含意要因となるテキスト中の表現と仮説の対を用いたテキスト含意認識

宇高 邦弘, 山本 和英
長岡技術科学大学 電気系
E-mail:{udaka,yamamoto}@jnlp.org

1 はじめに

テキスト含意認識とは、テキスト (text,T) および仮説 (hypothesis,H) と呼ばれる言語表現の対が与えられた場合に仮説が持つ文章の意味をテキストが含み得るか否かを自動的に判定するタスクを指す。以下に実際のテキスト含意認識の例を示す。

例 **1)** テキスト含意認識
T:アフリカの大統領はガーナのアクラで 2 日間のサミットを開きました。
H:アクラはガーナに位置しています。
含意判定:含意

テキスト含意認識は近年活発に研究されているタスクであり、過去の海外ワークショップでは第 1 回から第 3 回までの含意認識評価セットが公開されている [1][2][3]。

テキスト含意認識は自然言語処理における幅広いタスクにおいて様々な役割を果たす。例えば機械翻訳においてはテキストと仮説を翻訳前と翻訳後として考えることで翻訳精度の指標として使用でき、質問応答ではテキストを質問文とし仮説を質問の回答候補とすることで、回答を得る手法として応用することが可能である。過去に海外で公開された評価セットは各自然言語処理タスクの手法を用いて文章を加工したものを仮説とすることで構築されている。

上記の海外で公開された評価セットを観察すると、含意関係を持つテキストと仮説の対 (以下、T-H 対) の中にはテキスト中の一部の表現から仮説との含意関係を認識可能な場合がある。例えば上記のテキスト含意認識の例は第 1 回ワークショップで公開された評価セットに存在する T-H 対であり、テキスト中の「ガーナのアクラ」という表現から、仮説が正しいと認識可能である。このように過去に海外で公開された評価セット中の T-H 対について含意関係の有無を決める表現と仮説の対を抽出することで、実際の自然言語処理タスクに存在する含意関係を持つ表現対を獲得でき、これらを用いることで含意関係認識が可能であると考える。

本稿では過去に公開された海外のテキスト含意認識評価セット中に存在する含意関係を持つ対から抽出した含意要因となるテキスト中の表現と仮説の対について分析する。加えて抽出した含意要因となるテキスト中の表現と仮説の対を用いた日本語によるテキスト含意認識手法を提案する。具体的にはどのような表現を持つ T-H 対から含意要因となるテキスト中の表現と仮説の対を抽出することができ、抽出した対を用いて含意関係認識を行うことでどの程度の被覆率を持つか調べた。その結果本手法では 6 割の被覆率であり、抽出した対の多くが有効であった。

2 関連研究

1 節で述べた海外のテキスト含意認識の評価セットは Dagan et al.[4] の手法を元に構築、改良されている。Dagan et al. は新聞コーパスに存在する文をテキストとして使用し、質問応答、情報抽出、情報検索、複数文書要約、換言獲得、機械翻訳、文書読解の手法を用いてテキストを加工したものを仮説とすることで評価セットを構築している。これにより彼らの構築した評価セットについてシステムが正しく含意関係を認識可能ならば直接的に各自然言語処理タスクに応用可能であると言える。

日本におけるテキスト含意認識の評価セットおよびシステムを構築する研究としては、小谷ら [5,6] の研究がある。小谷らは推論要因を包含、語彙（体言）、語彙（用言）、構文、推論の 5 つに分類し、それぞれに下位分類を設け、1 つもしくは 2 つの要因から含意関係を認識することが可能な事例を作成することで網羅性の高い評価セットを構築した。この評価セットは一般公開されている。また小谷らは推論パターンを国語辞典などから自動獲得した類義表現を用いて構築することで含意認識を行った。加えて言語表現を述語項構造に正規化し、述語項構造を単位としてテキストと仮説間の一致を見ることで推論認識を行うシステムを構築している。述語や格要素の上位下位関係の認識や、否定表現および仮定法過去を用いた表現に対する例外処理も行っている。このシステムについて、構築した評価セットにおける節、補文、主語の変換、強調構文に属する事例を入力として用い、推論認識を行っている。しかし小谷らの構築した評価セットは 1〜2 つの要因のみから推論できる事例しか含まないため、各自然言語処理タスクにおける推論を全て扱っているとは言えない。

我々は各自然言語処理タスクに存在する推論を含む海外で公開されたテキスト含意認識の評価セットから含意要因となる表現と仮説の対を抽出し、抽出した対が海外の評価セットにおいてどの程度有効かを検証した。

3 含意要因となるテキスト中の表現と仮説の対の抽出

3.1 抽出方法

過去に海外のワークショップ (PASCAL1〜PASCAL3) で公開された 評価セット[(1)(2)(3)] 中に存在する含意関係を持つ全 T-H 対を観察し、含意関係を認識する要因となるテキスト中の表現とその表現から推論可能な仮説の対を人手で抽出した。日本におけるテキスト含意認識を行った研究は少なく、このように含意要因を抽出した研究はない。そのため PASCAL1〜PASCAL3 で公開された評価セットは全て英語で書かれているが人手で全て日本語に訳した後に対を抽出した。以下にその一例を示す。

例 **2)** 含意要因となるテキスト中の表現と仮説
T(原文):He also developed a breath test that could detect sulfur compounds emitted from the H. pylori.
H(原文):The H. pylori produces sulphur compounds.
T(訳):また、彼はピロリ菌から出る硫黄化合物を検出できる呼気検査を開発しました。
H(訳):ピロリ菌は硫黄化合物を作り出します。
抽出される対:ピロリ菌から出る硫黄化合物
→ ピロリ菌は硫黄化合物を作り出す

表 1:PASCAL1〜PASCAL3 の評価セットにおける各特徴ごとの対の数

	包含	「の」による名詞の結合	複合名詞	主語の交代	名詞-動詞	内の関係	述語含意	別名	同義語	名詞推論	言った-によると	を含む
PASCAL1	76	77	6	22	28	41	107	6	11	16	0	0
PASCAL2	66	73	22	28	46	47	72	2	6	42	1	0
PASCAL3	49	51	5	21	48	56	48	5	8	36	0	1

例 2 の抽出される対について、矢印の左辺はテキスト側から抽出された含意要因となる表現であり、右辺は仮説である。PASCAL1〜PASCAL3 で公開された評価セットには 2304 の含意関係を持つ T-H 対を含んでおり、個々の T-H 対を人手で観察することで 994 の含意要因となるテキスト中の表現と仮説の対を抽出した。

3.2 抽出した含意要因となるテキスト中の表現と仮説の対の分析

どのような特徴を持つ T-H 対から含意要因となる表現と仮説の対を抽出可能かを分析した。以下に分析結果および個々の特徴を持つ抽出した対の例を示す。

包含　テキスト中に仮説がそのままの形で存在している対である。

含意要因となる表現:アラビア語はコーランの言語であり
仮説:アラビア語はコーランの言語だ

「の」による名詞の結合　テキスト中に存在する<名詞>の<名詞>という表現から仮説の意味を表現している対である。「の」で結びつけられた 2 つの名詞は様々な意味関係を持っている。他にも である、による、にある、での、としての、への などで結びついている 2 つの名詞も、仮説の意味を表現している場合がある。

含意要因となる表現:法律学教授のジェシカ・リトマン
仮説:ジェシカ・リトマンは法律学教授だ

複合名詞　テキスト中に存在する複合名詞が仮説の意味を表している対である。以下の例は オリンピック大会 という表現から 開催される という動詞が推論されるため、後述する名詞-動詞の特徴も持っている。

含意要因となる表現:リレハンメルオリンピック大会
仮説:オリンピック大会はリレハンメルで開催される

主語の交代　テキスト中の表現と仮説の主語と目的語が交代した関係になっている対である。

含意要因となる表現:レノンはマーク・デヴィッド・チャップマンに殺された
仮説:マーク・デヴィッド・チャップマンはレノンを殺した

名詞-動詞　テキスト中に存在する名詞が仮説中の動詞を表している対である。以下の例では 出身 という名詞が仮説中の 生まれた という動詞を表している。

含意要因となる表現:アラバマ州ダンビル出身のジェシー・オーエンス
仮説:ジェシー・オーエンスはアラバマ州ダンビルで生まれた

内の関係　内の関係とは被修飾名詞と連体修飾節中の用言との間に格助詞を補うことで単文を作成できる関係である。[6] テキスト中の被修飾名詞と連体修飾節中の用言との間に格助詞を補ったものが仮説となる場合、その被修飾名詞と連体修飾節は仮説の意味を表している。

含意要因となる表現:バグダッドの中央にあるグリーンゾーン
仮説:グリーンゾーンはバグダッドの中央にある

述語含意　テキスト中の一部の表現における述語と仮説の述語が含意関係にあり、主語、目的語が同じ意味を持つ対である。

含意要因となる表現:1999 年に欧州通貨同盟は設立された
仮説:欧州通貨同盟は 1999 年に始まった

別名　テキスト中のある名詞が別の名詞の同義語であることを示しており、仮説がこれらの名詞について同義関係であることを言及している対である。

含意要因となる表現:BSE いわゆる狂牛病
仮説:狂牛病は BSE の別名だ

同義語　テキスト中の一部の表現と仮説の一部の表現が同義関係にある対である。以下の例では コスト と 費用 が同義関係であり、「の」で名詞が結ばれることで仮説の意味を表現している。

含意要因となる表現:紙のコストの上昇
仮説:紙の費用は上昇している

名詞推論　テキスト中の名詞から仮説中の名詞が推論される対である。以下の例では 未亡人 から 妻 という名詞が推論される。

含意要因となる表現:ジョン・レノンの未亡人のオノヨーコ
仮説:オノヨーコはジョン・レノンの妻だった

言った-によると　「言った」という動詞から「によると」という表現が推論できる対である。

含意要因となる表現:ロシアは世界で最も強力な経済の 1 つであるとプーチン氏は言った
仮説:プーチン氏によるとロシアは世界一強力な経済の 1 つだ

を含む　「を含む」で接続された 2 つの名詞がテキスト中にあり、仮説がこの 2 つの名詞について言及する対である。

含意要因となる表現:Lukoil、Zarubezhneft を含むロシアの大石油会社
仮説:Lukoil と Zarubezhneft はロシアの石油会社だ

以上の特徴を T-H 対が持っていた場合、その T-H 対から含意要因となるテキスト中の表現と仮説の対を抽出することができる。今回抽出した対はこれらの特徴を 1 つ〜最大 3 つまで持っている。

表 1 に、PASCAL1〜PASCAL3 の評価セットにおけるこれらの特徴を持つ対の数を示す。これを見ると、どの評価セットにおいても包含、「の」で結ばれた名詞、述語含意、

表 2:抽出した対を用いた含意認識結果

入力 T-H 対総数	2304
対を抽出した T-H 対総数	1012
対を抽出し 対に照合した T-H 対の数	633
対を抽出できず 対に照合した T-H 対の数	0
対を抽出し 対に照合しなかった T-H 対の数	380
対を抽出できず 対に照合しなかった T-H 対の数	1291

表 3:抽出した対の使用回数

使用回数 0 回	371
使用回数 1 回	612
使用回数 2 回	9
使用回数 3 回	1

内の関係の特徴を持つ対が多い。そのため、これらの特徴を有する T-H 対からは含意要因となるテキスト中の表現と仮説の対を抽出しやすいと考える。

4 含意要因となるテキスト中の表現と仮説の対を用いたテキスト含意認識

4.1 含意認識方法

抽出した対が含意認識においてどの程度有用かを調べるために含意判断を行った。以下に照合の手順を示す。

1. T-H 対及び抽出した含意要因となるテキスト中の表現と仮説の対を $CaboCha^{(4)}$ で構文解析
2. 構文解析結果から文節対を作成。この時に各単語の品詞を確認し内容語のみを残す
3. 入力された T-H 対の仮説から作成した文節対と、抽出した含意要因となる表現と対になる仮説から作成した文節対を照合する
4. 抽出した含意要因となる表現と対になる仮説から作成した文節対が全て T-H 対の仮説から作成した文節対と照合する場合、3 の処理をテキストと含意要因となる表現とで行う
5. 含意要因となる表現から作成した文節対が入力されたテキストから作成した文節対と全て照合した場合、入力された T-H 対が含意していると判定

T-H 対及び抽出した含意要因となるテキスト中の表現と仮説の対を構文解析し内容語のみを残して文節対を作成したのは文中の単語間の意味を保持し、機能語の些細な変化で含意関係を認識出来ない状況を考慮するためである。入力する T-H 対は前途の PASCAL1〜PASCAL3 に含まれる含意関係を持つ 2304 の T-H 対を使用した。残す内容語としては動詞、形容詞、名詞を選択した。

4.2 含意認識結果

表 2 に、今回抽出した含意要因となるテキスト中の表現と仮説の対を用いて含意認識した場合の結果を示す。表 2 を見ると、含意要因となる表現と仮説の対を抽出した T-H 対の約 6 割について抽出した対を用いて正しく含意認識できている。また、含意要因となる表現と仮説の対を抽出出来なかった T-H 対については抽出した全ての対に照合しなかった。反面、含意要因となる表現と仮説の対を抽出できたにもかかわらず抽出したどの対にも照合しなかった T-H 対が 4 割存在する。

表 3 に入力された T-H 対に対する抽出した含意要因となる表現と仮説の対の照合回数を示す。多くの対は 0〜1 の範囲で T-H 対に照合していたが、2 回や 3 回照合される対も稀に存在した。

5 考察

5.1 抽出した含意要因となるテキスト中の表現と仮説の対について

含意要因となるテキスト中の表現と仮説の対を抽出することのできる T-H 対の特徴について 4.2 節で述べたが、全体を通して言える傾向として出現する固有名詞が T と H で共通しているものが多いことが挙げられる。また、表 1 で示したように今回使用した評価セットには包含、「の」による名詞の結合、述語含意、内の関係のような特徴を持つテキスト中の含意要因となる表現と仮説の対が多く存在する。以上から、テキストと仮説が共通する固有名詞を多く含んでいる場合にテキスト中の含意要因となる表現と仮説の対を抽出しやすいと考える。

5.2 含意要因となるテキスト中の表現と仮説の対を用いた含意認識について

4.1 節の表 1 において、含意要因となるテキスト中の表現と仮説の対が抽出できたにもかかわらずどの含意要因となるテキスト中の表現と仮説の対にも照合しなかった T-H 対が 4 割存在した。これらは対の抽出時に含意要因として必要ない情報を取り除いて抽出したことが原因と考えられる。以下に例を示す。

例 **3)** 対が抽出可能であり、対と照合しなかった **T-H** 対
T:ブリタニカ百科事典によると インドネシアは 1 万 3670 の島から構成される 世界一大きい列島国家です。
H:1 万 3670 の島がインドネシアを構成しています。
抽出した対:
インドネシアは 1 万 3670 の島から構成される
→ 1 万 3670 の島がインドネシアを構成している

上の例において、下線部はこの T-H 対の含意関係を認識する上で不必要な表現である。この例のように個々の含意要因となるテキスト中の表現と仮説の対は含意関係を認識する上で不要な情報を除いて抽出した。この場合、T-H 対及び抽出した含意要因となるテキスト中の表現と仮説の対を構文解析するとテキスト中に存在する インドネシアは は 列島国家です に係る。しかし抽出した対のテキスト中の含意要因となる表現を構文解析すると インドネシアは は 構成される に係る。このため文節対を照合した場合、うまく照合出来ない。

また、表 2 から抽出した含意要因となるテキスト中の表現と仮説の対の多くは 1 度しか T-H 対と照合しなかった。これは PASCAL1〜3 の評価セットには多くの名詞、固有名詞、複合名詞が含まれており、照合した場合に固有名詞や複合名詞のままでは一致しないことが多いためである。そのため、抽出した含意要因となるテキスト中の表現と仮説の対はそのままでは抽出元となった T-H 対以外とは照合し辛く、他の含意認識評価セットには適用し辛い。以上から名詞、固有名詞、複合名詞を上位語などで汎化し、汎化した語で照合することで他の含意認識評価セットでも含意要因となるテキスト中の表現と仮説の対を用いて含意認識を行うことができると考える。しかし含意要因となるテキスト中の表現と仮説の対に含まれる全ての名詞、固有名詞、複

合名詞を汎化可能ではない。例えば以下の例である。

例 **4)** 名詞の汎化により含意関係を持たなくなる対の例

含意要因となる表現:ラオス政府のスポークスマンのヤング

仮説:ヤングはラオスの代表だ

仮説に含まれる 代表 は含意要因となる表現における スポークスマン から含意される。しかし スポークスマン を上位語に汎化した場合、この部分に他の職業や役職を表す名詞と照合し、仮説に含まれる 代表 を含意しない場合がある。そのため、含意要因となるテキスト中の表現と仮説の対に汎用性を持たせるためには個々の対を観察し、汎化可能な名詞、固有名詞、複合名詞を判別しなければならないと考える。

また、名詞の汎化だけでなく述語の含意関係についてもまとめることで、含意要因となるテキスト中の表現と仮説の対を用いてより多くの含意関係を認識可能であると考える。表 1 から、述語含意に属す T-H 対はどの評価セットにおいても多く存在している。そのため述語について含意関係の知識を集めることで取り扱うことのできる表現が増えると考える。

6 おわりに

本稿では、PASCAL1〜PASCAL3 で公開された含意認識評価セットから含意要因となるテキスト中の表現と仮説の対を人手で抽出した。抽出した含意要因となるテキスト中の表現と仮説の対を日本語に翻訳した含意認識評価セットに対して適用することで抽出した対の有用性を調べた。その結果 6 割の被覆率で抽出した対を適用することが可能であった。また抽出した対を分析し、どのような特徴を持つ T-H 対から含意要因となるテキスト中の表現と仮説の対を抽出可能であるかを調査した。これにより、内の関係や包含関係など 12 種類の特徴の存在を確認した。確認した 12 種類の特徴から、共通する固有名詞を多く含む T-H 対から含意要因となるテキスト中の表現と仮説の対を抽出しやすいと考える。

今後の課題として、含意関係を持たない T-H 対も含めた評価セットについてこれらの対を用いて含意認識を行う必要がある。抽出した対が T-H 対中に含まれているとしても、「ない」などの否定表現によって含意関係の有無は変わるため、含意関係を持たない T-H 対を含めた場合にどの程度含意認識可能かを調査する必要がある。加えて今回抽出した含意要因となるテキスト中の表現と仮説の対は多くの名詞、固有名詞、複合名詞を含むため汎用性に欠け、他の含意認識評価セットに適用し辛い。そのため抽出した含意要因となるテキスト中の表現と仮説の対に存在する固有名詞、複合名詞を適切な上位語に汎化させるなどで汎用性を高めることを考えている。

使用した言語資源及びツール

(1) RTE-1 datasets,Recognising Textual Entailment Challenge
http://pascallin.ecs.soton.ac.uk/Challenges/RTE/Datasets/

(2) RTE-2 datasets,Second Recognising Textual Entailment Challenge
http://pascallin.ecs.soton.ac.uk/Challenges/RTE2/Datasets/

(3) RTE 3 datasets,Third PASCAL Textual Entailment Challenge
http://pascallin.ecs.soton.ac.uk/Challenges/RTE3/Datasets/

(4) 構文解析器「南瓜」 ,Ver.0.52, 奈良先端科学技術大学院大学 松本研究室,
http://chasen.org/~taku/software/cabocha/

参考文献

[1] Ido Dagan,Oren Glickman and Bernardo Magnini. The PASCAL Recognizing Textual Entailment Challenge. *In Proceedings of the PASCAL Recognizing Textual Entailment Challenge*, 2005

[2] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro,Danilo Giampiccolo, Bernardo Magnini and Idan Szpektor. The Second PASCAL Recognizing Textual Entailment Challenge. *In Proceedings of the Second PASCAL Challenges Workshop on Recognizing Textual Entailment*, 2006

[3] Danilo Giampiccolo,Bernardo Magnini,Ido Dagan and Bill Dolan. The third PASCAL recognizing textual entailment challenge. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* , 2007

[4] 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 発表論文集,pp.260-263,2008.

[5] 小谷 通隆, 柴田 和秀, 黒橋 禎夫. 言い換え表現の述語項構造への正規化とテキスト含意認識での利用. 言語処理学会 第 15 回年次大会 発表論文集,pp.1140-1143,2009.

[6] 阿部川 武, 奥村 学. 日本語連体修飾節と被修飾名詞間の関係の解析. 自然言語処理,Vol.12,No.1,pp.107-123,2005.

[7] 寺村 秀夫. 日本語のシンタクスと意味, くろしお出版