論文概要からのタイトル自動生成の試み

安藤　一秋[†]　　新居　雅也　　溝渕　昭二[‡]
[†]香川大学工学部　　[‡]近畿大学理工学部

1　はじめに

論文のタイトルは，本文の内容を簡潔に表現した要約と考えられる．タイトルやヘッドラインの生成は，携帯端末表示や Web 検索結果の提示など，幅広い応用が考えられ，これまでにもさほど多くはないが，いくつかの方法が提案されている[1]-[5]．具体的には，重要文抽出に基づく要約と同じ流れで新聞記事のリード文からヘッドラインを生成する方法[3]や重要語を起点として，その語に先行，後続する語を再帰的に連結する方法[1][2]，本文とタイトルに共起する単語を基に言語モデルを作成し，タイトル生成する方法[4]，Support Vector Machine と言語モデルを用いる方法[5]などがある．

しかし，いずれの手法も，タイトルを構成する上で必要となる要素については言及していないため，情報が欠落したタイトルが生成される可能性がある．

そこで，本稿では，千田らの研究[6][7]を参考に，タイトルに必要な構成要素を手掛かりにして，論文概要からタイトルを自動生成する試みについて報告する．特に，本稿では，開発した技術の形態（システム，手法など）要素を起点に，開発した技術を提示するためのタイトル生成を行う．

2　主題要素

千田[6]らは，開発した技術をアピールするためにタイトルを分析することで，タイトルの構成要素を 6 種に大別し，出現頻度を基に任意要素と必須要素を定義した．以下に[6]から抜粋した構成要素を示す．

必須要素

対象：開発した技術の動作対象（何を）
動作：開発した技術の動作（どうする）
形態：開発した技術の形態（手法，システム）

任意要素

目的：開発した目的（何をどうするために）
方法：開発した技術の実現方法（どうやって）
長所：開発した技術の長所（いつ，どの様に）

以上の 6 種の要素に，新たに任意要素として「主旨」（「形態」要素をどうするといった内容を表す）を加えたものを，「主題要素」とよぶ．本稿では，主題要素を考慮することにより，以下のパターンで構成されるタイトル生成を目指す．

タイトル＝
{目的｜方法｜長所}・対象・動作・形態・{主旨}

タイトル例：
著者タイトルを利用した文書要約システムの試作

方法	著者タイトルを利用した
対象	文書
動作	要約
形態	システムの
主旨	試作

以上のパターンにより，開発した技術を提示するためのタイトルが生成される．

3　重要文の抽出

論文概要では，1 文または 2 文程度に主題を集約して記述することが多いと考えた．そこで，重要文抽出法により抽出された 1 文もしくは 2 文からタイトルを生成することにする．

3．1　タイトル分析と重要文抽出のための情報

重要文抽出には様々なテキスト中の特徴[8]が利用される．タイトルに関連する情報としては，タイトルに含まれる名詞を含む文，名詞と関連する文を重要文と見なす方法がある．本稿では，タイトル中の名詞ではなく，タイトルの構造的特徴に着目する．

まず，タイトルから構造的特徴を抽出するために，1975－2002 年度の情報処理学会技術研究報告（NL 研）1718 件のタイトルを分析した．各タイトルは形態素解析後に，名詞を中心に抽象化し，タイトル構造を抽出した．形態素解析には茶筌[10]を使用した．

[抽象化の例]：
著者／タイトル／を／利用／し／た／文書／要約／システム／の／試作
→ N を利用した N の N

実験により，1718 件から 860 件のタイトル構造が抽出された．異表記の統合はしていない．頻度の上位 5 件を表 1 に示す．

表 1　タイトル構造の頻度（上位 5 件）

タイトル構造	頻度
N の N	104
N における N の N	48
N による N の N	46
N を用いた N の N	41
N	40

タイトル構造には，「N における N」や「N による N」などの定型表現が数多く含まれていた．そこで，860 件の構造から手掛かりになり難いと考えられる「N の N」を除いた定型表現を抽出し，1718 件のタイトル中での頻度を調べた結果，67%（=1167/1718）のタイトルで定型表現が含まれていた．その上位 5 件を表 2 に示す．

表 2　部分構造の頻度（上位 5 件）

部分構造	頻度
における	258
による	237
を用いた	225
に基づく	151
のための	116

タイトル中に定型表現が約 70%含まれているということは，概要中の定型表現を含んでいる文は，タイトル生成に有用であると考えられる．そこで，2000－2002 年度の情報処理学会技術研究報告（NL 研）359 件から無作為に抽出した 115 件を利用して，タイトルでの定型表現の有無と概要中の重要文での有無を調査した．重要文は人手で判断した．結果を表 3 に示す．

実験により，74%（=85/115）の重要文には定型表現が含まれていることが判明した．また，タイトルに定型表現が含まれる場合は，76%（=65/85）で重要文にも定型表現が含まれ，逆に，タイトル中に含まれなくとも，60%（=18/30）で重要文に定型表現が含まれていた．以上より，重要文を抽出するための 1 つの情報として，定型表現が有効であると考える．しかし，重要文以外にも定型表現が含まれると考えられるため，他の情報も利用することにする．

表 3　タイトル・概要文と定型表現の関係

	タイトルに定型表現あり	タイトルに定型表現なし	合計
タイトルでの頻度	85	30	115
重要文での頻度	67	18	85

論文概要では，提案方法や主題を明示するために，“本論文では”や“提案する”などの手掛かり語を使うことが多い．本稿では，“を提案する”のような“提案”タイプと“本論文”のような“本～”タイプ，“そこで”の 3 種を手掛かり語として定義する．

重要文抽出の情報として，主題要素「形態」の導入を検討したが，今後のタイトル生成において，重要部の汎用性をもたせるために，本稿では利用しないこととした．「形態」以外の要素については，今後，更なる検討が必要である．

最終的に，重要文の抽出には，TF・IDF と定型表現，手掛かり語を利用する．

3．2　重要文の抽出

重要文抽出の流れを以下に示す．

Step1_1：【形態素解析】

入力された文書を形態素解析する．

Step1_2：【定型表現の固め処理】

文書に含まれる定型表現の固め処理（“を／用い/た”→“を用いた”）を行い，必要情報（品詞や得点など）を付加する．

Step1_3：【TF・IDF の計算】

文書 d_j の形態素 i の重み $w(i,j)$は，$f\_\max_j$ を d_j の最大形態素頻度とした場合，以下で計算する．

$$w(i,j) = TF(i,j) \times IDF(i) = TF(i,j) \times (\log(N/df_i) + 1)$$

$$TF(i,j) = \log(1 + f_{ij} / f\_\max_j)$$

Step1_4：【文の重要度計算】

文の重要度 $W(k)$は，形態素の重みの総和をとり，形態素数 $len(k)$で正規化した値に，定型表現の重み α_m と手掛かり語も重み β_n を掛ける．但し，定型表現，手掛かり語が存在しない場合は，α_m=β_n=1.0 とする．

$$W(k) = \alpha_m \beta_n \sum w(i,j) / len(k)$$

Step1_5：【並び替え】

重要度の基づいて並び替えを行う．（終わり）

α_m（1 or $1.05 \leqq \alpha_m \leqq 1.2$）は 1718 タイトル中の頻度を基に実験を行い決定した. β_n（1 or $1.3 \leqq \beta_n \leqq 1.5$）は定型表現とのバランス，重要度の最下位と最高位の関係を基に実験で決定した. 両方の実験には, 2000－2002 年度 NL 研の 359 件を利用した.

4　タイトル生成

主題要素の取り扱いについて述べる. まず,「形態」は，1975－2002 年度 NL 研の 1718 件のタイトルを調査し，諸分野の論文でも汎用的に利用できる“手法”，“システム”，“方法”，“法”，“方式”，“形式”，“ツール”と, 頻度が上位であった“アルゴリズム”，“モデル”の 9 種を利用する. 但し，“本”＋「形態」（本手法）や“提案”＋「形態」（提案手法）などはタイトル生成の起点としない.

「動作」は，動詞とする.「対象」は,「を」格を取る名詞に限定する.「目的」は,“のための”，動詞＋“ための”などの定型表現を利用する.「方法」も同様に,「による」,「を用いた」などの定型表現を利用する. 概要中の定型表現は活用も考慮する.「主旨」は，サ変動詞のみ扱う.「長所」は，限定して抽出しない.

以上に基づいたタイトル生成法を以下に示す.

Step2_1 :【前処理】
固め処理や情報の付加，不要情報の削除を行う.
Step2_2 :【文節区切り】
体言句と用言句に分かれるように拡張文節区切りを行う.
Step2_3 :【定型表現の処理】
定型表現の活用形を，タイトル頻度を基に決定した基本形に言い換えると共に，任意要素抽出のための情報（目的，方法）を付加する.
Step2_4 :【主題要素の獲得】
主題要素の獲得を行う.（終わり）

以下に，主題要素の獲得方法を示す.

Step3_1 :【形態要素の探索】
文末から文頭へ「形態」要素を含む文節 k を探索する. 文頭を 1 とする.
Step3_2 :【主旨要素の獲得】
k+1 の文節がサ変動詞（名詞）を含むなら，サ変語幹を「主旨」要素として獲得する.
Step3_3 :【形態要素の獲得】
文節 k から「形態」要素を獲得する. k 内で「形態」に先行する形態素数を数え，数が 0 なら Step3_4 へ，1 なら Step3_5(a)へ，2 以上なら Step3_6 へ.
Step3_4 :【動作要素の獲得】
文節 k から最も近い用言文節 x を探し,「動作」要素を獲得する. ない場合は Step3_7 へ.
Step3_5 :【対象要素の獲得】
｛(a)文節 *k-1*｜(b) 文節 *x-1*｝から前方で最も近い体言句を最も近い用言句までの範囲で探し,「対象」要素を獲得する. ない場合は Step3_6 へ.
Step3_6 :【任意要素の獲得】
文節 *k*-1 から Step2_3 で付与した情報を基に任意要素を獲得する. 但し，同じ属性の任意要素が出現する場合は，文節 k に近い方を優先する.
Step3_7 :【任意要素の対象の獲得】
任意要素に対する「対象」が必要な場合は, Step3_5 と同様な処理で獲得する.
Step3_8 :【タイトルの出力】
主題要素を整形してタイトルを出力する.（終わり）

5　実験と考察

5. 1　重要文抽出の精度

実験データとして，分析に利用した 2000－2002 年度 NL 研の 359 件の論文概要から「形態」を含む 20 件をランダムに抽出したもの（Data-A）と，分析に利用していないデータを 1999 年の論文概要から新たに 20 件（Data-B）用意した.

被験者 4 人が選択した重要文とシステムが 1 位で抽出した重要文が等しい割合で評価を行う. 比較対象は，A 法：TF・IDF のみ利用；B 法：TF・IDF と定型表現を利用；C 法：TF・IDF と手掛かり語を利用；D（提案）法：全て利用；である. IDF 値は，2000－2002 年度 NL 研の 359 件の概要から計算した. 表 4 に実験結果を示す.

表 4　重要文が等しい割合

手法	A 法	B 法	C 法	D 法
Data-A（%）	28.8	30.0	58.8	71.3
Data-B（%）	22.5	28.8	68.8	73.8
平均（%）	25.7	29.4	63.8	72.6

B 法と C 法を比較することで，定型表現に基づく方法が手掛かり語に基づく方法より精度がよいことがわかる. これは，定型表現が重要でない文にも出現していることが影響していると考えられる. 定型

表現と手掛かり語を考慮した提案（D）法の精度が最も高いことからは，手掛かり語を含んでいない重要な文を定型表現により抽出できたことを意味する．

5．2　タイトル生成法の精度

実験データは，5.1 と同じ 359 件から「形態」を含む 20 件[†]をランダムに抽出したもの（Data-A’）と，新たに 1999 年の論文概要から「形態」を含む 20 件（Data-B’）を用意した．それぞれのデータに対して，生成されたタイトルと論文概要，生成に利用された重要文を 4 人の被験者に提示し，以下の 4 段階でタイトルの妥当性を評価する．

(1)タイトルとして成立する；(2)タイトルして許容できる；(3)タイトルとして物足りない；(4)タイトルとして成立しない．

また，生成されたタイトルが評価 2～4 と判定された場合は，判定理由を以下から選択（複数可）してもらう．

（A）必要な構成要素が不足している；（B）不要な構成要素が存在する；（C）日本語として不適切である；（D）重要文が不適切である；（E）その他．

実験結果を表 5 に，判定理由を表 6 に示す．表 5 から，評価 1 に関しては Data-A’と B’に対して，ほぼ同等の精度であることが分かる．評価 1 と 2 を含めた場合は，Data-A’が約 75％で Data-B’が 68％と 7％の差が生じた．これは，表 6 より，その他に含まれる要素が影響したと考えられる．

表 5　評価結果

評価	1	2	3	4
Data-A’（%）	32.5	42.5	16.3	8.7
Data-B’（%）	31.3	36.3	23.8	8.8
平均（%）	31.9	41.0	20.1	8.8

評価 2～4 に判定した理由に関しては，表 6 から Data-A’と B’共に，「必要な構成要素が不足している」が最も多かった．本手法では，「形態」や「対象」要素への修飾は，定型表現と助詞“の”以外扱ってないため，その他の連体修飾が取れずに情報が欠落したと考えられる．2 番目に多かった理由は，「タイトル生成文の不適切」であった．これは，タイトル生成文の抽出精度が約 73％であることに起因すると考えられる．その他の理由の多くは，内容の重複であり，修飾節の取り扱い方に問題があった．

[†]本手法は形態要素を含まないタイトルを生成することができないが，359 文中の 76％（275 文）は「形態」を含んでいた．

表 6　判定理由

理由	A	B	C	D	E
Data-A’（%）	44.3	14.8	9.8	21.3	9.8
Data-B’（%）	48.5	7.6	9.1	18.2	16.7
平均（%）	46.4	11.2	9.5	19.8	13.3

また，著者タイトルと生成されたタイトルの形態素の一致度は約 36％（Data-A は 37.8％，Data-B は 34.1％）であった．本手法により，著者タイトルとは異なるが，開発した技術を提示するタイトルが約 73％の精度で生成できた．

6　おわりに

本稿では，タイトルの構成要素である「形態」に着目して，論文概要から，開発した技術を提示するためのタイトルを生成する方法について報告した．実験により，約 73％の重要文抽出精度で，約 73％のタイトルが妥当であると評価された．今後は複数文からの主題要素の獲得方法や「形態」を含まないタイトルの生成法の考案を行う．

参考文献

[1]松本賢司，伊藤山彦，谷田泰郎，柏岡秀紀，田中英輝，“重要名詞の共起情報を利用した表題生成”，言語処理学会第 7 回年次大会，pp.375-378，2001.

[2]松本賢司，伊藤山彦，柏岡秀紀，浦谷則好，“重要語の共起情報を用いた講演文の表題生成”，情報処理学会第 61 回全国大会，4T-02，2000.

[3]畑山満美子，松尾義博，大山芳史，白井諭，“日本語記事の重要情報に基づく英文ヘッドライン生成法”，言語処理学会第 5 回年次大会，pp.17-20，1999.

[4]M. Banko and V. Mittal and M. Witbrock，“Headline Generation Based on Statistical Translation”，Proc. of the 38th Annual Meeting of the Association for Computational Linguistics，pp.318-325，2000.

[5]廣嶋伸章，長谷川隆明，山崎毅文，“統計的手法に基づく Web ページからのヘッドライン生成”，情報処理学会技術研究報告，NL149-7，pp.45-50，2002.

[6]千田恭子，“開発した技術をアピールする表題のつけ方”，情報処理学会技術研究報告，NL145-14，pp.91-96，2001.

[7]千田恭子，篠原靖志，“読者の特性に応じた効果的な表題の作成方法”，言語処理学会第 8 回年次大会発表論文集，pp.212-215，2001.

[8]奥村学，難波英嗣，“テキスト自動要約に関する研究動向”，自然言語処理，Vol.6，No. 6，pp.1-26，1999.

[10]茶筅，http://chasen.aist-nara.ac.jp/