B3TB2078

卒業論文

談話解析における暗黙的な関係認識に関する研究

清野舜

2017年3月31日

東北大学
工学部 情報知能システム総合学科

談話解析における暗黙的な関係認識に関する研究*

清野舜

内容梗概

これまで，2 つの Argument（文）間に存在する “原因” や “対比” などの談話関係において，過去形や現在形などの時制情報が重要であると指摘されてきた．しかし，計算機に談話関係を自動推定させるタスクにおいて，談話と時制情報の数理的な関係性を報告した研究はない．そこで本論文では談話関係認識の性能向上に向けて，談話と時制変化の関わりについて分析する．具体的には，談話関係ごとの時制変化が起こる割合を調べ，その分布が談話関係によって偏るどうかを検証した．検証の結果，どの談話関係においても時制変化の割合は，さほど変わらないことがわかったが，一方で，特定の談話関係においては，時制変化の内訳に偏りが生じるとこがわかった．

キーワード

談話関係認識

*東北大学 工学部 情報知能システム総合学科 卒業論文, B3TB2078, 2017 年 3 月 31 日.

A Study on Implicit Relation Recognition for Discourse Analysis*

Shun Kiyono

Abstract

It has been reported that the tense information such as "past" and "present" play an important role on discourse relation that exists between two arguments. However, in the study field of automatic discourse relation recognition, the amount of the work on analysing the relationship between discourse and tense information is highly limited. Thus this paper conducts an analysis of the interaction between the discourse and the change in the tense between two arguments. Specifically, we researched the rate of the change in the tense for each discourse relation, and analysed whether their distribution differs by the discourse relation. As a result, although we found that the rate is not significantly different, the change within specific tense is likely to occur in some discourse relations.

Keywords:

Discourse Relation Recognition

*Graduation Thesis, Department of Information and Intelligent Systems, Tohoku University, B3TB2078, March 31, 2017.

Contents

List of Figures

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 [1]．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

2.1　言語学の時制への取り組み

言語学の分野では，Partee によって文の時制情報と代名詞が文の解釈にあたって似たふるまいを示すことが指摘され [2]，文の意味表現モデルは時制と代名詞に類似した構造を持つべきである，という主張がなされた．Partee は時制と代名詞を直示的 (deictic) な用法と照応的 (anaphoric) な用法に分けた議論を行い，例えば直示的な場合には，代名詞 “I” が指示する対象 (referent) と現在形が指示する時間軸上の点は共に明らかになりやすいのに対して，代名詞 “they” と過去形は指示対象が曖昧になりやすいことを定性的に示した．

Hinrichs は Partee の主張を元に，時制などの時間的情報を談話関係のモデル化に組み込む可能性を提案した．具体的には Kamp [3] のモデルを元に，3 種類の時制情報の組み合わせの取扱いを議論し，直示的な用法において，発話の言及するイベントが生じた時刻と，発話時刻は時制が決定する，などの知見を示した．

談話関係認識の性能向上のためには，これらの言語学からの知見が実際に大規模なコーパス上でも適用可能であるかを調べる必要がある．しかし直示的用法，照応的用法や，発話の言及先時刻の自動推定は非常に困難である．そのため，本研究では Penn Discourse Tree Bank [4] をコーパスとして用いて，容易に自動推定可能である「文の時制」に対象を絞り，統計的により踏み込んだ分析を試みる．

2.2　自然言語処理の時制への取り組み

時制情報の談話への影響は，言語学の分野からも指摘されてきた (節 2.1) が [5, 2]，これらの知見を応用した自然言語処理の研究は少ない．

談話関係認識の場合，伝統的には Pitler らの素性セット [6] を用いた素性ベクトルの作成，分類器の学習・推論によって取り組まれてきた．特に，既存研究では単語ペア素性の利用が顕著である [7, 8, 9]．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[10]，単語の分散表現 [11]，ニューラルネットを用いた手法 [12] などが提案されている．

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 [6]．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．

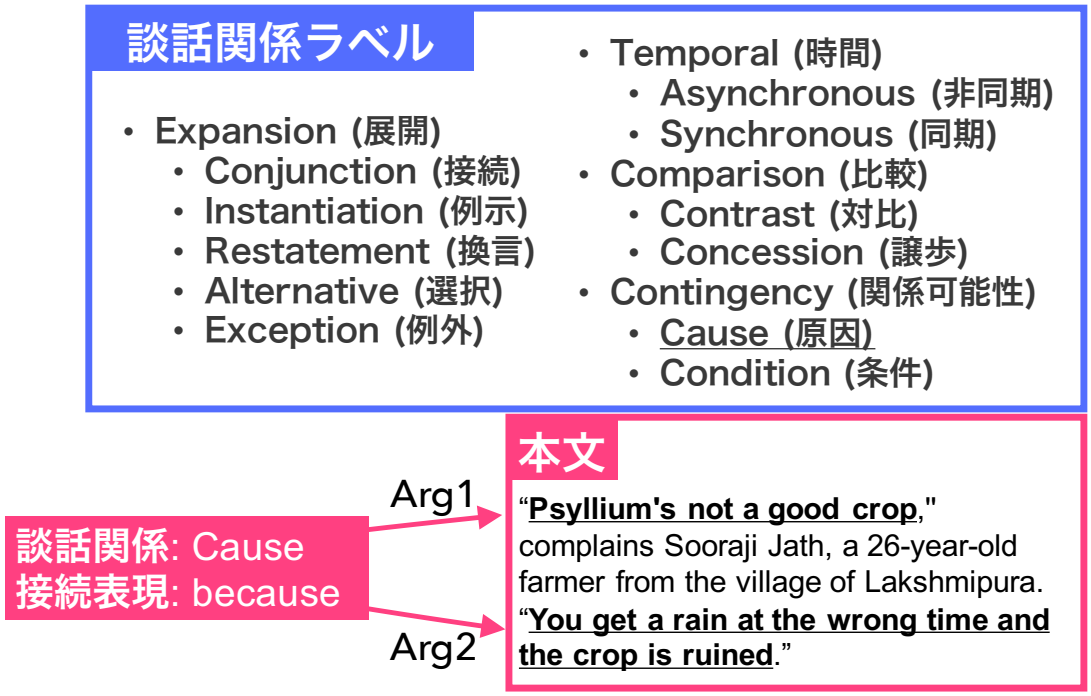

Figure 1: Penn Discourse Tree Bank のアノテーション構造

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) [4] 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 [13] 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 [14] 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 [6] 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〉・〈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？

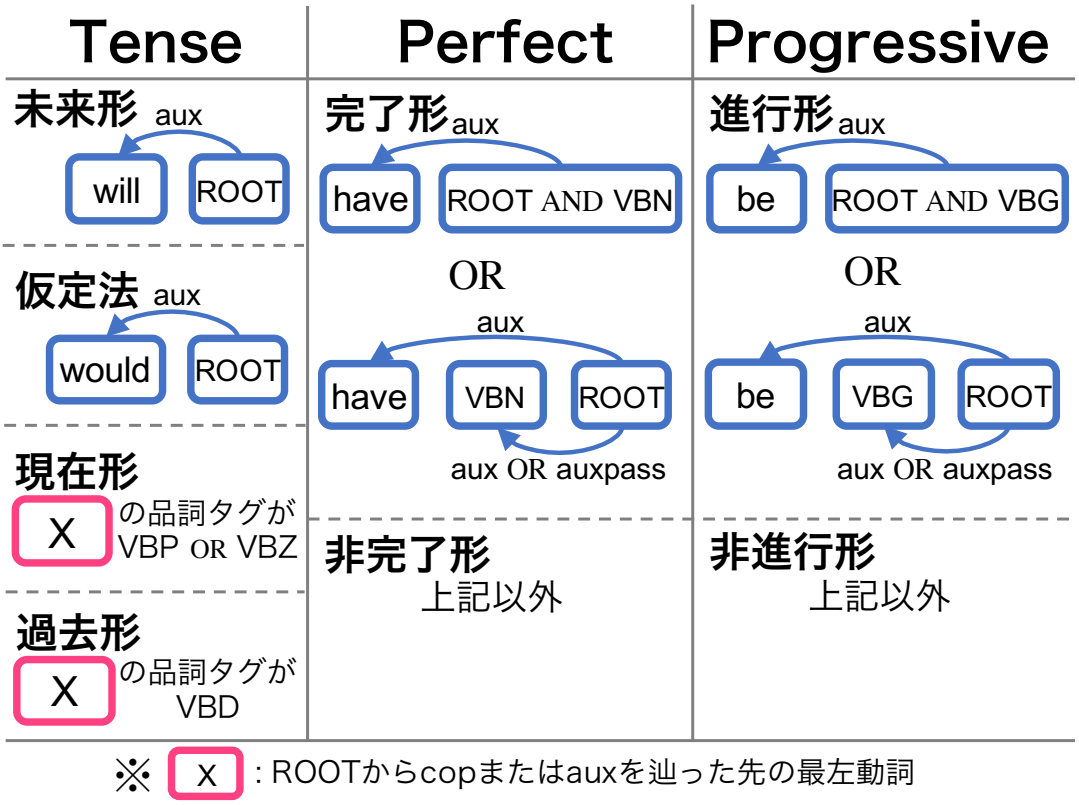

Figure 2: 時制情報 (Tense, Perfect, Progressive) の推定

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

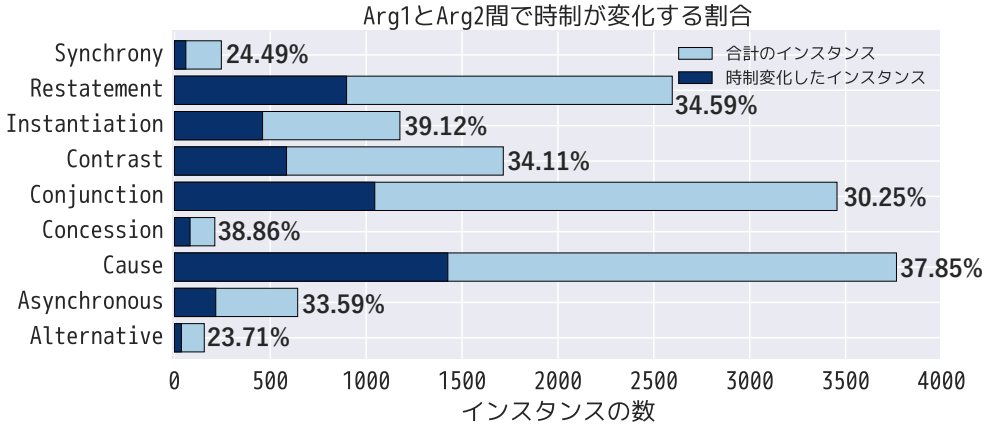

Figure 3: 各談話関係ラベルにおける時制変化の割合

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．

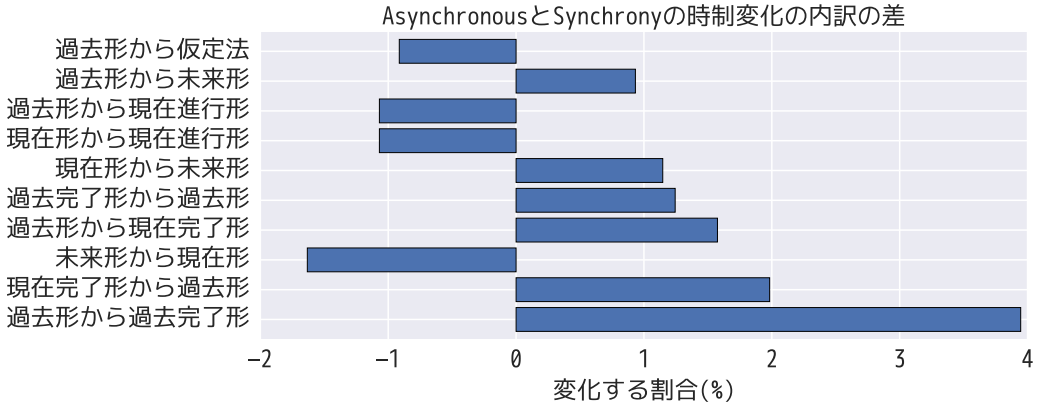

Figure 4: Asynchronous と Synchrony の時制変化の内訳の差

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．

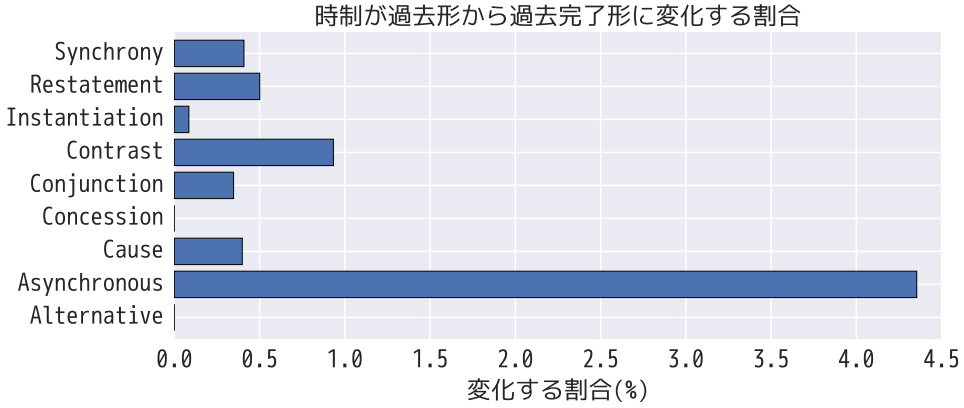

Figure 5: 過去形から過去完了への時制変化の割合

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．

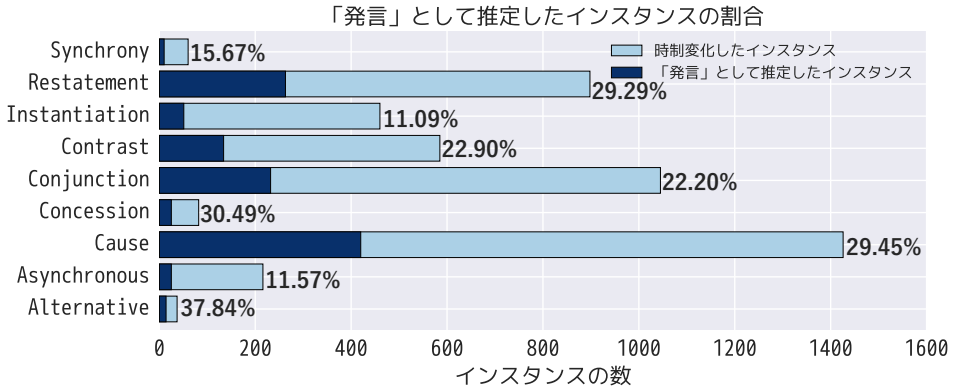

Figure 6: 「発言」として推定された Argument の割合

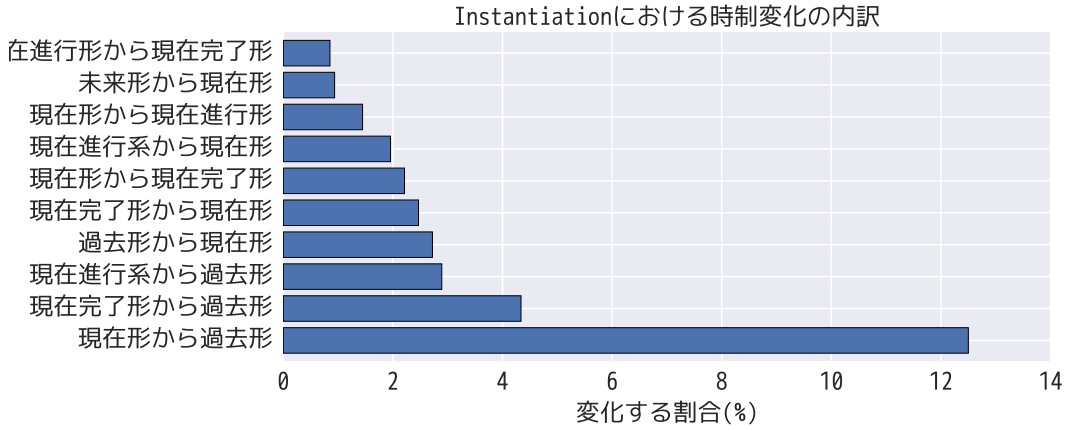

Figure 7: Instantiation の時制変化の分布

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 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

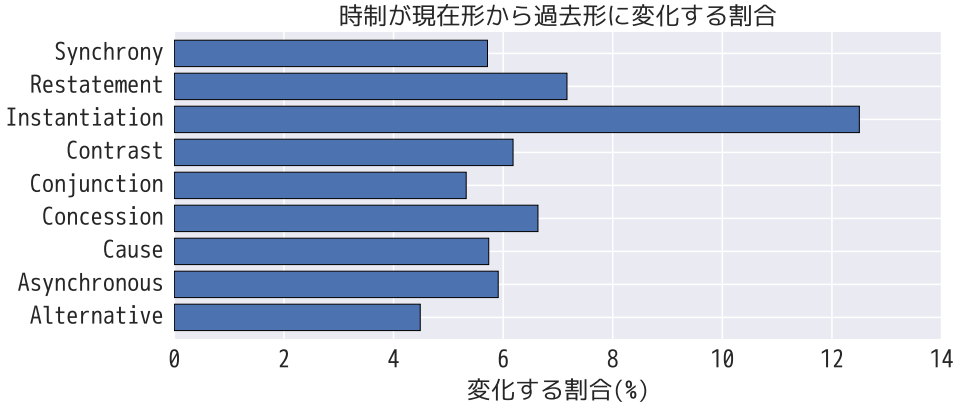

Figure 8: 現在形から過去形への時制変化の割合

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

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．故に，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

本研究を進めるにあたり、ご指導をいただいた乾健太郎教授、岡崎直観准教授, 田然研究特任助教に感謝いたします。また、日常の議論を通じて多くの知識や指摘をくださった乾・岡崎研究室の皆様に感謝いたします。

References

[1] Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. Easily identifiable discourse relations. In *COLING*, pp. 87–90, 2008.

[2] Barbara Hall Partee. Some structural analogies between tenses and pronouns in english. *The Journal of Philosophy*, Vol. 70, No. 18, pp. 601–609, 1973.

[3] Hans Kamp. Formal properties of 'now'. *Theoria*, Vol. 37, No. 3, pp. 227–274, 1971.

[4] Rashmi Prasad and et al. The penn disccourse treebank 2.0. In *LREC*, pp. 2961–2968, 2008.

[5] Erhard Hinrichs. Temporal anaphora in discourses of english. *Linguistics and Philosophy*, Vol. 9, No. 1, pp. 63–82, 1986.

[6] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *ACL and IJCNLP*, pp. 683–691, 2009.

[7] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*, pp. 343–351, 2009.

[8] Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. Using entity features to classify implicit discourse relations. In *SIGDIAL*, pp. 59–62, 2010.

[9] Joonsuk Park and Claire Cardie. Improving implicit discourse relation recognition through feature set optimization. In *SIGDIAL*, pp. 108–112, 2012.

[10] Attapol Rutherford and Nianwen Xue. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *EACL*, pp. 645–654, 2014.

[11] Chloé Braud and Pascal Denis. Comparing word representations for implicit discourse relation classification. In *EMNLP*, pp. 2201–2211, 2015.

[12] Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Implicit discourse relation detection via a deep architecture with gated relevance network. In *ACL*, pp. 1726–1735, 2016.

[13] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T Rutherford. The conll-2015 shared task on shallow discourse parsing. In *CoNLL*, pp. 1–16, 2015.

[14] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *ACL*, pp. 2442–2452, 2016.