分野連想語の出現位置に基づく話題分野の特定手法

1P-05

獅々堀 正幹† 岡田 真† 安藤 一秋†† 青江 順一†

†徳島大学 工学部 知能情報工学科

††香川大学 工学部 信頼性情報システム工学科

1．はじめに

複数の分野の話題が混在する文書から，話題のまとまり部分（パッセージ）を特定し，パッセージの話題分野を決定する技術は，様々な研究分野で重要な役割を担う技術である．本稿では，“投手”や“選挙”のように＜野球＞や＜政治＞といった特定の分野を連想できる単語（分野連想語[1]と呼ぶ）に着目し，この分野連想語を用いて，パッセージの範囲を決定し，そのパッセージの分野を特定する手法を提案する．

2．分野連想語

分野連想語とは，予め定義された分野体系内の各分野を連想できる単語のことである[1]．ここで，分野体系とは，各分野の上位下位関係を木構造で表現したものであり，葉に相当する分野を終端分野，葉以外の分野を中間分野と呼ぶ．この分野体系に従って予め文書データを分類し，各単語の集中率を計算することにより，分野連想語を決定する．図１に分野体系と分野連想語の例を示す．この分野連想語には，連想する分野の広さにより，以下のような水準に分類される．

水準１：唯一の終端分野のみを連想する．

水準２：同じ親分野をもつ終端分野の中で限られた複数の終端分野のみを連想する．

水準３：唯一の中間分野を連想する．

¥全体分野¥スポーツ¥テニス ⇒「シングルス，試合」
¥卓　球 ⇒「シングルス，試合」
¥相　撲 ⇒「横綱，試合，勝敗」
¥趣　味¥将　棋 ⇒「勝敗」
¥政　治¥選　挙 ⇒「勝敗」

図１　分野体系と分野連想語の例

水準４：複数の中間分野や終端分野を連想する．

水準５：特定分野を連想しない．

図１で例示した分野連想語に従うと，まず，「横綱」のように終端分野＜相撲＞を一意に連想する単語が水準１，「シングルス」のように同じ親分野内の複数の終端分野＜テニス＞，＜卓球＞を特定する単語が水準２，「試合」のように，一つの中間分野＜スポーツ＞を特定する単語が水準３，「勝敗」は，複数の終端分野＜趣味¥将棋＞，＜政治¥選挙＞や中間分野＜スポーツ＞を特定するので水準４となる．また，「場合」のように分野を特定しない単語は水準５である．

3．話題分野の特定法

本手法では，話題の継続性を計るため，話題の継続度αを計算する．継続度αは分野 Fj の分野連想語が連続して出現すると高まり，連続性が途絶えると衰退するように設定する．

3．1　継続度の計算方法

まず，話題の継続性が衰退する度合いとして，文 si での分野 Fj の衰退率($Decij$)を以下の式で定義する．

$$Dec_{ij} = -1 \times \left\{ \frac{\sum_{sk \in C_{i-1}} Freq(s_k, F_j) + Freq(s_i, F_j)}{num(C_{i-1}) + 1} \right\}$$

但し，$Freq(si,Fj)$は，文 si 内に出現する分野 Fj の分野連想語のポイント[1]であり，$Ci\text{-}1$ は，文 $si\text{-}1$ から遡って，分野 Fj の分野連想が連続出現した文集合，つまり，$Ci\text{-}1=\{si\text{-}n,\ldots,sk,\ \ldots,si\text{-}1\}$に対して，$Freq(sk,Fj) \neq 0$，$Freq(si\text{-}n\text{-}1,Fj)=0$ を満たす文集合である．また，$num(Ci\text{-}1)$は，文集号 $Ci\text{-}1$ の要素数 n を表す．

継続度は，新たな文を解析すると衰退し，その後，分野連想語が現れると上昇すると考え，以下の手順で継続度αij（文 si での分野 Fj の継続度）を求める．

A Retrieval Method of Relevant Passages Using Field Reminding Word's Locations, Masami Shishibori*, Makoto Okada*, Kazuaki Ando** and Jun-ichi Aoe*, *Tokushima University, **Kagawa University.

[1] 水準１，２，３，４のポイント値を順に 10，5，3，2 に設定した．

【 継続度α_{ij}の計算 】

手順１：$\alpha_{ij} = \alpha_{i-1j} + \rho \times Dec_{ij}$；

但し，$\alpha_{ij} < 0$の場合は，$\alpha_{ij}=0$ とする；

手順２：$\alpha_{ij} = \alpha_{ij} + Freq(si, Fj)$；

尚，パラメータρ（$0<\rho<1$）は，衰退率が継続度に影響する度合いとして定義される．

３．２　パッセージ特定アルゴリズム

本手法では，１文毎に継続度を計算し，以下の各処理により，パッセージを形成する．

（１）話題出現判定処理

本処理は，話題分野$Ftheme$が不定の場合に行われ，継続度α_{ij}が閾値γ_{th}を越え，かつ，最大となる分野が１分野に絞られた場合，Fjを$Ftheme$とする．

（２）話題転換処理

話題の転換[2]が起こった場合，隣接したパッセージ間の区切りを明確にする必要がある．本処理は，話題分野$Ftheme$の継続度α_{itheme}を越える継続度α_{ij}が現れた場合，話題分野$Ftheme$を分野Fjに転換する．

（３）話題継続処理

継続度α_{itheme}を越える継続度α_{ij}が存在しない場合，$Ftheme$の話題が継続しているとみなし，文siをパッセージに追加する．

４．評　価

４．１　実験データの内容

本実験では，3,102 個の分野連想語を準備した．図２に各分野に対する水準毎の連想語数を示す．実験データは，CD-毎日新聞’95 から各分野毎に無作為にN行選び，ランダムに 5 分野まとめたファイルを，Nの値を 5，10，15，20 と変化させながら各 30 ファイル作成した．尚，分野連想語を構築する際に用いた文書データから作成したクローズドデータセットと分野連想語を構築に用いなかった文書データから作成したオープンデータセットの２種類を用意した．

４．２　パッセージの特定精度

精度評価に用いる適合率Pと再現率Rは，出力パッセージと正解パッセージとが一致する文字数を$Pcon$，出力パッセージの文字数を$Pout$，正解パッセージの文字数を$Pans$とすると，$P=Pcon/Pout$，$R=Pcon/Pans$となる．

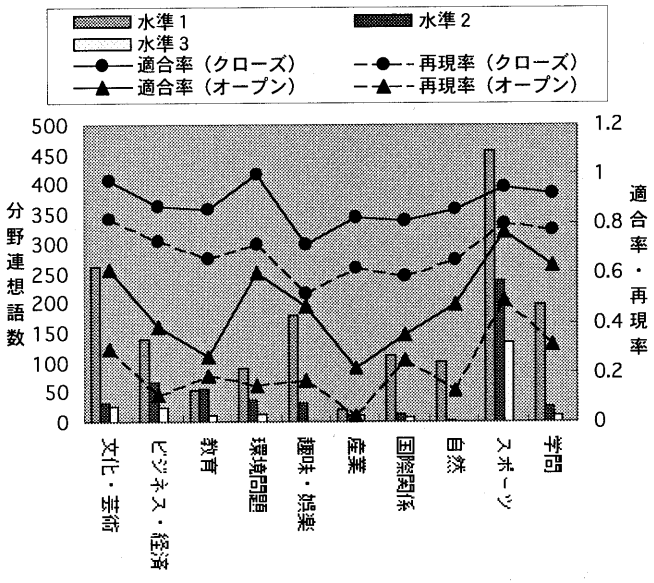

図２　分野連想語数と特定精度の関係

図２より，＜文化・芸術＞，＜スポーツ＞，＜学問＞等，水準１の分野連想語数が多い分野では，精度が高く，逆に，＜教育＞，＜産業＞等の水準１の分野連想語数が少ない分野では再現率が低かった．また，オープンデータセットに対する特定精度はクローズドデータに比べて低下しているが，＜スポーツ＞のように，質・量共に整った分野連想語が構築できている分野に関しては，精度の低下はさほど見受けられなかった．結論として，分野連想語が構築し易く，水準１の分野連想語が多い分野に対しては，本手法はかなり高い精度でパッセージを特定できると言える．

５．まとめ

本稿では，分野連想語の連続出現性から話題の継続度を計算し，パッセージを決定する手法を提案した．今回は，各水準に対して分野連想語のポイントやパラメータρを一意に決定していたが，コーパスデータからこれらの値を学習する機能も組み込んでいきたい．

参　考　文　献

[1] 辻孝子，泓田正雄，森田和宏，青江順一：複合語の分野連想語の効率的決定法，自然言語処理，Vol.7, No. 2，pp.3-26，2000.

[2] 内元清貴，小作浩美，井佐原均：対話型ネットニュースグループにおける話題転換記事の推定，言語処理学会第３回年次大会発表論文集，pp.377-380，1997.