

動詞項構造辞書への大規模用例付与

平野 徹[†] 飯田 龍[†] 藤田 篤[†]
乾 健太郎[†] 松本 裕治[†]

本論文では、述語項構造解析の精度向上のために必要となる大規模な項構造タグ付き事例を効率的に作成する方法について議論する。項構造タグ付き事例の効率的な作成方法にはさまざまな方法が考えられるが、本論文では大規模平文コーパスから抽出した表層格パターンの用例集合をクラスタリングし、得られたクラスタに項構造タグを付与することでタグ付与コストを削減する手法を提案する。提案手法では、(i) 表層格パターン同士の類似性と (ii) 動詞間の類似性という2種類の類似性を利用してクラスタリングを行う。評価実験では、実際に提案手法を用いて8つの動詞の項構造タグ付き事例を作成し、それを用いた項構造解析の実験を行うことによって、提案手法のクラスタリングの性能や、人手でタグ付き事例を作成するコストと項構造解析精度の関係を調査した。

キーワード: 項構造解析, 交替, 用例クラスタリング

Augmenting a Semantic Verb Lexicon with a Large Scale Collection of Example Sentences

TORU HIRANO[†], RYU IDA[†], ATSUSHI FUJITA[†], KENTARO INUI[†]
and YUJI MATSUMOTO[†]

In this paper, we propose a method of reducing the cost of annotating examples with argument structure in order to increase accuracy of argument structure analysis. First, a large raw corpus is parsed, and a large scale collection of example sentences is constructed from predicate-argument examples in the parsing results. Second, the collection of example sentences is clustered by using two similarities about verb. Finally, the acquired clusters are annotated with argument structure by human. We report preliminary experiments using our proposed method, and show that the method is effective in reducing the cost of annotating.

KeyWords: *Argument Structure Analysis, Alternations, Example Clustering*

1 はじめに

述語項構造とは述語とその項の間の意味的な関係を表現する構造の一つである。例えば、「彼が扉をひらく」という文中の述語「ひらく」の項構造は [agent, theme] のように表すことができる。agent, theme は項が述語に対してどのような意味的關係となっているかを表す意味役割である。また、所与の文章中の各述語に対して、(1) 述語が取り得る項構造のうち最も文の解釈

[†] 奈良先端科学技術大学院大学情報科学研究科, Graduate School of Information Science, Nara Institute of Science and Technology

に適った項構造を選択し、(2) その構造の各項に対応する要素を同定することで述語項構造を出力する処理を項構造解析と呼ぶ。例えば、文「彼が扉をひらく」を述語項構造解析する場合には、述語「ひらく」に対して図1に示すような項構造辞書から対応する項構造を選択し、入力文の格要素を各項に割り当てて構造 [agent:彼, theme:扉] を得る。項構造解析を高精度で実現すれば、「彼が扉をひらく」⇔「扉がひらく」のような交替に代表される表現の多様性を吸収でき、言い換えや情報抽出、質問応答などの自然言語処理技術を高度化できる。

述語の項構造に関する研究は、Filmore (1968) の格文法など古くから関心を集めている。これらの研究は、項構造辞書の作成、項構造タグ付きコーパスの作成、項構造解析の三つの研究に大別でき、項構造辞書作成の研究では、近年、Dorr (1997) によって項構造辞書作成の方法論が開発されている。この研究成果から大規模な項構造辞書を作成する基盤ができてきた。また項構造情報を含む詳細な動詞辞書 FrameNet (Baker, Filmore, and Lowe 1998) や項構造タグ付きコーパス PropBank (Kingsbury and Palmer 2002) も報告されている。項構造解析の研究は国際会議 CoNLL における Shared Task¹として取り上げられるなど関心が集まっており、近年提案されている主な手法は教師なし手法と教師あり手法に大別できる。現状では、PropBank のような項構造タグ付きコーパスが作成されたこともあり、教師あり手法の研究が盛んである。教師なし手法では、Lapata and Brew (1999) のように項構造辞書の下位範疇化の構造を利用して擬似的に訓練事例を作成する手法などが提案されているが、一般に解析精度が低い。これに対して、Gildea and Jurafsky (2002), Hacioglu ら (2003) や Thompson ら (2003) の提案する教師あり手法では、項構造タグが付与された学習コーパスから述語と文章中の要素との構文構造における位置関係などを素性として利用しており、教師なし手法よりも精度が高いという利点を持つ。しかし学習に用いるコーパス中の各述語に対し (i) 取り得る項構造と項構造辞書中の項構造の対応付け、および (ii) 選択した項構造の各項と文章中の要素の対応付け、という人手による項構造タグ付与作業が必要であるため作業コストが高いという問題がある。

そこで本研究では、項構造タグ付き事例を効率的に作成する方法について議論する。項構造タグ付き事例の効率的な作成方法にはさまざまな方法が考えられるが、本論文では、学習に用いるコーパス中の各述語に項構造タグを付与する過程で生じる類似用例への冗長なタグ付与作業の問題に着目する。具体的には、大規模平文コーパスから抽出した表層格パターンの用例集合をクラスタリングし、得られたクラスターに項構造タグを付与することでタグ付与コストを削減する手法を提案する。提案手法では、(A) 表層格パターン同士の類似性と (B) 動詞間の類似性という2種類の類似性を利用してクラスタリングを行う。評価実験では、実際に提案手法を用いて8つの動詞の項構造タグ付き事例を作成し、それを用いた項構造解析の実験を行うことによって、提案手法のクラスタリングの性能や、人手でタグ付き事例を作成するコストと項構造解析精度の関係を調査した。

¹ <http://www.lsi.upc.edu/~srlconll/>

動詞：ひらく			
語義1: 閉じていた物を開放状態にする。			
項構造1 [agent, theme]			
表層格パターン1	が(agent), を(theme)	表層格パターン2	が(theme)
用例	彼が扉を	用例	扉が
	部長が口を		傘が
	⋮		⋮
語義2: 会を催す。			
項構造2 [agent, theme]			
表層格パターン1	が(agent), を(theme)		
用例	出版社がサイン会を		
	デパートが展覧会を		
	⋮		

図 1 動詞項構造辞書の一例（「ひらく」について）

2 用例の類似性に基づく項構造タグ付与の効率化

入力文の各動詞の項構造を解析するタスクには、次の5種類の曖昧性の解消が必要である。

曖昧性(1) 「は」、「も」、「無格²」が兼務する表層格の曖昧性

「は」、「も」、「無格」が表層「ガ格」、「ヲ格」、もしくはそれ以外となる曖昧性。

曖昧性(2) 連体修飾節に関する曖昧性

連体修飾節の述語と被修飾名詞との間に格関係がない外の関係（例えば、「魚を焼くにおい」の「焼く」と「におい」と格関係がある内の関係（「魚を焼く男」の「焼く」と「男」）の曖昧性。さらに、内の関係における被修飾名詞の表層格の曖昧性。

曖昧性(3) 述語の取る項構造の曖昧性

動詞「話す」のように、項構造 [agent, theme, beneficiary] や [agent, beneficiary, content] などの複数の項構造を持つ述語が存在するために生じる曖昧性。例えば、「女性が悲鳴をあげる」と「彼が彼女にプレゼントをあげる」の動詞「あげる」のように、複数の語義を持つことによる曖昧性も含む。

曖昧性(4) 格の意味役割の曖昧性

² 無格とは、「明日東京に行く」の「明日」のように助詞のない格。

同じ述語の同じ表層格でも項構造中の意味役割が一意に決まらないという曖昧性。例えば「彼が扉をひらく」の表層格「彼」は項構造中の agent に対応するのに対し、「扉がひらく」の表層格「扉」は theme に対応する。また表層「二格」のように時間、目的地、場所などさまざまな意味役割を担う表層格がある。

曖昧性 (5) 接尾辞を伴うことによる格交替の曖昧性

受身「れる、られる」や使役「せる、させる」によって項構造の意味役割と表層格の対応関係が変化することによって生じる曖昧性。

この 5 種類の曖昧性のうち曖昧性 (1), (2) は、表層格レベルの問題であるため、河原ら (2002) の手法のように全自動で収集した用例を使って解消できる可能性がある。すなわち、必ずしも項構造情報を教示したデータが必要とならない。一方、曖昧性 (3), (4), (5) を解消するには述語の取り得る項構造と入力文とを照らし合わせるために、項構造情報のタグを付与したデータが必要となる。本論文では曖昧性 (3), (4) の解消に焦点をあて、項構造タグ付与作業の効率化を図る。具体的には、同じ項構造に対応する用例を自動的にひとまとめにすることでタグ付与作業のコスト削減を目指す。曖昧性 (5) は、本論文では取り扱わないが、各接尾辞で項構造の意味役割と表層格の対応関係に規則性があり本論文で扱う項構造解析手法を拡張し解消できると考えられる。

問題設定として、大量のタグなし用例と項構造辞書が与えられていると仮定する。ここで、用例とは係り受け解析結果から述語とその係り受けを取り出したものを指す。タグなし用例は、大規模な生コーパスの文書を係り受け解析することによって収集可能である。また、図 1 に示すような項構造辞書を仮定する。このような辞書は、Dorr の交替現象に基づいた大規模項構造辞書作成の研究により基盤ができているため、作成可能であると期待できる。この項構造辞書では、各述語に一つ以上の項構造が定義されており、各項構造に取りうる表層格パターンと小規模の用例が記述されているものとする。

以上の仮定のもとで、大規模用例集合を次の二つの類似性に基づいてクラスタリングする。

類似性 A: ある動詞について同じ項構造を持つ用例は、格の出現パターンとそれぞれの格要素が類似している。例えば、2 つの用例「女性が悲鳴をあげる」と「こどもが大声をあげる」で格の出現パターン「が、を」や対応する格の格要素(「女性」と「こども」、「悲鳴」と「大声」)が類似している性質をいう。

類似性 B: 意味的な類似性がある二つの動詞は、格の出現パターンとそれぞれの格要素が類似している。これは、Levin (1993) が提案する交替現象に基づく動詞分類の基本的な考え方であり、意味的に類似する動詞「公表する」と「発表する」の用例「大統領が 98 年に計画を公表する」と「首相が 10 月にプランを発表する」において、格の出現パターン「が、に、を」とそれぞれの対応する格の格要素「大統領」と「首相」、「98 年」と「10 月」、「計画」と「プラン」が類似するという性質をいう。

本論文では、類似性 A と類似性 B に基づいて用例をクラスタリングする．そしてクラスタに項構造を付与する作業はクラスタを代表する用例に対して項構造タグを付与し、それをクラスタの項構造タグと考える．つまり用例クラスタリングで得られたクラスタに対して人手でタグを付与するため、いかにして異なる項構造を持つ用例を 1 つのクラスタに含めることなく、できるだけ少数のクラスタに用例をまとめあげるかが課題となる．2.1 で類似性 A，2.2 で類似性 B に基づく用例クラスタリングについて述べる．

2.1 格要素ベースクラスタリング

格要素ベースクラスタリングでは、類似性 A を利用することで、格の出現パターンや格要素が類似する用例（例えば「女性が悲鳴をあげる」と「こどもが大声をあげる」）を自動的にまとめる．河原・黒橋 (2002) の大規模平文コーパスから表層格フレーム辞書を自動構築する研究は「動詞の用法を決定する重要な格要素は動詞の直前にくることが多く、動詞と直前の格要素をペアにして考えると動詞の用法はほとんど一意に決定される」という考えに基づいている．本論文でも、彼らと同様にこの考えに基づき、次の 3 つのステップでクラスタリングを行う．

- (1) 直前格とその要素が同じ用例のクラスタリング
- (2) 直前格が同じ用例のクラスタリング
- (3) 直前格が異なる用例のクラスタリング

各ステップで対象のクラスタ（用例）間の類似度を計算し、設定する閾値を超えるクラスタ（用例）の組みの中で最も類似度の高い 2 つのクラスタをマージするボトムアップクラスタリングを行う．また、クラスタ（用例）間の類似度として用例収集に用いた大規模平文コーパス中での用例の出現回数を考慮した類似度計算式（詳細は付録 A を参照）を用いる．この式は河原らによって提案されたものである．上の順序で段階的にクラスタリングを行うことで、常に動詞の直前格の出現回数が直前格以外の格の出現回数よりも多くなるように処理を進めることができる．その結果、類似度計算における動詞の直前格の重みが大きくなり、動詞の直前格を重視した用例クラスタリングが可能となる．

クラスタリングの各ステップを例を用いて説明する．最初に、ステップ 1 では例 1 に示すように直前格要素が同じ用例だけを対象にクラスタリングする．例 1 では、一つ目の用例と二つ目の用例の類似度が高くこれらをマージした結果を示している．

例 1)

$$\left. \begin{array}{l} \text{月:に 案:を 発表する} \\ \text{中旬:に 案:を 発表する} \\ \text{久しぶり:に 案:を 発表する} \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \{ \text{月, 中旬} \}: \text{に 案:を 発表する} \\ \text{久しぶり}: \text{に 案:を 発表する} \end{array} \right.$$

次に、ステップ2ではステップ1の結果を入力とし、例2に示すように直前格が同じ用例だけを対象にクラスタリングする。クラスタリングの例を例2に示す。

例2)

$$\left. \begin{array}{l} \{月, 中旬\}: 案: \underline{を} 発表する \\ 大統領: が \{年, 初め\}: 計画: \underline{を} 発表する \end{array} \right\}$$

⇒ 大統領: が {月, 中旬, 年, 初め}: に {案, 計画}: を 発表する

最後に、ステップ3ではステップ2の結果を入力とし、例3に示すように直前格が異なる用例だけを対象にクラスタリングする。このステップ3のクラスタリングによって、例3に示すような表層格の出現順が「に、を」「を、に」と順序が異なっている用例をマージ例えばできる。

例3)

$$\left. \begin{array}{l} 大統領: が \{月, 中旬, 年, 初め\}: に \{案, 計画\}: \underline{を} 発表する \\ \{プラン, 計画\}: \underline{を} \{年, 月\}: \underline{に} 発表する \end{array} \right\}$$

⇒ 大統領: が {月, 中旬, 年, 初め}: に {案, 計画, プラン}: を 発表する

正確に言えば、このクラスタリングは河原らが提案したクラスタリングとは異なる。本クラスタリング手法と河原らが提案した手法の主な異なりはステップ(1)の存在であり、河原らの手法にはステップ(2)と(3)しか存在しない。河原らの手法では動詞、動詞の直前格とその要素を最小単位と考えクラスタリングを開始する。つまり動詞、動詞の直前格とその要素が同じならば同じクラスタであると判断している。河原らは大規模平文コーパスからの表層格フレーム辞書自動構築を目的にクラスタリング手法を提案しており、このように判断することは大きな問題にならない。しかし、我々の目的は項構造を考慮したクラスタリングであり、動詞、動詞の直前格とその要素が同じならば同じクラスタであると判断することが大きな問題となる。そのためステップ(1)を導入した。これ以外にも異なる項構造を持つ用例やクラスタが誤ってマージされないようにいくつか工夫する必要がある。詳細は3.2節で述べる。

2.2 動詞ベースクラスタリング

動詞ベースクラスタリングでは、類似性Bに基づき、同じ項構造に対応する意味的に類似する動詞の用例(例えば、「大統領が98年に計画を公表する」と「首相が10月にプランを発表する」)を自動的にまとめる。具体的には、動詞 T の用例をクラスタリングするのに動詞 T と類似する動詞 $S = \{S_1, S_2, \dots, S_n\}$ の用例を利用する。動詞集合 S をどのように選択するか、また各動詞 S_i の用例をどのように利用するかにはさまざまな方法が考えられるが、今回は動詞を1つだけ用いることにしてクラスタリング手法を検討する。また、本クラスタリングは類似す

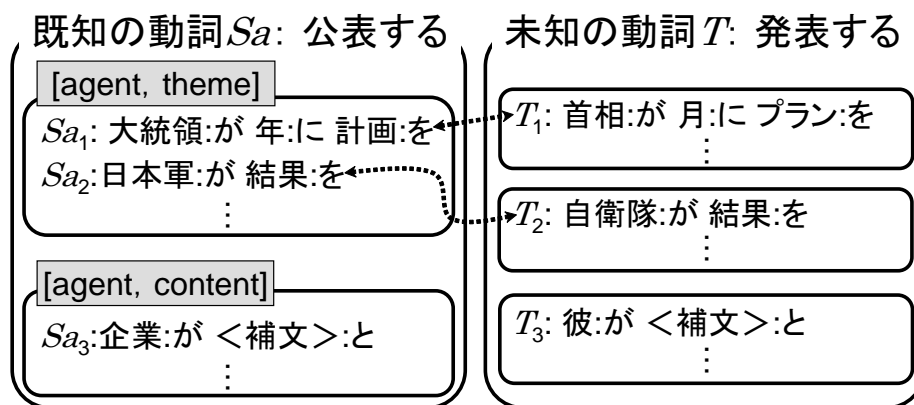


図 2 動詞の類似性に基づくクラスタリング

る動詞の一方の用例に項構造タグを付与した結果を利用する．つまり，すでに人手で項構造タグが用例に付与された動詞 S_a を用いて，動詞 S_a と類似する動詞 T の用例集合を次の3つのステップでクラスタリングする．なお，本クラスタリングの入力は格要素ベースクラスタリングの結果である．

- (1) 表層格が最も多い用例を動詞 T のクラスタを代表する用例として選択する．
- (2) (1) で選択された用例と動詞 S_a の用例の中で表層格パターンが同じでかつ最も類似する用例を対応付ける．
- (3) 動詞 T の異なるクラスタに属する用例が動詞 S_a の同じ項構造の用例に対応付けられた場合に，動詞 T 側の2つのクラスタをマージする．

まずステップ1では，格要素ベースクラスタリングによってできたクラスタにおいて，各クラスタを構成する用例で最も多くの表層格を持つ用例をそのクラスタを代表する用例とする．例えば，「首相が10月にプランを発表する」と「5月に計画を発表する」の用例で構成されるクラスタの例を考える．これらの用例の表層格はそれぞれ「が，に，を」と「に，を」であり，表層格数は前者が3，後者が2であるため，前者の用例をこのクラスタを代表する用例として選択する．図2では，ステップ1で動詞 T の1番目のクラスタの代表用例を T_1 ，2番目のクラスタの代表用例を T_2 として選択したとする．ステップ2で T_1 ， T_2 のそれぞれと最も類似度の高い用例を動詞 S_a 側の用例集合から選択し対応付ける．ここではそれぞれ用例 S_{a1} と用例 S_{a2} が対応付けられたとする．最後に，動詞 S_a 側で用例 S_{a1} ， S_{a2} が同じ項構造であるというタグ付与情報を利用して，1番目と2番目のクラスタをマージする．動詞ベースクラスタリングに

においても、用例間（動詞 T の用例と動詞 S_a の用例）の類似度計算式として、格要素ベースクラスタリングと同じ計算式を用いる。但し本クラスタリングにおいては、表層格パターンが同一の用例のみを類似度計算の対象とした。

動詞ベースクラスタリングでポイントとなるのはステップ 3 であり、本クラスタリングの狙いは次のとおりである。格要素クラスタリングでは対応する格の格要素の類似度だけを考慮しているが、実際には同じ項構造となる用例のなかには類似しない格要素を持つ用例も多い。つまり、述語とある意味的な関係を持つ要素の集合は、その要素自身が持つ意味だけでは表現することができないのである。例えば、図 2 の用例 T_1 「首相が 10 月にプランを発表する」と T_2 「自衛隊が結果を発表する」は格要素の類似度に基づいたクラスタリングでは、マージされなかった用例である。しかしながら、我々人間には同じ項構造であることがわかる。このような格要素の類似性だけではマージできない用例を、類似した他の動詞の用例に人間が項構造を付与することで与えた情報を利用してクラスタリングを行なうのが動詞ベースクラスタリングの狙いである。

3 評価実験

前節で述べたクラスタリングの効果および大規模な訓練事例を利用することがどの程度項構造解析の精度向上に寄与するかを調査した。まず、両実験で使用する実験データと訓練事例作成のための用例クラスタリングの設定、項構造解析モデルについて説明し、最後に実験結果を示す。

3.1 評価実験データ

4 つの動詞（「話す」、「発表する」、「発売する」、「増える」）を用いて評価した。これらの動詞は後述のテストデータのコーパス中に頻出する動詞であり、それぞれの動詞が持つ項構造数は「話す：5」、「発売する：1」、「発表する：3」、「増える：2」である。例として、表 1 に「話す」の項構造辞書の項目を示す³。この項構造辞書は IPAL 動詞辞書（情報処理振興事業協会技術センター 1987）を基に今回収集した用例を参考に、それらの格交替を考慮しながら作成したものである。この 4 つの動詞に対して、毎日新聞社の新聞記事でテストデータ作成に用いた 1ヶ月分を除いた、13 年分の新聞記事から用例異なりを除いた 8385 用例を抽出した（抽出条件は 3.2 の用例の収集を参照）。この用例をクラスタリングし、各クラスに項構造を対応付けて、項構造解析の訓練事例とした。

項構造解析実験のテストデータとして、上の 13 年分の新聞記事から抜き出したある月の新聞記事から対象動詞の用例を抽出し、用例異なりを除いた 220 用例にひとつずつ人手で項構造

³ その他の動詞の項構造辞書の項目は付録 B を参照

表 1 動詞「話す」の項構造辞書の項目

述語:話す	
語義:口に出して、ある事を人に知らせる。	
項構造 1 [agent, theme, beneficiary]	
表層格パターン	用例
が, を, に	彼が 用件を 先方に 話す。
が, について, に	彼が 事件について 警察に 話す。
が, まで, に	先生が そんなことまで 生徒に 話す。
から, について, に	私から 結婚について 親に 話す。
から, を, に	彼女から 将来のことを 彼氏に 話す。
項構造 2 [agent, beneficiary, content]	
表層格パターン	用例
が, に, と	彼が 母に 明日出発すると 話す。
項構造 3 [agent, beneficiary, theme, content]	
表層格パターン	用例
が, に, を, と	妻が 私に 映画を すばらしかったと 話す。
が, に, について, と	警察が 市民に 事件について 解決したと 話す。
語義:複数の者で会話、討議する。	
項構造 4 [agent, theme]	
表層格パターン	用例
が, を	政治家が 貿易摩擦問題を 話す。
が, について	先生らが いじめについて 話す。
語義:ある言語を用いる。	
項構造 5 [agent, theme]	
表層格パターン	用例
が, を	彼が 英語を 話す。

タグを付与したものをを用いる。ただし、2節で述べたように(1)「は」、「も」、「無格」が兼務する表層格の曖昧性、および(2)連体修飾節に関する曖昧性は表層格パターンを用いてある程度解析可能であると考えられるため、今回のテストデータについては「は」、「も」、「無格」の曖昧性は人手で解消し、連体修飾節の被修飾名詞は取り除き(3)、(4)の曖昧性の解消に焦点を当てて評価する。また、(5)の接尾辞を伴うことによる格交替の曖昧性については、文書中の約1割の述語が格交替を生じる接尾辞を伴って出現したが、訓練事例の抽出と同様に今回はそれらを除いて評価実験を行った。なお訓練事例、テストデータともに用例に項構造タグを付与する際、可能な項構造が複数あれば複数のタグを付与した。また項構造解析時には、3.3の項構造解析モデルで複数のタグを持つ用例が選択された場合、複数の項構造解析結果を出力する。表2に訓練事例、表3にテストデータ中の動詞とその項構造の出現回数を示す。各動詞の項構造番号は、表1や付録Bに示した動詞項構造辞書の項目番号と対応している。評価実験では、動詞「増える」の場合、システムが出力すべき項構造の数は142個(5+7+65×2)である(表3参照)。

表 2 訓練事例中の動詞とその項構造の出現回数

動詞	出現回数	項構造の種類	各項構造の出現回数 (項構造番号 : 出現回数)
話す	1867	5	1 : 105, 2 : 22, 3 : 133, 4 : 232, 5 : 62, 1+2+3 : 155, 1+3 : 194, 1+3+4 : 364, 2+3 : 214, 1+2+3+4+5 : 386
発表する	2822	3	3 : 60, 1+3 : 1286, 2+3 : 918, 1+2+3 : 558
発売する	635	1	1 : 635
増える	3061	2	1 : 138, 2 : 232, 1+2 : 2691

表 3 テストデータ中の動詞とその項構造の出現回数

動詞	出現回数	項構造の種類	各項構造の出現回数 (項構造番号 : 出現回数)
話す	57	5	1 : 4, 3 : 5, 4 : 8, 5 : 2, 1+2+3 : 4, 1+3 : 5, 1+3+4 : 10, 2+3 : 7, 1+2+3+4+5 : 12
発表する	76	3	3 : 4, 1+3 : 33, 2+3 : 24, 1+2+3 : 15
発売する	10	1	1 : 10
増える	77	2	1 : 5, 2 : 7, 1+2 : 65

3.2 用例クラスタリングの設定

用例の収集

項構造と対応付けるために収集できる用例には「は」、「も」、「無格」を伴った用例があるが、これらは兼務する表層格の曖昧性があるため訓練事例としては収集しない。また、格交替が生じる可能性のある接尾辞「れる」、「られる」、「せる」、「させる」、「たい」、「ほしい」、「もらう」、「いただく」、「くれる」、「くださる」、「やる」、「あげる」、「できる」を伴う述語の用例や連体修飾節の被修飾名詞も訓練事例としては収集しない。

格要素ベースクラスタリング

用例間の類似度の計算式は河原ら (2002) の提案した計算式⁴ を用いた。ただし、異なる項構造の用例を誤ってマージしないように次の制約を加えた。

- 一方の用例の格パターンが他方の用例の表層格パターンを包含する場合のみマージする。この制約により、用例「彼が結婚について両親に話す」と「私から結婚について両親に話す」を「彼が私から結婚について両親に話す」のように「話す」の agent が 2 つ存在するといった誤ったマージを避ける。
- ボトムアップクラスタリングの際に類似度を再計算する用例ペアとして各ステップの初めの類似度計算で設定した閾値を超えた用例ペアだけを対象にする。この制約により、直前格を重要視しすぎないように制御する。

⁴ 用例間の対応する格の割合と、対応する格要素の類似度による計算式。詳しくは付録 A を参照。

格要素ベースクラスタリングでは、3ステップのそれぞれで閾値を設定する必要がある。今回は 0.9, 0.85, 0.8 とした。

動詞ベースクラスタリング

4つの動詞「話す」、「発表する」、「発売する」、「増える」(図2における未知の動詞 T) の用例クラスタリングに「語る」、「公表する」、「売り出す」、「減る」(図2における既知の動詞 S) のタグ付き用例 4551 個を利用する。このタグ付き用例は格要素ベースクラスタリングの結果に人手で項構造タグを付与し作成したものである。

動詞ベースクラスタリングのステップ2の用例間の類似度計算には、格要素ベースクラスタリングと同じ類似度計算式を用いた。ただし、用例の表層格パターンが同一のもののみ対象とし、閾値は 0.85 に設定した。

3.3 項構造解析モデル

項構造解析モデルとしては、用例に対応する項構造候補の選択に最近傍法を用い、また用例間の近さを計算するのに Kurohashi ら (1994) が提案する計算方法を用いた。この計算方法は分類語彙表 (国立国語研究所 1993) を利用して名詞間の類似度を定め、また用例間の類似度に名詞間の類似度の和を用いている。項構造解析の処理を説明する。

- (1) 入力文の格要素とタグ付き用例の対応付けを行う
- (2) 対応付けられたそれぞれの格要素について、入力文の名詞とタグ付き用例との間の類似度を計算する。類似度の値は分類語彙表 (国立国語研究所 1993) における2つの名詞の分類コードの一致するレベルによって決定する。一致するレベルと類似度との関係を表4に示す。
- (3) 式1に従ってタグ付き用例と入力文の対応の評価値を計算し評価値の高い用例の持つ項構造から順に選択する。

$$\text{評価値} = \begin{cases} 0 & \text{if } l > n \\ \text{sum} \times \sqrt{\frac{1}{m}} & \text{otherwise} \end{cases} \quad (1)$$

n : 対応付けられた格要素数

l : $n +$ (入力文側の対応付けられない格要素数)

m : $n +$ (タグ付き用例側の対応付けられない格要素数)

sum : 対応付けられた格要素の類似度の和

表 4 黒橋らが提案した名詞間の類似度

レベル	0	1	2	3	4	5	6	一致
類似度	0	0	5	7	8	9	10	11

3.4 提案手法の有用性の評価実験

提案手法の有用性を示すために次の3点を経験的に明らかにする。

(a) 項構造タグ付与作業のコスト 訓練事例作成のために、クラスタリングの結果得られるクラスタに項構造タグを手で付与する必要がある。今回の実験では、クラスタに項構造を付与する作業はクラスタを代表する用例に対して項構造タグを付与し、それをクラスタの項構造タグと考える。そのためクラスタリング結果のクラスタ数でタグ付与作業のコストを評価する。

(b) タグ付与の品質 項構造タグ付与誤りを調査するために、人手で各用例に項構造タグを付与した。また提案手法によって得られたクラスタに項構造を付与する作業はクラスタを代表する用例に対して項構造タグを付与し、それをクラスタの項構造と考える。これらの結果を比較し、タグ付与誤りの割合でタグ付与の品質を評価する。

(c) 項構造解析精度 項構造解析は一文ごとに処理を行うため、省略などによって文脈を見なくては項構造を一意に決定することができない。省略のある入力に対しては可能な項構造解析結果を漏れなく出力することが望まれる。そこで項構造解析の評価尺度として、類似度の順で解析結果を出力していき、正解を漏れなく答えたときの精度で評価する。つまり、再現率が100%のときの精度⁵で項構造解析を評価する。なお評価は事例単位でなく、項構造単位で再現率と精度を計算した。例えば、入力「父が事件について話す」の項構造は、表1に示すように、項構造1[agent, theme, beneficiary]と項構造3[agent, beneficiary, theme, content]の可能性がある。このような入力に関して、項構造1と項構造3の両方の答えを出力するまでの解析結果の精度で評価する。ただし、タグ付き用例集合が入力文の正解を網羅していない場合は、その入力に対して可能なすべての項構造をシステムの出力とする。

上の(a)項構造タグ付与作業のコスト、(b)タグ付与の品質、(c)項構造解析精度について以下の3つの手法を用いて訓練事例を作成し比較実験を行なった結果を表5に示す。

(1) ベースライン

各用例に項構造タグを付与し訓練事例を作成。

⁵ 再現率 = 出力された正解項構造数 / 正解項構造数, 精度 = 出力された正解項構造数 / 出力された項構造数

表 5 提案手法の作業コスト、品質と項構造解析精度

	(a) [個]	(b) [%]	(c) [%]
(1) ベースライン	8385	0.00	99.3
(2) 格要素ベースクラスタリング	2745	0.31	97.7
(3) 動詞ベースクラスタリング	1505	0.70	97.3

(2) 格要素ベースクラスタリング

類似性 A (ある動詞について同じ項構造を持つ用例は、格の出現パターンとそれぞれの格要素が類似している) に基づいた用例クラスタリングの結果に項構造タグを付与し訓練事例を作成。

(3) 動詞ベースクラスタリング

格要素ベースクラスタリングに加え、類似性 B (意味的な類似性がある二つの動詞は、格の出現パターンとそれぞれの格要素が類似している) に基づいた用例クラスタリングの結果に項構造タグを付与し訓練事例を作成。

表 5 の (a) 項構造タグ付与作業のコストが示すように、格要素ベースクラスタリングによってベースラインの作業コストを約 3 分の 1 (2745/8385) に削減し、動詞ベースクラスタリングを用いてさらに半減 (1505/2745) できた。この結果より、用例のクラスタリングに一般的な名詞の類似度を用いた手法に加え、類似する動詞のタグ付き用例を利用して、未知の動詞の用例をさらにマージすることができたと言える。また、動詞ベースクラスタリングは格要素ベースクラスタリングと比べ、タグ付与品質では多少の低下が見られるものの、項構造解析の精度にはほとんど影響しなかった。この結果から、項構造タグ付与作業済みの動詞が増えると、ある動詞と意味的に類似する動詞が増加するので、より多くの動詞ベースクラスタリングが可能になり、項構造解析精度を保ったままタグ付与作業のコストをさらに削減できると考えられる。また、ベースラインと格要素ベースクラスタリングにおける項構造解析精度の差は、2.2 節で動詞ベースクラスタリングのねらいとして述べたように、述語とその項の間の意味的な関係を示す項構造の項となる要素の集合は、その要素自身が持つ意味だけでは表現することができないことを示す結果になっている。

3.5 用例規模の異なりによる評価実験

大規模な訓練事例を利用することがどの程度項構造解析精度の向上に寄与するかを評価するため、用例集合の規模を変化させたときの (a) クラスタ数と (c) 項構造解析精度の変化を調べた。

図 3, 4 は動詞ベースクラスタリングまで施したときの結果である。横軸は用例の規模 (年単位) であり、棒グラフはそれぞれ用例異なり数とそれをクラスタリングすることによって得られるクラスタ数を示している。また折れ線グラフは項構造解析精度 (図 3: テストデータの

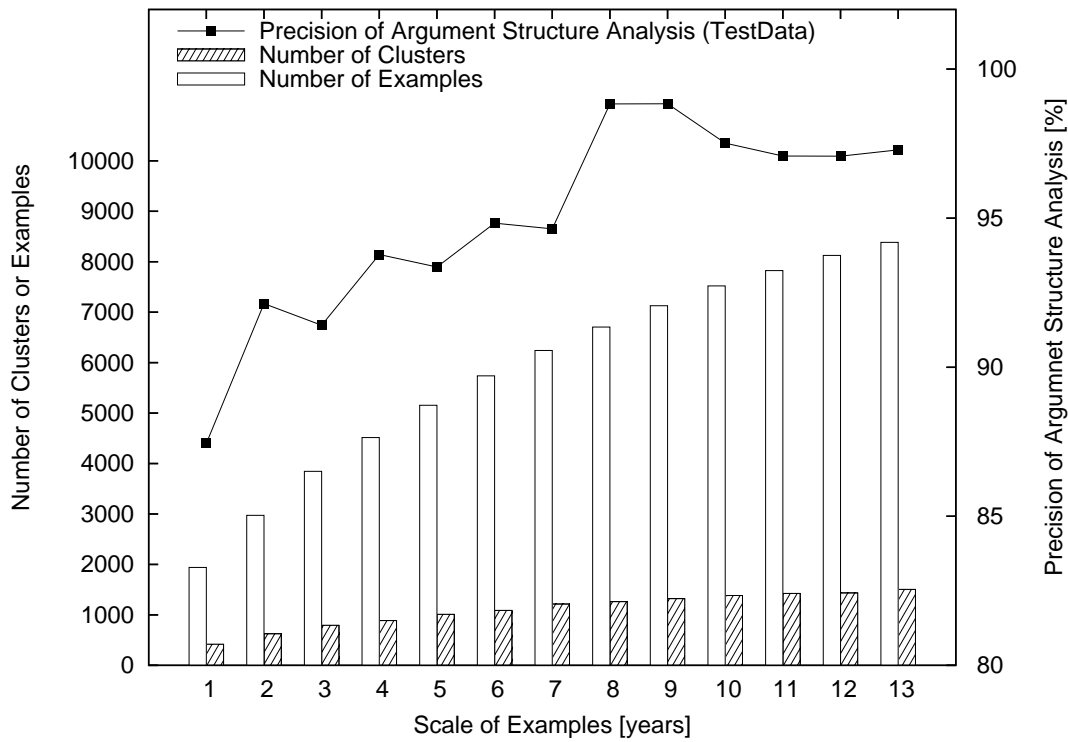


図 3 用例規模に対する用例異なり数とクラスタ数，項構造解析精度（テストデータ）

220 用例に対する結果，図 4：訓練事例の 8385 用例の 10 分割交差検定の結果）である．なお図 4 のクラスタ数，項構造解析精度は交差検定の結果の平均を示している．図 3, 4 を見ると，用例規模が増加すると収集できる用例異なり数は増加し続けているのに対し，得られるクラスタ数は収束しつつある．すなわち，用例規模を増やしたときに得られる新規の用例は収集済みの用例と類似している可能性が高いため，項構造解析の精度向上に寄与する保証は必ずしもない．しかし，異なるデータセットを用いた実験の結果，図 3, 4 に示すように，用例規模を増やせば項構造解析精度が向上するということが経験的に明らかになった．これは，項構造解析においてはクラスタの重心ではなく，最近傍の用例を参照しているためである．すなわち，用例規模を変化させたときのクラスタリングの結果に大きな違いがないとしても，クラスタの外延的定義はより緻密になっており，複数のクラスタ間の用例に対する類似度がより正確に見積もられていると解釈できる．

ちなみに「は」「も」「無格」の曖昧性を人手で解消せずに，Kurohashi ら (1994) の手法で項構造解析と表層格の曖昧性解消を行った結果，項構造解析精度は 90.8 % であり，人手で曖昧性を解消した結果と比べ約 7 % 低下している．この結果より，項構造解析精度を改善するに

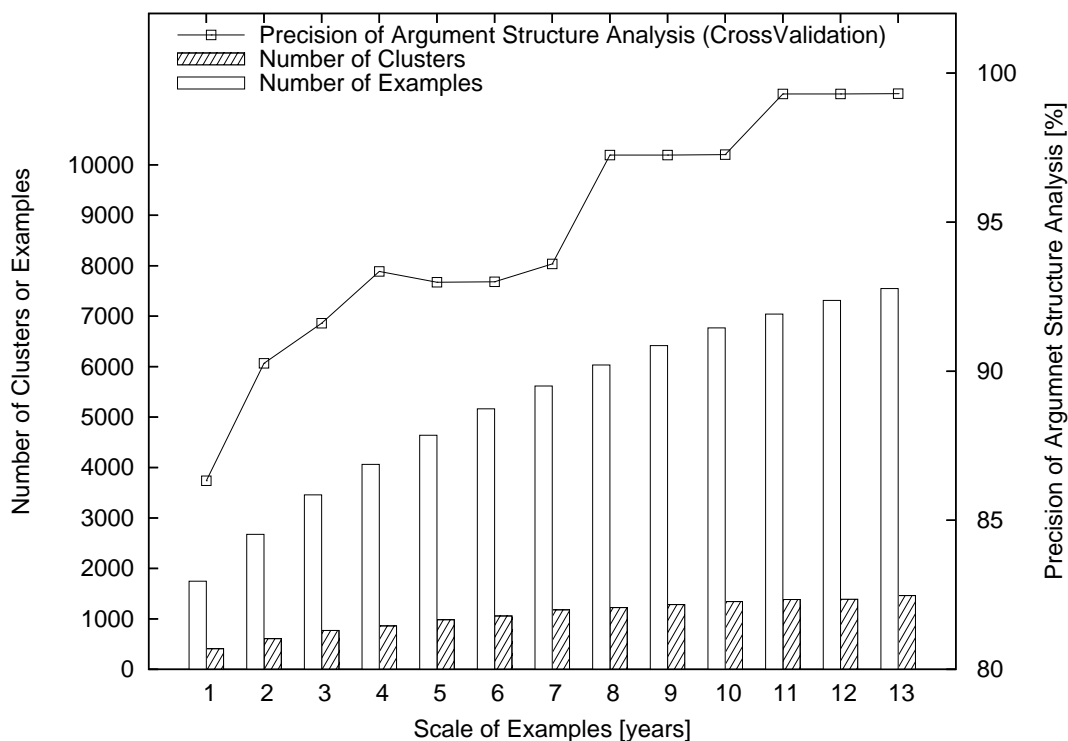


図 4 用例規模に対する用例異なり数とクラス数，項構造解析精度（交差検定）

はこの種の多義性解消の問題に取り組む必要があることが明らかになった。「は」、「も」、「無格」の曖昧性を解消するためには、単に用例を増やすだけでなく、ゼロ照応解析や連体修飾の解析との統合について検討すべきであると考えられる。

4 おわりに

本論文では、項構造解析の精度向上を目指して、用例と項構造を対応付けるタグ付与作業コストの削減について議論した。また、動詞に関する 2 種類の類似性に基づき用例をクラスタリングすることで作業コストを削減する一手法を提案した。提案手法を用いて実際にタグ付与作業を行ったところ、タグ付与の品質や項構造解析の精度を保ったまま、作業コストを約 2 割に削減できた。また、最近傍法を用いた項構造解析では、タグ付与された事例を増やすことで精度を向上できた。この結果より、解析精度の向上にはいかにコストをかけずに大規模な項構造タグ付き用例を収集するかが主要な問題であることがわかり、我々の提案手法はその問題の一つの解決策となっていることが確認できた。

ただし、動詞それぞれについて網羅的に大規模な用例を作成するためには、項構造解析精度

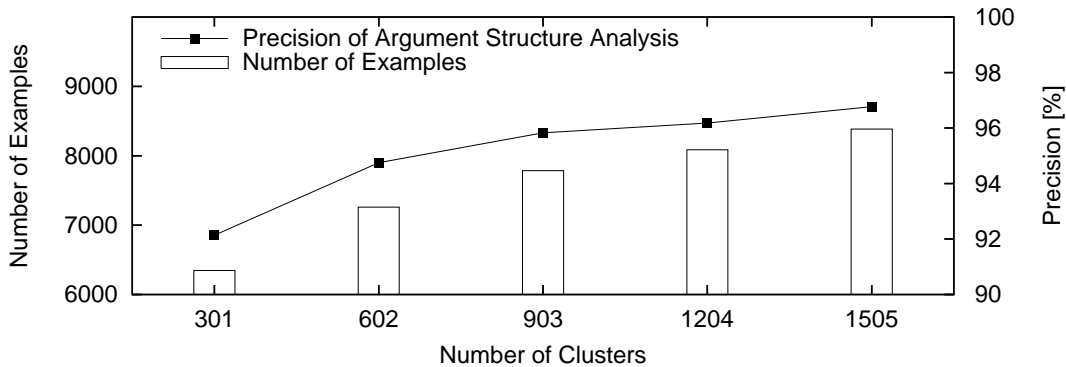


図 5 選択的サンプリングによる項構造解析精度

を低下させることなく選択的に用例をサンプリングするための評価尺度について検討したり、能動学習を採り入れるなど、さらなるコスト削減の方法を考える必要がある。そこで、提案した手法によって得られたクラスタすべてにタグを付与せず、クラスタを選択的にサンプリングし、一部の代表的なクラスタにタグを付与することで作業コストをさらに削減することを試みた。選択的サンプリングでは「いかに項構造解析に貢献するクラスタを選択するか」が課題となる。サンプルの選択基準は多々あるが、今回は動詞ベースクラスタリングの結果を利用し、サイズの大きいクラスタ（多くの用例から構成されているクラスタ）を優先した。3節の評価実験と同じ評価データ、クラスタリングの設定、項構造解析モデルで評価実験を行なった。結果を図5に示す。タグ付与を行ったクラスタ数（横軸）に対する、項構造解析精度とタグ付けされる用例数の変化を表している。結果から単にクラスタサイズに基づいてサンプリングするだけではさらなるコスト削減の方法として必ずしも良い結果が得られないことがわかった。今後、どのような基準でサンプリングするか、およびタグ付与作業の終了条件について検討を重ねる必要がある。

また項構造解析精度を向上させるには「は」「も」などの表層格の曖昧性解消や接尾辞を伴うことによる格交替の曖昧性の解消に今後取り組む必要がある。接尾辞を伴うことによる交替現象はある程度の文法規則によってとらえられると考えられるが、「は」「も」の曖昧性は単に用例を増やすだけでは解消できないことが今回の実験結果からわかる。この曖昧性を高精度で解析するためにはゼロ照応解析や連体修飾の解析も視野に入れた解析手法が必要であり、今後その問題にも取り組みたい。

参考文献

- Baker, C. F., Filmore, C. J., and Lowe, J. B. (1998). “The Berkeley FrameNet Project.” In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, pp. 86–90.
- Dorr, B. J. (1997). “Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring.” In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 139–146.
- Filmore, C. J. (1968). “The case for case.” In *Universals in Linguistic Theory*, pp. 1–88.
- Gildea, D. and Jurafsky, D. (2002). “Automatic labeling of semantic roles.” *Computational Linguistics*, **28** (3), 245–288.
- Hacioglu, K. and Ward, W. (2003). “Target Word Detection and Semantic Role Chunking using Support Vector Machines.” In *HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short papers*, pp. 25–27.
- 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (編) (1997). 日本語語彙大系: CD-ROM 版. 岩波書店.
- 情報処理振興事業協会技術センター (1987). 計算機用日本語基本動詞辞書 IPAL(Basic Verbs) 辞書編.
- 河原大輔 黒橋禎夫 (2002). “用言と直前の格要素の組を単位とする格フレームの自動構築.” *自然言語処理*, **9** (1), 3–19.
- Kingsbury, P. and Palmer, M. (2002). “From TreeBank to PropBank.” In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pp. 1989–1993.
- 国立国語研究所 (1993). 分類語彙表, 国立国語研究所資料集, 6 巻. 秀英出版.
- Kurohashi, S. and Nagao, M. (1994). “A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary.” *IEICE Transactions on Information and Systems*, **E77-D** (2), 227–239.
- Lapata, M. and Brew, C. (1999). “Using Subcategorization to Resolve Verb Class Ambiguity.” In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 266–274.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago Press.
- Thompson, C. A., Levy, R., and Manning, C. D. (2003). “A Generative Model for Semantic

Role Labeling.” In *Machine Learning: ECML 2003*, pp. 397–408.

付録

A 用例間の類似度計算

河原・黒橋 (2002) は以下の類似度計算式を提案した .

単語 e_1, e_2 の類似度 $sim(e_1, e_2)$ を , 日本語語彙大系 (池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林 1997) のシソーラスを利用して以下のように定義する .

$$\begin{aligned} sim_e(e_1, e_2) &= \max_{x \in s_1, y \in s_2} sim(x, y) \\ sim(x, y) &= \frac{2L}{l_x + l_y} \end{aligned}$$

ここで , x, y は意味属性を表し , $s_i (i \in 1, 2)$ は単語 e_i の意味属性の集合を表す⁶ . $sim(x, y)$ は意味属性の x, y 間の類似度であり , l_x, l_y はそれぞれ名詞意味属性の階層の根から x, y までの深さを表し , L は根から x, y までの共通している階層の深さを表す . 類似度 $sim(x, y)$ は 0 から 1 の値をとる .

用例 P_1, P_2 の格の一致度 cs は , P_1, P_2 に含まれるすべての格パターンに対する , P_1, P_2 の共通格に含まれている格パターンの割合とし ,

$$cs = \frac{\sum_{i=1}^n |E_{1cc_i}| + \sum_{i=1}^n |E_{2cc_i}|}{\sum_{i=1}^l |E_{1c1_i}| + \sum_{i=1}^m |E_{2c2_i}|}$$

と定義する . ただし用例 P_1 中の格を $c1_1, c1_2, \dots, c1_l$, 用例 P_2 中の格を $c2_1, c2_2, \dots, c2_m$, P_1 と P_2 の共通格を cc_1, cc_2, \dots, cc_n とする . また , E_{1cc_i} は P_1 内の格 cc_i に含まれる格用例群であり , $E_{2cc_i}, E_{1c1_i}, E_{2c2_i}$ も同様である . $|E_{1cc_i}|$ は E_{1cc_i} の頻度を表す .

用例 P_1, P_2 の共通格に含まれる格用例群の類似度 $sim_E(P_1, P_2)$ は , 格用例の類似度の和を正規化したもので ,

$$sim_E(P_1, P_2) = \frac{\sum_{i=1}^n \sum_{e_1 \in E_{1cc_i}} \sum_{e_2 \in E_{2cc_i}} |e_1| |e_2| sim_e(e_1, e_2)}{\sum_{i=1}^n \sum_{e_1 \in E_{1cc_i}} \sum_{e_2 \in E_{2cc_i}} |e_1| |e_2|}$$

とする .

用例 P_1, P_2 間の類似度は , 格の一致度 cs と P_1, P_2 の共通格の格用例群間の類似度の積とし , 次のようにして計算する .

$$\text{類似度} = cs \cdot sim_E(P_1, P_2)$$

6 日本語語彙大系では , 曖昧性を持つ単語を複数の意味属性に登録している .

B その他の動詞の項構造辞書の項目

表 6 動詞「発表する」項構造辞書の項目

述語:発表する	
語義:新しい事実や考えなどを、広く世間に知らせる。	
項構造 1 [agent, theme]	
表層格パターン	用例
が, を	研究者が論文を発表する。
項構造 2 [agent, content]	
表層格パターン	用例
が, と	政府が自衛隊を派遣すると発表する。
項構造 3 [agent, theme, content]	
表層格パターン	用例
が, を, と	政治家が結果を残念だと発表する。

表 7 動詞「発売する」項構造辞書の項目

述語:発売する	
語義:売り出すこと。	
項構造 1 [agent, theme]	
表層格パターン	用例
が, を	SONY が商品を発売する。

表 8 動詞「増える」項構造辞書の項目

述語:増える	
語義:数量が多くなる。	
項構造 1 [theme]	
表層格パターン	用例
が	映画館が増える。
項構造 2 [theme, goal]	
表層格パターン	用例
が, に	体重が 100kg に増える。
が, まで	川の水が腰の位置まで増える。
が, へ	予算が一千万円へ増える。

略歴

平野 徹: 2003 年和歌山大学システム工学部情報通信システム学科卒業。2005 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了。同年日本電信電話株式会社入社。現在に至る。自然言語処理の研究に従事。

飯田 龍: 2002 年九州工業大学情報工学部知能情報工学科卒業。現在、奈良先端科学技術大学院大学情報科学研究科博士後期課程在学中。情報処理学会学生会員。自然言語処理、特に照応解析の研究に従事。

藤田 篤 (正会員): 2005 年奈良先端科学技術大学院大学情報科学研究科博士

後期課程修了。京都大学情報学研究科産学官連携研究員を経て、2006年より名古屋大学大学院工学研究科助手。現在に至る。博士(工学)。自然言語処理、特にテキストの自動言い換えの研究に従事。情報処理学会、ACL各会員。乾 健太郎(正会員): 1995年東京工業大学大学院情報理工学研究科博士課程修了。同年より同研究科助手。1998年より九州工業大学情報工学部助教授。1998年~2001年科学技術振興事業団さきがけ研究21研究員を兼任。2001年より奈良先端科学技術大学院大学情報科学研究科助教授。2004年文部科学省長期在外研究員として英国サセックス大学に滞在。現在に至る。博士(工学)。自然言語処理の研究に従事。情報処理学会、人工知能学会、ACL各会員。

松本 裕治(正会員): 1977年京都大学工学部情報工学科卒。1979年同大学大学院工学研究科修士課程情報工学専攻修了。同年電子技術総合研究所入所。1984~85年英国インペリアルカレッジ客員研究員。1985~87年(財)新世代コンピュータ技術開発機構に出向。京都大学助教授を経て、1993年より奈良先端科学技術大学院大学教授。現在に至る。工学博士。専門は自然言語処理。情報処理学会、人工知能学会、日本ソフトウェア科学会、認知科学会、AAAI, ACL, ACM各会員。

(2005年11月16日受付)

(2006年2月22日再受付)

(2006年3月28日採録)