特徴的な意味内容を抽出する木構造マイニングのための日本語処理手法

坂尾 要祐　池田 崇博　佐藤 研治　赤峯 享
NEC メディア情報研究所
{y-sakao@bu, t-ikeda@di, k-satoh@da, s-akamine@ak}.jp.nec.com

1. はじめに

近年、大量のアンケートデータやコンタクトセンターのデータといった、日々大量に蓄積されるテキストから効率的に知識発見を行う枠組みとして、与えられたテキスト集合から特徴的な表現の抽出を行うテキストマイニング技術を使用することが盛んになっている。

近年では、テキストの構文構造を用いたマイニング手法も提案されている。工藤ら[1]は、木構造を簡略化した順列構造に対してマイニングを行う方法を提案している。この方法は、二段階までの係り受けを対象としたマイニングであり、大規模な特徴構造の抽出を行えないという問題があった。

工藤ら[2]は、さらに、テキスト集合を構文解析することにより得られた木構造を順序木に制限することで、Asai ら[3]の順序木の最右拡張によって効率よく全ての部分構造を枚挙する手法を提案している。しかしながら、この方法では、順序木を用いているため、意味は同じだが順序木としての構造が異なる表現を同一視して特徴構造の抽出を行えないという問題がある。また、抽出した木構造は、そのままでは意味が分かり難いという問題もある。

本稿では、日本語文のマイニング用の表現構造として、各節点が文節を表し、各枝が係り受け関係を表す依存構造木を採用し、順序木の最右拡張により部分構造を枚挙し、マイニングを行う。以下、このマイニング方法を木構造マイニングと呼ぶ。依存構造木の節点のラベルとしては、文節の代表単語の原型を採用する。木構造マイニングにおける前述の問題を解決するため、依存構造木の構造変形によって、意味は同じだが形が異なる木構造を順序木として同一視して木構造マイニングを行う手法を提案する。さらに、抽出された特徴構造の節点に対応する原文中の文節から文を構成し出力することで、分かりやすい結果を提示する手法を提案する。

以下では、まず順序木を用いた従来の木構造マイニング手法の問題点について詳細に述べ、本稿で提案する解決手法について説明する。その後、我々が開発したプロトタイプシステムについて説明し、最後にプロトタイプシステムの評価結果を示す。

2. 順序木を用いた木構造マイニングの問題点

従来手法による順序木を用いた木構造マイニングには、以下に挙げる二つの問題点がある。

⑴ 意味が同じで順序木としての構造が異なる表現を同一視することが困難

意味は同じでも順序木としての構造が異なる表現は、従来手法による木構造マイニングにおいて、同一視して特徴構造の抽出を行えない。意味は同じでも順序木としての構造の異なる表現の例として、図 1 に「若年層が高級車を購入」「高級車を若年層が購入」「若年層が購入する高級車」に対応する依存構造木を示す。依存構造木では、「若年層が高級車を購入」「高級車を若年層が購入」といった係り受けの順序の違いは兄弟節点の順序の違いとして表現され、「若年層が高級車を購入」「若年層が購入する高級車」といった係り受けの向きの違いは有向枝の向きの違いとして表現される。

⑵ 抽出結果を木構造のままユーザに提示しても分かり難い

木構造マイニングにより抽出された特徴構造は木構造で表現されるが、それをマイニングの結果としてユーザに提示しても意味を取り難い場合がある。特に、多くの節点や枝を持つ大きな特徴構造において、その傾向が顕著となる。

3. 提案手法

本稿では、前節で述べた 2 つの問題点に対応するため、木構造マイニングに対して以下の二つの手法を導入する。

⑴ 依存構造木の構造変形

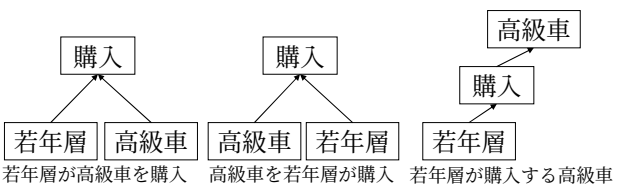


図 1: 意味は同じだが形の異なる木構造の例

意味は同じだが形が異なる依存構造木を順序木として同一視するために、以下の 2 つの構造変形を構文解析結果の依存構造木に対して行う。

兄弟節点のソート：係り受けの順序が異なる依存構造木を順序木として同一視可能とするために、兄弟関係にあるノードの兄弟順を一定の順序（辞書順など）に基づきソートする。図 2 に、「若年層が高級車を購入」「高級車を若年層が購入」の依存構造木の兄弟節点を辞書順にソートする例を示す。図 2 によると、2つの依存構造木の係り受けの順序の違いが兄弟節点のソートにより吸収され、2つの依存構造木を同一視できるようになったことが分かる。

なお、同じラベルを持つ兄弟が複数存在する場合は、その兄弟節点の全ての順列を実現する木構造を構築する。

ルート展開：係り受けの向きが異なる依存構造木を同一視可能とするために、依存構造木の各節点をルートとし、有向枝の向きを新しいルートの方に変更した依存構造木を構築する。この際、1つの木からそのノード数と同じ数の依存構造木が構築される。図 3 に、「若年層が高級車を購入」「若年層が購入する高級車」のルート展開の例を示す。ルート展開により、係り受けの向きの違いが吸収され、2つの依存構造木を同一視できるようになったことが分かる。

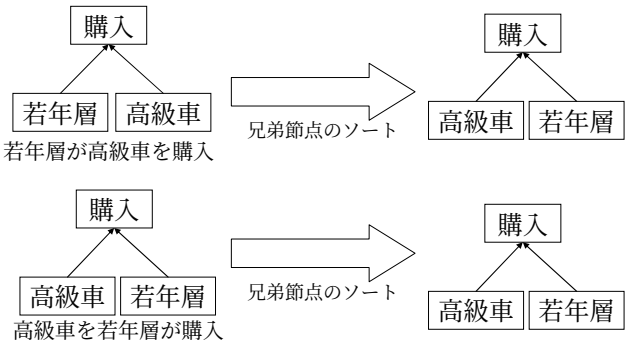


図 2： 兄弟節点のソートの例

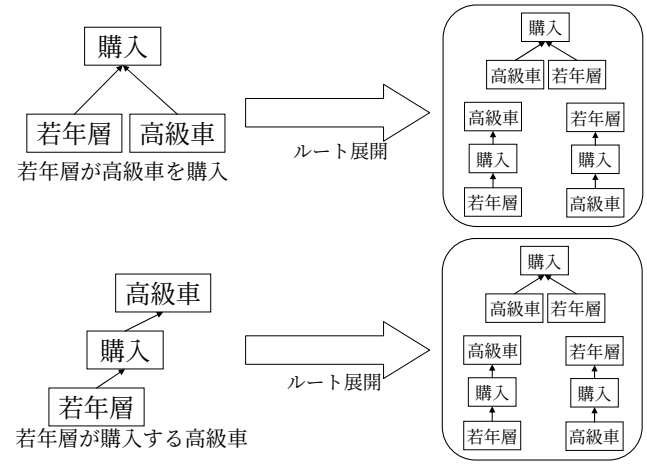


図 3： ルート展開の例

⑵ 抽出結果からの文の構成

木構造マイニングにより抽出された特徴構造について、特徴構造の各節点に対応する原文中の文節から文を構成し出力する。

具体的には、特徴構造を含んでいたテキスト集合中の各テキストに対し、特徴構造の各節点に対応する文節の表層を連結したものを出力文の候補とし、各候補のスコアを文節単位のbigramによって計算し、最もスコアの高い候補を特徴構造に対応する文として出力する。すなわち、出力文の候補 $\{W \mid w_1 \ldots w_n\}$からの出力文$S$の構成は、以下のように求められる。ただし、$w_1,\ldots,w_n$は$W$を構成する文節、$P(W)$は$W$の生起確率、$P(w_i \mid w_{i-1})$は文節$w_{i-1}$の後に文節$w_i$が続くbigram確率を表す。

$$S = \arg\max_W P(W) \approx \arg\max_W P(w_i \mid w_{i-1})$$

なお、文節単位の bigram は、マイニングの入力テキスト集合からあらかじめ構築しておいたものを用いる。

4. プロトタイプシステム

前節で述べた手法を導入した木構造マイニングを実装したプロトタイプシステムを開発した。プロトタイプシステムは、以下の手順で木構造マイニングを実行する。

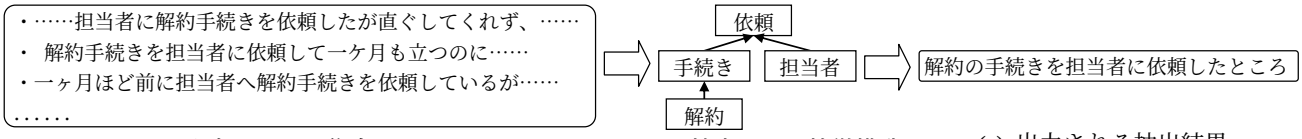
1) 構文解析：入力テキスト集合に対して構文解析を行い、依存構造木を構築する。

2) 依存構造木の構造変形：3 節の(1)で提案した構造変形を、構文解析で得られた依存構造木に適用する。

3) 特徴構造抽出：依存構造木から、Asai ら[2]の最右拡張により全ての部分構造の枚挙を行い、次に部分構造にスコア付けを行い、スコアの高いものを特徴構造として抽出する。部分構造のスコアとしては、森永ら[4]の情報量に基づくスコアを用いる。

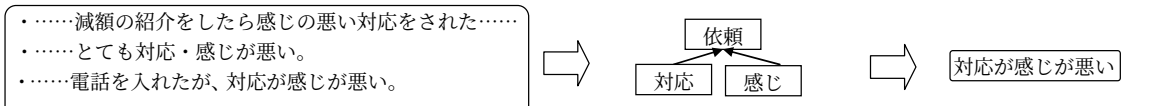
4) 包含構造の除去：意味的に同じ特徴構造や包含関係にある特徴構造が大量に抽出されるのを避けるために、無向グラフとみなした場合に同一となる特徴構造は一つだけ残して他を削除し、他の特徴構造に包含される特徴構造は除去する。

5) 特徴構造からの文の構成：**3** 節の(2)で提案した手法を用いて、抽出された特徴構造から文を構成し、ユーザに提示する。



(a) 入力テキスト集合　(b) 抽出された特徴構造　(c) 出力される抽出結果

図 4：抽出例 1



(a) 入力テキスト集合　(b) 抽出された特徴構造　(c) 出力される抽出結果

図 5：抽出例 2

5. 実験

2 種類のコーパスに対してプロトタイプシステムにより木構造マイニングを行い、プロトタイプシステムを評価した。

コーパスとしては以下の二つのテキスト集合を用い、それぞれのコーパスに対して 3 種類ずつ、計 6 種類の木構造マイニングを行った。

コーパス 1：営業レポート（56514 文）

コーパス 2：コンタクトセンター対応データ（4639 文）

実際の抽出結果の例を 2 点示す。

図 4 に示す抽出例 1 では、入力テキスト集合から、兄弟節点のソートの効果により「担当者に解約手続きを依頼したが」と「解約手続きを担当者に依頼して」のように、係り受けの順序が違う表現が同一視されて特徴構造が抽出され、各節点に対応する原文の文節を用いて「解約の手続きを担当者に依頼したところ」という文が構成されている。

図 5 に示す抽出例 2 では、入力テキスト集合から、ルート展開の効果により「感じの悪い対応」と「対応が感じが悪い」のように、係り受けの順序が違う表現が同一視されて特徴構造が抽出され、各節点に対応する原文の文節を用いて「対応が感じが悪い」という文が構成されている。

2 つのコーパスに対して行った計 6 種の木構造マイニングにより抽出された 2 節点以上からなる構造 1286 件に対して、構造変形を用いた特徴構造抽出の有効性、特徴構造からの文の構成の有効性の二つの観点から評価を行った。以下にそれぞれの評価結果について述べる。

(1) 構造変形を用いた特徴構造抽出の有効性

構造変形を用いた特徴構造抽出の有効性の指標として、正しく同じ意味の表現のみを同一視できているかどうかを示す適正抽出率と、特徴構造を含むテキストから漏れなく抽出が行われているかを示す抽出被覆率を用いて評価を行った。

適正抽出率：抽出された特徴構造に対応する原文のうち、特徴構造が表す内容が原文中に実際に書かれていたものの割合。

抽出被覆率：特徴構造が表す内容が書かれていた入力テキストのうち、実際にその特徴構造が抽出されたものの割合。

なお、特徴構造に対応する原文に特徴構造の内容が実際に書かれていたか否かの判定は、人手で原文を読むことにより行った。また、特徴構造の内容が書かれていた入力テキストは、検索システムを用いてテキスト集合に対し特徴構造から構成された文をクエリとする検索を行い、検索された文の中から人手で収集した。

コーパス別に求めた適正抽出率と抽出被覆率を、表 1 に示す。

この結果の適正抽出率の値は、抽出された各特徴構造について、特徴構造に対応する原文のうち、平均 90%以上にその内容が実際に書かれていたということを示している。また、抽出被覆率の値は、各特徴構造について、特徴構造の表す内容が書かれていたテキストのうち、平均 80%以上から実際に特徴構造が抽出されているということを示している。このことから、プロトタイプシステムは実用上、十分正確にテキスト集合からの特徴構造の抽出を行っていることが分かる。

なお、特徴構造が表す内容が書かれているにもかかわらず、特徴構造が抽出されなかったテキストには、同義表現を同一視できていないものが多く見られた。（例：「WEBメールにログイン出来ません。」から「Web Mailにログインできません」が抽出されない）

表 1：適正抽出率、抽出被覆率

コーパス	適正抽出率	抽出被覆率
1	91.4%	82.5%
2	90.5%	84.4%

表 2：有意抽出結果率、抽出結果可読化率

コーパス	(A) 抽出結果可読化率	(B)有意味抽出結果率	(A)-(B)
1	97.9%	87.8%	10.1%
2	95.3%	83.9%	11.4%

⑵ 特徴構造からの文の構成の有効性

特徴構造からの文の構成の有効性の指標として、構成された文がどの程度正しく意味が取れる文となっているかを示す抽出結果可読化率を以下のように定義し、これを用いて評価を行った。

抽出結果可読化率：抽出結果から構成された文のうち、文意が通るものの割合。

また、抽出結果の意味がユーザに理解できたしても、マイニング結果としては十分な情報を含まないものが存在する可能性がある。そのような抽出結果の割合を調べるために有意味抽出結果率という指標を以下のように定義し、これを用いて評価を行った。

有意味抽出結果率：抽出結果から構成された文のうち、マイニング結果として十分な意味を持つものの割合。

ユーザに意味が理解できてもマイニング結果としては無意味な抽出結果の割合は、(抽出結果可読化率－有意味抽出結果率) で求められる。

抽出結果から構成された文の文意が通るかの判定、抽出結果から構成された文にマイニング結果としての意味があるかの判定は、人手で行った。

コーパス別の抽出結果可読化率および有意味抽出結果率を、表 2 に示す。

この結果は、木構造マイニングのユーザにとって、抽出結果のうち 95%以上が可読であることを示している。このことから、文の構成法が可読性の高い文を出力していることが分かる。さらに、表 2 の (抽出結果可読化率－有意味抽出結果率) の値を参照すると、抽出結果のうち 10%以上が可読であるがマイニング結果として意味がないということが分かる。このことは、本手法では部分的な構造を特徴構造として抽出しているが、無意味な特徴構造は少なかったことを示している。

文意が取れないと判定された抽出結果を確認したところ、誤った構文解析により生成された依存構造から抽出された結果がほとんどであることが確認できた。また、可読ではあるがマイニング結果としては意味がないと判定された抽出結果を確認したところ、これらの抽出結果には以下のような傾向が確認された。

複合名詞の分断：特徴構造に複合名詞の一部しか含まれていないために、マイニング結果として無意味な文が出力される。(例：「メールにてメールを送ろうとすると」を「WEBメールにてメールを送ろうとすると、メッセージが送信されませんでした。」から抽出)

単体で大きな意味を持たない用言の抽出：特徴構造に単体で大きな意味を持たない用言が含まれるにも関わらず、その用言を十分に修飾する情報が含まれていないために、マイニング結果として無意味な文が出力される。(例：「あり危険だ」を「支社窓口の入口付近に段差があり危険だ。」から抽出)

6. おわりに

本稿では、同じ意味で順序木として構造の異なる木構造を同一視し、分かりやすい結果を提示する、木構造マイニングの実現方法ついて述べた。順序木として構造の異なる木構造を同一視するために依存構造木の構造変形を導入し、また、特徴構造を分かりやすく提示するために特徴構造と入力テキスト集合を用いた文の構成を導入した。この手法を実装したプロトタイプシステムの開発を行い、営業レポートとコンタクトセンターの対応データに対して適用することで本手法の評価を行った。その結果、90%以上の適正抽出率と 80%以上の抽出被覆率を示し、依存構造木の構造変形を用いた特徴構造の抽出が、マイニングの実用上有効であることが示された。また、実験結果は、95%以上の抽出結果可読化率と 80%以上の有意味抽出結果率を示し、特徴構造からの文の構成が、マイニングの実用上十分にユーザに分かりやすい出力を提示していることが示された。

今後は、同義表現の同一視による抽出漏れの削減を行っていく予定である。また、複合名詞の一部を含む抽出結果には複合名詞の残りの部分を付加する、特徴構造中の用言の十分な修飾情報を含まない抽出結果は出力しない、等の手段による、意味のない抽出結果の削減を行っていく予定である。

参考文献

［1］ 工藤拓, 山本薫, 坪井祐太, 松本祐治: 言語情報を利用したテキストマイニング, 情報処理学会 自然言語処理研究会 SIGNL-148, pp.65-72, 2002

［2］ 工藤拓, 松本裕治: 半構造化テキストの分類のためのブースティングアルゴリズム, 情報処理学会 知能と複雑系研究会 SIGICS-135, pp.58-63, 2002.

［3］ Asai et al.: “Efficient substructure discovery from large semistructured data”, Proc. SDM’02, pp.158−754, 1997.

［4］ 森永聡, 有村博紀, 池田崇博, 坂尾要祐, 赤峯享, 部分順序木枚挙を利用した特徴無順序木/自由木構造抽出 ―テキストマイニングへの応用―, 第7回情報論的学習理論ワークショップ IBIS2004, pp.106-111, 2004.