Vol.2009-NL-193 No.9
2009/9/29

談話の顕現性を考慮した重要語抽出とその応用

飯田　龍[†1]　徳永健伸[†1]

本稿では文献[3]で提案した顕現性の観点から文章中の語の重要度を求める手法を提案する．この手法では，文章の顕現性を考慮した語のランキングを行い，その順序をもとに語の重要度を決定する．提案手法が出力する重要度と既存の重要度の指標を比較するため，語の重要度の情報が重要な手がかりとなる自動要約の課題を対象に評価を行った．Text Summarization Challenge 2 の評価データを対象に重要文抽出と重要箇所検出の問題について，tf-idf をベースラインとしてどちらの重要度の情報が各問題の解析に役立つかを調査した．評価実験の結果より，提案手法の重要度の情報を tf-idf のような既存の重要度と組み合わせることで，重要文抽出と重要箇所検出の各問題で精度が向上する可能性があることを示す．

Salient Word Extraction in Discourse and its Application

Ryu Iida[†1] and Takenobu Tokunaga[†1]

This paper proposes a method to estimate the word significance based on the word salience in discourse, which was introduced by Iida et al.[3]. The method ranks words in a text in order of their salience in discourse and estimates the word significance based on these saliency ranking. The effectiveness of the proposed significance metric was evaluated in comparison to tf-idf through automatic text summarization application using the Text Summarization Challenge 2 dataset. The experimental results demonstrate the potential of our approach to improve performance in both tasks.

1. はじめに

文章中に出現する単語の重要度の情報は情報検索や自動要約などさまざまな自然言語処理の分野で利用されている．語の重要度に関しては特に tf-idf[5] が自動要約などの応用処理で一般的に用いられており，各応用処理において精度向上に貢献することが示されている．この tf-idf は該当文章内に出現する語の頻度の情報をもとに重要度を決定するという特徴を持つ．しかし，日本語などの言語では主題となる表現が頻繁に省略され，重要語とすべき語が頻出するとは限らないため，文章中の出現頻度をもとに重要度を求める tf-idf が必ずしも適切な重要度の値を出力するとは限らない．

このように tf-idf では単純に文章中の bag-of-words の特徴を捉えて重要度を求めるのに対し，我々が文献[3]で導入した手法では，談話の顕現性に基づく語の順序付けを行う．この手法では，Walker[8] の議論をもとに，文章中の顕現性の高い語をキャッシュに保持し，その中だけを探索するという枠組みを提案し，これを機械学習に基づく手法として実現することにより照応解析の探索範囲を解析精度を保ちつつ削減することに成功した．この手法では，ある名詞句が以降の文脈で代名詞化もしくは省略される場合には，その名詞句は現在の文で顕現性が高いという仮説に基づき，人手でタグ付与された照応関係を手がかりとすることで顕現性の観点から文章中に出現している語の順序付けを行う．これにより，現在参照している文までに出現した語の集合を以降の文脈でどの語が主題となり得るかという観点で順序付けした結果を得ることができる．このような顕現性の観点に基づく語の順序情報は tf-idf だけでは捉えることができない重要度を求めるための手がかりとみなすことができ，この手がかりに基づく重要度と tf-idf の情報を組み合わせることで応用処理の精度を向上させることが期待できる．そこで，本稿ではこの予備調査として自動要約，特に重要文抽出と重要箇所検出の 2 つの課題を対象に顕現性に基づく重要度と tf-idf の比較を行い，それぞれの重要度にどのような違いがあるかを調査した結果について報告する．

本稿では，まず 2 節で我々が以前提案した教師有りの談話の顕現性に基づき名詞句をランキングする手法を説明し， 3 節でこの手法を語の重要度として利用するための一例を示す．次に 4 節では評価実験として重要文抽出課題と重要箇所検出課題について各手法がどのように自動要約に貢献するかを調査した結果について報告する．さらに 5 節で提案手法と関連する既存研究を紹介し，最後に 6 節でまとめる．

2. 談話の顕現性を考慮した重要語抽出

我々は文献[3]で顕現性に基づき語を順序付けする手法として，談話状況の遷移とは独立に文章中に出現する各語の顕現性を推定するモデル（以後，**静的モデル**）と，談話の各状況に応じて動的に顕現性の情報を更新するモデル (以後，**動的モデル**）の 2 種類の方法を提案

†1 東京工業大学 大学院情報理工学研究科
Tokyo Institute of Technology, Department of Computer Science

した．

これら 2 種類のモデルでは共通にセンタリング理論[2] などの談話理論で導入されている「談話のある時点の発話において顕現性の高い表現が次の発話で代名詞化される，つまりそのような場合には照応関係となりやすい」という特徴を利用している．各モデルではこの特徴を捉えるために，あらかじめ人手でタグ付与された照応関係を訓練事例として利用することで，語の顕現性の高さを学習する．具体的には，NAIST テキストコーパス[11] にタグ付与された文間のゼロ照応関係を訓練事例としてランカーを作成し，そのランカーを利用して文章中に出現する名詞句を顕現性の観点で順序付けする．この節では静的モデルと動的モデルの詳細についてそれぞれ説明する．

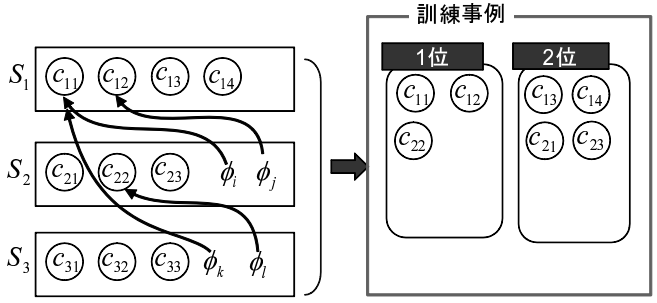
2.1 静的モデル

静的モデルでは，入力となる文章が与えられた場合に文章中のすべての名詞句を一度にランキングする．このランキングを行うモデルを作成するにあたり，まず文章中のどの表現が顕現性が高いかという情報が必要となる．ここでは,「後の文脈中のゼロ代名詞から指される表現はその文章の主題である」という仮説に基づき，訓練事例中の文章内である名詞句がその後方の文に出現するゼロ代名詞と照応関係になる場合には顕現性が高い，それ以外の名詞句は顕現性が低い，という 2 値の情報を文章中のすべての名詞句に付与する．この顕現性の高さの情報をもとに，顕現性が高い名詞句を 1 位，それ以外の名詞句を 2 位と人手でランク付けすることで学習に利用する訓練事例を作成し，この半順序関係が付与された文章集合を Ranking SVM[4] を用いて学習することにより，与えられた名詞句の集合に対して全順序を出力するランカーを得る．

静的モデルの訓練事例作成について，図 1 に描かれた状況を例に説明する．図 1 の文章は S_1 から S_3 の 3 文で構成されており，ゼロ代名詞 ϕ_i と ϕ_k の先行詞が c_{11}，ϕ_j の先行詞が c_{12}，ϕ_l の先行詞が c_{22} である．この状況で後続文脈のゼロ代名詞と照応関係にある先行詞候補である c_{11}，c_{12}，c_{22} を 1 位，それ以外の候補を 2 位として順序学習を行う．

2.2 動的モデル

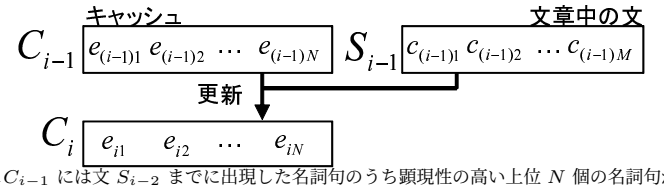
静的モデルが文章全体の顕現性を静的に捉えようとするのに対し，動的モデルでは談話の各状況，つまり現在どの文を処理しているのかという情報を反映してその状況における既出の名詞句の顕現性を推定する．この談話の状況に応じた処理を実現するために，動的モデルでは図 2 に示すキャッシュという概念を導入し，現在参照している文における名詞句の顕現性を見積る．このキャッシュにはこれまでに出現した名詞句のうち顕現性が高い上位 N 個の名詞句が保持されており，この N 個の名詞句と現在参照している文に出現する M 個の名詞句から新たに顕現性が高い上位 N 個を決定することでキャッシュの情報を更新する．この処理を文章の最初から最後まで繰り返すことにより，文章中のある文における既出の名詞句の顕現性の情報は，各文に対応するキャッシュに保持されることになる．このキャッシュの更新を実現するためにはどの語をキャッシュに保持するかという問題を考える必要があるが，ここではこの問題を静的モデルと同様に教師有りの名詞句のランキング問題として考える．ただし，動的モデルでは静的モデルが実現しているように文章中のすべての名詞句を一度にランキングするのではなく，キャッシュ内に出現している N 個の名詞句と現在参照している文に出現している M 個の名詞句を顕現性の観点からランク付けする．このランク付けを行うランカーを作成する際には，静的モデルと同様にタグ付与されたゼロ照応関係の情報と Ranking SVM を利用する．ただし，動的モデルの場合は，名詞句の顕現性に関して「ある文においては高く見積る必要があるが以降の文脈では低くなる」といった顕現性の遷移を扱う必要がある．このため，静的モデルのように文章から 1 つの訓練事例集合を作成するのではなく，文章の各文ごとにその文における既出名詞句の顕現性を反映した訓練事例集合を作成する．



c_{ij} は文 S_i に出現した名詞句を表し，ϕ_i はゼロ代名詞を表す．

図 1 静的モデルの訓練事例の作成

図 3 に描かれた状況を例に動的モデルの訓練事例作成について説明する．まず 1 文目から訓練事例を作成する場合には，名詞句 c_{11} と c_{12} が後方文脈のゼロ代名詞から指されており，この 2 つの名詞句の顕現性が高いと考える．このため，c_{11} と c_{12} を 1 位に，それ以外の c_{13} と c_{14} を 2 位にランク付けしたものを訓練事例とする．2 文目については，1 文目に出現している c_{11} がさらに後方の文脈でもゼロ代名詞として出現しているため，この名詞句はこの文でも顕現性が高いと考える．また，c_{22} についても後方のゼロ代名詞 ϕ_l と照応関



キャッシュC_{i-1} には文 S_{i-2} までに出現した名詞句のうち顕現性の高い上位 N 個の名詞句が保持されており，一方 S_{i-1} には M 個の名詞句が出現している．ここで，e_{ij} はキャッシュC_i に保持された名詞句であることを表し，c_{ij} は文 S_i に出現している名詞句であることを表す．

図 2 動的モデルにおける顕現性の検出

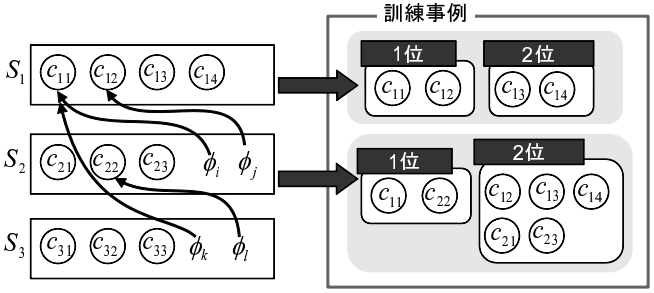


図 3 動的モデルの訓練事例の作成

係にあるため，顕現性が高い．一方 c_{23} は後方のどのゼロ代名詞からも指されないため顕現性が低いとみなす．同様に前方文脈に出現した c_{12}，c_{13}，c_{14} についても後方文脈で言及されることはないので，これら 3 つの表現はこの文では顕現性が低いと考える．以上の顕現性の高低に従い，図 3 に示すように 2 文目についての訓練事例として c_{11}，c_{22} を 1 位，それ以外の名詞句 c_{12}，c_{13}，c_{14}，c_{21}，c_{23} を 2 位の事例として訓練事例集合を作成する．このように談話の各文における顕現性の序列を学習することによりどのような状況で顕現性の遷移が起こるかを捉えられる可能性がある．動的モデルの訓練事例作成の手続きの詳細については図 4 にまとめる．

2.2.1 素　　性

静的モデルと動的モデルで利用するランカーを作成する際に利用する素性を表 1 にまとめる．この中で格助詞の情報は，センタリング理論[2] における現在の発話の顕現性を捉えるための手がかりとして Walker ら[6] をはじめ多くの研究者に利用されている[7]．例えば，

```
Function makeTrainingInstances (T: input text)
    C := NULL // set of preceding candidates
    S := NULL // set of training instances
    i := 1; // init
    while (exists s_i) // s_i: i-th sentence in T
        E_i := extractCandidates(s_i)
        R_i := extractRetainedInstances(E_i, T)
        D_i := E_i \ R_i
        r_i := extractRetainedInstances(C, T)
        R_i := R_i ∪ r_i
        D_i := D_i ∪ (C \ r_i)
        S := S ∪ {⟨R_i, D_i⟩}
        C := C ∪ E_i
        i := i + 1
    endwhile
    return S
end
Function extractRetainedInstances (S, T)
    R := NULL // init
    while (elm ∈ S)
        if (elm is anaphoric with a zero-pronoun located
            in the following sentences of T)
        R := R ∪ elm
        endif
    endwhile
    return R
end
```

図 4 動的モデルの訓練事例作成の手続き

Walker ら[6] のゼロ照応解析の手法では，“は > が > に > を > その他” のような序列で顕現性の高さを定義し，各発話で最も顕現性の高い候補を次の発話のゼロ代名詞の先行詞として解釈している．つまり，このような助詞の情報や “について” や “に関して” など主題を表す手がかり表現を素性として導入することで，現行発話内の候補の相対的な顕現性の高さを捉えることが可能になると考えられる．また，接続表現の情報は文章の首尾一貫性を捉える手がかりであり，談話の構造を解析するための素性として利用されている．動的モデルにおける名詞句のランキングについても，現行の発話から先行文脈までの接続表現の情報を導入することで，例えば，逆説的に文がつながる場合には主題が遷移するなど，主題の遷移が起こる際の談話の流れを近似的に捉えることが可能になると考えられる．

3. 顕現性に基づく重要度の算出と要約への応用

2.1 と 2.2 で説明した静的モデルもしくは動的モデルは文章中の名詞句を顕現性の観点からランク付けするだけであり，そのままでは語の重要度のスコアとしては利用できない．そ

表 1 キャッシュモデルで利用する素性

素性	説明
POS	naist-jdic ⋆1の定義に従う C の品詞.
IN_QUOTE	C が引用の中に含まれる場合は 1．それ以外は 0.
BEGINNING	C が文章の最初の文に含まれる場合は 1．それ以外は 0.
CASE_MARKER	C を含む文節に出現する “は”，“が” などの助詞の情報.
DEP_END	C が最後の文節の係り元である場合は 1．それ以外は 0.
CONN*	C と Z の間に出現する接続表現の集合．各表現は 2 値素性として利用される.
IN_CACHE*	C がキャッシュ内に保持されている場合は 1．それ以外は 0.
SENT_DIST*	C と Z の文単位での距離.

C は先行詞候補，Z は対象となるゼロ代名詞を表す．“*” が付いた素性は動的モデルでのみ利用される.

表 2 重要文抽出に利用する素性

素性	説明
POSITION	S が文章全体のどの位置に出現したかを 0〜1 で正規化した値.
LENGTH	S の長さ（S が何文節で構成されるか）.
CONN	S に出現する接続表現.
CASE_MARKER	S に出現する助詞.
PROP_NOUN	S に固有名詞-人名，組織，地域，一般が出現する場合は 1．それ以外は 0.
TFIDF	tf-idf に関する文の重要度.
STATIC	静的モデルに基づく文の重要度.
DYNAMIC	動的モデルに基づく文の重要度.

S は分類対象となる文を表す．CONN と CASE_MARKER は表現ごとに出現するか否かの 2 値素性とする． 4.1 の実験では TFIDF，STATIC，DYNAMIC の 3 つの値を利用する場合としない場合で比較を行う.

こで，本研究では各モデルが出力する順位の逆数を各名詞句の重要度のスコアとした．また，動的モデルの場合は，ある名詞句に対して談話の各文ごとに異なった順位を出力するため，各文における順位の逆数の総和をその名詞句の重要度とした．各モデルの重要度を求める式を式 1 と式 2 にまとめる．ここで，rank_i は静的モデルが出力する語 w_i の順位を表し，また rank_{ji} は語 w_i の文 S_j における順位を表す．$|S|$ は該当文章に含まれる文の総数である.

$$\mathrm{score}_s(w_i) = \frac{1}{\mathrm{rank}_i} \tag{1}$$

$$\mathrm{score}_d(w_i) = \sum_j \frac{1}{\mathrm{rank}_{ji}} \qquad (1 \leq j \leq |S|) \tag{2}$$

また， 4.1 で後述する重要文抽出の問題では，式 3 に定義する文の重要度を用いて重要度の高い順に指定された文数の文の抽出を行う.

$$\mathrm{importance}(S_i) = \sum_j \mathrm{score}(w_i) \qquad (w_i \in S_i) \tag{3}$$

さらに，各文から得られる他の情報を併用した教師有り手法に基づく重要文抽出についても評価を行う．ここでは，それぞれの重要度を素性として導入することで結果がどのように変化するかについて調査する．本研究で導入する教師有りの重要文抽出モデルを作成する際は，どの文が重要文として抽出されるかがタグ付けされた文章のうち，重要文として抽出すべき文を 1 位，それ以外の文を 2 位にランク付けし，その半順序を表 2 に定義された素性を用いて Ranking SVM[4] で学習する． 4.1 の実験で利用する Text Summarization Challenge 2（TSC2）[9] のデータセットには 10%，30%，50%の 3 種類の要約率が設定されているため，それぞれの要約率ごとにランカーを作成する．また，実際に重要文を決定する際には，学習の結果得られたランカーを利用して入力文章中の文集合をランキングする．次に，各要約率ごとにあらかじめ設定された文数だけランキングの結果の上位から抽出することで重要文抽出を実現する.

⋆1 http://sourceforge.jp/projects/naist-jdic/

4. 評 価 実 験

提案モデルが出力する重要度が高い値を持つ語がどのようなものであるか，またその語の情報が応用処理に役立つかを調査するために，本稿では要約を例に既存研究で一般的に利用されている tf-idf との比較を行った．日本語の要約評価用データセットである TSC2 のデータを利用し，重要文抽出と重要箇所検出の課題について提案手法と tf-idf でどちらが要約に貢献できるかを調査した.

静的モデルと動的モデルで利用するランカーの学習には NAIST テキストコーパス[11] 中の 287 記事に出現している 699 の文間ゼロ照応関係を利用した.

実験で利用する tf-idf の値は式 4 に従って計算する．idf の計算には日経新聞 1991 年から 2000 年までの記事を形態素解析した結果を利用した．tf-idf については，既存の重要文抽出の手法で利用されているように，文中に出現するすべての名詞の tf-idf の値の和をその文の重要度として重要文を抽出する.

$$\text{tf-idf}(w_{ij}) = \frac{w_{ij}}{\sum_k w_{ik}} \cdot \log \frac{|D|}{|d : d \ni w_{ij}|} \tag{4}$$

ここで，w_{ij} は文章 T_i に出現する語を表し，$|D|$ は総文章数，$|d : d \ni w_{ij}|$ は w_{ij} を含む文章の総数を表す.

Vol.2009-NL-193 No.9
2009/9/29

静的モデルや動的モデルは各文章に出現する語についての重要度であるため，tf と同様に idf の値で重み付けが可能である．そこで，式 1 や式 2 に示した静的モデルと動的モデルの重要度の値を idf の値で重み付けした結果についても比較対象とする．また，idf の重み付けの良さについて考察するため，tf のみの場合も比較対象に加える．

実験で利用する TSC2 の 180 記事は MeCab[⋆1] と CaboCha[⋆2] を用いて形態・構文解析され，3 節で示した教師有り手法で用いる素性はこの解析結果を利用して抽出した．また，動的モデルに関しては各文でキャッシュに保持する顕現性の高い語の上限を決める必要があるが，本実験ではこの上限値を設定せず，文章中の各文では毎回現在の文に出現している名詞句全体と前方文脈に出現しているすべての名詞句を対象にランキングを行った．

4.1 重要文抽出における語の重要度の比較

重要文抽出の課題では，3 節で示した談話の顕現性を捉える動的モデルや静的モデルから算出する語の重要度と tf-idf をもとに得られる重要度を比較することでどちらが重要文抽出に役立つかを調査した．この際，各要約率に関して文章ごとに抽出すべき文数が指定されているため，それに従い重要度の高い文から順に抽出し，各文章についての正答率の平均で評価を行う．まず，各重要度のみを利用して重要文を抽出した場合の実験結果を表 3 に示す．この実験では文章の最初から指定された文数の文を抽出する lead 法をベースラインとする．表 3 に示した結果より，要約率が高い場合，つまり文章の主題を優先的に抽出しなければならない場合には，tf-idf よりも静的モデルもしくは動的モデルを用いて抽出したほうが結果が良いことがわかる．これは，提案する手法が tf-idf では捉えることができない，文章には頻出していないが主題となっている重要語を捉えられているためだと考えられる．さらに動的モデルでは，重要語として出現した語が次の文でも重要であるかという tf-idf では考慮することができない特徴を捉えることが可能である．動的モデルでは，例えば，重要文中に出現している主題を表す語が，次の文では省略されて出現していないが継続してその語について記述されているかを見積ることができ，他のモデルと比較してより適切に重要文を選択することができたため，さらに良い結果を得たと考えられる．逆に要約率が低い場合には，tf-idf もしくは tf に基づく重要度を利用した場合の結果が良くなっている．これは，要約率が低い場合は文章の主題だけでなく関連する話題も補足的に抽出されるが，そのような主題とは直接的に関連しない箇所を提案するモデルでは捉えることができなかったためだと考えられる．一方，tf-idf の場合は主題となる文やその補足的な文に横断的に出現する語を頻度に基づいて重要語として認定し，その結果提案モデルを上回る結果を得たと考えられる．このように，要約率が高い場合と低い場合では捉えるべき特徴が異なるため，提案する重要語の尺度と tf-idf を適切に使い分けることで最終的な正答率が向上すると考えられる．そこで，次に述べる教師有り手法については tf-idf と静的モデルもしくは動的モデルの 2 つを素性として利用した場合についても評価を行った．

表 3 重要文抽出の実験結果（教師無し）

手法 \ 要約率	10%	30%	50%
lead 法	0.260	0.412	0.553
tf	0.279	0.428	**0.616**
tf-idf	0.277	**0.440**	0.609
静的モデル	0.258	0.383	0.570
静的モデル-idf	0.299	0.380	0.570
動的モデル	**0.328**	0.432	0.585
動的モデル-idf	0.316	0.432	0.590

教師有り手法の場合も正答率の平均で評価を行う．この実験でのベースラインモデルでは，表 2 に示した素性集合のうち，TFIDF，STATIC，DYNAMIC を利用せずに学習したランカーを用いて重要文を抽出する．このベースラインモデルに対し，TFIDF，STATIC，DYNAMIC の素性を加えた場合にどのように結果に影響するかを調べた．この結果を表 4 に示す．この結果より，どの重要度の情報を加えた場合でもベースラインモデルより良い結果を得ていることがわかる．また，要約率が低い場合には表 3 の結果と同様に tf-idf の情報が有効であることがわかる．一方，要約率が 10%の場合は静的モデルの情報が重要文抽出に貢献していることがわかる．これは，動的モデル場合はこのモデルが捉える局所的な談話状況の更新の情報は近似的には表 2 で導入した格助詞の情報で捉えることができるため，tf-idf と同程度の結果であるのに対し，静的モデルが文章中の名詞句全体を一度にランク付けして得られる大域的な順序の情報はその文章における語の重要度を大域的に決定している可能性があり，その情報が最終的な教師有り手法の結果に貢献したと考えられる．ただし，この実験では重要度の値を素性として利用しているだけであり，各モデルが出力する重要度の情報をうまく利用しているとはいえない．今後は，例えば Clarke ら[1] や富田ら[12] が提案しているように，要約に関する手がかりを整数計画問題に導入し，その制約の上で重要度を最大化する文を抽出するといった試みが考えられる．また，今回導入した静的モデルと動的モデルの重要度は各モデルが出力する順序の逆数というヒューリスティックなものであり，この値に

⋆1 http://mecab.sourceforge.net/
⋆2 http://sourceforge.net/projects/cabocha/

表 4 重要文抽出の実験結果（教師有り）

手法 \ 要約率	10%	30%	50%
ベースライン	0.320	0.434	0.604
+tf-idf	0.334	**0.463**	**0.626**
+静的モデル-idf	**0.341**	0.430	0.607
+動的モデル-idf	0.330	0.429	0.601
+tf-idf +静的モデル-idf	0.331	0.460	0.624
+tf-idf +動的モデル-idf	0.330	0.460	0.624

ついても重要文抽出の事例を訓練事例とすることで最適な値を見積ることが可能だと考えられるが，これらそれぞれについての詳しい調査については今後の課題としたい．

4.2 抜粋箇所の包含率の調査

重要箇所の抽出についても提案手法の重要度と tf-idf を比較する．TSC2 の評価データには，文の単位だけでなく文章の重要箇所が抜粋率 20%と 40%で抜粋されたデータも含まれている．例えば，図 5 の (a) の文章は重要箇所を抜粋率 20%で抜粋された結果，(a) の太字で書かれた箇所が正解として抜粋されている．つまり，重要箇所抽出の課題についてはこの太字で書かれた語句に高い重要度を付与することで，最終的に抜粋の精度に貢献できると考えられる．そこで，この実験では，それぞれの語の重要度に基づき N 語を抽出した場合に，どのくらい抜粋の中の語を選択できているか，また N 語のうちどのくらいの語が抜粋の中に含まれるか，という再現率と精度の関係を調べることによりどの重要度が重要箇所抜粋に役立つかを調査する．この実験では，tf-idf，静的モデルの出力を idf で重み付けした重要度，動的モデルの出力を idf で重み付けした重要度の 3 つについて比較を行う．

図 6 に各抜粋率の評価データについて抽出語数を変動させて描いた再現率-精度曲線を示す．図 6 のうち (a) が 20%の抜粋率，(b) が 40%の抜粋率に関する結果を表している．この結果より，どちらの抜粋率の場合も tf-idf では特に重要度の高い語は抜粋内に出現しているが，それ以外の表現を同定できないため，図 6 では再現率が上がるにつれて極端に精度が低下していることがわかる．これに対し，静的モデルもしくは動的モデルでは tf-idf の値が最も高い表現は検出できないが，格助詞などの手がかりをもとに tf-idf の値が低いが抜粋の中に含まれる語を適切に検出できるため，再現率が大きい点では tf-idf より精度が高いことがわかる．この結果からも tf-idf の上位の語と提案手法で検出した語を組み合わせることで重要箇所抜粋についても精度が良くなる可能性が示されており，この組み合わせについては今後検討したい．

最後に，tf-idf と動的モデルが上位にランク付けした語の具体例を図 7 に示す．ここで示す各尺度で抽出された結果はそれぞれ抜粋の中に含まれている表現である．この例では「発生」や「装置」などの語が tf-idf について上位にランク付けされているのに対し，文章中に出現している助詞「は」を伴う 3 つ「受け皿」という表現を動的モデルは上位にランク付けしており，各尺度が異なる手がかりをもとに相補的な関係で抜粋内の表現を選択していることがわかる．

(a) 原文章:

結核予防ワクチンであるＢＣＧに、日本人とタイ人に特徴的なエイズ・ウイルス（**ＨＩＶ**）**の遺伝子の一部を組み込んだエイズワクチンを**、国立予防衛生研究所と味の素中央研究所のグループが**開発**、マウス実験などで**免疫力を高める効果を確認した**。近く国内で初めて、サルを使った感染予防実験を開始する。アジアを中心に**広く途上国で使える可能性がある**。予研エイズ治療室の本多三男室長らは**ＨＩＶの「急所」が外被たんぱくのＶ３ループ部分らしいという**最近の米国の**研究成果を応用**。日本人感染者に共通するＶ３ループ部分のＨＩＶ遺伝子配列を決定し、タイ人感染者に特徴的なＨＩＶ遺伝子配列を使った組み換えＢＣＧも作製した。ワクチンでエイズ感染を防ぐには、ＨＩＶに感染した細胞を見つけて異物として排除するＴリンパ球と、ＨＩＶそのものを攻撃する抗体を増やさなければならない。マウスとモルモット各五匹で免疫効果を別々に実験したところ、マウス全例でＴリンパ球の活性が高まり、モルモットでは二匹で抗体が大量に増えたことを確認。予研グループは「有望な結果が得られた」と判断した。感染防止力を調べるサルの実験は、予研霊長類センター（茨城県つくば市）で一月から実施する予定だ。新ワクチンはウイルスそのものではないため、発病する危険はないとされ、主体となるＢＣＧも安全性が確立されている。新生児にも接種でき、**エイズ母子感染の防止に役立つという**。山崎修道・予研所長は「アジアを対象にしたワクチンを一日も早く実用化したい」と話している。

(b) 20%に圧縮された文章:

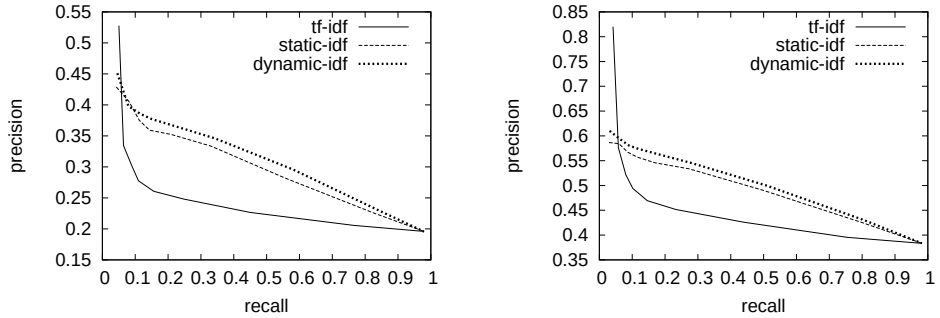
結核予防ワクチンである**ＢＣＧ**に、**ＨＩＶ**の**遺伝子**の**一部**を組み込んだ**エイズワクチン**を、**開発**、**免疫力**を高める**効果**を**確認**した。広く**途上国**で使える**可能性**がある。**ＨＩＶ**の「**急所**」が**外被たんぱく**の**Ｖ３ループ部分**らしいという**研究成果**を**応用**。**エイズ母子感染**の**防止**に役立つという。

図 5 重要箇所抽出の例

5. 関 連 研 究

今回導入した顕現性に基づく重要度と同様に，文脈に依存して重要度の値を求める手法が存在する．例えば，原ら[10] は「特定範囲の多くで同時に出現する語どうしは関連性が高い」「特定範囲に出現する語は，その範囲に出現する全語数が少ないほど重要性が高い」という 2 つの仮定に基づいて語の重要度を定義している．この手法では，文章が段落などの意味のある M 個の範囲 A_i（$1 \leq i \leq M$）に分割できる状況を想定する．ここで，文章における語 w の重要度 C_r は，式 5 で定義されるように，w が範囲 A_i に出現している場合にはその範囲に出現している語の総数 N_i の逆数をその範囲のスコアとし，全ての範囲におけるスコアの平均で定義される．

Vol.2009-NL-193 No.9
2009/9/29



(a) 抜粋率:20%　　(b) 抜粋率:40%

static-idf と dynamic-idf はそれぞれ静的モデルと動的モデルの結果を idf で重み付けした重要度を用いた結果を表す.

図 6　各モデルの重要箇所の包含率

東京・ＪＲ新宿駅のトイレ個室内のトイレットペーパーケースに時限式の青酸ガス**発生装置**が仕掛けられた事件で、**発生装置**の下に敷かれていた 受け皿 は、プラスチック板を接着剤で張り合わせた「手製皿」だったことが八日、警視庁捜査本部の調べで分かった。捜査本部は、犯行グループが**発生装置**を安定させるために、ケースのサイズに近いプラスチック板を入手し加工したものとみて、入手ルートを追跡している。調べでは、ケースは高さ八〇センチ、幅二〇センチ、奥行き一五センチ。受け皿 はケースの下側に付属するホルダーに入れたペーパーの上に置かれ、その上に**発生装置**が載っていた。受け皿 は、弁当箱のような形で、深さ約三センチ。プラスチック板を接着剤で張り合わせ、ケースの幅、奥行きに合うサイズに加工していたことが分かった。時限**装置**のモーターに接続されたカッターが**希**硫酸入りのビニール袋を破り、**希**硫酸とシアン化ナトリウムが化学反応を起こすと青酸ガスが**発生** する仕組みだった。しかし、時限**装置**はセットされた時間に作動して袋 は破れたものの、**希**硫酸が流れ過ぎたため、化学反応がうまくいかず、ごくわずかの青酸ガスしか**発生**しなかったらしい。また同じ日に見つかった営団地下鉄日比谷線茅場町駅の**発生装置**も同じ 仕組み。しかし、カッターが**希**硫酸入りの袋とほとんど接触せず、袋は破れなかった。「手製皿」を使わず、時限**装置**の固定が不十分だったためとみられる。

太字箇所が tf-idf の値が高い語を表し，下線部が動的モデルの出力する重要度が高い箇所を表す.

図 7　重要箇所抽出の結果の例

$$C_r(w) = \frac{1}{M}\sum_{i}^{M}\frac{\alpha_i}{N_i} \tag{5}$$

ここで，α_i は A_i に w が出現している場合には 1，それ以外には 0 をとる．つまり，この手法では特定範囲内にどのくらい密にある語が出現するか，またどのくらい範囲横断的にある語が出現しているかの両方を評価したものだと考えることができる．ただし，この尺度では idf に相当する情報は考慮されていないため，頻出する非自立名詞などが高い重要度を持つ可能性がある．

また，Clarke ら[1] が提案した語 w の重要度 $I(w)$ を式 6 に示す.

$$I(w) = \frac{l}{N}\cdot f_i \log\frac{F_a}{F_i} \tag{6}$$

ここで，f_i と F_i は出現文章中もしくはコーパス全体における w の出現頻度を表し，F_a はコーパス中のすべての語の出現頻度の総和を表す．また，l は語 w が出現する文 s の統語構造の中でどのくらい深い節に出現しているかを表し，N は s における節の深さの上限値を表す．つまり，ここで導入されている重要度 $I(w)$ とは，tf-idf に類似する重要度の値をその語が統語構造上深い位置に出現しているほど大きな値をとるように重み付けをしていることに相当する．彼らはこの重要度の情報を文圧縮の問題に利用しているが，必ずしも統語構造における深さが圧縮の際に残したい重要箇所と対応するとは限らない． 4.2 に示した結果から，本研究で提案した顕現性に基づく重要度の情報は tf-idf や式 6 のような単純な頻度情報に基づく重要度では捉えることができない省略された語についても重要度を求めることができるため，tf-idf の代わりに利用する，もしくは tf-idf と併用するなどが考えることができ，この重要度の情報を Clarke が提案する要約の枠組みに導入することで精度が向上することが期待できるが，詳細な評価実験については今後の課題としたい.

6. おわりに

本稿では，文献[3] で提案された談話の顕現性に基づく名詞句のランキング手法を概観し，その手法を語の重要度として利用する手法を提案した．この手法の有効性を調査するために，語の重要度の情報が解析時に重要な手がかりとなる自動要約の課題を例に，tf-idf のような既存の重要度の尺度と比較を行った．TSC2 で定義された重要文抽出と重要箇所抽出の 2 つの課題について評価を行い，それぞれの課題で提案手法の情報が tf-idf と相補的に利用可能であることを示した.

今後の課題としていくつかの方向性が考えられる．まず，今回提案した重要度の情報を要約以外の応用処理に利用することが考えられる．例えば，文章生成の課題，特に照応表現の生成については顕現性に基づく重要度が精度を向上させる見込みがある．また，Web のような大規模文書集合からあるクエリに関して情報抽出を行う場合にも顕現性の情報が利用可能だと考えられる．情報抽出の処理において，入力クエリに関してすべての文書から情報を抽出するのではなく，入力クエリの表現が文章中で顕現性が高くなっている場合のみから

情報抽出を行うことで，最終的な情報抽出の精度が向上する可能性がある．

また， 4.1 でも述べたように，今回提案した重要度では単純にランキングの逆数を重要度の値としており，明示的にどの語が重要であるかを同定することは行っていない．しかし，情報検索のような応用処理によってはランキングの情報よりも重要語はどれであるかを明示的に同定しているほうがより利用しやすいと考えられる．そのため，この顕現性の観点に基づく重要語であるかを判別するために，どの表現が重要でないかという情報が付与された自動要約のデータセットも訓練事例として利用し，文章中で重要語判別のための分類器を構築するという立場も考えられ，具体的にどう実現すべきかについては今後詳細を考えたい．

参 考 文 献

1) Clarke, J. and Lapata, M.: Global Inference for Sentence Compression An Integer Linear Programming Approach, *Journal of Artificial Intelligence Research*, Vol.31, pp.399–429 (2008).
2) Grosz, B.J., Joshi, A.K. and Weinstein, S.: Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics*, Vol.21, No.2, pp. 203–226 (1995).
3) Iida, R., Inui, K. and Matsumoto, Y.: Capturing Salience with a Trainable Cache Model for Zero-anaphora Resolution, *Processing of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP 2009*), pp.647–655 (2009).
4) Joachims, T.: Optimizing Search Engines Using Clickthrough Data, *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 133–142 (2002).
5) Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, Vol.24, No.5, pp.513–523 (1988).
6) Walker, M., Iida, M. and Cote, S.: Japanese discourse and the process of centering, *Computational Linguistics*, Vol.20, No.2, pp.193–233 (1994).
7) Walker, M., Joshi, A.K. and Prince, E.(eds.): *Centering Theory in Discourse*, Oxford Univ. Press (1997).
8) Walker, M.A.: Limited attention and discourse structure, *Computational Linguistics*, Vol.22, No.2, pp.255–264 (1996).
9) 奥村学，福島孝博：TSC 2（Text Summarization Challenge 2）の目指すもの，情報処理学会情報学 (基礎研究会報告) FI-63-5，pp.33–39 (2001).
10) 原正巳，中島浩之，木谷強：テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出，情報処理学会論文誌， Vol.38, No.2, pp.297–309 (1997).
11) 飯田龍，小町守，乾健太郎，松本裕治：NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション，情報処理学会研究報告 (自然言語処理研究会) NL-177-10, pp.71–78 (2007).
12) 富田紘平，高村大也，奥村学：重要文抽出と文圧縮を組み合わせた新たな抽出的要約手法，情報処理学会研究報告 (自然言語処理研究会) NL-189-3，pp.13–20 (2000).