6N-01

共起関係と係り受け関係を用いた文書グラフによる重要語抽出法の提案

Proposal of Keyword Extraction Method with Document Graph based Co-occurence and Dependency Tree

今井 智宏 †
Tomohiro Imai

望月 久稔 †
Hisatoshi Mochizuki

1. はじめに

ビッグデータにおいて，さまざまなタスクの基盤技術である重要語抽出を取り扱う．ある文書の特徴を表す語を重要語として，これを抽出することにより，文書分類や評判分析への応用が期待できる．

日本語の構造に係り受け関係があり，これを考慮することで，従来使用される頻度に限らず，日本語の特徴に則した解析ができると考える．そこで，言語的特徴として共起関係と係り受け関係を取り入れた文書グラフの構築法を提案する．このグラフを解析することで特徴ベクトルを算出し，重要語を抽出する．

2. 関連研究

まず，着目した構造である係り受け関係について説明し，続いて重要語抽出として，頻度に着目した手法とグラフ問題に帰着した手法を説明する．

2.1. 係り受け関係

日本語の特徴を表す構造として，係り受け関係がある．これは，ある文節が他の文節に係るという形式により，日本語文の構造を捉える．例に，京都大学の JUMAN [5]，KNP [6] で「彼は自然言語を研究する.」を解析した結果を図 1 に示す．係り受け関係を考慮することで，各文節の重要度を抽出できると考えられる．

2.2. 重要語抽出

重要語を抽出する方法として，頻度情報を用いた Salton [1] の $tf\text{-}idf$ がある．文書中に多く出現し，他の文書に少なく出現する語ほど重要であるとして語にスコア付けする．また，頻度に加えて，名詞の連接頻度に着目することで複合名詞の抽出を実現した研究に，中川ら [2] の FLR がある．

一方，松尾ら [3] は，語の共起関係から文書グラフを構築して，グラフ構造上で各節点を結ぶ上で重要な節点に大きくスコア付けした．同様に文書をグラフとして捉えて解析する手法に，Hassan ら [4] の $TextRank$ がある．この手法は，共起関係に基づいた文書グラフを構築し，べき乗法によって特徴ベクトルを抽出する．Hassan らは頻度に基づいた手法よりも精度が高いことを示しており，今後の課題に言語の性質に基づいた文書グラフの構築をあげている．

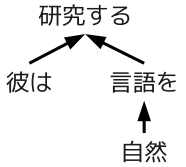

図 1: 係り受け関係

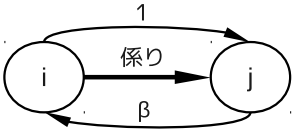

図 2: 係り受けグラフ

3. 共起関係と係り受け関係による文書グラフの構築

日本語の特徴に基づいた特徴ベクトルを抽出するために，共起関係に基づいた共起グラフと係り受け関係に基づいた係り受けグラフを構築し，これらを併合することで文書グラフを構築する．これをべき乗法で解析することにより，各文節にスコア付けする．次に，文節のスコアを各語のスコアとし，複数存在する場合は最大値を取って，さいごに，重要語を抽出する．

共起関係を「同じ文に出現した文節同士の関係」と定義する．共起関係にある文節同士は意味的なつながりがあると考えられ，以下に示した共起グラフ $\mathbf{C}$ はこの文節同士を双方向に接続する．接続する文節はグラフ上で 1，そうでない場合は 0 となる．例文「1:彼は 2:自然 3:言語を 4:研究する」から構築した共起グラフ $\mathbf{C}$ を式 (1) に示す．$(1,4)=1,(4,1)=1$ で「1:彼は」と「4:研究する」の共起を表す．同様に同じ文に出現した文節同士が双方向に接続していることがわかる．

$$\mathbf{C}(i,j)=\mathbf{C}(j,i)=\begin{cases}1 & if\ (i,j)\ are\ CoOccurrence,\\ 0 & otherwise.\end{cases}$$

$$\mathbf{C}=\begin{pmatrix}0&1&1&1\\1&0&1&1\\1&1&0&1\\1&1&1&0\end{pmatrix} \tag{1}$$

次に，係り受けグラフ $\mathbf{D}$ を構築する．文節 i が文節 j に係るとき，i は j を修飾すると捉えた関係を図 2 に示す．$\mathbf{D}(i,j)=1,\mathbf{D}(j,i)=\beta,0\leq\beta\leq 1$ である．β が小さいほど係る側と係られる側の差は大きくなり，i から j に，より多くスコアが流れ込む．

ここで，ある文節が係りを受けた数をスコア流入数と定義する．提案手法はスコア流入数が大きい文節ほどスコアが大きくなる．また，グラフを構築するときに，係り受け関係にある文節の品詞に着目する．文の意味を構成する文節を体言，用言，それ以外をその他とする．この順に重要であると考えて，優先度が高いとする．優先度の高い文節が優先度の低い文節に係る場合，係り受けグラフ上のリンクを逆にする．よって，優先度の高い文節ほど大きくスコアが割り当てられる．

そして，構築した 2 つのグラフを合わせて，文書グラフを構築する．2 つのグラフの荷重を定めるパラメータ $\gamma(0\leq\gamma\leq 1)$ を導入して，文書グラフ $\mathbf{T}$ を式 (2) に示す．γ が小さいほど，係り受けの影響が大きい．

$$\mathbf{T}=\gamma\mathbf{C}+(1-\gamma)\mathbf{D} \tag{2}$$

†大阪教育大学, Osaka Kyoiku University

さいごに例文の文書グラフ $\mathbf{T}$ を式 (3) に示す．左は共起グラフ $\mathbf{C}$，右は係り受けグラフ $\mathbf{D}$ で，$(2,3) = 1, (3,2) = \beta$ で「2:自然」が「3:言語を」へ係る．

$$\mathbf{T} = \gamma \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} + (1-\gamma) \begin{pmatrix} 0 & 0 & 0 & \beta \\ 0 & 0 & 1 & 0 \\ 0 & \beta & 0 & \beta \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad (3)$$

4. 評価

重要語の抽出精度について適合率，再現率，F 値で評価する．形態素解析器に JUMAN [5] を，構文解析器に KNP [6] を使用する．実験データとして CD-毎日新聞データ集 95 版 [7] の記事から無作為に選択した 1,000 件を使用する．提案手法に *idf* が付加可能であるため，これを提案+*idf* として評価する．提案，提案+*idf* のパラメータはもっとも精度の高かった $\beta = 0, \gamma = 0.2$ とする．また，*tf-idf*，*FLR*，松尾らの手法（以降 *SW*）と，*TextRank* は提案手法において共起関係のみを考慮したものと同様であるため，パラメータ $\beta = 0, \gamma = 1$ の値を *TextRank* による抽出として，これに *idf* を付加したものも加えて評価する．

4.1. 係り受け関係の有用性の検証

提案手法はスコア流入数が大きい文節を重要語として抽出するため，これらが解を多く含めば，高い精度の重要語抽出が期待できる．図 3 は平均的な文節数の文書におけるスコア流入数と解の包含率に関するグラフである．実線は解の包含率，破線はサンプル数を表す．サンプル数が比較的多い，スコア流入数 4 までについて，解の包含率とスコア流入数との相関係数は 0.936 と高い正の相関を示した．よって，係り受け関係を導入して重要語抽出を改善できると考えられる．

4.2. 重要語抽出の精度比較

各手法の平均の適合率，再現率，F 値を表 1 に示す．適合率において，*SW* がもっとも高い値を示した．しかし，*SW* は文書グラフ上で重要となる語にのみスコアを与えるため，極端に低い再現率となった．また，再現率が非常に高かった *tf* がもっとも高い F 値を示した．これは，同じスコアになる語が非常に多く存在したため，同率で抽出した結果，ほとんどの語を抽出する事例が多数存在したためである．

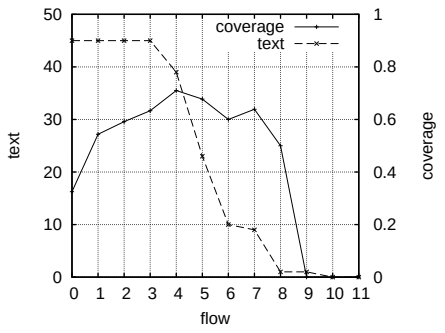

図 3: 平均的な文節数の集合におけるスコア流入数と解包含率

表 1: 適合率・再現率・F 値

手法	適合率	再現率	F 値
tf	**0.495**	**0.997**	**0.661**
idf	**0.586**	**0.595**	**0.590**
tf-idf	**0.575**	**0.588**	**0.582**
FLR	**0.517**	**0.682**	**0.588**
SW	**0.621**	**0.116**	**0.195**
TextRank	**0.508**	**0.538**	**0.523**
TextRank+idf	**0.577**	**0.578**	**0.578**
提案	**0.558**	**0.564**	**0.561**
提案+idf	**0.605**	**0.605**	**0.605**

SW と比べて，提案手法ではすべての語にスコアを与えるため，同じことは起こり得ない．また，たとえ文書中のすべての語の頻度が 1 であったとしても，係り受け構造上の差異によって，スコアに差が生じるため，*tf* のようにすべての語が抽出されることは起こりにくい．*TextRank* と比較すると，係り受け関係を導入した提案手法の方が，より精度が高いことがわかった．

5. おわりに

提案手法は，共起関係と係り受け関係を導入した文書グラフを構築して，特徴ベクトルを抽出することで重要語を抽出した．そして，文節の役割を考慮して抽出することが，重要語抽出に有効であることを確認した．

今後は，係り受け関係にある文節間の品詞の組み合わせと解の包含率の相関についてより詳細に分析して，これをグラフの構築に導入することでさらに改善する．

参考文献

[1] Gerard Salton, Michael J.McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1984.

[2] 中川裕志，湯本紘彰，森辰則，出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，vol.10, No.1，pp.27-46，言語処理学会，2003.

[3] 松尾豊，大澤幸生，石塚満，Small World 構造に基づく文書からのキーワード抽出，情報処理学会論文誌，vol.43，No.6，情報処理学会，2002.

[4] Samer Hassan, Rada Mihalcea, Carmen Banea, Random-Walk Term Weighting for Improved Text Classification, Workshop on TextGraphs, at HLT-NAACL 2006, pp53-60, New York City, 2006.

[5] JUMAN，http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN，2014.

[6] KNP，http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP，2014.

[7] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html，CD-毎日新聞データ集 95 版.