文節間限定関係に基づく文間弱対立関係認識

大西真輝† 水野淳太† 福原裕一† 渡邉陽太郎† 乾健太郎†
東北大学†
true0024@gmail.com, {junta-m, fukuhara, yotaro-w, inui}@ecei.tohoku.ac.jp

1 はじめに

含意関係認識課題は，与えられた 1 組の文対に対して，一方から他方が推論可能である場合を「含意」，2 文が同時に成立しえいない場合を「矛盾」，それ以外を「その他」に分類する課題である [1, 2]．しかし，この技術をウェブ上の文に対して適用すると，分類不可能な事例が多いことが報告されている [3]．大木らは，その中でも，条件付きで含意／対立するような関係を「弱対立」関係と呼び，その定義および認識手法について研究を進めてきた [4]．

例えば，図 1 に示す文対を考える．以下，H は仮説を，T は他方のテキストを示す．この例において，赤色部の「毎食後摂取」という条件を満たす場合のみ，T は H を含意する．大木らは，テキスト中で仮説側の内容に相当する部分に対する付加情報（図 1 の赤色部）の有無に基づいて，弱対立認識を行った [4]．付加情報の有無は，人手で整備した辞書を用いて，その情報が条件や程度を示しているかに基づいて認識された．しかしながら，付加情報と呼ばれる情報には，どのような種類のものがどの程度存在するのか，また，その統語的な特徴はどのようなものかなどについては言及されていない．

以下では，付加情報を「限定節」，その付加対象を「被限定節」と呼び，この関係を同定する課題を「限定関係認識」と呼ぶ．ここで言う「限定」は，仮説中のモノの性質や，コトの成立が，テキスト側で制限されていることを指す．図 1 の例では，「虫歯を防ぐ」という事象の成立条件が，「毎食後摂取」するという条件によって制限されている．

本研究の目的は，限定関係認識に含まれる問題を明らかにすることである．まず，大木らの作成した 155 個の弱対立関係文対を対象として，限定関係の分類を行った．次に，分類結果に基づいて，限定関係認識器を構築した．評価実験では，まず，155 文対を対象として，限定関係認識の性能を評価した．次に，オープンテストとして，言論マップコーパス [5] に含まれる弱対立関係文対に対して，限定関係認識および弱対立認識実験を行った．実験結果を分析した結果，助詞の曖昧性に起因する誤りが多数派であることが明らかになった．

2 文節間の限定関係

前述の通り，文間の弱対立関係を認識するためには，文節間の限定関係を認識することが肝要である．しかし

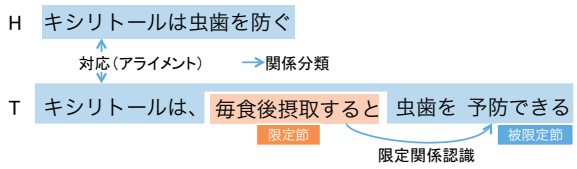

図 1: 文間弱対立関係認識の例

ながら，限定関係を構成する文節間の構造や，その文法的特徴は明らかではない．そこで，大木らによって作成された弱対立文対に対して，文節間の限定関係を付与し，その文法的特徴について分類・分析を行った．

分析の結果を表 1 に示す．まず，限定節と被限定節の間の構造について，その多くは直接の係り受け関係にあることが分かった．一方で，直接の係り受け関係にない場合は，大きく 2 つに分けられる．1 つは照応表現やゼロ照応を介している場合であり，もう 1 つは被限定節側の名詞や述語が，他の名詞や述語と並列構造を構成している場合である．これらの事例を解析対象とするには，照応解析や並列構造解析を行う必要があるため，本研究では直接係り受けの関係にある文節対のみを認識の対象とする．

次に，限定節の文法的特徴について述べる．ここで，限定節とは，被限定節となる事象や実体を修飾し，その文節に含まれる情報が，他の類似概念との対比を示唆するような文節のことを指す．すなわち，対比されている他の概念に対する限定情報となっている場合に，限定節となる．例えば，以下の文において，下線 a は類似概念（ここでは類似物質）との対比を示唆せず，文全体の主題を示していると解釈できる．一方で，下線部 b は，下線部 a と対比して，波線部の事象に対する効果の存在を示している．すなわち，波線部の事象が成立するためには，下線部 b によって限定する必要がある．

(1) トルマリン$_a$ は殆どマイナスイオンを発生しないが，製品 X$_b$ は特殊な技術によりマイナスイオンを発生させる．

しかしながら，ある情報が対比を示唆するかどうかは，統語的特徴からでは判断できない．特に例 (1) に示される「は」は，その文の情報からでは，主題と対比のいずれを示すかを判断することは難しい．弱対立関係認識においては，1 組の文対を認識対象とし，テキスト中で仮説と内容的に対応する部分を主題であると考える．以下の例において，H_1 に対しては，T の下線部 b,c のいず

れも限定節となる．しかし，H_2 に対しては，下線部 c のみが限定節となる．H_2 に，下線部 a と同等の情報が含まれているためである．また，下線部 a は，H_1，H_2 のいずれにも含まれているため，主題として扱われ，限定節にはならない．

(2) T トルマリンは$_a$ 静止状態では$_b$，わずかに$_c$ マイナスイオンを発生させる
H_1 トルマリンはマイナスイオンを発生させる
H_2 トルマリンは静止状態であればマイナスイオンを発生させる

以下の節では，限定関係を構成する文法要素のうち，出現頻度の高いものについて述べる．

2.1 とりたて助詞

とりたて助詞と呼ばれる語群は，副助詞・係助詞に分類される要素であるが，文献によって解釈の異なる概念である [6, 7, 8]．野口ら [9] によると，とりたて助詞は，統語的な特徴によってまとめられる概念ではなく，意味解釈機能の特徴によってまとめられる概念である．とりたて助詞を含む文節は，他の情報と対比され，とりたてられていることから，限定節となる可能性が高い．

以下の例において，下線部は波線部の事象が成立するための特定の状態を示しており，任意の状態と対比され，ある状態に限定されていることを示している．文法的には，状態を示すデ格が,「は」によってとりたてられている.「は」は主格を表す場合もあるが，T では「トルマリン」が主格となっており，文頭から数えて 2 つ目以降の「は」は，対比を示している可能性が高いと考えられる．

(3) H トルマリンはマイナスイオンを発生する
T トルマリンは，常温・静止状態では，ほんの少しのマイナスイオンしか発生しません.

2.2 接続助詞「と」

「と」は，肉と魚のように「本来、異なる存在を合一させる」プロトタイプ的意味のほかにも，引用，条件を表すなど様々な用法が存在する [10]．例えば，以下の下線部は，並列の「と」ではなく条件を示す「と」である．

(4) H トルマリンはマイナスイオンを発生する
T 特許取得ローラーローラーのトルマリンが水に接するとマイナスイオンを発生

上の例に見られるように，用言の終止形「接する」などに「と」が付いている，終止形＋「と」の後に文の切れ目がある場合，他の事象が考えられるため，条件，対比を示しやすい．しかし,「後で掃除すると，太郎は言うが信用できない」の場合にも，限定節に「と」があるが，これは条件を示しているとは言えない．このような事例は限定節か否かの判定が難しい．

2.3 助詞「の」

「名詞 1 ＋「の」＋名詞 2」は,「名詞 1 ＋「の」」が他の要素に対する対比を示す場合，名詞 2 の限定節になり，そうでない場合は，名詞 2 の限定節になりにくい傾向がある.「太郎の家」「花子の写真」のような場合,「太郎の」「花子の」は，それぞれ「家」「写真」との関係を示すものであり，対比要素が考えにくい．一方，下の例のような「高濃度キシリトール配合の」であれば，高濃度でないキシリトールなど，対比要素が容易に示唆できる．

表 1: 限定関係の分類結果

限定関係	事例数
とりたて助詞	184
ノ格	112
数量・程度表現	93
格要素を伴う条件	67
接続助詞	46
形容詞による条件	37
その他	56
合計	595

(5) H キシリトールは虫歯を予防することができる
T う蝕予防効果を十分に発揮させるためには，高濃度キシリトール配合のガムかタブレットを 1 日 3 回 3ヵ月以上続ける必要があります

上の例では，下線部が他の要素に対比する対比要素を容易に示唆できるため，波線部の限定節として働く．

2.4 数量表現

数量表現は，言及された数量の他の要素，いわゆる対比要素が容易に示唆できる．そのため，数量表現を含む文節は，ほぼ対比を示す限定節であると判断できる．本研究で用いたデータ 155 事例においては，数量表現はその対象の性質や成立を限定していた．

(6) H キシリトールは虫歯予防に役立つ
T フィンランドでは，50 %以上キシリトール含有のものが、う蝕予防に効果があるとされています

上の例では，下線部「50 %以上」の他の要素としての対比要素を容易に考えることができる．このことからも，数量表現を含む文節は，限定節であると判断できる．

3 限定関係の自動認識

本稿では，弱対立関係の認識のための手法として，ルールベースの手法，および機械学習に基づく手法の二種類のアプローチを提案する．

3.1 ルールベースシステム

提案するルールベースシステムは，まず二文についてそれぞれ前処理を行った後，パターンに基づいて限定関係になっている文内の箇所を同定する．次に，アライメント情報を利用して，弱対立関係の判定をおこなう．

3.1.1 前処理

入力された二文について，まず日本語形態素解析器である MeCab [11] を用いることによって，各文を構成する形態素に分割し，それらの品詞情報を得る．次に，日本語依存構造解析器である CaboCha [12] を用いて各文節の係り先を判定することで，文の依存構造を得る．弱

対立関係の判定には，二文間のアライメント情報を利用する．これには，水野らの局所構造アライメント [13] の基準に基づき，人手で付与したアライメント情報を利用する．

3.1.2 ルールに基づく弱対立認識

弱対立の認識をおこなうために，開発用のデータセット 200 文対から語彙表現のリストと構造的なパターンを人手で作成した．これを利用して，入力された文対が弱対立であるかどうかを判定する．弱対立認識の手順は以下の通りである．

手順 1 依存構造上で係り元，係り先の関係になっている文節のペアを全て抽出し，手順 2 に進む．

手順 2 文節のペアが限定節・被限定節であるかどうかの判定をおこなう．手順 1 で求めた係り元について，下記の「条件を表す語」および「制限された状況を表す語」のリストに含まれる語が文節内に存在しているかどうかを判定する．存在している場合，その係り元は限定節であるとし，手順 3 に進む．

手順 3 二文間の関係認識をおこない，「同意・対立・その他」の 3 値のいずれかに分類する．分類の結果，「その他」に分類された事例に関しては弱対立とは認識されない．

手順 4 手順 2 で限定節があるとされ，手順 3 で同意または対立と分類された事例について，限定節の係り先がアライメントされている場合，その限定節を弱対立とする．

条件を表す語: 場合，たら，時，とき，は，なら，と，で等

制限された状況を表す語: 限り，かきり，だけ，しか，初めて，こそ，(ため) には等

3.1.3 ルールベースシステムの問題点

前述の通り，ルールベースシステムでは，とりたて助詞や限定節を形態素の表層情報のみから判断している．そのため，例えば，限定関係を構成する対比の「は」以外の，主題を表す「は」についてもルールによって限定関係があると誤認識されてしまうなどの問題が生じ，結果として適合率が下がってしまうという現象が起こる．とりたて助詞は端的に統語的特徴から判断することはできず，意味解釈的な側面を多く含むが，ルールを用いてその曖昧性を解消することには限界がある．我々は機械学習手法を採用することで，このとりたて助詞の曖昧性の問題を解決する．

3.2 機械学習に基づく手法

我々の提案する，機械学習に基づく手法は，依存構造上で親子関係になっている二文節を入力とし，限定関係を持つものを正例，そうでない事例を負例としてモデルを学習し，未知の事例の限定関係の認識をおこなう．この手法は，3.1.2 節の中で，手順 2 を機械学習に置き換えたものに相当する．文節間限定関係認識の機械学習モデルの構築には，Classias[1] を利用し，L2 正則化 L1 損失 SVM を，Pegasos アルゴリズムを用いて学習した．

[1] http://www.chokkan.org/software/classias/index.html.ja

- 次の単語のいずれかが限定節に含まれているか（単語ごとに区別）：として，とした，では，には，れば，にも，でも，での，ときは，による，により，によって，から，しか，的に
- 次の単語のいずれかが限定節の最後にきているか（単語ごとに区別）：に，と，で，も，の，は，ば
- 限定節に数量表現が含まれているか，程度表現が含まれているか
- 限定節に条件名詞（とき，時，限り，かぎり，場合，ばあい）が含まれているか
- 限定節に数量表現か程度表現が含まれているか，程度表現か条件名詞が含まれているか，数量表現か条件名詞が含まれているか，数量表現か程度表現か条件名詞が含まれているか
- 限定節に限定名詞が含まれているか，限定副詞が含まれているか，限定助詞が含まれているか，これらのいずれかが含まれているか
- 限定節が文頭であるか
- 限定節の形態素「と」の前に動詞があるか
- 限定節に「を」か「や」か「が」が含まれているか
- 限定節が 2 度目のガ格を含んでいるか
- 限定節の最後の品詞が助詞であるか，助動詞であるか，助詞・助動詞のいずれかであるか，また，その品詞は基本形であるか
- 被限定節に助詞か助動詞が含まれているか，または形容詞か副詞を含んでいるか
- 被限定節が 2 度目のガ格を含んでいるか
- 被限定節の最後の品詞が助詞であるか，助動詞であるか，助詞・助動詞のいずれかであるか，動詞であるか，動詞であれば基本形であるか
- 被限定節が文末であるか
- 被限定節がアライメント部分であるか

表 2: 弱対立関係認識のための素性一覧

3.2.1 学習事例の作成

3.1 節で用いたデータと同様の 200 文対に対し，依存構造上で親子関係になっている全ての文節対を抽出する．この際，限定関係が人手で付与された文節対を正例，それ以外を負例として用いる．また，弱対立認識において，アライメントされている文節は限定文節にはならないため，該当する事例を除いた．

3.2.2 素性

提案する，機械学習に基づく文節間限定関係認識モデルに利用する素性を表 2 に示す．

4 評価実験

評価実験では，1) 二文節間が限定・被限定の関係であるかを判定する節間限定関係認識，2) テキスト T と仮説 H が与えられた時に，二文が弱対立の関係になっているかを判定する弱対立関係認識，の二種類の実験をおこなうことで，手法の有効性を検証する．

4.1 節間限定関係認識の評価実験

表 3: 限定関係認識の実験結果

	ルールベース	機械学習ベース
Recall	0.804(320/398)	0.610(247/398)
Precision	0.475(320/674)	0.645(247/382)
F-score	0.5969	0.6260

まず，3 節で述べたルールに基づく手法，および機械学習に基づく手法を用いて，節間限定関係認識の結果を示す．本実験では，弱対立関係を持つ 155 文対を用いた．それぞれの手法を用いた限定関係認識の結果を表 3 に示す．ルールに基づく手法は，再現率は 8 割と高い数字が得られているのに対し，適合率は 5 割に満たない．これは 3.1.3 節で述べたとりたて助詞の曖昧性によるもので

あると考えられる．それに対し，機械学習に基づく手法は，ルールに基づく手法と比較し，再現率は劣るものの，適合率が 6 割まで向上し，F 値では，3 ポイント程度の改善が得られている．

4.2 文間弱対立関係認識の評価実験

表 4: 機械学習を用いた弱体対立関係認識の実験

	ルールベース	機械学習ベース
Recall	1.000(143/143)	0.755(108/143)
Precision	0.225(143/635)	0.263(108/410)
F-score	0.3676	0.3906

次に，前述の 155 文対とは異なる，20 種類のクエリからなる文対に対して弱対立認識実験を行った．それぞれのシステムの弱対立認識実験の結果を表 4 に示す．ルールに基づく手法は再現率は 100 %であるが，適合率が 2 割程度と非常に低い結果となった．一方，機械学習ベースシステムは再現度は低下するものの，適合率が若干向上し，結果として，ルールに基づく手法と比較して高い F 値が得られた．しかし，依然として性能は限定的であり，実用レベルには達していないのが現状である．

4.3 考察

ルールベースシステムおよび機械学習に基づく手法の問題は，適合率が非常に低いことにある．以下は，ルールベース，機械学習共に，誤って正例と判断された事例である．

(7) *T* ダイエットのほか、便秘解消、冷え性改善、体脂肪率を下げる（脂肪燃焼）などの効果があると、バナナ酢が注目されています。
H バナナ酢はダイエットに効果がある

ここでは,「効果があると」の接続助詞「と」が，条件として判断されたため，誤って正解と判断された．しかし，ここでの「と」は条件を表してはおらず，限定関係にはなっていない．また，以下の例では，

(8) *T* お肌にハリを保ち、シワのないなめらかな肌を維持する主な成分がコラーゲンです
H コラーゲンは肌に良い

「シワのない」の「シワの」が，限定節として認識されてしまったが，これは，対比の意味として用いられているわけではないため，限定の関係ではない．これも，助詞の曖昧性に起因する問題である．このような，助詞の曖昧性をどのように解消すればよいかは，今後の大きな課題の一つである．

5 おわりに

本稿では，弱対立認識のために，係り受け関係にある 2 文節に対して，限定関係が存在するかを判断する限定関係認識課題を設計した．1 組の文対が与えられた時に，テキスト側で，仮説側の文が内容的に対応する部分が被限定節となるとき，2 文は弱対立の関係にあると判断できる．

評価実験では，まず，155 個の弱対立関係文対に対して，限定関係の認識を行った．その結果，表層情報に基づくルールベース手法と比較して，機械学習に基づく手法の方が適合率の面で改善され，F 値で 3 ポイント程度改善された．次に，20 種類のクエリからなる文対に対して，弱対立関係認識実験を行った．その結果，機械学習ベースの手法はルールベース手法より改善されたものの，認識性能は限定的であった．誤分類された事例を分析した結果，助詞の曖昧性を解消することが重要であるということが分かった．

謝辞

本研究は，東北大学工学部 情報知能システム総合学科「Step-QI スクール」の支援を受けた．

参考文献

[1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Prox. of the First PASCAL Machine Learning Challenges Workshop, Vol. 3944, pp. 177–190, 2005.

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

[3] 大木環美, 村上浩司, 水野淳太, 増田祥子, 乾健太郎, 松本裕治. 文間の限定関係認識：課題設計および分析と予備実験. 言語処理学会第 16 回年次大会発表論文集 D3-1, 2010.

[4] Megumi Ohki, Eric Nichols, Suguru Matsuyoshi, Koji Murakami, Junta Mizuno, Masuda Shouko, Kentaro Inui, and Yuji Matsumoto. Recognizing confinement in web texts. In Proc. of IWCS 2011, pp. 215–224, 2011.

[5] 水野淳太, Eric Nichols, 渡邉陽太郎, 村上浩司, 松吉俊, 大木環美, 乾健太郎, 松本裕治. 言論マップ生成技術の現状と課題. 言語処理学会 第 17 回年次大会, 2011.

[6] 益岡隆志, 野田尚史, 沼田善子. 文の階層構造からみた主題ととりたて. くろしお出版, 1995.

[7] 奥津敬一郎, 沼田善子, 杉本武. いわゆる日本語助詞の研究. 凡人社, 1986.

[8] 寺村秀夫. 日本語のシンタクスと意味 III. くろしお出版, 1991.

[9] 野口直彦, 原田康也. とりたて助詞の機能と解釈- 量的解釈を中心にして-. 郡司隆男 (編)『日文研叢書, Vol. 10, , 1996.

[10] 森田良行. 基礎日本語辞典. 角川書店, 1989.

[11] 工藤拓, 山本薫, 松本裕治. Conditional Random Fields を用いた日本語形態素解析. 情報処理学会自然言語処理研究会 SIGNL-161, pp. 89–96, 2004.

[12] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[13] 水野淳太, 後藤隼人, 渡邉陽太郎, 村上浩司, 乾健太郎, 松本裕治. 文間関係認識のための局所構造アライメント. 情報処理学会自然言語処理研究会 SIGNL-196, pp. 1–8, 2010.