ニュースからのトピック構造の抽出法とその対話的ニュース提供への適用
An Extraction method of Topic Structures from News Documents and its Application to Interactive News Providing Systems

東原 智幸† Tomoyuki Higashihara
渥美 雅保† Masayasu Atsumi

1. まえがき

インターネット上の文書数の増大によりユーザ自身の求める文書を検索することが難しくなっている．近年，その問題を解決するため，ユーザが興味を示す文書をシステムが自動的に検索し，提示する研究が多く行われている[1][2]．また，ブログなどの文書から話題(トピック)を抽出するサービスも提供されている．それに伴って，ブログからのトピック抽出について研究されており，[3]では文書ベクトルに基づいて，クラスタリングを行い，得られたクラスタをトピックとしている．その際，興味や文書の表現としてtf・idf法が多く用いられているが，文の係り受け構造や格情報など情報が利用できない点に問題がある．本研究では，ユーザの興味に基づくニュース提供システムの構築のためにトピック構造を抽出する方法について提案する．トピック構造の抽出は，ニュースを形態素解析[4]・構文解析[5]で得られた格構造の係り受け関係から抽出される．また，そのトピック構造をニュース提供に適応した場合のその有効性について検討する．

2. トピック構造

トピック構造は，ニュース中の格構造集合の内，係り受けの関係から主題と推測される格構造集合の部分集合である．

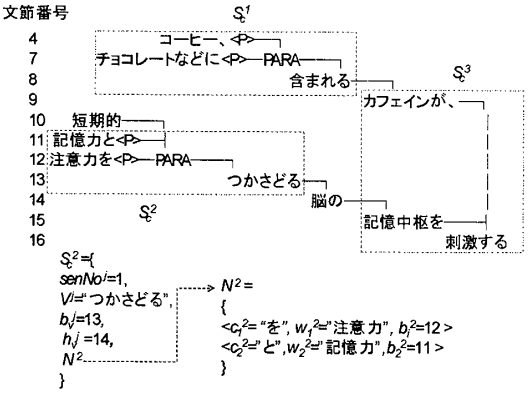

図 1 構文解析結果と格構造(一部略)

ニュース中のj番目の格構造S_c^j(図 1)は

$$S_c^j=\{senNo^j, V^j, b_v^j, h_v^j, N^j\} \quad (1)$$

と表現される．$senNo^j$は，ニュース中の文番号，V^jは動詞，b_v^jは動詞の文節番号，h_v^jは動詞の係り先文節番号である．

N^jは動詞V^jに係るM個の名詞集合で，表層格c_m^j，その格の要素(単語または複合語)w_m^j，その要素の文節番号b_m^jの組の集合からなり，

$$N^j=\{<c_1^j,w_1^j,b_1^j>,\ldots,<c_m^j,w_m^j,b_m^j>,\ldots,<c_M^j,w_M^j,b_M^j>\} \quad (2)$$

と表される．

例えば，ニュース中の 1 番目の文章に“カフェインが記憶中枢を刺激する”が存在する場合は，格構造は{1,“刺激する”,16,-1,{<“が”, “カフェイン”,9>,<“を”,“記憶中枢”, 15>}}と表現される．動詞の係る文節がない場合には，係り先文節番号は-1 となる．

2.1 トピック構造の抽出

ニュース n からトピック構造は(Ⅰ)～(Ⅴ)の手順で抽出される．

(Ⅰ)n 中の格構造集合の抽出

形態素解析・構文解析を 1 文ごとに行い，格構造集合$ScSet_n$を抽出する．

$$ScSet_n=\{S_c^1,\ldots,S_c^j,\ldots,S_c^J\} \quad (3)$$

ここで，Jは格構造の数である．S_c^jの動詞V^jに係る名詞w_l^jの抽出では，V^jに直接係っていない文節でも，w_l^jと並列関係または同格関係ある名詞もN^jに追加する．

(Ⅱ)S_c^j内の文節間距離$dist(S_c^j)$を次式により計算する

$$dist(S_c^j)=\sum_{m=1}^{M} abs(b_v^j-b_m^j) \quad (4)$$

$dist(S_c^j)$は，動詞に係る文節の数が多く，より離れた文節から係っているほど値が大きくなる．

(Ⅲ) S_c^jを修飾しているS_c^kの$dist(S_c^k)$を$dist(S_c^j)$に追加する

S_c^kのV^kの係り先番号h_v^kとS_c^jの動詞V^jに係る名詞w_m^jの文節番号b_m^jが同じ，つまり S_c^kによってw_m^jが修飾されている場合に，$dist(S_c^j)$に$dist(S_c^k)$を追加する．

$$dist(S_c^j)=dist(S_c^j)+\sum_{k\ s.t.\ (h_v^k=b_m^j)} dist(S_c^k) \quad (5)$$

(Ⅳ) 文節間距離$dist(S_c^j)$の上位n個をトピック構造として抽出する．

3. トピック構造抽出評価

3.1. 評価方法

以下の手順でトピック抽出法について評価を行う．

(1) トピック抽出を行い文章にする
(2) ニュース文と抽出された n 個のトピック構造をユーザに提示
(3) 抽出されたトピック構造がニュースの主題であるかどうかを，A（主題である），B(どちらかといえば主題)，C(どちらかといえば主題ではない)，D(主題ではない)の 4 段階で評価する

3.2. 評価結果

2005 年 12 月 2 日～5 日の goo ニュース 30 件を使用した．ニュースに含まれる格構造数の平均は，6.4 個である． n を 1～3 に変更し評価を行った．図 2にニュース本文，抽出されたトピック構造，評価の例を示す．

表 1 評価の割合(%)

	A	B	C	D
n=1	13.3	36.7	20	30
n=2	13.3	46.7	30	10
n=3	20	60	13.3	6.7

n が大きくなるにつれて，主題を表すトピックが抽出される割合が大きくなっていくことが表 1から確認できる．

3.3. 考察

トピック構造として抽出されなかったものの特徴として以下のものがあげられる．

・ 強調構文：“～のは，A だ。”

強調構文の場合，ある部分を強調するため，強調される部分が A の位置に移動する．例えば “グランプリに選ばれたのは B だ”の文章の場合，“グランプリに選ばれた”は格構造として抽出されるが，“B”の部分は抽出されない．そのため不十分な格構造となってしまい，トピック構造も意味が通じないものになる．

・ 名詞述語，形容詞文

動詞の格構造を抽出しているため，“A は B だ”のような名詞文や“A は B より高い”などの形容詞文は選択されない．

†創価大学大学院工学研究科情報システム工学専攻

・“の”格ある文

“Aは，運動を解禁するBの検討を始めた”の文章の場合，格構造として“Aは、検討を始めた”，“運動を解禁する”が抽出される．“検討”に係る動詞節がある場合には，式(5)を用いて文節間距離値が変更されるが，“の格”を介して係る場合に変更されない問題がある．“の格”で連体修飾される単語がN^{j}に含まれる場合に、その文節をN^{j}に追加することで解決できる．

・複文構造

“A が殺害された事件で，B を所持していたことがわかった”のような複文の場合，抽出される格構造は，(a)“A が殺害された”，(b)“B を所持していた”，(c)“事件で，ことがわかった”の３つである．トピック構造として(c)が選択されるが，トピック構造内の単語が“こと”のような形式名詞であったり，“事件”のように抽象的な単語が含まれている場合に，全体としての意味が不十分になることがある．この場合には抽出するトピック構造数 n の値を大きくすることで，意味的な情報が増え解決できると期待される．

・構文解析の間違い

構文解析の間違いにより，本来ならばトピック構造に含まれない単語が入ってしまい文章の意味が不明になってしまうこともある．

【タイトル】
カフェインに短期的記憶力を向上させる作用＝豪研究者
【本文】
オーストラリアの大学の研究者チームは３０日、コーヒー、紅茶、ソフトドリンク、チョコレートなどに含まれるカフェインが、短期的記憶力と注意力をつかさどる脳の記憶中枢を刺激する働きがある、との研究結果を発表した。
【 nBest=1 】評価 C
① 研究者チームは研究結果を発表した
【 nBest=2 】評価 C
① 働きがある、との
② 研究者チームは研究結果を発表した
【 nBest=3 】 評価 B
① カフェインが、記憶中枢を刺激する
② 働きがある、との
③ 研究者チームは研究結果を発表した

図 2　トピック構造抽出と評価の例

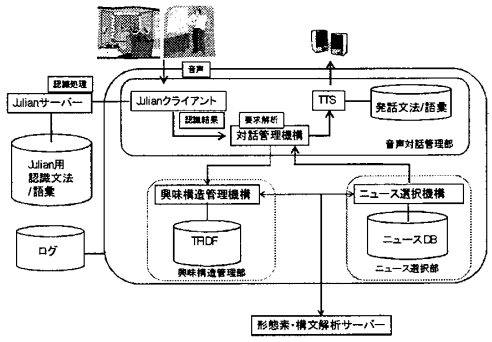

図 3 システム構造

4. トピック構造のニュース提供への適用

トピック構造をニュース提供に適用した場合の有効性について検討を行う．

4.1. システム構成

ニュース提供システム(図 3)[6][7]は，ユーザの要求を処理する音声対話部，ユーザの興味を管理する興味構造管理部，ユーザに提供するニュースを検索するニュース選択部に分かれる．興味構造管理部は，ユーザが興味をもった構造を管理する．提供するニュースを選択する際に，興味構造をニュース選択部に渡す．ニュース選択部では，取得した興味構造と類似度の高いニュースを選択し，音声対話管理部を介してユーザに提供する．類似度の高いニュースが存在しない場合には，新規のニュースから順次ユーザに提供する．提供したニュースに対するユーザの評価を，音声対話管理部より受け取り，興味構造管理部へ伝える．その評価により興味構造の分類を行う．

4.2. 現状のシステム

現状のシステムにおける問題点について述べる．

まず，ニュース提供においては，ニュース本文をユーザに提供する際，一度にすべて提供している．音声における提供においては，多くの情報を一度に提供してしまうと，ユーザが関心を示した部分だけを拾えないことや，あらかじめ知っていることなどを繰り返して提供してしまうなどの問題が生じる．

また，提供したニュースに対してユーザが興味を示した場合に，どの部分に興味を示したかが不明になる．

4.3. トピック構造導入で期待される効果

トピック構造を音声ニュース提供に適用する方法として以下の３つの方法が挙げられる．

(1)　中心話題提供

トピック構造を提供することにより，ニュースの中心的な話題からユーザに提供できるため，ユーザのニュースに対して理解しやすいと考えられる．

(2)　対話的提供

提供したトピック構造のある単語に対してユーザから質問があった場合，その単語と係り受け関係のある部分についてさらに提供することで，ユーザが関心を持つ部分を提供できる．また，ユーザが知っていることは省略して提供しないなど対話的な提供が行える．

(3)　興味の取得

提供したニュースに対してユーザが興味を示した場合に、その興味は，ニュースの中心話題であるトピック構造に向けられている可能性が高いと考えられる．そのため，ユーザの興味の学習や分類にトピック構造を用いることが有効である．また，対話的に提供を進めていくことで，どの単語にユーザの興味があるのかを追跡を行うことが可能である．

5. まとめ

ニュースからのトピック抽出法について提案し，評価を行った．また，トピック抽出をニュース提供に適用した場合の可能性について考察を行った．今後の課題としては，長いニュースにおける抽出度の検証，格構造と tf 値を用いたトピックの抽出法の提案，格構造間の意味的な距離を考慮した抽出法の提案が挙げられる．

参考文献

[1] 河合，熊本，田中：印象と興味に基づくユーザ選好のモデル化手法の提案とニュースサイトへの応用，日本知能情報ファジィ学会誌，vol.18 No.2，pp.173-183，2006.

[2] 野美山，紺谷，渡辺他：個人適応型情報検索システム－個人の興味を学習する階層記憶モデルとその協調的フィルタリングへの適用－，情報学基礎研究会報告，Vol.96，No70，pp.49-56，1996.

[3] 戸田，黒田，福田，石川：ブログにおける多視点からのトピック抽出手法の提案，DEWS2008(第19回データ工学ワークショップ)，B4-2，2008.

[4] 黒橋：日本語形態素解析システム JUMAN ver. 5.1，東京大学大学院情報理工学系研究科，2005.

[5] 黒橋：日本語構文解析システム KNP ver. 2.0，東京大学大学院情報理工学系研究科，2005.

[6] 東原，三吉，渥美：ウェブニュース提供のための自己組織化関係ネットワークと格重み付き単語頻度ベクトルを用いたユーザの興味構造表現，2D-1，p.2-63～p.2-64，2008.

[7] 東原，三吉，渥美：スマートホームにおける音声ニュース提供システムアーキテクチャの構築，FIT2008(第 7 回情報科学技術フォーラム)，E-025，p.193-p.195，2008.