E-007

構文グラフ集合を用いた Key Semantics マイニング

Mining Key Semantics Using Dependency Graphs

森永　聡†　　有村　博紀‡　　池田　崇博†　　坂尾　要祐†　　赤峯　享†
Satoshi Morinaga　Hiroki Arimura　Takahiro Ikeda　Yosuke Sakao　Susumu Akamine

1. はじめに

構文グラフの集合から一定の統計的性質を満たす部分構造をとりだすことにより，テキストデータにおける特徴的な記述内容 (Key Semantics) を抽出するシステム SurveyAnalyzer V5 を報告する．

近年，コールセンタやＷＥＢコンテンツのデータに対するテキストマイニング技術の適用が盛んになっているが，中でも，与えられた文書群における特徴的な表現を抽出する技術は，基本機能として重要なものとなっている．

この抽出技術として，Yamanishi and Li[3]は文書群を形態素解析した上で，各単語の特徴度を情報量基準に基づき計算し，その値の高いものを抽出することを提案していた．一方，工藤[4]は文書群を構文解析した上で，Asai ら[1]の最右拡張でその部分構造を枚挙し，それらに対して特徴度を計算することを提案している．

しかしながら，工藤の方法は構文解析結果を順序木（兄弟間に順序が定義されている木）とみなしているため，語順等が異なる表現（「メールを社外に送る」「社外にメールを送る」「社外に送ったメール」など）を同じものと見なして特徴度を計算することが出来ず，記述内容が同じでも表現が異なると特徴として抽出できないという問題があった．また，互いに包含関係にある表現（「英語のメールを社外に送る」「英語のメールを送る」「メールを社外に送る」，，，など）が大量に出力されてしまう場合が多いことや，そもそも木構造データとして結果が出力されている，といったことに起因する抽出結果の可読性の低さの問題も，一般的に存在していた．

本稿ではこれらの問題点を解決するものとして，構文解析の結果を，兄弟間に順序が定義されていない無順序木（「メールを社外に送る」と「社外にメールを送る」を同一視）や無向グラフ（「社外に送ったメール」も同一視）と見なして特徴部分構造の抽出を行い，それらに対して包含関係などによる取捨選択を行った上で，相当する日本語の表現を再生成したものを出力するシステム SurveyAnalyzer V5 を報告する．

本システムが出力するのは，語順や係り受けの順序/向きの点で表現としては異なっていても，記述されている内容として特徴的であるというものである．以下では，この（表現が異なっても）特徴的である記述内容のことを Key Semantics とよぶことにする．

2. Key Semantics のマイニング

SurveyAnalyzer V5 は，与えられた正例および負例の文書集合に対して，正例の Key Semantics を抽出するシステムである．構文解析部，入力フィルタ群，特徴部分木抽出部，出力フィルタ群，日本語再生成部からなっており，処理は１）正/負例文書群の構文解析，２）解析結果の入力フィルタによる変換，３）変換結果を順序木とみなした特徴部分木の抽出，４）抽出結果の出力フィルタによる取捨選択，５）相当する日本語の再生成の順で行われる．

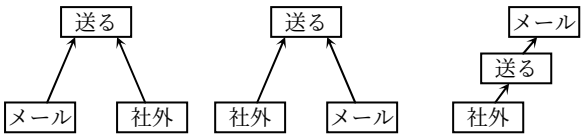
３では順序木とみなした特徴抽出が行われているが，２で用いるフィルタの種類によって，システム全体として構文解析結果を順序木/無順序木/無向グラフのどれとみなして特徴抽出するかを変更できるようになっている．

また，各処理の結果として出力される中間データ類は，全て元の文書のＩＤとも紐付けされており，特徴度の計算や日本語再生成に必要な情報が得られるようになっている．

以下，各処理に関して説明する．

2.1 構文解析

テキストデータを形態素解析・構文解析し，依存構造木を構築する．構文グラフとして，本稿では依存構造木を用いる．依存構造木は，各ノードが文節を表す順序木で，「社外にメールを送る」「メールを社外に送る」のような係り受けの順序の違いが兄弟ノード間の順序の違いに対応し，「社外にメールを送る」「社外に送ったメール」のような係り受けの向きの違いがエッジの向きの違いに対応する（図 1）．特徴抽出時に，付属語による表現の違いを吸収できるように，文節内の自立語を原形で表記したものをノードのラベルとする．付属語が表す内容は，適宜，属性値に縮約してノードに付加する[5].



1）メールを社外に送る　　2）社外にメールを送る　　3）社外に送ったメール

図 1：　依存構造木による文の表現

2.2 入力フィルタによる変換

構文解析結果を変換する入力フィルタには兄弟ソートフィルタとルート展開フィルタの二種類がある．

兄弟ソートフィルタ：入力木の兄弟関係にあるノードの兄弟順をそのラベルのアルファベット順にソートした順序木を出力する．特に，同じラベルを持つ兄弟が複数ある場合は，その全ての順列を実現した順序木を出力する（例えば，同ラベルの３兄弟が存在したら，6 つの順序木が出力される）．

ルート展開フィルタ：入力木の各ノードに対して「無向グラフと見なした場合に入力木と同一になる順序木で，そのノードをルートとするもの」を出力する．一つの入力木に対して，そのノード数と同じ数の順序木が出力される．

後段の特徴部分木の抽出は順序木として行われるので，どちらの入力フィルタも用いなければ，システム全体としては構文解析結果を順序木とみなした特徴抽出が行われる．

† 日本電気株式会社, NEC Corporation

‡ 北海道大学, Hokkaido University

一方，兄弟ソートフィルタのみを用いた場合は全体として無順序木と見なした特徴抽出，ルート展開フィルタと兄弟ソートフィルタをこの順に用いた場合は全体として無向グラフと見なした特徴抽出が行われる．

2.3 特徴部分木の抽出

入力フィルタによる変換結果の集合（以下Ｓと記述）に対して順序木としての特徴部分木抽出を行う．まずＳにおいて正例文書に紐付けされている木の集合に関して，Asaiら[1]の最右拡張によってその全ての部分木の枚挙を行う．次にそれらに対する特徴度の計算とその値の大きいものの抽出を行う(枝刈等に関する説明は割愛する).

部分木Ｔの特徴度 G(T)としては，Yamanishi and Li[3]による単語の特徴度の計算式 G(T)=ESC-(ESC1+ESC0)をそのまま用いる．ただし

ESC=A+L*Sqrt((A+B)*Log(A+B))

ESC1=D+L*Sqrt((C+D)*Log(C+D))

ESC0=(A-C)+L*Sqrt((A+B-C-D)*Log(A+B-C-D))

であり，A および B は正例および負例の文書数，C および D はＴと紐付けられている正例および負例の元文書数，Lは定数である．この式は，木の空間に一様分布の重みを与えたときのＥＳＣで情報利得を定義したことに相当する[2].

この値がユーザ指定の閾値を超えている部分木や，この値でソートしてユーザが指定した順位以内に入るものが特徴部分木として抽出される．

2.4 出力フィルタによる取捨選択

特徴部分木は出力フィルタによって取捨選択された上で日本語再生成される．出力フィルタには以下の三つがある．

同一グラフ削除フィルタ：特徴部分木の集合から，無向グラフと見なして同一のものは一つを残して削除する．

包含木削除フィルタ：特徴部分木の集合から，他の木に（順序木として）包含されるものは削除する．

下位包含木削除フィルタ：特徴部分木の集合から，特徴度が自分より低くない他の木に（順序木として）包含されるものは削除する．

これらのフィルタを用いることで，互いに包含関係にある特徴部分木等を整理することが可能である．特にルート展開フィルタを用いた場合は，無向グラフとして同一の部分木が大量に抽出されるので，同一グラフ削除フィルタは必須となる．

2.5 日本語再生成

削除されなかった特徴部分木に対して，構文グラフ構築とフィルタ処理によって失われた情報の復元を行い，相当する日本語の再生成を行う．本稿では，元文書中の典型的な文に合わせられるように，元文書の表現を利用して日本語文を生成する．

具体的には，与えられた部分木に紐付けられている元文書中の各文に対して，その部分木に対応する表層表現を連結したものを出力文の候補とし，その中から日本語として出現しやすいものを単語 bigram に従って統計的に選択する．すなわち，出力文の候補 $\{W \mid w_1 \ldots w_n\}$ に対する出力文を S とすると，

$$S = \underset{W}{\operatorname{argmax}}\, \mathrm{P}(W) \approx \underset{W}{\operatorname{argmax}} \prod_i \mathrm{P}(w_i \mid w_{i-1})$$

なお，日本語の bigram 確率は，入力文書集合を含む大規模なコーパスより予め求めておいたものを用いる．

3. 実験

社内メールシステムのヘルプデスクにおける実データを分析した結果を報告する．入力データは特定期間の受付内容の記録である．特に電話で受付けたもので対応が５分以下で済んだもの(274 件)を正例文書集合，５分以上かかったもの(592 件)を負例文書集合とした．

無向グラフとして特徴抽出を行い，下位包含木と重複の両削除フィルタを通した結果は以下の通りであった（特徴度上位 10 件，固有名詞は匿名化した）．

01: 開設
02: いつから
03: パスワードがわからなくなってしまいました
04: ○○○○利用停止を行うにはどうすれば良いですか
05: 新入社員の○○○○アドレスは
06: どうやって
07: 確認変更システムにてアドレスを変更したのですが
08: 構内作業者のパスワードを
09: 障害は
10: 社員確認変更システムにて変更したアドレスは

無向グラフと見なすことで語順等の違いが吸収され，7位のような長い Key Semantics も抽出された．また，出力フィルタは 4 位以下がその部分グラフで占められることを抑制していた．さらに７位などは，木構造で結果出力されると人間では元の文が想像できなかったが，このようなものに対しても適切に日本語再生成が行われていた．

結果の可読性は非常に高く，正例文書群の特徴を概観できるような出力となっている．「ヘルプデスクで簡単に対応が済んでしまう受付内容」に関する知見を容易に得ることができ，この実験から本システムは知識発見に有効であるということができた．

4. まとめと今後の課題

表現としては異なっていても記述内容として特徴的なもの (Key Semantics) を抽出するシステム SurveyAnalyzer V5 の報告を行った．また，実データを用いた実験で知識発見に有効であることも検証した．

今後は，同義表現の同一視や共起分析といった，形態素レベルでの特徴抽出システムでは既に広く使用されているような機能に関しても実現していきたい．

参考文献

[1] Asai et. al., “Efficient substructure discovery from large semi-structured data”, Proc. SDM'02, pp.158-174, SIAM, 2002.

[2] K. Yamanishi, “A decision-theoretic extension of stochastic complexity and its applications to learning”, IEEE Trans. On Information Theory, vol. 44(4), pp.1424-1439, 1998.

[3] K. Yamanishi and H. Li, “Mining open answers in questionnaire data”, IEEE Intelligent Systems, Sept/Oct, pp.58-63, 2002.

[4] 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.163-169, 2004.

[5] 佐藤研治 他: CRM 分野へ向けた日本語処理機能のミドルウェア化, 言語処理学会第 9 回年次大会, pp.109−112, 2003.