談話関係認識のための時制情報の分析

清野 舜　田 然　渡邉 研斗　岡崎 直観　乾 健太郎
東北大学
{kiyono, tianran, kento.w, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

文章において，文や段落，節などの基本単位 (Argument) は意味的なつながり（談話関係）を持っている．談話関係の例として〈換言〉や〈対比〉,〈同期〉,〈非同期〉,〈条件〉,〈原因〉などがある．2 文間の談話関係を推定するタスクを談話関係認識と呼ぶ．

2 文間にはしばしば「談話マーカ」と呼ばれる手がかり表現が存在し，これを利用することで談話関係を高精度で推定できる [11]．例えば下記 (1) の場合，談話マーカが接続詞 “So” であることから，Arg1 と Arg2 は談話関係〈原因〉であるとわかる：

(1) **Arg1:** We’re standing in gasoline.
Arg2: So don’t smoke.

このように談話マーカが明示される関係を Explicit な談話関係と呼ぶ．一方，下記の (2) では Arg1 と Arg2 の間には談話マーカ “Previously” が省略されているため，談話関係〈非同期〉を推定することは難しい：

(2) **Arg1:** The trial begins today in federal court in Philadelphia
Arg2: (Previously) the government’s assertions of the cover-up were made in last minute pretrial motions

このような談話マーカが存在しない関係を Implicit な談話関係と呼び，本研究ではこちらの自動推定に取り組む．

Implicit な談話関係を推定する手法の一つとして，イベントの時制情報を用いることが考えられる．例えば上記の (2) の場合，Arg1 の動詞句 “begins” が現在形であるのに対して，Arg2 のイベント動詞句 “were made” は過去形であることから，談話関係〈非同期〉だと推定できる．この例のように，時制情報が談話関係認識の重要な手がかりになることがしばしばある．本研究では談話関係認識の性能向上に向けて，いくつかの談話関係に着目し，談話と時制変化の関わりについて考察をする．

2 関連研究

時制情報の談話関係への影響は，以前より言語学の分野からも指摘されてきたが [4, 8, 9]，これらの知見を応用した自然言語処理の研究は少ない．本研究では Penn Discourse Tree Bank [12] を用いて，統計的な側面からより踏み込んだ分析を試みる．

談話関係認識の素性として，既存研究は単語ペア素性を主に利用してきた [5, 6, 7]．これは Arg1 と Arg2 に含まれる単語から Bag-of-Words 的に素性を作る手法である．近年では単語ペア素性の持つスパースネス問題を解決するため，Brown Cluster[13]，単語の分散表現 [2]，ニューラルネットを用いた手法 [3] などが提案されている．

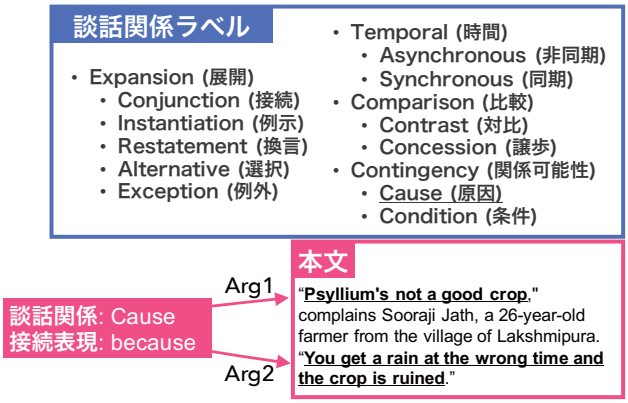

図 1: Penn Discourse Tree Bank のアノテーション構造

時制情報を談話関係認識一般に適用した例として，Piter らの研究がある [10]．Pitler らは単語ペアの他，感情極性単語の数，時間/数量/割合表現など多数の素性を用いて分類器を作成した．その際，談話関係によって時制の割合が異なることを予想し，各文の時制を素性として用いることを提案した．具体的には，各文の主動詞の品詞タグが文の時制を表すと仮定し，素性を作成した．しかしこの素性単体での性能は報告されていない．また主動詞の品詞タグだけでは，文の時制を正確に推定することはできない．例えば “will” や “would” などの助動詞を考慮できず，仮定法と未来形が同じ時制として扱われてしまう．また受動態の文の場合，時制の違いに関わらず主動詞の品詞タグには VBN が付与されるため，現在形や過去形を区別できない．本研究では文の依存構文木の係り受けパスを用いることで助動詞や受動態を考慮しながら時制情報を推定し，時制変化を調べる．

3 時制の変化と談話関係

3.1 使用したデータ

本研究では Penn Discourse Tree Bank (PDTB) [12] を分析対象のデータとして用いた．PDTB は Wall Street Journal の記事に対して談話関係のアノテーションを付与したコーパスである (図 1)．

各談話関係は二つの項 (Argument) 間に接続表現と共に定義されており，接続表現が付与される項を Arg2，もう片方を Arg1 と呼ぶ．接続表現は明示される場合と明示されない場合に分けられ，それぞれ Explicit な談話関係と Implicit な談話関係に対応する．本研究では Implicit な談話関係を分析の対象とした．

PDTB には談話関係ラベルが 3 段階の階層構造で定義

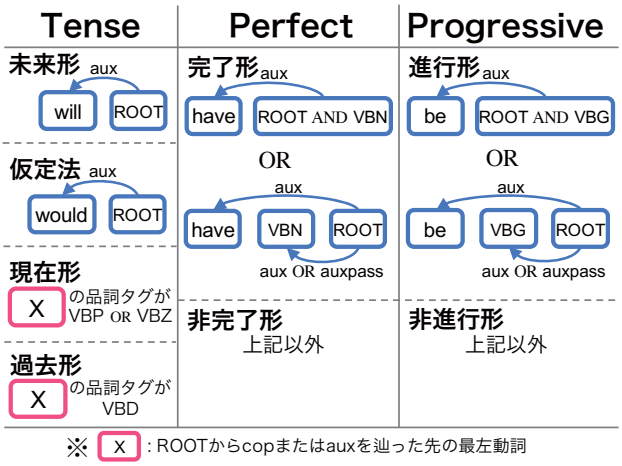

図 2: 時制情報 (Tense, Perfect, Progressive) の推定

されており，それぞれクラス, タイプ, サブタイプと呼ぶ．クラスは《Temporal》,《Contingency》,《Comparison》と《Expansion》の 4 つに大別されており，その下により詳細なラベルが定義されている．クラスレベルでの分類は粒度が荒く，各クラスに対して直観を持つことは難しい．そのため今回はタイプレベルでの分析を行った．なお関係ラベル〈Condition〉と〈Exception〉については，訓練データの数が著しく少ないため分析から除外した．PDTB を CoNLL2015 Shared Task [14] の設定にならい，訓練・開発・テストデータに分割した．

3.2 時制情報の推定

Argument の時制情報を推定するために，まず構文解析器 SyntaxNet [1] を用いて，PDTB 全体の品詞タグと依存構文木を得た．次に各依存構文木の根を起点とした係り受けパスと品詞タグを探索するルールベース手法 (図 2) を用いて，以下の 3 種類の時制情報を推定した．

1. Tense: 現在形/過去形/未来形/仮定法 の 4 値
2. Perfect: 完了形/非完了形 の 2 値
3. Progressive: 進行形/非進行形 の 2 値

Tense と Perfect，Progressive の組み合わせで，各文の時制を 16 通りに分類した．Argument が文の一部の場合は，Argument を含む文の時制情報を用いた．

3.3 時制変化が起こる割合

各談話関係について，Argument 間で時制が変化する割合について調べた（図 3）．ここではグラフ横の数値が大きいほど，その談話関係では高い割合で時制が変化していることを意味する．図 3 より時制変化の割合はどの談話関係においても約 35%付近であることがわかった．そのため，仮に時制変化の有無を分類器の素性として用いても，談話関係認識の性能向上につなげることは困難だと予想できる．談話関係認識に時制情報を活用するためには，時制の変化が談話に与える影響を，各談話関係ごとに詳細に分析することが必要である．

また，談話関係によらず一定の割合で時制変化が生じる事実は，各談話関係に対する素朴な直観と一致しない場合もある．例えば [10] の中で,《Expansion》では時制変化が生じにくく,《Contingency》,《Temporal》では時制変化が生じやすいと予想されたが，図 3 から《Expansion》に入る〈Restatement〉·〈Instantiation〉·

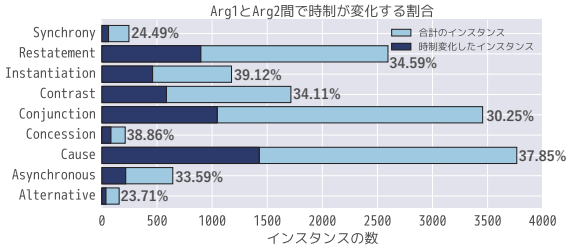

図 3: 各談話関係ラベルにおける時制変化の割合

〈Conjunction〉のどれも高い割合で時制変化が生じている．我々の疑問を以下にまとめた．

1. **Asynchronous と Synchrony:** どれも時間的関係を表すラベルであり，時制情報と相関が高いと予想されるが,〈Asynchronous〉と〈Synchrony〉で時制変化の割合に約 10%の差があるのはなぜか？
2. **Restatement:** これらは Arg1 の内容を Arg2 が言い換える場合に付与されるラベルである．単なる言い換えなのになぜ約 30%のインスタンスで時制が変化するのか？
3. **Instantiation:** これらは Arg1 の事象の詳細を Arg2 が述べる場合に付与されるラベルである．時制変化の割合で見て一番高い数値を示している．それがどんな変化なのか？
4. **Concession:** これらは片方の Argument の示唆する事象をもう一方の Argument が否定する場合に付与されるラベルである．時制変化の割合では 2 番目に高く，その内訳はどうなのが？
5. **Alternative:** これらは Arg1 と Arg2 の事象が代替関係にある場合に付与されるラベルである．時制変化が最も低い割合を示すことと関連があるのか？

なお，談話関係〈Cause〉や〈Conjunction〉は事例の絶対数は多いが，(1) 時制変化の割合が特に大きいわけではなかったこと (2) 一般的に〈Cause〉の分類は因果関係知識を必要とし,〈Conjunction〉の分類は新・旧情報の分析が必要なため，時制変化との関連が強くないと予想されるから，分析の対象から除外した．

4 分析

4.1 Asynchronous と Synchrony

PDTB のアノテーション定義によれば，〈Asynchronous〉は “時間的に連続した（重ならない）” 事象を扱う一方で,〈Synchrony〉は “時間的に重なり合った” 事象を扱っている．図 3 から,〈Asynchronous〉では約 30%の事例が時制変化を伴うのに対して,〈Synchrony〉では比較的に時制変化の割合が小さく，約 20%の事例だけが時制の変化を伴った．この差はどのような時制変化で生じているのか．考えられるすべての時制変化において，両者の割合の差の上位 10 件を図 4 に示した．これより，より多くの時制変化は，“時間的に重ならない” と判断されやすいことがわかり，中でも “過去形から過去完了形” の変化は “時間的に重ならない” ことを示す強い指標であることがわかった．一方，“未来形から現在形” や “現在形から現在進行形” のような変化は “時間的に重なり合う” と判断されやすい．

談話関係認識のタスクに向けて，上記 “過去形から過去完了形” の変化は各談話関係ラベルでどれだけ起こっ

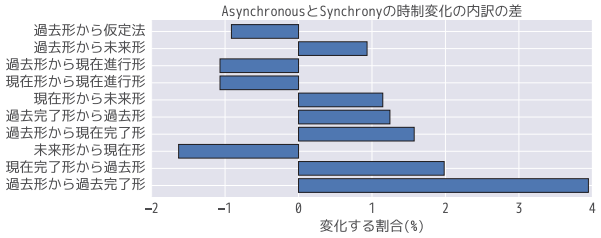

図 4: Asynchronous と Synchrony の時制変化の内訳の差

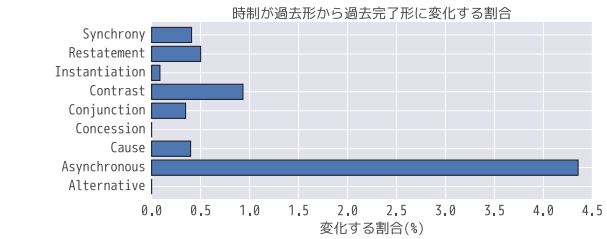

図 5: 過去形から過去完了への時制変化の割合

ているのかを調べた（図 5）.〈Asynchronous〉では約 4%の事例がこの遷移を生じたのに対して，他の談話関係では 1%以下の事例でしか生じていない．これより，過去形から過去完了形への遷移は〈Asynchronous〉に特有の現象であり，識別する手がかりになりうることがわかった．

4.2 Restatement

談話関係〈Restatement〉では時制変化が生じにくいという直観に反して，約 34.6%の事例で時制変化が起こった．そこで解析結果を確認したところ，多くの事例で Argument の時制と，それを含む文の時制が一致していないことがわかった．これが原因で，実際には同じ時制を持っている Argument ペアが，時制変化として判断されている事例が多く見受けられた．具体的には，文と Argument 間の時制の不一致の多くは，以下に示す例文 (3) のように，文がある人物の発言を含んでいる場合に生じる．

(3) “We would have to wait until we have collected on those assets before we can move forward,” he said.

例文 (3) では “he” の発言の中身が Argument であり，これ自体は仮定法の時制を持っている．しかし文全体の時制は過去形であり，時制は一致しない．

この現象が起こる割合を調べるため，文が発言を含み，かつその中身が Argument として定義されている事例を抽出した．なお，文の係り受けパスを用いて ccomp (clausal complement) ラベルのエッジが文から Argument の内部に張られている場合を発言だとみなした．その結果を図 6 に示した．図 6 より，一定数（500 個）以上の事例を含む談話関係ラベルに対象を絞ると，〈Restatement〉が高い割合で「発言」と推定される Argument を含むことがわかった．このことから，時制の不一致は〈Restatement〉が高い時制変化を示した要因の一つだと言える．

4.3 Instantiation

〈Instantiation〉では約 40%のインスタンスが時制変化を伴った（図 3）が，その内訳について図 7 に示した．

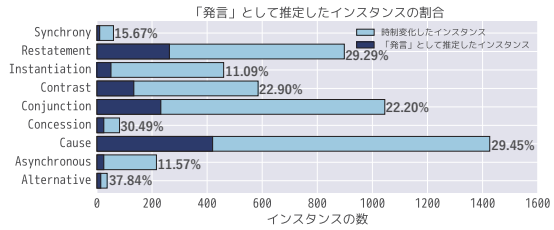

図 6: 「発言」として推定された Argument の割合

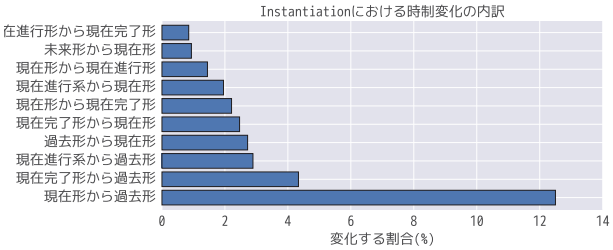

図 7: Instantiation の時制変化の分布

図 7 より〈Instantiation〉で最も多く生じた時制変化は現在形から過去形への変化（約 16%）だとわかる．この結果は，〈Instantiation〉において「Arg1 で一般的な事柄を述べたあと，Arg2 でそれらの具体例を述べる」際に，現在形から過去形への変化が生じやすいことを示唆していると考えられる．具体例を (4) に示した．

(4) **Arg1:** Gene-splicing now is an integral part of the drug business

Arg2: Genentech’s 1988 sales were $335 million, both from licensing and its own products

ここで Arg1 は “Gene-splicing” が，製薬業界において重要な位置を占めると述べているが，この時点では具体的な理由については言及しておらず，あくまでも現在形を用いて一般的に述べているに過ぎない．Arg2 で具体的な企業名 (“Genetech”) の過去の売上を過去形で提示することで，Arg1 の内容についての具体化が行われていると考えられる．

各談話関係ごとの現在形から過去形への変化の割合を図 8 に示した．この図から，〈Instantiation〉における現在形から過去形への変化は，他の談話関係ラベルよりも比較的に生じやすいことがわかる．そのため，談話関係認識の素性として有用である可能性がある．

4.4 Concession

〈Concession〉は時制変化の割合が全クラス中で 2 番目に大きい値を示した（図 3）.〈Concession〉ラベルがついた談話関係は，特定の事象に対して Arg1 が予想される出来事を示唆し，Arg2 がそれを打ち消すという構造を持っている．時制変化 (38.86%) の内訳に着目すると，過去形から現在形への遷移 (8.5%) と現在形から過去形への遷移 (6.6%) が大きな割合を占めていた．

例えば (5) の場合では，過去形から現在形への遷移が生じている．

(5) **Arg1:** Consumers Power Co., now the main unit of CMS Energy, ran into financial problems over its $4.2 billion Midland nuclear plant

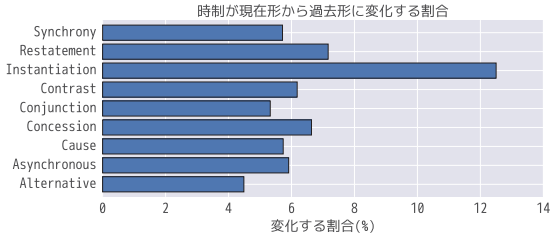

図 8: 現在形から過去形への時制変化の割合

Arg2: CMS is nearly done converting the Midland plant to a gas-fired cogeneration facility at a cost of $600 million

Arg1 が過去の出来事 (“ran into financial problems”) を述べることで，“CMS” への悪影響を予想させるが，Arg2 が現在形で実際には影響が無いことを述べそれを否定しているとわかる．

また (6) の場合では，現在形から過去形への遷移が生じている．

(6) **Arg1:** In Sidhpur, it is almost time to sow this year’s crop

Arg2: Many farmers, too removed to glean psyllium’s new sparkle in the West, have decided to plant mustard, fennel, cumin, fenugreek or castor-oil seeds

この例の Arg1 では「もうすぐ psyllium の種まきの季節であること」が述べられており，これは「pyllium の種まき」イベントが生じることを示唆しているが，Arg2 では「Mustard, fennel, cumin など他の作物の種まきをすると決めた」と述べられることにより打ち消されている．ここで，Arg2 は過去形を用いて「実際に起きた出来事」に言及することで,「Arg1 から予想できる出来事」を否定していると考えられる．

これらの例からわかるように,〈Concession〉を認識するためには，例えば「財政問題は悪影響を及ぼしやすい」ことや「もうすぐ psyllium の種まきの季節であること」が「pyllium の種まき」を示唆すると理解する必要があり，そのためには現実世界についての事前知識が必要である．時制変化のみを用いて〈Concession〉を認識することは難しいと考えられる．

4.5 Alternative

〈Alternative〉は時制変化の割合が全クラス中で最も小さい値を示した（図 3）．これは〈Alternative〉において,「特定の事象に対して二つの Argument が同じ時間軸上の点から言及する」ことが多いからだと考えられる．具体例を (7) に示した．

(7) **Arg1:** he won’t be paying for it

Arg2: The donations will come out of the chain’s national advertising fund, which is nanced by the franchisees

ここでは「何かの資金が支払われる」という事象について,「“he” が払う」ことと「“donation” から支払われる」の二つの並行する選択肢が述べられている．支払うという行為について，二つの選択肢は同じ時間軸上の点に属するため，同じ時制を持っていると考えられる．

5 おわりに

本研究では各談話関係ごとの時制変化が起こる割合を調べ,「時制変化の割合の分布は特定の談話関係に偏る」という直観が必ずしも正しくないことを示した．そのため，時制変化の有無を単純に分類器の素性として用いても，Implicit な談話関係認識の性能向上は困難と思われる．しかし，特定の談話関係に絞った分析の結果，〈Instantiation〉や〈Asynchronous〉など一部には，時制変化の内訳に偏りが生じることがわかった．今後はこれらの知見を利用して，Implicit な談話関係認識の性能向上を目指す．

謝辞 本研究は，文部科学省科研費 15H01702，15H05318，および JST，CREST の支援を受けたものである．

参考文献

[1] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins. Globally normalized transition-based neural networks. In *ACL*, pages 2442–2452, 2016.

[2] C. Braud and P. Denis. Comparing word representations for implicit discourse relation classification. In *EMNLP*, pages 2201–2211, 2015.

[3] J. Chen, Q. Zhang, P. Liu, X. Qiu, and X. Huang. Implicit discourse relation detection via a deep architecture with gated relevance network. In *ACL*, pages 1726–1735, 2016.

[4] E. Hinrichs. Temporal anaphora in discourses of english. *Linguistics and Philosophy*, 9(1):63–82, 1986.

[5] Z. Lin, M.-Y. Kan, and H. T. Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*, pages 343–351, 2009.

[6] A. Louis, A. Joshi, R. Prasad, and A. Nenkova. Using entity features to classify implicit discourse relations. In *SIGDIAL*, pages 59–62, 2010.

[7] J. Park and C. Cardie. Improving implicit discourse relation recognition through feature set optimization. In *SIGDIAL*, pages 108–112, 2012.

[8] B. H. Partee. Some structural analogies between tenses and pronouns in english. *The Journal of Philosophy*, 70(18):601–609, 1973.

[9] B. H. Partes. Nominal and temporal anaphora. *Linguistics and Philosophy*, 7(3):243–286, 1984.

[10] E. Pitler, A. Louis, and A. Nenkova. Automatic sense prediction for implicit discourse relations in text. In *ACL and IJCNLP*, pages 683–691, 2009.

[11] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi. Easily identifiable discourse relations. In *COLING*, pages 87–90, 2008.

[12] R. Prasad and et al. The penn disccourse treebank 2.0. In *LREC*, pages 2961–2968, 2008.

[13] A. Rutherford and N. Xue. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *EACL*, pages 645–654, 2014.

[14] N. Xue, H. T. Ng, S. Pradhan, R. P. C. Bryant, and A. T. Rutherford. The conll-2015 shared task on shallow discourse parsing. In *CoNLL*, pages 1–16, 2015.