実テキストにおける含意関係認識と因果関係認識のための評価データ構築手法の提案と評価

理学専攻・情報科学コース 金子貴美　(指導教員：戸次大介)

1　はじめに

自然言語の高度な意味処理に関する研究が、近年重要視され、盛んに行われるようになってきた。その高度な意味処理の領域を二分する主要タスクとして「含意関係認識[1]」と「因果関係認識」がある。これら両タスクにおける認識精度を向上させるには、それぞれ適した正解データを与える必要がある。そのことを踏まえ、本研究では、「含意関係認識」と「因果関係認識」のための日本語評価データ構築手法を提案し、各提案手法に従って構築した正解データを評価し、本手法の妥当性を考察する。

2　基本含意関係に分解した含意関係認識日本語評価データの構築

2.1　提案手法

含意関係を基本含意関係に分解し、含意関係認識のための日本語評価データを作成する方法を考案した。基本的に書き換えていく度に意味が広くなるよう書き下し、t1 を徐々に t2 に近づけるよう変換するとした。また、2 文の関係を自動認識させるために必要な基本文関係が書き出せるよう、t1 と t2 の書き換えは関係の繋がりを考慮して行うこととし、個々の現象にフォーカスし分析できるよう、書き下した各ペアでは、複数の関係が成立しないように書き換えることとした。紙面の都合により、含意関係が成立する場合の書き換えのみ、以下に例を示す。

(1) 含意関係が成り立つ場合の例
- ラベル＝ Y (含意関係成立)
- t1:川端康成は、「雪国」などの作品でノーベル文学賞を受賞した。
- t2:川端康成は「雪国」の作者である。

この 2 文のペアを以下のように書き下し、関係ラベルを付与した。

(1') (1) を書き下した例
- [**フレーズ:含意・前提**]　川端康成は「雪国」などの作品の作者である。
- [**集合・リスト**]　川端康成は「雪国」の作者である。
- [**一致**]　川端康成は「雪国」の作者である。

2.2　評価と考察

2.2.1　既存研究との比較評価

RITE-2 BC サブタスク[2]のデータセットの一部に本手法を適用した。173 組のデータを 1 名で書き下し、2 名のアノテータがラベル付けを行った。173 組中、112 組のデータは本手法の策定に用い、残りの 61 組について一致率と平均書き換え数、平均書き換え時間を算出した。今回扱ったデータにおける一致率は 0.83 であった。一致率は、以下の式で算出した。

$$\text{一致率} = \text{ラベル一致数}/\text{全体数}$$

先行研究 [1] の一致率の算出方法は Dice 係数[4]であるため異なっているが、参考までに比較すると、本研究の一致率は [1] の、書き下した文毎の一致率 0.78 より若干高い値となっており、本手法の分類の妥当性は [1] と大差ないと考えられる。書き換え数は全体で 212 回、平均 3.48 回であった。これは [1] の平均 2.98 回よりも若干高い値となっている。また、「Y」ラベルの付与された組は平均 3.62 回、「N」ラベルの付与された組は平均 3.31 回で、これらも共に、[1] の平均 3.03 回、平均 2.80 回よりも若干高い値となっている。日本語であることや、複雑なデータが多いことを考慮すると、本手法は [1] 相応の妥当性があるといえる。書き換え時間については、平均 12 分/1 組であった。[1] の書き換え時間が平均 15 分/1 組であることから、本手法は [1] と同程度の実用性をもつと考えられる。

表 1[5]に [1] と本研究の関係ラベルの対応を示す。

表 1: 既存手法 [1] と本手法における関係ラベルの分布

関係ラベル	文のペア数					
	既存手法 [1]			本手法		
	total	Y	N	total	Y	N
同義・類義	25	22	3	45	45	0
上位・下位	5	3	2	5	5	0
含意・前提	-	-	-	44	44	0
全体・部分	7	4	3	1	1	0
フレーズ： 品詞の変換	9	9	0	1	1	0
共参照・照応	49	48	1	3	3	0
語順入れ替え	-	-	-	15	15	0
態の変化	7	5	2	7	7	0
修飾句削除	25	15	10	42	42	0
主辞削除	6	6	0	1	1	0
並列・従属節	5	4	1	14	14	0
集合・リスト	1	1	0	3	3	0
同格	3	2	1	1	1	0
関係節	1	1	0	8	8	0
時間の推論	2	1	1	1	1	0
空間の推論	1	1	0	1	1	0
数の推論	6	0	6	0	0	0
暗黙の関係	7	7	0	18	18	0
その他の推論	40	26	14	2	2	0
単語・フレーズの不一致	3	0	3	27	0	27
モダリティの不一致	1	0	1	1	0	1
時間の不一致	-	-	-	1	0	1
空間の不一致	-	-	-	0	0	0
数の不一致	-	-	-	0	0	0
“Demonymy”	1	1	0	-	-	-
“Statements”	1	1	0	-	-	-
合計	205	157	48	241	212	29

[1] の分類のうち、“Demonymy”, “Statements” は日本語の場合はあまり問題とならなかったため、本研究では除外し、一方で「語順入れ替え」「含意・前提」を追加した。表 1 より、[1] は本研究より「共参照・照応」が多く見られることや、本研究では [1] と比べ「含意・前提」「語順入れ替え」が頻発していることが分かる。これらから、言語の違いが関係ラベルの分布や分類の違いに現れていると推察される。

[1]含意関係認識とは、与えられた 2 文の間に含意関係が成り立つかどうかを自動認識する技術であり、文 t1 と t2 の間の含意関係とは、人間が「t1 が正しいとき、t2 も (ほぼ) 正しい」と判断する関係のことである。

[2]http://www.cl.ecei.tohoku.ac.jp/rite2/doku.php

[4]Dice 係数とは、2 つの集合の共通要素数を各集合の要素数の平均で割ったもので、アノテーションの一致率計算にも用いられる。

[5][1] では、「単語」と「フレーズ」を分けずに分類しているため、表 1 では分けずに分類している。

2.2.2 関係ラベル毎の精度評価

RITE-2 フォーマルランにおいて、15 チームが、本データを各自のシステムに認識させ、精度を算出した。表 2 に関係ラベル毎の平均精度一覧を示す。

表 2: 関係ラベル毎の平均精度と人間による誤分類数

関係ラベル	平均精度 (%)	文のペア数	人間による誤分類数
語順入れ替え	89.6	15	4
修飾句削除	88.8	42	0
集合・リスト	88.6	3	0
時間の推論	85.7	1	1
関係節	85.4	8	2
並列・従属節	85.0	14	2
単語：上位・下位	85.0	5	1
フレーズの不一致	80.1	25	0
態の変化	79.9	7	2
単語：同義・類義	79.7	9	6
主辞削除	78.6	1	2
暗黙の関係	75.7	18	2
フレーズ：同義・類義	73.6	36	9
共参照・照応	70.9	3	1
フレーズ： 含意・前提	70.2	44	7
単語の不一致	69.0	2	0
単語：全体・部分	64.3	1	1
フレーズ： 品詞の変換	64.3	1	0
同格	50.0	1	1
空間の推論	50.0	1	1
その他の推論	40.5	2	2
モダリティの不一致	35.7	1	0
時間の不一致	28.6	1	1
合計	-	241	41

「語順入れ替え」、「修飾句削除」、「集合・リスト」が平均精度が 90%程度と高い精度を出した。一方、70%以下という低精度のラベルはすべて出現回数が 3 回未満であった。ラベルが 3 回以上出現したものはすべて 70%以上の精度を出しているため、ラベルの出現頻度が精度に関連している可能性がある。出現頻度に関係なく正確に認識するシステムを構築するにはどうすべきか、を今後検討する必要があると考えられる。

また、表 2 の「人間による誤分類数[7]」より、人により判断が揺れやすく、誤分類が多いラベルの精度が、全体的に低くなっていることが分かる。このことから、人によりラベルの判断が揺れるケースと、計算機の判断が揺れるケースは相関関係があると考えられる。

3 SDRT に基づく日本語談話関係アノテーション

3.1 提案手法

文の主節と従属節、等位接続、およびその中間的なもの（日本語の連用接続など）と連続する 2 文に対して、1 組のイベントにつき、それぞれ 1 つの時間関係と談話関係を付与することとした。例文 (2a) への関係ラベルの付与結果は以下の (2a') のようになる。

(2) a. 風が吹いた。剥がれた張り紙が飛んで行った。
a'. **[Precedence(π1,π2),Cause(π1,π2)]**
$_{\pi 1}$風が吹いた。$_{\pi 2}$剥がれた張り紙が飛んで行った。

紙面の都合により、以下に談話関係のみを示す。

3.1.1 談話関係

談話関係は、SDRT[2] を元に表 3 の 8 種を用意した。

(3) a. **[Subsumption(π1,π2),Elaboration(π1,π2)]**
$_{\pi 1}$コース料理を頂いた。$_{\pi 2}$まず食前酒を飲んだ。
b. **[Precedence(π1,π2),Narration(π1,π2)]**
$_{\pi 1}$東京駅に行った。$_{\pi 2}$新幹線に乗った。

[7]アノテータ 2 名により付けられたラベルが異なった文の数を関係毎に示している。

関係ラベル	説明
Alternation(A,B)	「A か B」のように、論理の「∨」の関係と対応するもの。
Consequence(A,B)	「A ならば B」のように、論理の「→」の関係と対応するもの。
Elaboration(A,B)	(3a) のように、B が A の詳細を説明する用法。 イベント B はイベント A に包含される。
Narration(A,B)	(3b) のように、トピックに繋がりがあるもの。
Cause(A,B)	前件が後件の事柄の原因・理由となる用法。
Account(A,B)	前件が後件の判断の根拠になる用法。
Contrast(A,B)	対句法などのように、A と B が類似した意味構造と対照的な 意味を持つもの。および、「A だが B」のような逆接の用法。
Parallel(A,B)	A と B が類似した意味構造を持つが、A と B がテーマを 共有していたり、類似の意味を持つもの。

表 3: 談話関係一覧

3.2 評価と考察

本手法に基づき、試験的に 34 文を実際に注釈付けた。その 34 文を対象に分析した結果を述べる。

時間順序の判断が難しいケースの 1 つに根拠用法の場合のものがある。以下に例を示す。

(4) **[Precedence(π1,π2), Account(π1,π2)]**
$_{\pi 1}$11 月 23 日に大学の文化祭がある。$_{\pi 2}$文化祭当日までに、出し物の準備をしなければならない。

(4) の場合，π_1：「11 月 23 日に大学の文化祭がある。」という文は π_2：「文化祭当日までに、出し物の準備をしなければならない。」の文よりも未来の予定について述べたものであるが、発話者が π_1 の事実を認識したのは、π_2 の発話をする前である。このように、ときに認識視点を考慮すべきケースが存在する。したがって、認識視点を考慮すべきか否かの判断を計算機が行えるよう、認識視点の情報もアノテーションする必要性があると考えられる。

4 まとめ

まず最初に、含意関係を基本含意関係・非含意関係にブレイクダウンした日本語評価データの構築方法を考案し、本手法に従って日本語評価データを構築した。構築したデータの一致率、平均書き換え数、書き換え時間を既存研究と比較し、本手法の妥当性を示すと共に、言語による違いが関係ラベルやアノテーション結果の違いに現れることを指摘した。また、関係毎の認識精度の分析結果から、出現頻度と精度の関連性、人間の誤分類率と精度の関連性を示した。

次に、SDRT[2] を元に、談話関係と時間関係ラベルを付与した、因果関係認識のための日本語評価データ構築手法を提案し、本手法に従って 34 文にアノテーションを行った。注釈付けしたデータを分析した結果、認識視点を導入する必要があることが分かった。

参考文献

[1] Bentivogli, L., Cabrio, E., Dagan, I, Giampiccolo, D., Leggio, M. L., Magnini,B. : Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. In Proceedings of LREC 2010, Malta (2010).

[2] Asher, Nicholas and Alex Lascaridas：Logics of Conversation: Studies in Natural Language Processing, Cambridge University Press. (2003).