述語と項の分布類似度を利用した非明示的な根拠帰結関係の同定*

林 賢吾　　飯田 龍　　徳永 健伸
東京工業大学 大学院情報理工学研究科
{khayashi,ryu-i,take}@cl.cs.titech.ac.jp

1 はじめに

近年，Web に記述された意見情報の抽出の研究が盛んに行われている [1]．この研究分野では主に，Web ページが肯定的な (もしくは否定的な) 意見を述べているといった分類問題を解く，もしくは文章中から〈評価対象，属性，評価〉といった対を抽出する課題を解いている．このような既存の意見抽出の問題では意見の有無を扱っているため，これに加えてなぜその意見が述べられたかという根拠も同時に抽出することにより，その意見の信頼性を容易に判断可能になる．このような背景から，本研究では文章中に出現する原因・理由・動機・根拠・目的といった談話関係を根拠帰結関係とみなし，この関係の自動同定の問題に取り組む．例 (1) a. に示すように，根拠帰結関係の一部は「ため」のような明示的な接続表現を伴っており，このような手がかり表現が出現している場合には容易に関係を同定できる．

(1) a. 頭を強く打った$_{\text{根拠}}$ため 間もなく死亡した.$_{\text{帰結}}$
b. 頭を強く打って$_{\text{根拠}}$間もなく死亡した.$_{\text{帰結}}$

これに対し，例 (1) b. のようなテ形接続や連用中止の場合にも同様に根拠帰結関係を同定する必要があるが，この場合は手がかり表現がないため，明示的な接続表現を伴う場合と比べて同定が困難になる．そこで，本研究では，特に接続表現を伴わない非明示的な根拠帰結関係の同定に有効な情報を調査する．

本稿では，まず 2 節で関連する談話関係の研究を紹介し，3 節で本研究で利用する評価用データについて説明する．4 節では既存の事態間関係知識獲得の手法を採用し，根拠帰結関係の同定に役立つ資源の構築可能性について議論する．5 節では 4 節で明らかになった問題点を解決するために機械学習に利用する素性の改良について説明し，6 節でその素性を用いて行った評価実験について述べる．最後に 7 節でまとめる．

2 関連研究

文章中の談話セグメント間の談話関係の同定に関して，近年 Penn Discourse TreeBank (以後，PDTB) [2] をはじめさまざまなタグ付きコーパスが作成されている．その中でも特に PDTB は隣接する談話セグメントに対象を限定し，網羅的に談話関係の付与を試みている．例えば，PDTB では例 (2) のように，接続表現 “After” に関して arg1 と arg2 の間に TEMPORAL:Asynchronous:succession という談話関係が付与されている．

(2) *After*$_{\text{TEMPORAL:Asynchronous:succession}}$ [$_{\text{arg2}}$ adjusting for inflation] the Commerce Department said [$_{\text{arg1}}$ spending didn’t change in September]

また，例 (3) のように，明示的に接続表現が記述されていない場合であっても定義された談話関係を持つ場合にはタグが付与される．この例では，arg1 と arg2 の間に接続表現 “in particular” が入ると仮定して， 結果的に，EXPANSION:Instantiation という談話関係がタグ付けされている．

(3) He says [$_{\text{arg1}}$ he spent $300 million on his art business this year.] Implicit = *IN PARTICULAR*$_{\text{EXPANSION:Instantiation}}$ [$_{\text{arg2}}$ A week ago, his gallery racked up a $23 million tab at a Sotheby’s auction in New York buying seven works, including a Picasso.]

Wellner ら [3] は，これらの関係の自動同定を試みており，一般に明示的に接続表現が出現している場合に比べ，接続表現が出現していない場合には同定の精度が低下することを報告している．このため，接続表現を伴わない場合に問題を限定し，どのような言語的な手がかりが有効かの調査も進められている [4, 5]．彼らはセグメント間の統語的な関係や各セグメントに出現する語の組み合わせなどを学習することで，ある談話関係にあるか否かの分類を行う問題を解いている．例えば，Pitler らは，arg1 に出現する語と arg2 に出現する語の任意の対のうち，分類に貢献するか否かをあらかじめ information gain を用いて分類し，分類に貢献しそうな対のみを素性として利用することで同定精度が向上することを報告している [4].

3 根拠帰結関係の評価用データの作成

我々の根拠帰結関係同定の予備調査 [6] では，Web テキストを対象に根拠帰結関係の同定を行ったため，形態素・係り受け解析のレベルで解析誤りが生じ，厳密に精度を求めることが困難であった．本研究では，形態素・係り受け関係が正しく付与された京都大学テキストコーパス [7] を対象に根拠帰結関係を付与し，そのデータを対象に評価を行う．

人手タグ付与の指針としては概ね文献 [6] を参考にした．タグ付与の対象となる根拠と帰結の該当箇所は，原因・理由・動機・根拠・目的の談話関係で出現している箇所とし，タグ付与の範囲はそれぞれの談話セグメント

*Detection of implicit evidence-conclusion relations using distributional similarity of predicates and arguments
Kengo Hayashi, Ryu Iida, and Takenobu Tokunaga
Tokyo Institute of Technology

の最右の文節とした．これらの関係となる場合の多くは「ため」,「ので」,「から」といった接続表現を伴う．例えば，例 (4) では，文節「得たい」と「提示した」を根拠帰結の関係としてタグ付与する．

(4) 新党は国民の指示を 得たい根拠ため，国民が望む政策を 提示した帰結．

また，接続表現を伴わない場合についても上記の関係として認められると作業者が判断した場合にはタグを付与する．例えば，例 (5) では,「追放され」のように連用中止で根拠が提示されており，明示的な接続表現は存在しないが,「政，官，財のリーダーが追放される」ことが「未経験の若い人たちがトップに立った」ことの根拠だと判断できるのでタグを付与する．

(5) 政，官，財のリーダーが 追放され根拠，未経験の若い人たちがトップに 立った帰結．

文章中の任意の 2 つの文節をタグ付与の対象とすると，判断すべき文節の組の数が膨大となり作業の揺れが生じる可能性がある．また，作業対象を同一文内に限定しても，談話セグメントに対する認定の基準の揺れが生じる．このため，本研究では作業対象を係り受け関係にある 2 つの述語対に限定し，それらの間の根拠帰結関係のタグ付与を行った．

作業は作業者 1 人が京都大学テキストコーパスに出現する 12,911 文節対を対象に行い，3,683 文節対に対して根拠帰結関係を付与した．このうち,「ため」「ので」「から」「おり[1]」のような明示的な接続表現を伴う場合が 906 事例，それ以外が 2,777 事例であった．このことから，根拠帰結関係は，明示的に接続表現を伴わない場合の方が多く出現しており，非明示的な根拠帰結関係の同定は重要な問題であることがわかる．

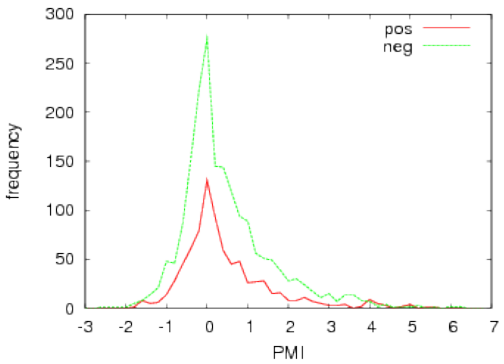
4 明示的な手がかり表現を利用した根拠帰結関係同定のための資源作成

文献 [8, 9, 10] などに代表される事態間関係の資源構築の手法では，主に (i) 大規模テキストコーパスから欲しい事態間関係となる動詞対が抽出できるようなパタンをあらかじめ作成する，もしくは自動収集し，(ii) 収集した動詞対インスタンスを用いて相互情報量などの共起尺度を計算することで動詞対のスコア付けを行っている．この枠組みを根拠帰結関係の知識獲得に利用するため，係り受け関係にある動詞対のうち「ため」,「ので」,「から」の 3 種の根拠帰結関係となる手がかり表現を伴う事例を収集し，収集した動詞対を式 (1) の自己相互情報量を用いてスコア付けすることにより根拠帰結関係となる動詞対を得る．

$$MI(v_i, v_j) = \log \frac{p(v_i, v_j)}{p(v_i)p(v_j)} \quad (1)$$

ここで，$p(v_i, v_j)$ は接続表現「ため」,「ので」,「から」を伴う動詞対のうち，動詞 v_i が係り元として，動詞 v_j が係り先として同時に出現する確率である．また，$p(v_i)$ ($p(v_j)$) は接続表現「ため」,「ので」,「から」を伴う動詞対のうち，係り元 (係り先) に動詞 v_i (v_j) が現れる確率である．本実験では上記 3 種の接続表現を伴って出現する動詞対のうち，係り元，係り先の動詞それぞれが 10 回以上出現している場合を対象に評価を行う．

[1]我々の事前調査の結果,「ため」「ので」「から」に加えて「おり」や「おらず」が伴う場合，根拠帰結関係になっていることが分かった．そのため，非自立語「おる」を伴う文節対は本稿で対象とする自動同定の対象からあらかじめ除外した．



pos: 根拠帰結関係にある事例，neg: それ以外の事例．

図 1: 述語対の自己相互情報量とその頻度

表 1: 明示的な接続表現を伴う動詞対 (上位 10 件)

根拠側の述語	帰結側の述語	自己相互情報量
連れ帰る	現す	8.95
選択する	分裂する	8.60
発砲する	応戦する	8.50
分類する	集計する	8.36
目撃する	眺める	8.31
占う	見逃せる	8.31
鑑定する	採取する	8.19
光る	気が付く	8.13
出動する	免除する	8.05
見落とす	逆転する	8.03

この手法の有効性を調査するために，毎日新聞 1991 年から 1994 年，1996 年から 2002 年までの新聞記事を対象に「ため」,「ので」,「から」の接続表現を伴う動詞対を収集し，202,863 対の共起事例を得た．これを利用し，自己相互情報量の値を計算することで，動詞対のスコア付けを行った．さらに，3 節で示した根拠帰結関係タグ付きコーパスを利用し，スコア付けされた結果の有効性を確認する．タグ付けされた動詞対のうち，明示的な接続表現を伴わない動詞対を対象に自己相互情報量の値を計算し，根拠帰結関係にある事例とそれ以外の場合それぞれについてどのような自己相互情報量の値を取るかを調査した．自己相互情報量の値のある範囲について，その値を取る動詞対の頻度を根拠帰結関係にある場合とそれ以外に分けて描いたグラフを図 1 に示す．この結果からわかるように，接続表現をもとに収集した動詞対の情報は，根拠帰結関係を区別するための良い指標とはいえない．このため，6 節で報告する評価実験でこの情報を素性として利用しても精度が向上することは期待できない．

新聞記事から収集した動詞対のうち，自己相互情報量の値が最も大きい上位 10 件を表 1 に示す．この表からわかるように,「選択する–分裂する」といった述語対は述語単体の組み合わせを見ても根拠と帰結の関係にあるかを判断することができない．この対は具体的には例 (6)

のような文脈で出現しており,「共闘の道を選択する」ことで「後援会が分裂する」という関係にある．

(6) 菅原氏は，参院選で民社党県連が社会党と共闘の道を選択した$_{\text{根拠}}$ために，自分の後援会が分裂した$_{\text{帰結}}$ことなどを理由に離党を申し出ていた．

つまり，根拠帰結関係のような粒度で事態間の関係を扱うためには，述語単体ではなく述語の項も同時に考慮して知識を獲得する必要があると考えられる．

5 述語と項の分布類似度を用いた根拠帰結関係の同定

述語対の根拠帰結関係の同定問題を機械学習に基づく 2 値分類問題として解く．述語対の機能語などを素性としたものをベースラインモデルとし，それにいくつかの素性を加えてどの程度分類性能が良くなるかを確かめる．ベースラインモデルは具体的には以下の 3 つを素性としている．

- 根拠 (帰結) 文節に出現する機能語
- 根拠文節と帰結文節の間に出現する文節内の機能語
- 根拠文節と帰結文節の間に出現する文節内の接続詞

根拠 (帰結) 文節に出現する機能語の情報は，例えば根拠文節側に「〜しても」「〜すれば」のような機能語が出現していることで，一般的に根拠とならないことを捉えるために有効である．また，Lin ら [5] では構文的な位置関係を解析のための手がかりとしているが，根拠文節と帰結文節の間に出現する文節内の機能語と接続詞の情報を加えることで根拠文節から帰結文節へどのように話題が遷移したかを近似的に導入できると考えられる．

このベースラインモデルに加えて影響を調査する素性として，述語と項の関係を捉える素性を 5.1 で説明する．

5.1 述語と項の分布類似度

近年，述語と項を合わせた句レベルの意味表現を生成的に計算して求める手法が提案されている [11, 12, 13]. これらの手法では，局所文脈の語との共起ベクトルでそれぞれの語の意味を表現し，このベクトル間でベクトル和や積を計算することで句の意味を生成する．この手法では，述語とその項のうち一つを組み合わせることにより句の意味を生成し，その結果を選択選好の知識獲得などに利用しているが，一般に述語は複数の項を取り，それら全体が述語の意味の曖昧性を解消するため，これらの手法を述語項構造全体の意味の表現に利用するためにはさらなる手法の洗練が必要となる．本研究では，これらの手法を適用する代わりに，根拠 (帰結) 文節に出現する述語とその項の情報を分類に利用する．ただし，タグ付きコーパスに出現する語をそのまま素性として利用すると，各語の出現数が少ないため，過学習する可能性がある．このため，あらかじめ大規模コーパスから述語と係り受け関係にある項の共起情報を収集し，その共起行列を pLSI [14] で次元圧縮した結果を素性として利用する．具体的には，名詞と格助詞の対 nc と述語 v の共起行列を次元圧縮し，得られた隠れクラス z への帰属確率分布 $p(z|v)$ や $p(z|nc)$ を素性として利用する．帰属確率分布 $p(z|v)$ は述語 v の中で同じ nc と共起する述語について類似した確率分布となる．このため，単純に語を素性とする代わりにこの確率分布を素性とすることで，典型的に根拠 (帰結) の述語 (項) となりやすい表現を学習できる可能性がある．根拠側と帰結側に出現する述語では，それぞれ異なる表現が特徴として現れると考えられるため，根拠側と帰結側で異なる素性として扱う．また，項に関しては任意格よりも必須格の情報が分類に貢献することが考えられるため，本手法ではガ格，ヲ格，ニ格の 3 つの格を対象に素性を抽出する．したがって，例えば隠れクラスの個数が 1,000 である場合，根拠側と帰結側を区別し，述語とそのガ格，ヲ格，ニ格から素性を抽出するため，最終的に 8,000 次元の素性集合を利用することになる．

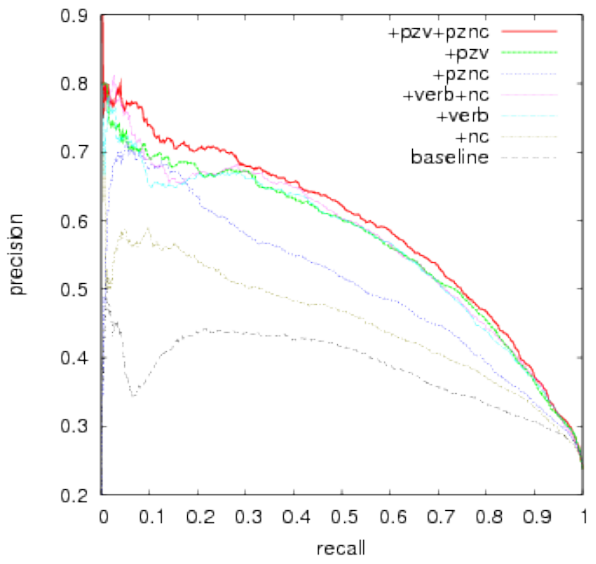
6 評価実験

5 節に示した素性を利用することでどの程度根拠帰結関係を同定できるかを調査するために，3 節で説明したタグ付きコーパスを対象に 5 分割交差検定により評価を行った．コーパス中の明示的な接続表現を伴わない述語対 11,693 事例から同定したい 2,777 事例をどれだけ正しく選択できるかを，再現率と精度で評価する．

学習には Support Vector Machine[2] を使用し，カーネルに線形カーネル，パラメタ c はデフォルト値を用いた．5 節に示したように，ベースラインとして機能語などの情報を素性として利用するモデルを用意し，このモデルに 5.1 の述語と項の分布類似度の素性を加えた場合の精度の変動を調査した．この素性を利用するため，毎日新聞 1991 年から 1994 年，1996 年から 2002 年までの新聞記事を，CaboCha[3] により形態素解析を行い，係り受け関係にあった述語[4] と格要素の組を抽出し，隠れクラスの個数を 1,000 として次元圧縮を行った．また，次元圧縮したことの有効性を調査するために述語の見出し語や項の見出し語をそのまま素性として加えた場合と比較を行った．結果の再現率–精度曲線を図 2 に示す．この結果より，述語と項の情報，特に述語の情報 $p(z|v)$ を素性として加える (+pzv) ことで大きく精度が向上した．さらに，述語と項の両方の帰属確率分布の情報を加える (+pzv+pznc) ことでさらに精度が向上していることがわかる．また，次元圧縮をした場合，特に根拠 (帰結) 箇所の格要素の見出し語を利用した場合 (+nc) と根拠 (帰結) 箇所の格要素の確率分布を利用した場合 (+pznc) の結果を比較することで，項の表現を次元圧縮することにより分類性能が向上していることがわかる．

一方，根拠 (帰結) 箇所の述語と項の見出し語を利用した場合 (+verb+nc) と根拠 (帰結) 箇所の述語と項の確率分布を利用した場合 (+pzv+pznc) を比較することで述語と項両方を次元圧縮した場合の精度の差がわかる．図 2 より両方を次元圧縮した場合でもそのまま見出し語を素性として利用した場合と比べ，精度が向上することがわ

[2] http://svmlight.joachims.org/
[3] http://chasen.org/ taku/software/cabocha/
[4] 本実験では述語として動詞のみを抽出対象とした．



baseline:ベースラインモデル
+pzv:根拠 (帰結) 箇所の述語の確率分布を素性に加える
+pznc:根拠 (帰結) 箇所の格要素の確率分布を素性に加える
+verb:根拠 (帰結) 箇所の述語の見出し語を素性に加える
+nc:根拠 (帰結) 箇所の格要素の見出し語を素性に加える

図 2: 根拠帰結関係同定の再現率–精度曲線

かった．

具体的にどのような表現が分類に貢献したかを調査するために，評価データ全体を利用して分類モデルを作成し，そのモデルから各素性の重みを計算した．また，評価用の事例中の述語対を収集し，根拠側と帰結側の述語とその項それぞれの帰属確率分布のそれぞれの値とその確率値に対応する重みを掛け合わせて最終的にその述語，項のスコアとする．このスコアが大きいほど根拠帰結関係の分類の際に特に影響する表現となる．これらの表現のうち根拠側，帰結側と述語，項の組み合わせのそれぞれについて上位 20 件を表 2 にまとめる．表 2 より，根拠側の項には「地価:が」や「株価:が」のような変動して何かに影響しそうな表現が上位にきているのに対し，帰結側の項には何らかの結果「反響:を (呼ぶ) 」や「不発:に (終わる) 」といった結果に関連する表現が上位にきており，この結果 $p(z|nc)$ を加えることで精度が向上したと考えられる．

7 おわりに

本稿では，明示的な接続表現を伴わない場合の根拠帰結関係の同定の問題において，項と述語の分布類似度の情報を利用した解析手法を提案した．項と述語の共起関係をもとに次元圧縮した結果を素性として利用することで，同定精度の向上が見られた．述語と項の組み合わせをどのように表現するかについては谷塚ら [11] が提案した手法を採用するなど改善の余地があり，今後さらに検討したい．また，本稿では係り受け関係にある述語対のみを対象に根拠帰結関係の同定を試みたが，実際には係り受け関係にない場合や異なる文に出現する関係についても同定する必要がある．同一文内に出現する場合には文内の構造，異なる文に出現する場合については談話構造などさまざまな手がかりを考慮しながらこの問題に今後取り組みたい．

表 2: 分類に貢献した述語と項の表現（上位 20 件）

根拠側の項	根拠側の述語	帰結側の項	帰結側の述語
到着:が	記念する	罪:に	捜査する
軽視:に	化す	不発:に	捜索する
飽き:が	傷付ける	物別れ:に	起訴する
安定:に	優勝する	憶測:を	負傷する
学期:が	否認する	反響:を	手配する
地価:が	放つ	天然記念物:に	死傷する
株価:が	敗退する	短命:に	難航する
物価:が	衰える	最低:を	逮捕する
訴訟:を	傷つける	空手形:に	遅れる
新雪:が	終える	最高:を	浮かべる
雨:が	気付く	違反:に	衰える
迷い:に	しのびなく	声援:を	化す
体重:が	めざす	窮地:に	送検する
今日:に	陥る	後遺症:に	死亡する
人口:が	気づく	二の足:を	優勝する
繁栄:に	下りる	空回り:に	決裂する
体面:に	祝う	悪影響:を	受賞する
機運:が	難航する	不全:に	焼死する
カウントダウン:が	成功する	容体:が	運休する
流言:が	整う	物議:を	早まる

参考文献

[1] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[2] E. Miltsakaki, R. Prasad, A. Joshi, and B. Webber. The penn discourse treebank. In *Proccedings of the Language Resources and Evaluation Conference*, pp. 2237–2240, 2004.

[3] B. Wellner and J. Pustejovsky. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 92–101, 2007.

[4] E. Pitler, A. Louis, and A. Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 683–691, 2009.

[5] Z. Lin, M. Kan, and H. T. Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 343–351, 2009.

[6] 飯田龍, 乾健太郎, 松本裕治. 根拠情報抽出の課題設計と予備実験. 言語処理学会第 15 回年次大会発表論文集, pp. 817–820, 2009.

[7] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115–118, 1997.

[8] V. Pekar. Acquisition of verb entailment from text. In *Proceedings of Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL06)*, pp. 49–56, 2006.

[9] D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[10] S. Abe, K. Inui, and Y. Matsumoto. Two-phased event relation acquisition: Coupling the relation-oriented and argument-oriented approaches. In *Proceedings of The 22nd International Conference on Computational Linguistics (COLING)*, pp. 1–8, 2008.

[11] 谷塚太一, 飯田龍, 徳永健伸. 格要素間の依存関係を考慮した選択選好モデル. 情報処理学会 自然言語処理研究会報告 NL-193, 2009.

[12] K. Erk and S. Pad'o. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 897–906, 2008.

[13] J. Mitchell and M. Lapata. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pp. 236–244, 2008.

[14] T. Hoffman. Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR*, pp. 50–57, 1999.