2006－ＮＬ－174（13）
2006／7／28

濃縮還元型文要約モデルの検討

池田 論史，牧野 恵，山本 和英
長岡技術科学大学電気系 〒940-2188 新潟県長岡市上富岡町 1603-1
E-mail:{ikeda,makino,ykaz}@nlp.nagaokaut.ac.jp

現在、様々な文要約の研究が行われている。これらの研究は、原文より重要な部分を抜き出すことで文要約を行っている。そのため人間が行う要約のような原文にない表現を用いた要約文になることはない。我々は人間が要約を行うように、原文から必要な表現を抜き出し、その表現を用いて文を作成することで、より自然な要約文の生成を目指した。そこで、人間と同様に原文から要約に必要な表現を抽出し（濃縮）、その表現に機能語を補完することで文を生成する（還元）要約モデルの提案及び検討を行なった。要約に必要な表現の抽出にはSVMを用い、文の生成には原文情報を用いずに他のコーパスによる統計情報を用いて行なった。結果として80%の要約率で要約した際に可読性の評価で36%、意味の評価で45%の正解率を得た。

キーワード 自動要約，単語抽出，文生成

Examination of Select-and-Complement Summarization Model

Satoshi Ikeda , Megumi Makino , Kazuhide Yamamoto
Department of Electrical Engineering,Nagaoka University of Technology
1603-1 Kamitomioka,nagaoka-shi,Niigata,940-2188 Japan
E-mail:{ikeda,makino,ykaz}@nlp.nagaokaut.ac.jp

There has been many works on summarization of the sentence. Most of these them make a summary by extracting words from the original sentence. Therefore, the summary don't have expressions not in the original sentence. In contrast, there are expressions not in the original in manual summarization. Consequently, our aim here is to generate a summary in a natural way that a human does. That is, we propose a summarization method in which we extract terms necessary to make a summary from original sentence, and generate a sentence from those terms. We used SVM for extracting terms necessary to make summarization, and statistical information from general corpus to generate the sentence. The experimental result showed the performance of readability is 36% and meaning identification is 45% under the summary rate of 80%.

key words automatic summarization , extract term , generate sentence

1 はじめに

現在、様々な文要約の手法が研究されている。例えば、堀ら[4]は、単語重要度を最大にし、かつ日本語として自然な部分単語列の抽出を動的計画法によって行っている。田中ら[8]は、文書要約において重要語を決定し、その単語の必須格等を要約要素語として抽出して文としている。その文について不要部を削除することで要約を行なっている。これらは原文に存在する単語のみを用いて要約を行っている。しかし、人間が要約を行う際には、自立語の言い換えのみではなく、機能語の変更も行い可読性の高い要約を行っている。例1や例2は実際にWebのニュース記事と人手で要約した文の一部である。この例から人間は、自立語の言い換えのように機能語についても言い換える場合があることが分かる。

例 1) … イラクで行方不明 になっていた 米民間人ら …
→ … イラクで行方不明 の 米民間人ら …

例 2) … ダイエー について、支援を決定した場合でも、…
→ … ダイエー への 支援が決定した場合でも …

さらに、これらの例で言い換えられている「になっていた → の」や「について → への」は常に同じ意味で使われるわけではなく、前後の文脈から人間には同じ意味で使われていると判断が可能であり、常に言い換えできるわけではない。文を要約するということは、使用単語数を減らしているということである。そのために、文の構造が変化することや前後の単語とのつながりが変化し同じ機能語を使うことができない状況もある。例 3 はその例の 1 つである。

例 3) … は宇宙関連事業を統合し、2 つの新会社を設立 することで 基本合意した。
… は宇宙関連事業を統合し、2 つの新会社設立 で 基本合意した。

例 3 では「新会社を設立する」という文節を「新会社設立」という複合名詞で言い換えている。そのために文の構造が変わり、機能語「ことで」を「で」と言い換えることが必要となる。

以上のことから、現在の要約手法では要約できない、または良い要約ができない文が存在する。

そこで、人間が要約を行う際にどのような手順で行っているかを考える。人間が要約する方法はいくつか考えられる。その中には、現在の自動要約の手法で行われているように、原文の単語をそのまま用いて、必要の無い単語のみを省いていくことで要約する方法もある。しかし、先に述べたように、その手法では上手く要約できない場合もある。その場合には、要約に必要な単語をいくつか抜き出し、その単語群を用いて文を生成することによって要約を行うこともある。そこで我々は、自動要約でもこの方法を採用することが可能であると考え、原文から単語を抜き出し (濃縮)、その単語群より文を生成する (還元) ことで要約を行う手法を提案する。

関連研究として、単語から文を生成する研究を挙げる。内元ら [9]、肥塚ら [2] は { 国, 政策, 発足 } のような 3 つの単語からの文生成を行っている。これらの研究は、単語から文を生成することを目的としており、要約を目的としているわけではないので原文との意味的な一致のような評価をしていない。また、要約では任意の単語数での文生成が必要となるが、これらの研究では単語が多い場合においても文を正しく生成できるかについては言及されていない。そのうえ、後の単語を全て係り先の候補とし、全ての係り受け候補について計算しているので単語数の増加とともに処理時間が指数的に増加することが予想される。

我々は順序付きの単語群からの文の生成を試みた [5]。これは 2 単語間に入りやすい単語をコーパスから学習することにより行なった。しかし文全体のつながりを考えずに生成を行なったので、良い結果が得られてはいない。

文を生成するという問題はその他にも様々な問題に応用できる。廣嶋ら [3] は文章から必要な内容語及び機能語を抜き出し、それらを連ねることで文章のヘッドラインを作成している。他に文生成が利用可能な状況としては失語症患者との意志の疎通や第 2 言語での作文支援等が考えられる。これらは両者とも単語は出てくるが、その単語を用いて文の作成を行えない状況である。またテキストマイニングで取り出された単語より日本語の再生成を行うことも検討されている [6, 7]。

2 濃縮還元型文要約モデル

本稿で提案する手法は、原文より単語群を抽出する単語抽出部と、単語抽出部で抽出した単語群より文を生成する文生成部の 2 つの処理部からなる。提案手法の処理の流れを図 1 に示す。それぞれの処理部について以下の節で述べる。

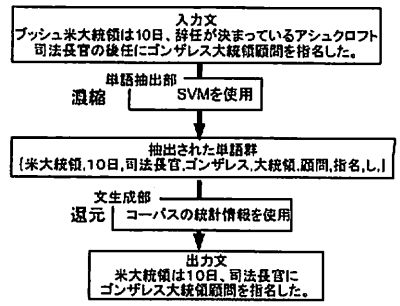

図 1: 提案手法の処理の流れ

2.1 単語抽出部

単語抽出部では原文から要約に必要な単語群を抽出する。

2.1.1 要約に必要な単語

まず、要約に必要な単語について考える。人間が文を読むときに最低限必要な情報は内容語である。これは、人間が日本語を速読しようと試みると機能語部分を読み飛ばすことから分かる。このことより人間は内容語があれば機能語を補完して文意を把握することが可能であると考える。これは、日本語は内容語があれば文を再構築することが可能であることを示している。また、形容詞や副詞は要約文を作成する際に省かれることが多いので、今回はこれらの品詞を用いないこととした。これらのことから単語抽出部で抽出する単語は名詞、動詞のみとした。ここで扱う単語の単位は形態素解析[1]を行い、1 形態素を 1 単語としたものである。接尾辞[2]や接頭詞はそれぞれ前後の形態素とまとめて 1 単語とする。また固有表現、数の表現については一般化処理を行っている。固有表現は CaboCha(5) の NE タグを用いて同定している。

2.1.2 単語群の抽出

単語群の抽出は、ある単語が要約文に必要か否かの 2 値の分類であると考え SVM(Support Vector Machine) を用いて行なった。ここで SVM の識別平面からの距離を考慮することで任意の数の単語群を抽出することが可能であると考えた。よって本手法では任意の要約率で要約が可能である。SVM のカーネルとして線形カーネルを使用し、素性として、判定する単語と原文でのその前後 2 単語、合計 5 単語の表層形及び品詞[3] と判定する単語の *idf* 及び *tf* の全 12 種類を用いた。ここで *idf* の計算には日本経済新聞 2000 年度版 (2) を用い、1 記事を 1 ドキュメントとした。

2.2 文生成部

文生成部では 2.1 節で抽出された単語群から要約文を生成する。文の生成は単語群に機能語を補うことで行う。また、抽出された単語群は原文の出現順と同じ順序で用いることとした。本来は要約する際に、単語の並び替えも考慮する必要がある。これは人手で要約を行うと原文とは違う単語の並びで要約文を作成することがあるためである。しかし今回は問題の簡単化のために単語の並び替えを行わない。今後、単語の順序を入れ換えて要約することも視野にいれているので、文を生成する際に原文の構文構造は用いることができない。なぜならば、人手で要約を行なった結果、要約文中の単語が原文とは違う順序で使われる際には原文と要約文で構文構造が異なる。

機能語を補完する箇所は単語の間全てとする。また、補完の際には補完しないという可能性も考慮する。ε を補完するとい

(1)形態素解析には ChaSen(4) を用いた。
(2)接尾辞は形態素解析において第 2 階層が接尾の形態素としている。
(3)品詞は ChaSen で解析された第 3 階層までを用いている。

う処理で補完しないという選択肢を用意する。例4のような補完によって文が生成される場合、「安全」と「検査」の間にはεを補完している。「旅客」と「安全」の間や「安全」と「検査」の間のように機能語を補完する可能性のある箇所を補完箇所と呼ぶ。

例4){ 旅客, 安全, 検査, 一部, 簡素化, する, 方向, 検討, する }
→ 旅客の安全検査を一部簡素化する方向で検討する。

文の生成は以下の手順で行う。

1. 補完候補の出力
2. 機能語の決定

以下の節でこれらについて詳しく述べる。

2.2.1 補完候補の出力

ここでは補完箇所に補完する補完候補の出力について述べる。補完候補の出力は補完箇所の前後の単語を用いてコーパスを検索することで行う。例えば前後の単語が「安全」と「検査」の時は、「安全+(機能語)+検査」となる全ての機能語を補完候補として出力する。ここでの機能語は例5の「からの」のように連続する機能語もまとめて1つの機能語として扱う。また「安全検査」のように間に機能語が何も無い状態でコーパスに出現した場合はεを補完候補とする。つまり、補完しないということを補完候補に入れる。

例5) 政府 からの 要請を受ける。

この処理で補完候補が見つからない場合、つまりこの例で「安全+(機能語)+検査」という表現がコーパスに存在しない場合、「安全+(機能語)」でコーパスを検索する。この処理でも「安全+(名詞)」のような表現がコーパスに存在する時は、εを補完候補とする。

また、2.1.1節で一般化された単語についてはコーパス内でも一般化を行い、一般化された状態で検索する。これらの処理で補完候補が出力されない場合、補完は行わない。

2.2.2 機能語の決定

2.2.1節で出力した補完候補から補完する機能語を決定する。補完語の決定は、ラベル付与問題をもとに行なった。図2はラベル付与問題の例である。ラベル付与問題はこのように観測$\mathbf{x}$が与えられたときに、確率が最大となるラベル列$\mathbf{y}$を見つける問題である。この問題は、式(1)で求めることが出来る。

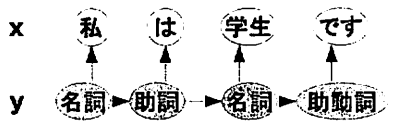

図2: ラベル付与問題の例

$$\mathbf{y} = \underset{\mathbf{y}\in\Sigma_y^T}{\operatorname{argmax}} \prod_{t=1}^{T} P(x_t|y_t)P(y_t|y_{t-1}) \qquad (1)$$

T : ラベルの数
$P(x_t|y_t)$: 出力確率
$P(y_t|y_{t-1})$: 遷移確率

補完する機能語を決定するときにはこのラベル付与問題のラベルを補完候補として考える。この例を図3に示す。図3のように観測$\mathbf{x}$を2.1節で抽出した要約に必要な単語列とし、付与するラベルy_iは単語x_iの後に補完される機能語となる。

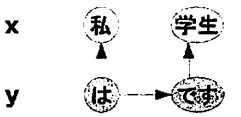

図3: 機能語付与問題の例

また、機能語の付与はラベル付与問題と違い、後方のつながりも機能語を補完する際に考慮する必要が考えられる。そこで後方の連接確率もスコアとして導入する。そこで式(2)で付与する機能語の推定を行う。

$$\mathbf{y} = \underset{\mathbf{y}\in\Sigma_y^T}{\operatorname{argmax}} \prod_{t=1}^{T} P(y_t|x_t)P(x_{t+1}|y_t)P(y_t|y_{t-1}) \qquad (2)$$

T : ラベルの数
$P(y_t|x_t)$: 前方の連接確率
$P(x_{t+1}|y_t)$: 後方の連接確率
$P(y_t|y_{t-1})$: 遷移確率

3 評価実験

本手法の妥当性を測るために、実際に実験を行なった。SVMの学習データとして原文と要約文の対(以下要約対と呼ぶ)が必要となる。そこで、要約文としてNikkei-gooの行っているサービスである日経ニュースメール(1)を利用した。これは新幹線の電光掲示板のニュースで用いられている記事をメールサービスとして配信しているものでありニュース文が非常に短い表現になっている[4]。日経ニュースメールの例を例6に示す。また対応する原文として日本経済新聞社のWebページであるNIKKEI NET(3)のニュース記事を用いた。これらの記事は同じ新聞社のニュース記事なので記事のタイトルが全く同じ記事が存在する。また新聞記事の特徴として、重要な情報が最初にあるということがいえる。そこで、これらのタイトルが同じ記事の1文目を要約対として利用する。ここで集めた要約対は3316対である。このうち無作為に3300対取り出し33分割交差検定を行うことでこれらの要約対を学習データ及びテストデータとして用いる。また、補完候補の出力及び各種連接確率を求めるために日本経済新聞2000年度版(2)を用いた。SVM学習にはTinySVM(6)を用いた。

例6)
タイトル:米、移民の融和策検討へ新組織
本文:米大統領は合法移民を社会に溶け込ませる方策を
検討する新組織設置の大統領令を発表。
英語教育拡充などの政策を立案へ。

3.1 要約率

本手法では、2.1節で抽出する単語群の数で要約率を可変にできると考える。しかし要約率を指定することはできない。そこで、要約率を指定するために、単語を抽出する割合を変更したときの要約率の変移について調べた。ここで要約率は文字単位で求めた。その結果を図4に示す。

この結果に最小二乗法による線形近似を行うと式(3)となる。この式を用いることで本手法は要約率を任意に設定することが可能になる。

$$y = 1.08x + 3.50 \qquad (3)$$

y : 要約率 [%]
x : 抽出単語割合 [%]

[4]我々はこのような文を新幹線要約と呼び、その特徴について観察を行っている[10]。

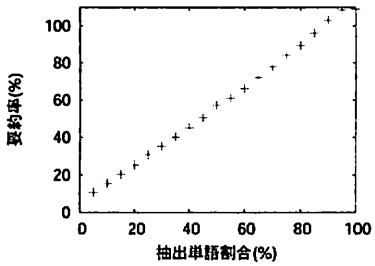

図 4: 単語抽出率における要約率の変移

3.2 人手による評価

実際に本手法要約率が約 80%になるように抽出単語割合を 70%で要約を行なった。要約した 100 文について、3 人の被験者が独立に評価を行なった。評価は生成された要約文が日本語として正しいか (評価 1) と、原文と比較して意味が保持されているか (評価 2) という 2 点について行なった。その結果を表 1 に示す。

表 1: 正解の人数を変えたときの正解率

正解とした評価者数	≥1	≥2	=3
可読性の評価 (評価 1)	80%	36%	10%
意味同一性の評価 (評価 2)	78%	45%	13%

表 1 より、人による揺れがかなり大きいことが分かる。
ここで、正解例を例 7 に示す。例は原文、要約文の順で示す。

例 7) 英下院で 15 日、イングランドとウェールズ での猟犬を
使ったキツネ狩りを禁止する法案が賛成多数で
可決した。
→ 猟犬を使ったキツネ狩りを禁止する法案を可決した。

3.3 先行研究との比較

本手法の有用性の確認のために先行研究との比較を行う。比較する先行研究は堀らの手法 [4] である。評価尺度として ROUGE スコアを用いた。この結果を表 2 に示す。人手での評価との相関を見るために先行研究のアルゴリズムで要約率を 80%に指定して要約を行なった文を 3.2 節で評価を行なった 3 人の被験者が評価した結果とともに記す。ここで、ROUGE は要約率によって値が変化するので要約率の一致を調べておく。本手法で要約した 100 文の要約率を計算し平均をとると 77%であった。堀らのアルゴリズムは実装の際に簡易化した。このために、要約率を単語単位でしか指定できない。比較のために要約した際には、単語単位で要約率 80%とした。本手法の要約率は文字単位で出しているので、堀らの手法で要約した文についても文字単位での要約率を求めた。要約した文 100 文の平均を計算すると 75%となった。このことから本手法と堀らの手法の要約率がほぼ等しいことが確認できた。

表 2: 先行研究との比較

	ROUGE-1	ROUGE-2	人手の評価	
			可読性	意味
本手法	0.62	0.43	36%	45%
先行研究	0.71	0.52	53%	21%

表 2 から人手の評価では可読性においては先行研究の方が優れており、意味の評価では本手法の方が優れていることが分かる。ROUGE スコアでは n-gram を使用しているために、可読性の評価では人手の評価と同様に先行研究の方が良い結果であるが、意味同一性の評価をみると人手の評価とは逆になっている。

4 考察

4.1 単語抽出の精度

2.1 節で提案した単語抽出の精度について考える。単語の抽出数を変更したときの単語抽出の精度を求めると図 5 となった。ここで示す結果は要約対 3300 対を用いた 33 分割交差検定によって得られた精度の平均値である。これより抽出精度は原文の名詞、動詞の数の 60%を抽出したときがもっとも良い精度になっていることが分かる。また、図 5 にはベースラインとして、抽出単語の決定に $tf \cdot idf$ を用いて重要度が高い単語から抽出した場合の F 値も描いている。この結果を見ると本手法は $tf \cdot idf$ より高い精度で単語の抽出が行われていることが分かる。

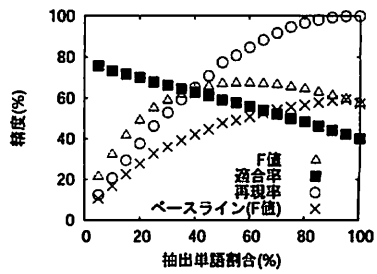

図 5: 提案手法での単語抽出割合における抽出精度の変移

4.2 単語抽出について

実際に要約を行なった、単語抽出数 70%において抽出した単語群について観察した。

抽出した単語群から人手で文の生成を行なった。文生成は被験者 2 人 {A,B} が独立に行い、被験者には単語群のみを提示し原文は提示していない。また、文生成を行う際には機能語または ε を補完するという本手法と同じ手法で行なった。生成に使用した単語群は 3.2 節で人手で要約の評価を行なった 100 文から抽出したものである。人手で生成した 100 文を別の被験者 1 人が原文と見比べて意味の保持について評価を行なった。人手で生成を行っているため、可読性については間違っていないと考える。被験者が文を生成できない単語群も存在した。これらの集計を表 3 に示す。

表 3: 2 人の被験者による文生成の結果

被験者	A	B
文生成可能	59 文	65 文
意味同一性の評価	40 文	38 文

表 3 より、被験者 A、B ともに約 4 割の文が生成不可能であるとしている。生成できた文についても約 4 割が原文と比較して意味が変わっている。このことから、本手法で抽出した単語群から文生成を行なうことは、人手でも困難であることが分かる。このことから単語抽出部を改善する必要性が大きいことが分かる。

被験者両方が文生成できた文は 56 文で、そのうち 26 文は意味同一性の評価も正解であった。また被験者両方が文生成不可能とした文は 20 文であった。被験者 1 人のみ文生成ができた文は 24 文である。以上の結果から、文生成可能であるかは人によって大きな揺れがあることが分かる。

ここで、被験者 2 人とも意味が保持できなかった文について

観察する。それぞれの例は、原文、抽出した単語群、人手の生成の順に示してある。人手による生成は1人分のみ例示する。

例8)日銀が2日発表した11月の資金供給残高（マネタリーベース、月中平均）は、前年同月比4.9％増の109兆4035億円だった。
→{ 発表, し,11月, 資金, 供給, 残高, マネタリーベース, 月中, 平均, 前年, 同月比,9％増 }
→ 発表した11月の資金供給残高はマネタリーベースの月中平均が前年同月比の9％増。

例8は必要な単語が抽出されていない例である。人手で生成した文を見ると「発表した」から文がはじまり、「誰が発表した」かという単語、つまりこの例では「日銀」が要約に必要な単語として抽出されていない。このために、文を生成した際に意味情報が少なくなるうえに、日本語としても違和感のある文になっている。また、同じような生成の失敗に「～によると」という文において、「よる」だけ抽出されてしまいどこからの情報であるという情報が抜け落ちた例も見られた。

例9) 米マクドナルドは14日、1株当たりの年間配当金を15セント（38％）増やし、同55セントにすると発表した。
→{ 米マクドナルド,14日, 1株, 当たり, 年間, 配当金, 増やし,する, 発表, し }
→ 米マクドナルドは14日1株当たりの年間の配当金を増やしすると発表した。

例9では下線部「増やしする」の「増やし」を原形の「増やす」にし、その直後の「する」を削除して「増やす」とすることにより正しい文にすることが可能である。本手法では必ず出現形で使うこととしているが、このように原形に直した方が良い場合もある。また、単語の順序を入れ換えることを考えると、原形のみではなく活用形も考える必要がある。

また、「する」はサ変名詞とまとめてサ変動詞として扱った方が良い可能性がある。しかし、本稿でサ変名詞と「する」をまとめずに別々に扱った理由として例10のような要約が存在するからである。サ変名詞の「買収」と「する」をまとめて「買収する」として扱うと例10のような要約は出来ない。

例9は「する」の前にサ変名詞がない例である。しかし、サ変名詞があるときと同様に前の「55セントに」がないと意味が取れなくなってしまう。このように「する」について特殊な処理をする必要があると考える。

例10)A社は14日、同業のB社を買収することで合意した。
→A社は14日、B社の買収で合意した。

被験者が2人とも生成することが出来なかった単語群の例について原文と一緒に示す。例はそれぞれ原文、抽出した単語群、人手でできる範囲で生成した文の順である。また人手でできる範囲で生成した例には1人分のみ例示する。

例11) ソニーと米投資会社による名門映画会社メトロ・ゴールドウィン・メイヤー（MGM）の買収計画が難航していることが28日明らかになった。
→{ 米, 投資, 会社, 映画, 会社, MGM, 買収, 計画, 難航, し, いる, こと,なっ}
→ 米投資会社は映画会社MGMの買収計画が難航していること なっ

例11は「明らかになった」の「なっ」の部分のみ抽出したので、文を生成できなくなった例である。この「なる」や「する」のように他の語とまとめて意味をなす語については単体で抽出された場合には生成できないことがある。

また被験者1人のみが生成できなかった単語群には例12, 例13のような例があった。

例12) エジプト政府は認可制だった新政党の設立を届け出制にし、選挙開票の監督を内務省から独立した選挙管理委員会に移す法改正案を人民議会に提出した。
→{ 政府, 認可制, 新政党, 設立, 届け出制,し, 選挙, 監督, 独立, 選挙管理委員会, 移す, 法, 改正案, 人民議会, 提出 }
→ 政府, 認可制, 新政党の設立届け出制, し, 選挙の監督の独立選挙管理委員会は移す法の改正案を人民議会に提出

例13) IBMなど米情報技術（IT）大手が顧客企業から人事部門の業務を受託する事業を拡大する。
→{ 米, 情報, 技術, 大手, 企業, 人事, 部門, 業務, 受託, する, 事業, 拡大 }
→ 米情報技術の大手企業の人事部門が業務受託, する, 事業拡大

例12は、「する」の変化形の「し」の部分で生成が出来なくなっている。これは生成するときに「届け出制」と「し」の間に「に」を補完し「届出制にし」とすれば、文が生成されるのだが、人手では生成しにくいということである。これは人間は一度部分的に補完を行い、文全体を見たときにその補完が間違っている場合でも、一度行なった補完と違う補完を行いにくいということである。また、例13においても同様に実際は生成が可能である。

また、人手で生成すると単語数が多い方が得られる情報が多いために、原文の意味が保持されやすい傾向が見られた。

4.3 文生成についての考察

2.2節で提案した文生成の精度について考える。文生成部のみの精度をみるために、要約文の全ての名詞、動詞を用いて文生成を行なった。その評価を3人の被験者がそれぞれ独立に評価しその結果を表4に示す。このときの評価指標は2つで可読性の評価と意味の評価である。また、以前に我々が行なった文生成[5]の精度も示す。これより、被験者の過半数が正解とした場合を正解とすると、可読性の評価及び意味の評価ともに約60%の精度で文が生成されていることが分かる。

表4: 文生成部のみの精度

正解とした評価者数		≥1	≥2	=3
可読性の評価 (評価1)	本手法	93%	61%	33%
	先行研究	77%	53%	33%
意味同一性の評価 (評価2)	本手法	83%	58%	20%
	先行研究	46%	23%	15%

ここで、先行研究[5]との相違点について述べる。先行研究では機能語を補完するか否かの判断及び、助詞「の」の補完の判断を最初にSVMを用いて行っていた。本手法ではこれを他の機能語と同じように決定している。また補完する機能語を求めるためのスコアが先行研究では単語連接スコアとして3-gramを使用し単語遷移スコアとして5-gramを使用している。また、文全体を見るのではなく2単語間のみを見て判断している。この部分について本手法では確率をともに2-gramで求め、文全体を考慮して補完する機能語を決定している。先行研究よりも少ない情報で文の生成を行なっているにもかかわらず表4から先行研究よりも優れていることが分かる。

文生成部のみで不正解だった例について観察する。

例 14) 中国石油大手の中国海洋石油が米石油大手ユノカルの買収を検討、買収総額は１３０億ドル
→{ 中国, 石油, 大手, 中国, 海洋, 石油, 米, 石油, 大手, ユノカル, 買収, 検討, 買収, 総額, １３０億ドル }
→ 中国に対し石油大手の中国の海洋石油の米の石油大手のユノカル買収を検討買収総額は１３０億ドルである。

例 14 をみると、助詞「の」が連続で多く補完されているのが分かる。これは新聞記事においても助詞「の」が 2 つなら連続して入る場合がある。今回のスコアの計算において機能語の遷移確率は前の機能語から次の機能語と機能語を 2 つしか考慮していない。このために、この例のように助詞「の」が多数連続してもスコアが小さくなることがない。そのために助詞「の」が連続したものと考える。これに対応するためには、スコアを大局的なスコアで計算する必要がある。

4.4 文生成時における単語数について

文を生成する際に、単語群の単語数が多い場合の方が生成の精度が悪いという傾向がみられた。これは日本語としての可読性の評価、原文と比較した意味の評価の双方にいえることである。

この原因として 4.3 節でも述べたように遷移確率が前の機能語と現在の機能語の 2 つだけしか考慮してないこと、加えて、連接確率も 2-gram のみ考慮しているため、文を大局的にみるスコアがないことがあげられる。提案手法ではラベル付与問題とすることで文全体を考慮するような式を導入している。しかし、大局的なスコアの導入をしていない。このために局所的には正解でも大局的に不正解になることが多くあった。これは局所的に正しければスコアが小さくなるということがないためである。そのために、この問題が起こると考える。この解決のためには、スコア計算を行う際のスコアにも大局的にみたスコアを導入する必要がある。

実際に大局的にみて文生成ができていないことを示すものとして、人手で文生成を行なったときには本手法で行なったときとは逆の傾向がみられる。つまり人手で文生成を行って不正解となるときは、単語数が少ないときの方が多いということである。人間は文全体を考えて文生成を行っている。単語数が多いことは情報が多いということなので、単語数が多い方が正しく文を生成しやすいということである。このことからも本手法では大局的に文の生成が行えていないことが分かる。

5 まとめ

原文から単語群を抜き出して、その単語群から文を生成することで要約文を作成するモデルを提案した。単語群の抽出には SVM を用い、文の生成は他のコーパスから様々な確率を計算しスコアを求めることで行なった。その結果、要約率 80% で要約を行なった際の精度は可読性の評価で 36%、意味の評価で 45%であった。

単語抽出部では人間でも文が生成できない単語群が多数抽出されたので、今後人手で抽出を行う際にどのような基準で単語を選択しているかといった調査や手法の改善が必要になることが分かった。文生成部においては、先行研究 [5] よりは文の生成精度が向上しているが「の」が連続で補完される等の問題もある。そこで、より大局的な単位でみた生成が必要となる。

本手法では、単語抽出部と文生成部が完全に独立している。しかし人間は文を生成するときのことも考えて単語を選択しているので、独立した処理とするのではなく、文を生成しやすい単語の抽出方法といった処理を考えることで精度の向上が可能ではないかと考える。

謝辞

本研究の一部は, 科学研究費補助金 基盤 (a)「円滑な情報伝達を支援する言語規格と言語変換技術」課題番号 16200009 によって実施した.

使用した言語資源及びツール

(1) 日経ニュースメール,NIKKEI-goo, http://nikkeimail.goo.ne.jp/

(2) 日本経済新聞全記事データベース 2000 年度版, 日本経済新聞社.

(3) NIKKEI NET, 日本経済新聞社, http://www.nikkei.co.jp/

(4) 形態素解析器 "ChaSen",Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

(5) 係り受け解析器 "CaboCha",Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/

(6) SVM 学習ツール "TinySVM",Ver.0.09, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/TinySVM/

参考文献

[1] Chin-Yew Lin, "Looking for a Few Good Metrics : ROUGE and its Evaluation, "Proc. of the 4th NTCIR Workshops, pp.1–8,Tokyo,Japan,2004.

[2] 肥塚真輔, 岡本紘幸, 斎藤博昭, "サポートベクタマシンを用いたキーワードからのテキスト生成, " 言語処理学会第 10 回年次大会発表論文集,pp.409–412,2004.

[3] 廣嶋伸章, 長谷川隆明, 奥雅博, "Web ページのヘッドライン生成のための統計的要約, " 言語処理学会論文誌「自然言語処理」, Vol.12 No.6,pp.113–128,2005.

[4] 堀智織, 古井貞熙, "単語抽出による音声要約生成法とその評価, " 電子情報通信学会論文誌, Vol.J85-D-II, pp.200–209,2002.

[5] 池田諭史, 沢井康孝, 山本和英, "文生成のための機能語の補完, " 言語処理学会第 12 回年次大会発表論文集,pp.208–211,2006.

[6] 森永聡, "– テキストマイニング技術の動向 – Key semantics マイニング、動的トピック分析による Knowledge Organization, " 日本行動計量学会第 33 回大会発表論文抄録集,pp.370–373,2005.

[7] 坂尾要祐, 池田崇博, 佐藤研治, 赤峯享, "特徴的な意味内容を抽出する木構造マイニングのための日本語処理手法, " 言語処理学会第 10 回年次大会発表論文集,pp.73–76,2005.

[8] 田中信彰, 面来道彦, 野口貴, 矢後友和, 韓東力, 原田実, "意味解析を踏まえた自動要約システム ABISYS, " 言語処理学会論文誌「自然言語処理」,Vol.12 No.1,pp.143–164,2006.

[9] 内元清貴, 関根聡, 井佐原均, "キーワードからのテキスト生成, " 言語処理学会第 8 回年次大会発表論文集,pp.375–378,2002.

[10] 山本和英, 池田諭史, 大橋一輝, "新幹線要約のための文末の整形, " 言語処理学会論文誌「自然言語処理」,Vol.12 No.6,pp.85–112,2005.