

大学入試センター試験を題材とした含意関係認識技術の評価

宮尾祐介
国立情報学研究所
yusuke@nii.ac.jp

嶋英樹
カーネギーメロン大学言語技術研究所
hideki@cs.cmu.edu

金山博
日本アイ・ビー・エム株式会社東京基礎研究所
hkana@jp.ibm.com

三田村照子
カーネギーメロン大学言語技術研究所
teruko+@cs.cmu.edu

1 はじめに

含意関係認識とは、二つの文 t_1, t_2 が与えられ、その間に含意関係が成り立つかどうかを判定する自然言語処理タスクである。含意関係とは、 t_1 の言明を真とした時、 t_2 の言明も真であると推論できるという関係である。含意関係認識は文レベルの意味的同値性を判定することを目指しており、様々な意味解析技術を統合する枠組みであると同時に、質問応答や自動要約などのアプリケーションの高精度化に寄与する重要な技術であるため、近年さかんに研究が行われている [1]。

本研究では、大学入試センター試験¹から正誤を問う設問を抽出し、含意関係認識の評価データを開発した。正誤を問う設問とは、選択肢として文がいくつか与えられ、その中から正しい言明（あるいは誤った言明）を一つ選ぶものである。正しい言明に対しては Wikipedia にその根拠となるテキストが存在すると仮定し、その文を t_1 、選択肢の文を t_2 として文ペアを作成した。一方、誤った言明に対しては、Wikipedia から矛盾する文あるいはキーワードが重なっているが含意関係の無い文を抽出して t_1 とした。

本研究で開発した含意関係認識評価データは、評価型ワークショップ NTCIR-9 の RITE タスクにおいて提供され [4]²、6 チーム 16 システムが本データを利用した。本稿では、データ開発方法の詳細と、参加システムの評価結果について報告する。

2 正誤を問う問題と含意関係認識

大学入試センター試験では、選択肢として文がいくつか与えられ、その正誤を問う問題が多く出題される。例えば、図 1 の問題では、各選択肢の言明が歴史的事実として正しいかどうかを判断することが要求される。したがって、教科書や Wikipedia など事実を記述したテキストには、これらの選択肢の正誤を判断する根拠となるテキストが存在すると考えられる。実際、

下線部 (3) (ギリシア) の地域について述べた文として誤っているものを、次のうちから一つ選べ。

1. 前 5 世紀に、ギリシアの諸ポリスはアケメネス朝ペルシアと戦った。
2. 5 世紀には、ギリシアは西ローマ帝国の支配下にあった。

(2009 年度センター試験 世界史 A)

図 1: 正誤を問う問題の例

...紀元前 5 世紀にアケメネス朝（ペルシア帝国）が地中海世界に進出してくると、各ポリスは同盟を結び、これに勝利した（ペルシア戦争）。...395 年にローマ帝国が東西に分裂した後は、ギリシャ地域は東ローマ帝国に属した。...

図 2: 図 1 に関連する Wikipedia の記述

Wikipedia には図 2 に示す記述が存在し、図 1 の問題では 1 が正しく、2 が誤りであることが分かる³。

このような問題回答のプロセスは、根拠となるテキストが選択肢の文を含意するかどうかを判定する問題に帰着することができる。上記の例からは、Wikipedia の文と選択肢の文をペアにして図 3 のような含意関係認識評価データを作成することができる。逆に、このデータに対して含意関係が正しく判定できれば、試験問題に対して正答することができる。試験問題への回答という実世界タスクを対象とすることで、知識を用いて問題に回答する際に必要な意味理解能力を直接評価・分析することができる。また、含意関係認識の性能を試験の正答率という直感的に分かりやすい指標で評価することができるという利点がある。

3 含意関係認識評価データの開発

3.1 データ開発のながれ

データ開発は、以下の 3 つのページをウェブブラウザで閲覧しながら作業を行った。

試験問題ファイル 試験問題を閲覧する。選択肢と正

³題意は「誤っているものを選ぶ」なので問題の正解は 2 である。

¹<http://www.dnc.ac.jp/>

²<http://artigas.lti.cs.cmu.edu/rite/Main.Page>

ラベル: Y
t1: 紀元前 5 世紀にアケメネス朝（ペルシア帝国）が地中海世界に進出してくると、ギリシャの各ポリスは同盟を結び、これに勝利した（ペルシア戦争）。
t2: 前 5 世紀に、ギリシアの諸ポリスはアケメネス朝ペルシアと戦った。
ラベル: N
t1: 395 年にローマ帝国が東西に分裂した後は、ギリシャ地域は東ローマ帝国に属した。
t2: 5 世紀には、ギリシアは西ローマ帝国の支配下にあった。

図 3: 含意関係認識評価データ

答が見やすいように、正答の選択肢は赤色、誤答の選択肢は黄色で表示される。

Wikipedia 検索ツール Wikipedia の全文検索を行う。全文検索エンジン SEPIA を使用した。

作業シート 作成したデータを記入する。Google docs のスプレッドシートを利用した。

データ開発手順は以下のとおりである。

1. **データ作成適否の判断** 与えられた試験問題ファイルの各設問について、含意関係認識のデータを作るのに適しているかどうかを判断する。データを作るのに適さない問題は捨てて次の問題に移る。
2. **含意関係ラベルの付与** 各選択肢に対し、含意関係ラベル（Y または N）を決定し、作業シートに記入する。含意関係ラベルは、その言明が真か偽かによって決定されるため、問題の題意、および当該選択肢が正答か否かから一意に定まる。正しい言明を選ぶ問題では、正答はラベル Y、誤答はラベル N となる。誤っているものを選ぶ問題の場合は逆となる。
3. **t2 の作成** 各選択肢の文を作業シートにコピーし、t2 とする。この時、そのままの文では意味を為さない場合は、情報を足すなど適宜編集を行う。
4. **関連するテキストの検索** 各 t2 について、その正誤を判断する根拠となる文章を Wikipedia から検索し、作業シートにコピーする。関連する文が見つからない場合は、その選択肢は捨てる。
5. **t1 の作成** 検索した文章を要約・編集して一文にし、作業シートに記入する。

以下では、ステップ 1 とステップ 4 について、およびステップ 3 と 5 の編集作業について詳述する。

3.2 データ作成適否の判断

本研究は、Wikipedia のテキストから含意関係で正答を導くことができる設問を対象としている。実際の試験問題にはこの条件から外れた設問も存在するため、まず問題のタイプを分類し、含意関係認識に帰着できない設問はデータ作成に使用しなかった。予備調査に

おいて、今回対象とした科目では主に以下の 4 つの問題タイプが見られた。

含意関係認識に帰着できる問題 選択肢が文であり、その正誤を問う問題は、含意関係認識に帰着できると判断してデータ作成の対象とした。

選択肢が文ではない問題 NTCIR RITE は文どうしの含意関係認識を対象としているため、選択肢が文ではない設問は対象外とした。典型例として、ある出来事の起きた年代を問う問題がある。そのような問題は、含意関係認識ではなく factoid 型質問応答に帰着すべきものである。

非テキスト情報を参照している場合 問題中で与えられた非テキスト情報（図、絵、グラフ、表など）を参照して初めて解くことができる問題は、テキストの含意関係に帰着できないため対象外とせざるを得ない。

読解問題 問題中で文章が与えられその説明や解釈を問う問題は、Wikipedia から正答を導くことができない。読解問題は、与えられた文章を t1 とすることで含意関係認識に帰着できる可能性があるが [2, 6]、今回は対象外とした。

3.3 関連するテキストの検索

含意関係ラベルと t2 に基づき、t2 の正誤を判断する根拠となるテキストを Wikipedia から検索する。含意関係ラベルが Y の時は、t2 が正しい言明であると判断する根拠となるテキスト、すなわち t2 を含意するテキストを検索する。Wikipedia 中にはそのようなテキストが複数存在する場合があるが、そのうち任意の一つを採用する。また、t2 を含意するために複数の文が必要である場合には、必要な文を全て採用する。

含意関係ラベルが N の場合は、Wikipedia 中のどのテキストからも t2 が含意されないことになる。これには 2 つのケースが考えられる。一つは、Wikipedia に t2 と矛盾するテキストが存在する場合、もう一つは、矛盾するテキストも含意するテキストも存在しない場合である。どちらのケースが当てはまるかは事前に分からないため、本研究では以下のアプローチを採った。

- t2 と矛盾するテキストを検索する。そのようなテキストが見つかったら、それを採用する。
- 矛盾するテキストが見つからない場合は、t2 に含まれるキーワードが多く含まれるテキストを検索して採用する。

3.4 t1, t2 の編集

t1, t2 は基本的には選択肢・Wikipedia の文をそのまま抽出する。しかし、元の文は文脈の中に置かれているため、そのまま抽出すると正誤を判断すべき言明として意味をなさない場合がある。例えば、図 2 のテキ

表 1: 含意関係認識の精度

	精度	適合率	再現率	F 値
IBM-1	0.722	0.696	0.569	0.626
IBM-2	0.674	0.637	0.475	0.544
IBM-3	0.584	0.489	0.376	0.425
JAIST-1	0.622	0.537	0.564	0.550
JAIST-2	0.652	0.596	0.464	0.522
JAIST-3	0.652	0.596	0.464	0.522
JUCS	0.520	0.426	0.492	0.456
Kyoto-1	0.593	0.600	0.017	0.032
Kyoto-2	0.656	0.620	0.414	0.497
Kyoto-3	0.656	0.620	0.414	0.497
LTI-1	0.602	0.586	0.094	0.162
LTI-2	0.654	0.593	0.492	0.538
LTI-3	0.667	0.639	0.431	0.515
TU-1	0.649	0.676	0.276	0.392
TU-2	0.115	0.692	0.099	0.174
TU-3	0.113	0.680	0.094	0.165
TU-2*	0.718	0.692	0.600	0.643
TU-3*	0.704	0.680	0.567	0.618
全て Y	0.410	0.410	1.000	0.581
全て N	0.590	0.000	0.000	0.000
ランダム	0.500	0.410	0.500	0.450

表 2: 各科目における正答数と正答率

	世 A	世 B	日 A	日 B	現社	政経	合計	正答率
IBM-1	15	13	6	6	11	10	61	0.56
IBM-2	11	11	6	5	10	10	53	0.49
IBM-3	8	10	3	1	9	6	37	0.34
JAIST-1	5	10	3	2	11	12	43	0.40
JAIST-2	5	9	7	4	10	5	40	0.37
JAIST-3	4	13	7	7	10	10	51	0.47
JUCS	4	13	3	1	8	4	33	0.31
Kyoto-1	6	6	3	2	4	3	24	0.22
Kyoto-2	6	14	5	5	12	10	52	0.48
Kyoto-3	6	14	5	5	12	10	52	0.48
LTI-1	5	8	1	2	4	2	22	0.20
LTI-2	7	12	3	5	7	6	40	0.37
LTI-3	7	12	10	4	11	7	51	0.47
TU-1	9	8	2	4	3	4	30	0.28
ランダム	6	6	4	3	5	3	27	0.25
対象問題数	24	24	16	12	20	12	108	
全問題数	33	36	36	24	36	38	203	

ストでは、「各ポリス」がギリシャのものであることは明示されておらず、文外参照によってそれが分かる。このような場合は、抽出した文を適切に編集することとした。図 3 に示したデータでは、*t1* の文に「ギリシャの」という句が追加されている。

また、*t2* の根拠となるテキストが必ずしも一文ではなく複数文に分かれて記述されている場合がある。RITE は一文どうしの含意関係を判定するタスク設定であるため、このような場合は関連箇所をまとめて一文に要約することとした。

4 含意関係認識システムの評価

センター試験問題の予備分析から、世界史 A (世 A)、B (世 B)、日本史 A (日 A)、B (日 B)、現代社会 (現社)、政治・経済 (政経) の 6 科目は含意関係認識に帰着できる問題が多いと判断し、これらを本研究の対象データとした⁴。本節の評価実験では、2009 年度の試験問題から作成したデータ (499 ペア) を開発用データ、2007 年度の問題から作成したデータを最終評価用データ (442 ペア) とした。データ作成には計 315.5 人時を要した。

本研究で開発した含意関係認識データは、NTCIR-9 RITE の大学入試サブタスクにおいて提供された。国内外より 6 チームが参加し、各チーム最大 3 システムまで結果を提出できるため、最終的に 16 システムが本データによる評価対象となった。参加チームは日本 IBM (IBM)、北陸先端大学院大学 (JAIST)、Jadavpur University (JUCS)、京都大学 (Kyoto)、カーネギーメロン大学 (LTI)、東北大学 (TU) である。参加システムの詳細は、NTCIR-9 予稿集を参照されたい [3]。

⁴日本史 A、B は一部共通問題が出題されるため、日本史 B のうち日本史 A と共通の問題 (各年度 12 問ずつ) は対象外とした。

表 1 に含意関係認識の精度を示す⁵。適合率、再現率、F 値はラベル Y に対する値である。ベースラインとして、全てラベル Y を出力、全て N を出力、ランダムにラベルを出力するシステムの精度を示す。多くのシステムがベースラインを大きく超える精度を達成しているが、F 値で見ると IBM-1 以外はベースラインに達していないことが分かる。

本研究の枠組みでは、これらのシステムの性能を試験の正答率という形で評価することができる。試験問題は複数の選択肢から一つを選択する形式のため、システムが出力したラベルとその確信度を用いて問題に対する回答を生成した。正しい言明を選ぶ問題の場合は、システムが 1 つの選択肢のみにラベル Y を出力した場合はそれを回答とし、複数の選択肢に Y を出力した場合はそのうち確信度が最も大きいもの、全て N と出力した場合は確信度が最も低いものを回答とした。誤った言明を選ぶ問題の場合は逆となる。

各システムを試験問題に対する正答率で評価した結果を表 2 に示す。これによると、高精度なシステムは 5 割程度の正答率を達成しており、ベースラインをはるかに上回ることが示された。表 1 の結果ではベースラインに対する優位性があまり明らかではないが、試験問題に対する正答率を評価することで、現在の含意関係認識技術は応用アプリケーションにおいて有意な寄与ができる性能を持つことが明らかとなった。

5 データ開発の問題点

今回のデータ開発において目立った問題点を報告する。

⁵TU-2、TU-3 はデータの一部を対象としてラベルを出力しているため、ラベルを出力しなかったペアを不正解とした場合 (TU-2、TU-3) と精度計算から除外した場合 (TU-2*, TU-3*) の評価結果を掲載した。ラベルが無いペアからは試験問題の回答を生成できないため、表 2 の評価は行わなかった。

年代と出来事の正しい組合せを選択する問題 センター試験では、年代が指定され、その年代に起きた出来事を選択する問題がある。この時、同じ出来事を記述したテキストを t_1 として採用してしまうと、誤った言明では必然的に年代が異なる。したがって、 t_1 と t_2 の年代の一致を見るだけでラベルが予測できてしまうペアが少なからず存在した。実際、IBM のシステムでは、時間表現の一致を見る素性を入れたところ含意関係認識精度が 5 ポイント向上したと報告されている。

Wikipedia の文章の編集 正誤判断の根拠となるテキストが複数文に渡る場合、本研究の方針では関連する箇所をまとめて一文に要約することとしたため、大幅な編集が行われたケースが見られた。このように作成したデータでは、含意関係が判定しやすく編集されている可能性がある。特に、政治・経済や現代社会では歴史的事実ではなく抽象概念に関する言明が多く、根拠となる文が複数に渡るケースが多く見られた。

6 関連研究

現在までの英語や日本語の含意関係認識評価データ [2, 7, 5] では、質問応答や複数文書要約などの応用に組み込まれる要素技術を想定し、それらの使用事例をシミュレートして評価データが作成されている。本研究では言明の正誤判断と含意関係認識が直接関係していることに着目しており、含意関係認識の新たな評価枠組みと考えられる。試験問題を用いるため、誤答の選択肢を利用することで負例を人為的に作成・抽出する必要がなく、また含意関係ラベルが自動的に決定できるという利点がある。さらに、含意関係認識の性能を試験の正答率という直感的に分かりやすい指標で評価することができる。

読解問題は含意関係認識を直接評価する仕組みととらえることができる [2, 6]。人間のためのタスク設定を利用する点が共通しているが、本研究の枠組みでは、試験が測定しようとしている能力と言語処理タスクとして評価している能力が本質的に異なる。本研究が対象とした問題は人間の知識の測定を目的としており、知識から各言明が導けることは自明と想定している。したがって、本研究で評価している含意関係認識は人間にとっては自明な場合が多い。つまり、人間にとっては非常に簡単な認識問題を、言語処理システムがどれだけ再現できるかを試していると言える。

7 おわりに

本稿では、大学入試センター試験の正誤を問う設問を利用し、含意関係認識技術の評価する手法について報告した。正誤を問う問題は、Wikipedia などのテキストから選択肢の文を含意できるかどうかを判定するタスクととらえられる。そこで、センター試験の選択肢と Wikipedia 中の関連するテキストをペアとし、含

意関係認識データを開発した。世界史 A, B, 日本史 A, B, 現代社会, 政治・経済の計 6 科目の 2009 年度、2007 年度の試験問題を使用し、計 941 文ペアのデータを作成した⁶。本データは NTCIR-9 RITE 大学入試サブタスクにおいて提供され、6 チーム 16 システムが本データを用いて評価された。評価結果から、高精度のシステムは 5 割以上の正答率で試験問題に回答することができ、ベースラインを大きく上回ることが示された。本研究は含意関係認識が知識を問う問題への回答に適用できることを示しており、テキストから知識を獲得する有用な手法であることを示唆している。

今回のタスク設定では、根拠となる文章の検索や文外参照に関する編集は作業者が人手で行っており、試験問題を完全に自動で解いているわけではない。より現実の設定に近づけるため、次回の NTCIR RITE では、 t_1 を明示的に与えず、与えられたテキスト集合から t_2 を導くことができるかどうかを判定するタスク設定を検討している。

参考文献

- [1] I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, Vol. 38, pp. 135–187, 2010.
- [2] I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, Vol. 3944 of *LNCS*. 2006.
- [3] D. Ishikawa, H. Joho, N. Kando, T. Kato, T. Sakai, M. Sugimoto, E. Sumita, T. Akiba, S. Geva, F. Gey, I. Goto, B. Lu, H. Shima, E. Tang, and A. Trotman, editors. *Proceedings of the 9th NTCIR Workshop Meeting*, 2011.
- [4] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of NTCIR-9 RITE: Recognizing inference in text. In *Proc. NTCIR-9*, 2011.
- [5] 宇高邦弘, 山本和英. 複数の客観的手法を用いたテキスト含意認識評価セットの構築. 言語処理学会第 17 回年次大会, 2011.
- [6] 笠原要, 平博順, 永田昌明, 柴田知秀, 黒橋禎夫. 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. 言語処理学会第 17 回年次大会, 2011.
- [7] 小谷通隆, 柴田和秀, 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, 2008.

⁶本データは、NTCIR より公開予定である。