-1143, 2008.