述語項構造に基づく言語情報の基本単位のデザインと可視化

齋藤 純 † 坂口 智洋 ‡ 柴田 知秀 ‡§ 河原 大輔 ‡ 黒橋 禎夫 ‡§
† 京都大学工学部 ‡ 京都大学大学院情報学研究科 § 科学技術振興機構 CREST
{saito, sakaguchi}@nlp.ist.i.kyoto-u.ac.jp
{shibata, dk, kuro}@i.kyoto-u.ac.jp

1 はじめに

計算機によるテキスト解析の基本は形態素解析と構文解析である。形態素解析ではテキストを意味をもつ最小限の単位 (形態素) に分割する。構文解析ではテキスト中の句の統語関係を推定する。大規模コーパスの利用や機械学習手法の進展によりこれらの解析精度は向上しており、多くのツールが構築され広く用いられている [6][8][11]。

しかし、テキストマイニングや情報分析など、事象や行為を扱う言語処理アプリケーションにとっては、形態素や句は粒度が小さすぎる。そのため、まずテキストを形態素解析器や構文解析器を用いて形態素や句のレベルで解析した後、各解析結果を解釈し目的に合った粒度に再構成する必要がある。これまでのアプリケーション構築の際には、この処理を各々で開発する必要があり、コストが大きいという問題があった。例えば述語項構造解析器 KNP[11] は図 1 のような出力形式であり、小さい粒度の情報が多いためこれを適切に解釈し再構成するのは容易ではない。

言語処理のアプリケーションで広く利用されている単位の一つが述語項構造である。述語項構造は述語とその項からなる構造であり事象や行為を表す単位として用いられることが多い。しかし、事象や行為を表すには情報が不十分なことも多く、各アプリケーションで独自に修正・拡張する必要がある。

(1) 昨日言語処理学会の懇親会に登録をした。

例えば文 (1) の述語は「した」であるが、実質的な意味を担うのは事態性名詞「登録」であり、「した」を事象として扱う利点は少ない。また項を句単位で考えると「登録」のニ格は「会」あるいは「懇親会」となってしまうが、これでは事象を理解するには情報が不十分であり、より前方の修飾句も含める必要がある。また、述語項構造間の関係は談話関係解析器などを用いて推定する必要があり、アプリケーションとして事象や行為を扱うためには多くの処理が必要となる。

```
雨が降ったので昨日の祇園祭は中止になったそうだ。
# S-ID:1 KNP:4.14-CF1.1 DATE:2018/01/17 SCORE:-40.76235
* 1D <文頭><ガ><助詞><体言><一文字漢字><係:ガ格><区切:0-0><格要素><連用要素><正規化代表表記:雨/あめ?雨/う><主辞代表表記:雨/あめ?雨/う>
+ 1D <文頭><ガ><助詞><体言><一文字漢字><係:ガ格><区切:0-0><格要素><連用要素><名詞項候補><先行詞候補><正規化代表表記:雨/あめ?雨/う><解析格:ガ>
雨 あめ 雨 名詞 6 普通名詞 1 * 0 * 0 "代表表記:雨/あめ 漢字読み:訓 カテゴリ:抽象物" <代表表記:雨/あめ><漢字読み:訓><カテゴリ:抽象物><正規化代表表記:雨/あめ?雨/う><品曖><ALT-雨-う-雨-6-1-0-0-"代表表記:雨/う 漢字読み:音 カテゴリ:抽象物"><品曖-普通名詞><原形曖昧><文頭><漢字><かな漢字><名詞相当語><自立><内容語><タグ単位始><文節始><文節主辞><名詞曖昧性解消>
が が が 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* 5D <時制-過去><用言:動><係:連用><レベル:B+><区切:3-5><ID:〜ので><提題受:20><連用要素><連用節><動態述語><正規化代表表記:降る/ふる><主辞代表表記:降る/ふる>
+ 6D <時制-過去><用言:動><係:連用><レベル:B+><区切:3-5><ID:〜ので><提題受:20><連用要素><連用節><動態述語><節機能-理由><正規化代表表記:降る/ふる><用言代表表記:降る/ふる><格関係0:ガ:雨><格解析結果:降る/ふる:動1:ガ/C/雨/
```

図 1: 述語項構造解析器 KNP の出力例

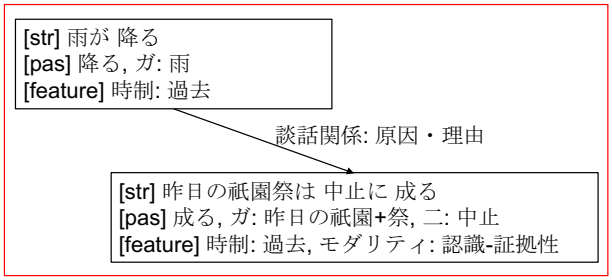

図 2: 「雨が降ったので昨日の祇園祭は中止になったそうだ。」という文のイベントグラフ。黒枠はイベント、赤枠は文を表し、矢印はイベント間関係を表す。

本研究ではこのような問題を統一的に解決することを目指し、言語情報の基本単位を導入する。これを**イベント**と呼ぶ。イベントは、述語項構造をベースにしているが、上記の問題に対処したもので、事象や行為などを扱うのに適した単位となっている。さらに、テキストをイベント (ノード) とイベント間関係 (エッジ) からなるグラフへと変換する枠組みを提案する。本稿ではこのグラフを**イベントグラフ**と呼ぶ。この枠組みの導入により、アプリケーション構築の際は形態素や句のレベルからではなく、事象や行為のレベルからテ

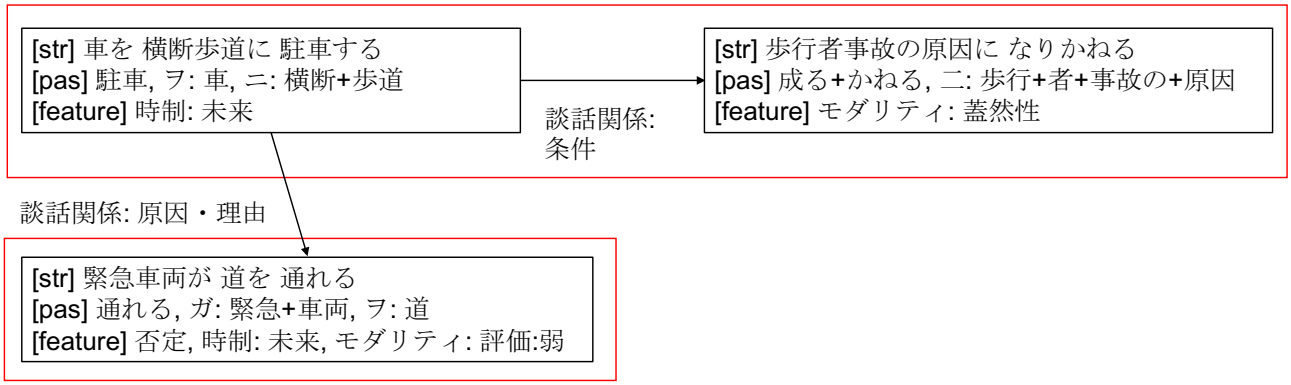

図 3: 「車を横断歩道に駐車すると、歩行者事故の原因になりかねません。また、緊急車両が道を通れないこともあるでしょう。」という 2 文からなるテキストのイベントグラフ。文内だけでなく、文をまたいだイベント間にもエッジが張られている。

キストを処理することが可能となる。

イベントは、テキスト中の述語とその概念を構成する項、そしてモダリティやテンスなどの述語の機能情報からなる構造である。イベント間に談話関係や係り受け関係がある場合はそれらがエッジとなる。

(2) 雨が降ったので昨日の祇園祭は中止になったそうだ。

図 2 は文 (2) をイベントグラフに変換したものである。この文には「降る」と「なる」の 2 つの述語があり、それぞれがイベントを構成する。図中の [str] はイベントを構成する述語や項の文字列、[pas] は述語と項の格関係、[feature] はテンスや否定、モダリティなど述語の機能情報を表す。またこれらのイベントの間の「談話関係: 原因・理由」というエッジは、イベント間の関係を表している。

このようなテキストのイベントグラフへの変換は、複数文からなるテキストに対しても同様に行うことができる (図 3)。

2 関連研究

述語項構造解析に関する研究はこれまでに多く行われてきた。代表的なものでは、Gildea ら [3] が述語に対する項の適切な意味役割を自動的に付与するタスクを提案しており、これまで継続的に研究されている。日本語では河原ら [11] が構文・格解析の統合的確率モデルによる高精度な述語項構造解析手法を提案しており、KNP と呼ばれるツールとして広く利用されている。しかし、これらの研究では事象単位での処理を想定していないため、これらの解析結果からイベントを抽出するのはコストがかかる。

文同士・節同士の意味的なつながりは談話関係と呼ばれ、多くの研究がなされている。英語では、専門家による談話関係タグ付きコーパスとして Penn Discourse Treebank [9] や RST-DT[2] が存在する。日本語では、河原ら [12] がクラウドソーシングを活用して談話関係タグ付きコーパスを構築し、談話関係解析器を開発した。本研究ではこの談話関係解析器を用いて、イベント間の関係を推定する。

文の意味表現として、Banarescu ら [1] の抽象的意味表現 (Abstract Meaning Representation, AMR) がある。AMR は一文に対して意味を持った概念を取り出してグラフ構造で表現するものである。しかし、AMR は文章を扱っておらず、またグラフ構造中の各部分が文中のどの単語に対応しているかという対応関係が一般に付与されていない点が本研究とは方向性が異なる。

3 イベントグラフ

3.1 イベントグラフの定義・構成

イベントグラフは、イベントとそれらの関係からなる。イベントは、イベントグラフのノードに対応し、述語項構造をベースとした、事象や行為を表す基本単位である。本研究ではアプリケーションでの利用を想定し、述語であっても意味の重要性や事象性が希薄なものはイベントとしない。例えば次のような述語はイベントを構成しない。

1. 副詞的な形容詞
   (例) 「大きく表示する」の「大きい」
2. 単独では意味が弱い述語
   (例) 「登録をする」「登録を行う」の「する」や「行う」

3. 機能的な述語

(例) 「美しいと思う」の「思う」。「美しい」の構成するイベントに属し、推量・伝聞を表す機能情報として扱う。

イベントは、テキスト中の述語とその概念を構成する項、述部のモダリティやテンスなどの機能情報からなる。項が複合名詞である場合や連体修飾されている場合は、それらの句・節の全体を項とする。例えば文 (1) において「登録」のニ格は「言語処理学会の懇親会」となる。

イベント間の関係は、文法的関係と意味的関係の 2 種類からなる。文法的関係は、述語間の係り受け関係であり、これに基づきイベント間にエッジをはる。意味的関係は、原因・理由や目的、条件などの談話関係であり、談話関係が認定されたイベント間にエッジをはるとともに、その談話関係をエッジのラベルとする。あるイベント間に両方の関係が存在する場合は談話関係を採用する。

3.2 イベントグラフの表現形式

様々な言語処理アプリケーションにおいてイベントグラフを利用するために、イベントグラフの表現形式を策定した。この表現形式を**イベントグラフフォーマット**と呼ぶ。イベントグラフフォーマットは、簡便な key-value 形式を採用しており、3.1 節で述べたイベントグラフの構成要素に加え、形態素情報や統語情報を含んでいる (図 4)。従って、言語処理アプリケーションは、形態素解析器や構文解析器の出力結果を直接扱う必要はなく、イベントグラフフォーマットのみを扱えばよい。

本研究では、イベントグラフフォーマットを JSON 形式で実装し、テキストをイベントグラフフォーマットに変換する変換器を作成した。変換器はまず、入力テキストに対して JUMAN++と KNP を用いた形態素・構文・省略解析を適用し、さらに河原ら [12] の手法を用いた談話関係解析を行う。その後、これらの解析結果と 3.1 節の定義を対応付けるルールを用いてイベントグラフフォーマットに変換する。

4 イベントグラフを用いたアプリケーション

本節では、これまで形態素や句を用いて構成していた事象を、イベントグラフを用いることで容易に実装できた 2 つの応用例を示す。

```
"event": {
   "str": "雨が 降る",
   "pas": {
       "predicate": {
           "str": "降る",
           "repname": "降る/ふる",
           "sentence_id": 1,
           "token_id": 1
       },
       "argument": {
           "ガ": {
               "str": "雨が",
                       ⋮
   "str": "昨日の祇園祭は 中止に 成る",
   "pas": {
       "predicate": {
           "str": "成る",
           "repname": "成る/なる",
                       ⋮
```

図 4: 図 2 の文のイベントグラフフォーマット (一部)

4.1 タイムライン構築

長年に渡り蓄積されてきた大量のテキストの比較や集約、分析を目的として、テキストの時間情報解析に関する様々な研究が行われている。中でも、あるトピックに関してテキスト横断的に事象と時間情報を対応付けたものはタイムラインと呼ばれ、情報集約の観点から自動生成に関する研究が行われている [7]。

これらの先行研究では述語を事象そのものと見なして扱うことが多いが、タイムラインをアプリケーションとして考えると、事象や行為の主体となるエンティティなども不可欠な情報である。本稿で提案したイベントはこのような事象を一単位として扱うことを目的としているため、この枠組みを用いることで図 5 に示すようなユーザが事象について理解・想像するに必要十分な情報を容易に提示することが可能となる。

4.2 因果関係に基づく意見集約と可視化

近年、Web ページや新聞記事、SNS などの大規模なテキストから知識を取り出す試みが盛んに行われている。とりわけ因果関係抽出は対話システムなどへの応用が期待され、盛んに研究が行われている [5][4]。本研究では、三澤ら [10] が構築した消費者の意見を収集したコーパスから因果関係をもつ意見の集約と可視化を行った。

コーパスをイベントグラフに変換した後、イベント間に「原因・理由」の談話関係があるイベントペアを抽出し、クラスタリングを行った。図 6 にその一例を示す。イベントグラフの導入により、句単位の解析結果使用時と比較して、意見抽出や因果関係抽出を独自

時間	事象	文書作成日	文書・文ID
1980-01 〜 1995-01-05	Aさんは、ここ十五年ほど、正月には必ず木曽駒ケ岳を訪れていた。	1995-01-05	950105230 005
1995-01-02	Aさんらは二日早朝に高速バスで東京から現地入り。	1995-01-05	950105118 014
1995-01-02 〜 03	Bさんによると、一行はその夜は雪洞を泊まった。	1995-01-05	950105230 006
1995-01-03	Bさんらは三日午前中に木曽駒ケ岳に登頂。	1995-01-05	950105230 009
	Bさんらは全員で昼食を済ませた。	1995-01-05	950105230 010
	一行はBさんは「四日は宝剣岳に登り、下山すると話していたのだが……」	1995-01-05	950105230 011
1995-01-04	四日午前十一時十分ごろ、 中央アルプス・木曽駒ケ岳近くの千畳敷カールで雪崩が 発生。	1995-01-05	950105003 000
	この日の捜索を断念、同県警では、午後五時、捜索を打ち切った。	1995-01-05	950105003 004
1995-01-05	事故で、長野県警は五日午前六時四十五分から、駒ケ根署員や山岳救助隊員約五十人で捜索を再開。	1995-01-05	950105169 000

図 5: ニュース記事から生成されたタイムライン例

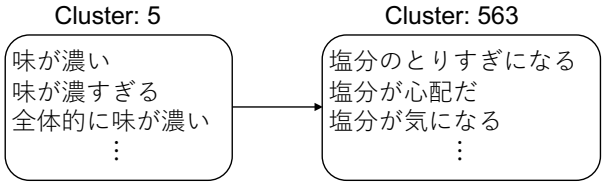

図 6: 因果関係を持つ意見クラスターの例

に判定・実装する必要がなくなり、処理の手間が大きく軽減された。

5 おわりに

本研究では、事象レベルでのテキスト解析において基本単位として利用できる、述語項構造に基づく意味構造「イベントグラフ」を提案した。これを利用することで、これまで形態素や句のレベルから行っていた処理を事象のレベルから簡単に行うことが可能となる。

今後はより多くのアプリケーションによる利用を想定し、イベントに対してより幅広い情報を付与したいと考えている。例えば現在イベントに付与されている情報は統語的なものが多いが、イベントの項の情報に対して Wikification を行うなどより多くの意味情報を付与することでアプリケーションの多様な用途に対応したい。

参考文献

[1] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[2] Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania, 2002.

[3] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, Vol. 28, No. 3, pp. 245–288, 2002.

[4] Christopher S. G. Khoo, Syin Chan, and Yun Niu. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of 38th Annual Meeting of the ACL, Hong Kong, 2000*, pp. 336–343, 2000.

[5] Christopher S. G. Khoo, Jaklin Kornfilt, Robert N. Oddy, and Sung Hyon Myaeng. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, Vol. 13, No. 4, pp. 177–186, 1998.

[6] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of EMNLP*, pp. 230–237, 2004.

[7] Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, and Ruben Urizar. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 778–786, Denver, Colorado, June 2015. Association for Computational Linguistics.

[8] Graham Neubig, 中田陽介, 森信介. 点推定と能動学習を用いた自動単語分割器の分野適応. 言語処理学会第 16 回年次大会 (NLP2010), 東京, 2010.

[9] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC*, 2008.

[10] 三澤賢祐, 成田和弥, 田内真惟人, 中島正成, 黒橋禎夫. 定量調査のための意見調査コーパス構築への取り組み. 言語処理学会 第 23 回年次大会, pp. 1014–1017, 2017.

[11] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[12] 河原大輔, 町田雄一郎, 柴田知秀, 黒橋禎夫, 小林隼人, 颯々野学. 2 段階のクラウドソーシングによる談話関係タグ付きコーパスの構築. 情報処理学会研究報告, 2014.