論　文

重要文抽出によるWebページ要約のためのHTMLテキスト分割

砂山　渡[†]　井山　晃洋[††]　谷内田正彦[††]

HTML Texts Segmentation for Web Page Summarization by Using a Key Sentences Extraction Method

Wataru SUNAYAMA[†], Akihiro IYAMA[††], and Masahiko YACHIDA[††]

あらまし　検索エンジンにおいて検索結果として表示される情報は，目的の情報を素早く獲得する上で重要なものである．特に，検索結果の各 Web ページの要約文は，各 Web ページの内容を知る上で重要であるとともに，ユーザが入力した検索語が各 Web ページ内でどのように使われているか，すなわち検索語と各 Web ページとのかかわりを知るために有効である．しかし，従来の検索エンジンにおける検索結果の要約文は，Web ページの冒頭部分のテキストが抜き出されて検索語が含まれていなかったり，検索語を含んでいても文の途中で切れていて文として不完全で，文脈や Web ページの内容を把握できないという問題点がある．そのため文を単位とした要約の出力が望まれるが，HTML テキストにおいては，句点を含まない，文以外の記述が数多く含まれているため，そのまま文を単位とした重要文抽出システムによって要約文を提供することは困難である．そこで本論文では，各 Web ページのソースを文に相当する意味の切れ目において分割する HTML テキスト分割システムを提案する．また，本システムにより生成されるテキストが，Web ページの要約生成に有効に働くことを実験により検証した．

キーワード　テキストセグメンテーション，重要文抽出，Web ページ要約，検索エンジン

1. まえがき

インターネットが普及するにつれ，様々な情報がネット上にあふれ，存在している．しかしその情報量は多く，情報の獲得と理解に多大な労力を要するようになってきた．そのため，Web ページを要約する技術が求められている．Web 文書に対する要約には以下の二つが挙げられる．

1. 各 Web ページの内容をまとめて内容理解を助ける，各 Web ページに固有の要約
2. 検索結果の概要として，各 Web ページの取捨を検索語に関連して判断できる動的な要約

本論文においては，この後者を題材として取り扱う．この後者の要約においては，検索語が Web ページ中でどのような位置付けで現れているかを明確にすることが重要であり，そのためには，検索語を含みかつ前後の意味のつながりがある部分を抽出することが適当と考えられる．特に，検索エンジンの出力に用いるための要約は，1 画面中にできるだけ多くの検索結果を表示したいというインタフェースの制約から，およそ 3 文程度（120 字から 150 字）の量にまとめる必要があり，この限られた量の中で，できるだけ意味のつながりのあるテキストを出力することには意味がある．

そこで本論文では，HTML テキストを意味のつながりのある一つの単位（セグメントと呼ぶ）に分割するシステムを構築する．セグメントはおよそ文と同程度の意味情報を含むように生成する．これによって，既存の重要文抽出システムを用いて，セグメントを単位として重要セグメントを抽出し，意味のつながりがある Web ページ要約を得ることが可能となる．

以下の **2.** で関連研究について述べ，**3.** で HTML テキストをセグメントに分割する HTML テキスト分割システムについて述べる．**4.** で提案システムが適切な分割を行っているか実験的に検証し，**5.** で HTML テキストを分割した結果得られるセグメント集合から，

[†] 広島市立大学情報科学部，広島市
Faculty of Information Sciences, Hiroshima City University, Hiroshima-shi, 731–3194 Japan
[††] 大阪大学大学院基礎工学研究科，豊中市
Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560–8531 Japan

重要セグメントを抽出する実験を行う．最後に，**6.** で本論文を締めくくる．

2. 研究背景

検索エンジンが提示する Web ページの要約文には，ページの冒頭部分を提示したり [1]，検索語の前後一定文字を提示するもの [2] がある．これらは意味の区切りを明確に扱っておらず，意味の離れた内容のものが無理に連結されていたり，文の一部分のみが表示されていたりするため，検索語にかかわる Web ページの内容を知る要約としては不十分となる．

HTML テキストを要約する先行研究 [3], [4] では，確率モデルや，構文解析と TFIDF 法 [5] を組み合わせた手法により Web ページの要約を行っている．しかし，HTML テキスト中の文のみを対象として，短い単語のリンクや項目が並べられている個所を対象とせずに，文章固有の要約を生成している．そのため，検索エンジンの検索結果として提示する場合，検索語を含む動的な要約が望まれ，検索語がリンク集に含まれている場合などに適切な要約を得られない．

そこで本研究では，文以外のテキスト部分においても文と同程度の長さの意味の区切りを設け，従来の重要文抽出システムが扱える入力形式に HTML テキストを変換する．HTML テキストを文分割することの重要性は既に指摘 [6] されているが，それに対処した上で重要文を抽出するシステムは提案されていない．また，文単位ではなく段落やフォーム，項目といったユニットと呼ばれる単位に，Web ページを分割する研究 [7] や，タグ情報を用いて Web ページ中の各テキストに意味付けを行う研究 [8] もある．しかし，分割されてできるテキストのユニットが大きな段落になるなど，検索エンジンが出力する要約に用いるにはテキストの量が多いという問題点がある．また，各ユニットの大きさが不均一であるため，重要文抽出システムへの入力にも適さない．

従来の重要文抽出による文章要約システムにおいては，各文の評価値を文に含まれる単語の評価値の総和で与えたり，文に含まれる単語間の関連を調べる [9], [10] ことによって実現している．また，ユーザが入力した検索語を多く含む文を抽出する要約 [11] も存在する．しかし，これらのシステムでは文の区切りである句点が文章に含まれていることが前提になっているため，自由な形式で書かれている HTML テキストに対しては適用することができない．文間で TFIDF 法を適用して他の文で使われていない特徴的な単語を抽出したり，文内の単語間の共起関係から，文章の主張を表す単語を取り出す KeyGraph [12] を用いるためにも，文章中における意味の区切りである句点が必要であり，この句点の果たす役割は大きい．

そこで本論文では，通常の文章のように必ずしも句点が含まれていない HTML テキストから，検索結果に用いられる大きさのテキストを生成するために，文の大きさを基本として意味の区切りである句点を挿入し，HTML テキストを分割するシステムを提案する．

3. システム構成

現在，Web ページは HTML 言語を用いて記述されており，この HTML テキスト（表 1）は実際にブラウザに表示される「テキスト」（図 1）と，文字属性やレイアウト情報が書かれた文字列である「タグ」（表 1

表 1　Web ページの HTML テキスト
Table 1　Sample HTML source text.

```
<HTML><HEAD><TITLE>Web ページ要約を考えるページ </TITLE></HEAD>
<BODY BGCOLOR="#aaddff"><HR><CENTER>
<IMG SRC="photo.jpg" align="right" width="13%">
<H1>Web ページ要約を考えるページ </H1>
</CENTER>
<H3><FONT color="green"> 今週の言葉（または念頭）</FONT></H3>
誰もやらずに誰がやる、俺がやらずに誰がやる
<TABLE border="1" cellspacing="10"
cellpadding="10" width="100%">
<TD width=200>
<FONT size=4 color="red"> 研究関連 </FONT>
<BR>
<A HREF="search.html"> 研究内容 </A><BR>
<A HREF="lib.html"> 辞書 </A><BR>
<A HREF="link.html"> リンク集 </A><BR>
<HR>
<FONT size=4 color="red"> 演劇関係 </FONT>
<BR>
<A HREF="workshop.html"> 演劇ワークショップ
</A> <BR>
<A HREF="TR.html"> 所属劇団「TR」</A><BR>
<A HREF="search_play.html"> 研究と演劇（文章理解）
</A><BR>
</TD><TD>
修士課程の <A HREF="～xxxxx/"> 私 </A> は研究を頑張っています。<BR>
テーマは、「雑多な Web ページの要約」です。<BR>
従来の自動要約システムは文章を要約するものでした。
<BR> けれども、Web ページは文章ばかりではないため、
<BR> 文章以外の部分も文章として見立てて、<BR>
文章要約システムで要約することを試みます。<BR>
</TD></TABLE>
</BODY></HTML>
```

図 1　入力 Web ページのサンプル
Fig. 1　Sample Web page for input.

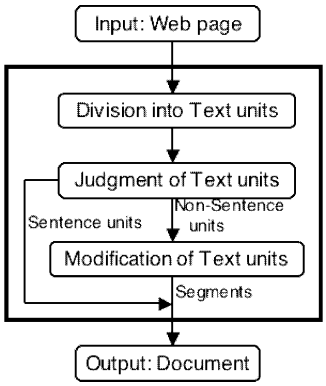

図 2　HTML テキスト分割システム
Fig. 2　HTML texts segmentation system.

の，図 1 に出力されているテキスト以外）とで構成される．本章では，タグと句点を用い，HTML テキスト中の「テキスト」部分を通常の文に準えてセグメントに分割し，「テキスト」を「セグメント集合」に変換する HTML テキスト分割システムについて述べる．

HTML タグは，本来 Web ページの文書構造や意味構造を反映するためのものであるが，改行のために，本来段落を作るための P タグが使われるなど，必ずしも正しい使われ方をしているとは限らない．そのため，各タグ本来の意味は利用せず，ブロックレベル要素と呼ばれる明確な意味の区切りがあると期待されるタグの集合を，意味の区切りの指標としてのみ利用する．この意味の区切りさえも正しくない場合もあるが，検索エンジンの任意の結果を閲覧する場合において，そのような Web ページの割合は少ないと考えられる．

本システムが出力するセグメント集合に含まれる各セグメントは次の三つのいずれかとなる．

1. 句読点を用いて書かれた文
2. 句読点はないが人間は文とみなすテキスト
3. 意味の切れ目において分割されたテキストで，句読点を用いて書かれた文と同程度の意味情報を含む長さのもの

3.1 HTML テキスト分割アルゴリズム

図 2 に，HTML テキスト分割システムの全体構成を示す．システムは，Web ページ（HTML で書かれたソーステキスト）を入力とし，この Web ページを明示的な意味の区切りが存在する個所で分割する．ここで分割される一つのテキストをテキストユニットと呼び，各テキストユニットを，**3.1.3** で後述する「文ユニット」と「非文ユニット」とに分けた後，「非文ユニット」を「文ユニット」と同程度の自立語を含むまとまりに整形して出力とする．以下で，システムの各部の詳細について述べる．

表 2　HTML ソースの情報
Table 2　Information of HTML source text.

Source	<HTML>	<HEAD>	<TITLE>	Web
Info.	Tag	Tag	Tag	Noun

ページ	要約	を	考える	ページ
Noun	Noun	Postpositional Pariticle	Noun	Noun

</TITLE>	</HEAD>	...
Tag	Tag	...

3.1.1　HTML ソーステキストの入力

入力となる Web ページは，表 2 のように形態素解析 Chasen [13] によって単語ごとに分割した上で，単語ごとにタグ若しくは品詞の情報を与える．

3.1.2　テキストユニットへの分割

HTML テキストを明示的な意味の区切りが存在する個所で区切った，部分テキスト（テキストユニットと呼ぶ）へと分割する．この分割には，HTML のタグ情報と句点記号「。」とを用いて，以下の 3 通りの個所を区切ることで行う．

1. ブロックレベル要素が存在する個所
2. 複数のリンクタグが存在する個所
3. 句点が存在する個所

まず 1 に関して，HTML のタグ（要素）にはブロックレベル要素とインライン要素とがあり，ブロックレベル要素と呼ばれるタグ（表 3）はブラウザでの表示の際に改行を伴い，段落を構成できる要素として定義されている．そこでまずテキストユニットへ分割する最初の手順として，ブロックレベル要素が存在する個所で HTML テキストを分割する．

次に 2 に関して，分割された各テキスト内において，

表 3　ブロックレベル要素
Table 3　Block-Level elements.

ADDRESS, BLOCKQUOTE, CENTER, DIR, DIV DL, FIELDSET, FORM, H1, H2, H3, H4, H5, H6 HR, ISINDEX, MENU, NOFRAMES, NOSCRIPT, OL P, PRE, TABLE, UL, DD, DT, FRAMESET, LI TBODY, TD, TFOOT, TH, THEAD, TR

複数の A 要素（リンクタグ）が連続して存在する場合，リンク集とみなして，一つひとつのリンクを切り離すために，各 A 要素が存在する個所でテキストを分割する．これは，リンクは単独で一つの意味をもっていると考えられるため，複数のリンクには別々の意味が存在すると解釈しそのリンクの間に，意味の区切りを与えている．

最後に 3 に関して，文章中の意味の区切りとなる句点「。」によってテキストを分割する．これら三つの方法で分割して得られるテキストユニットに対して，次節で述べる文判定を行う．

3. 1. 3　テキストユニットの文判定

HTML テキストを分割して得られた各テキストユニットが，文であるかどうかを判定する．そこでこの判定条件を得るための予備実験を行った．予備実験に用いたデータは，検索エンジン Excite [14] を用いて，平仮名「あ」「い」で検索した各検索結果上位の Web ページを対象とし，それらに含まれる句点を含む文 1000 件と，それ以外のテキストをブロックレベル要素（表 3）のタグで区切って得られる文以外のサンプル 1000 件とした．ただし，ここで用意した文のサンプルとは，主語と述語が助詞や助動詞でつながっており，かつ句点を含むという基準で選んだ．また，Web ページ中の各単語の品詞情報の獲得には，形態素解析 Chasen [13] を用いた．

図 3 に，文及び文以外の各サンプルが含む，自立語数の累積割合を示す．すなわちグラフの縦軸は，横軸の数以下の自立語を含むサンプル数の全サンプル数に対する割合を表す．この図をもとに，例えば自立語数が 7 個以上のテキストを「文ユニット」と判定すると，図中横軸の値が 6 の個所における縦軸の値より，94%の文以外を除くことができるが，同時に 21%の文も除かれることが分かる．そこで，複数の条件設定によってこの精度を上げるために，文としての性質にかかわる条件として次の三つを取り上げる．

1. 意味のある単語（自立語）を一定数以上含む
2. 付属語（助詞，助動詞）を全単語数に対して一定以上の割合で含む
3. 付属語を自立語の数に対して一定以上の割合で含む

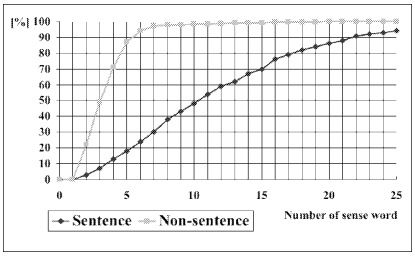

図 3　文と文以外が含む自立語数の累積割合
Fig. 3　Accumulation rates of the number of sense words in sentences and non-sentences.

表 4　パラメータの探索範囲
Table 4　Search range of parameters.

Condition	Range	Partition
C1	1–10	1
C2	0.4–0.7	0.03
C3	0.2–0.3	0.01
C4	0.2–0.3	0.01
C5	0–0.1	0.01

これらの性質を具体的にしたテキストの文判定の条件を以下のように実験結果より設定した．

C1.　自立語の数が 7 個以上
C2.　自立語数の全単語数に対する割合が 0.64 以下
C3.　付属語数の自立語数に対する割合が 0.22 以上
C4.　助詞の数の自立語数に対する割合が 0.26 以上
C5.　助動詞数の自立語数に対する割合が 0.06 以上

上記五つの条件（Condition）に含まれる数値は，各条件について表 4 で表される範囲（Range）の，刻み幅（Partition）ごとの文及び文以外の分布と，五つの条件のうちで，どれをいくつ以上満たす必要があるかを網羅的に調べた上での，最適な値として決定した．すなわち，上記の値による五つの条件のうち三つ以上の条件を満たすものを「文ユニット」，それ以外を「非文ユニット」と判定したときに，最も高い精度（正解判定数 ÷ サンプル数）として 95.2%（文を「文ユニット」と判定する精度：97.5%，文以外を「非文ユニット」と判定する精度：92.8%，の平均値）が得られた．

この条件によると，たとえ句点を含んでいたとしても条件を満たさなければ「非文ユニット」と判定され，逆に句点を含まなくても，条件を満たしていれば「文ユニット」と判定される．この句点の有無によらない文判定によって，句読点を含まない Web ページに対

表 5　文判定結果の例
Table 5　Example results of sentence judgement.

Judgement	Text Unit
Non-Sent.	Web ページ要約を考えるページ
Non-Sent.	今週の言葉または念頭
Sentence	誰もやらずに誰がやる、俺がやらずに誰がやる
Non-Sent.	研究関連
Non-Sent.	研究内容
Non-Sent.	辞書
Non-Sent.	リンク集
Non-Sent.	演劇関係
Non-Sent.	演劇ワークショップ
Non-Sent.	所属劇団「TR」
Non-Sent.	研究と演劇（文章理解）
Sentence	修士課程の私は、研究を頑張っています。
Sentence	テーマは、「雑多な Web ページの要約」です。
Sentence	従来の自動要約システムは文章を要約するものでした。
Sentence	けれども、Web ページは文章ばかりではないため、文章以外の部分も文章として見立てて、文章要約システムで要約することを試みます。

応するとともに，1 単語 + 句点のようなテキストを文以外と判定できる．

表 5 に図 1 のテキストに対して，文判定を行った結果を示す．リンクである「研究内容」「辞書」「リンク集」などに加えて，ページタイトルの「Web ページ要約を考えるページ」や「今週の言葉（または念頭）」といった記述が「非文ユニット（Non-Sent.）」と判定されている．また，図 1 右下の文章部分と「誰もやらずに誰がやる、俺がやらずに誰がやる」といった，句点を含まない文らしい記述を「文ユニット（Sentence）」と判定できている．

3.1.4　テキストユニットの整形

前項の文判定により「非文ユニット」と判定されたテキストユニットを，適当な長さのセグメントへと整形する．整形にはテキストユニットの分割と結合の二つがある．長いテキストユニットを分割する理由は，長さを限定しない場合に際限なく長くなるテキストユニットが存在するためであり，長すぎるテキストは要約に用いることができないからである．また，短いテキストユニットを結合する理由は，短いテキストユニットは文に含まれる単語を評価する重要文抽出システムによって，重要文として抽出されにくく，仮に選ばれても前後のつながりが分からなくては，Web ページ中での検索語の出現理由を知るという要約の目的を果たさないためである．そこで，図 3 より 90%以上の文が含む自立語の数は 22 個未満であることから，条件 C1 と合わせて 7 個から 21 個の自立語を含むようにテキストユニットを整形する．ただし，ブロックレベル要素のタグが存在する個所では，段落の区切りがあり意味が大きく離れるとして必ず区切るため，ブロックレベル要素のタグの間に 7 個未満の自立語しか存在しない場合は，自立語の数が 7 個未満のセグメントができることがある．

表 6　HTML テキスト分割結果の例
Table 6　Example of HTML text segmentation.

Web ページ要約を考えるページ。
今週の言葉または念頭。
誰もやらずに誰がやる、俺がやらずに誰がやる。
研究関連—研究内容—辞書—リンク集。
演劇関係—演劇ワークショップ—所属劇団「TR」—研究と演劇（文章理解）。
修士課程の私は，研究を頑張っています。
テーマは、「雑多な Web ページの要約」です。
従来の自動要約システムは文章を要約するものでした。
けれども、Web ページは文章ばかりではないため、文章以外の部分も文章として見立てて、文章要約システムで要約することを試みます。

テキストユニットの結合は隣接するテキストユニット間にブロックレベル要素が存在しない場合，すなわち，もともと句点やリンクタグで区切られたテキストユニットを再結合することによって行う．これによって，段落を構成するブロックレベル要素を超えて意味のかけ離れたユニット間の結合を避け，近くに存在する意味が近いテキストや，並列関係にあるテキストを結合させる．

また，結合されてできたセグメントが 22 個以上の自立語を含む場合，セグメント内の自立語の数が 22 個未満になるまで，ブロック中のリンクや改行などのタグが存在する場所で二つに分割する．このとき，分割候補のリンクや改行などのタグが複数存在する場合，それらに優先順位を付ける明確な基準がないため，情報が片側に偏らないように自立語の数が最も均等に別れる位置で分割する．ここで各セグメントの区切りを明確にするために，セグメントの終わりに句点がない場合には句点「。」を挿入する．この挿入は，重要文抽出システムが句点によって一文の終わりを判断できるようにするためである．表 6 に表 5 のテキストユニットのセグメントへの整形結果の例を示す．リンク集のテキストユニットがブロックレベル要素のタグを境に結合されている（結合は「—」で表している）．このように，整形されたセグメントの集合を HTML テキスト分割システムの出力として出力する．

4. HTML テキスト分割結果の評価

本システムによって生成されるセグメントが，どの程度意味の切れ目において分割されているかを確かめる実験を行った．実験に用いた Web ページは，検索エンジン Google [2] に「あ」を検索語として与えたときの検索結果上位 20 件である．この 20 ページを本システムへの入力として与えたところ，合計 2251 個のテキストユニットに分割された後，1098 個のテキストユニットが結合，11 個のテキストユニットが分割され，最終的に 1164 個のセグメントが出力された．

実験においてはまず，連続する 6 個所の，セグメントの切れ目またはテキストユニットの結合個所（**3.1.4** のテキストユニットの整形処理により結合された個所）を各 Web ページの一部としてランダムに抽出して，各個所の前後のテキストを実験提示用データとして用意した．そこで 20 人の大学生及び大学院生に，提示したセグメントの切れ目が意味の切れ目として適切か否か，セグメント内のテキストユニットが結合された部分においては，意味の異なるものが無理に結合されていないかを，Web ページを参照しながら「適切（OK）」と「不適切（NG）」の 2 値で回答してもらった．また，テキストユニットが分割された全 11 個所についても同様に，分割された個所が意味の切れ目として適切か否かを判断してもらった．表 7 に実験結果を示す．ただし，表中の Rate は全回答に対する「適切（OK）」の割合を表し，Std. は Rate の被験者 20 人に対する標準偏差を表している．

まず，各セグメントが意味の切れ目において分割がなされているかを評価した結果（Segments）は，適切と判断された割合が 82.3%で標準偏差も 10.0%と安定して高い値が得られている．この結果を **3.1.2** で述べた HTML テキストの分割方法の違いでまとめたものを，表 8 に示す．最も多かった分割方法はブロックレベル要素のタグがある個所での分割（Block-level）で，その存在する割合（Share）は 7 割を超え，分割の適切さも 90%と高い値となった．このことから，ブロックレベル要素を用いた分割は，意味の区切りを作る上で大きな役割を果たしていることが確認できる．次に多かったのは約 2 割を占める句点（Period）であるが，その適切さは 6 割程度にとどまっている．これは，句点が必ずしも意味の切れ目とはならないことを示しており，句点のみでの分割では明確な意味の区切りを作るのには不十分であることが分かる．また，連続するリンクタグを用いた分割（Link）は，句点とほぼ同様の効果を上げていることが確認できる．これら三つの分割方法の組合せにより，ブロックレベル要素で大きく意味を区切り，句点とリンクタグによってより細かく意味を区切ることで，文の長さに近いセグメントを意味の区切りにおいて分割して生成できている．

表 7 意味によるセグメントの評価
Table 7 Evaluation of segments by meaning.

	OK	NG	TOTAL	Rate (%)	Std. (%)
Segments	757	163	920	82.3	10.0
Combination	940	540	1480	63.5	18.2
Division	169	51	220	76.8	25.2

表 8 セグメントの断点評価
Table 8 Break-points evaluation of segments.

	OK	NG	TOTAL	Rate (%)	Share (%)
Period	113	67	180	62.8	19.6
Link	50	30	80	62.5	8.7
Block-level	594	66	660	90.0	71.7
TOTAL	757	163	920	82.3	100.0

同様にテキストユニットの分割（Division）も，その絶対数が少ないものの 76.8%という値が得られており，意味の切れ目での分割がおよそ実現されている．

テキストユニットの結合（Combination）に関しては，適切な割合が 63.5%であるという結果が得られた．これは，短いテキストユニットに前後の情報を増すための結合であるため，意味のつながりに関して不十分な点が残るものの，なお 6 割以上の意味のつながりが確認されるため，本結合によるセグメント単位での重要文抽出への効果が期待できる．

5. HTML テキスト分割システムと Web ページ要約

HTML テキスト分割システムの出力であるセグメント集合が，重要文抽出システムへの入力として，有効であるか否かを確かめる実験を行った．評価は 5 種類の後述のテキストを重要文抽出システムに与え，出力される要約文を比較することで行った．用いた重要文抽出システムは展望台システム [15] であり，検索語を観点とした重要文を抽出する(注1)ため，Web ページ検索における要約文の改善を目的とする本システムに合っている．

しかし，今回は要約システムの評価ではなく要約シ

(注1)：展望台システムは，検索語と複数の文で共起しやすい単語とに評価値を与え，文に含まれる単語の評価値の合計が文の評価値となり，評価値の高い文を重要文として出力する．

ステムへの入力を与える HTML テキスト分割システムの評価が目的であるため，重要文の絶対的評価は行わず，要約生成に適切な入力テキストの比較評価を行う．すなわち，重要文抽出システムは他のシステムとの置換が可能であり，既存若しくは将来現れる重要文抽出システムと本システムとを併用した場合の，本システムの効果を実験により検証する．

具体的には，Web ページから次の 5 種類のテキストを生成し，それらのテキストを重要文抽出システムに入力して 5 種類の重要文を得ることで行う．今回用いた重要文抽出システムは入力文章中の句点で区切られる各テキストを 1 文として認識し，この実験では最重要文のみを抽出させた(注2)．

1. HTML ソーステキスト（Html_text）：Web ページの HTML ソースからタグを除去したテキスト
2. テキストユニット（Text_unit）：ソーステキストをテキストユニットへ分割し，各ユニットの終端に句点を挿入したテキスト
3. 文ユニット（Sentence）：テキストユニットの文判定によって「文ユニット」と判定されたテキストのみを取り出し，各ユニットの終端に句点を挿入したテキスト
4. 非文ユニット（Non_sentence）：テキストユニットの文判定によって「非文ユニット」と判定されたテキストのみを取り出し，各ユニットの終端に句点を挿入したテキスト
5. セグメント（The System）：HTML テキスト分割システムの出力として生成されたセグメント集合

5.1 Web ページ要約例

図 1 の Web ページから生成された前節の 5 種類のテキストから，「演劇」という単語を要約の観点として，最重要文を抽出した結果を表 9 に示す．

もとの Web ページ中では「演劇」という単語はリンク中に現れており，1. のソーステキストから抽出された重要文では，リンクの前後にある文章とリンクがすべて連結されて出力されている．そのため，ユーザは出力された重要文の中から再び「演劇」という話題に関連する部分を特定して意味を理解する必要があり，ユーザの負担が大きくなる．2. のテキストユニットからの重要文では，「演劇」という単語を含んでいるものの，短すぎて前後のつながりが不明で Web ページを要約できているとはいいがたい．3. の文ユニットを扱った場合には，「演劇」という単語が出現しないため，「演劇」という単語に関連する単語の評価値をもとに重要文が抽出されているが，検索語そのものを含まないため，内容を十分に把握することができない．4. の非文ユニットの出力も 2. の出力と同様に短すぎて意味が掴めない．5. の HTML テキスト分割システムによる出力においては，「演劇」を含む前後のリンク部分のみが連結して出力されており，ユーザは Web ページ中の演劇に関連した部分を過不足なく知ることができる．

表 9　重要文抽出結果の例
Table 9　Example of topic sentences extraction.

Text	Topic Sentence
1.	Web ページ要約を考えるページ今週の言葉（または念頭）誰もやらずに誰がやる、俺がやらずに誰がやる研究関連研究内容辞書リンク集演劇関係演劇ワークショップ所属劇団「TR」研究と演劇（文章理解）修士課程の私は、研究を頑張っています。
2.	演劇関係。
3.	けれども、Web ページは文章ばかりではないため、文章以外の部分も文章として見立てて、文章要約システムで要約することを試みます。
4.	研究と演劇（文章理解）。
5.	演劇関係—演劇ワークショップ—所属劇団「TR」—研究と演劇（文章理解）。

5.2 要約比較実験

本実験によってテキスト分割方法の違いによる要約の違いを検証する．すなわち，本論文で提案するのは要約システムではなく HTML テキスト分割システムであり，本システムの出力が他のソーステキストに比べて要約生成に適切であることを示す．

実験に用いた Web ページは検索エンジン Google [2] を用いて，「ワールドカップ」「インターネット」「知能」「ロボット」の 4 語をそれぞれ検索語として与えた際に得られた検索結果上位の Web ページから，上で述べた 5 種類のテキストを用いた要約がすべて同じとなる場合を除いた各 10 件の合計 40 件の Web ページである．実験は男女 20 名の大学生及び大学院生に対して行い，実験インタフェースとして，図 4 のように，上部フレーム内に問題を選択するためのリンクを提示し，中央フレーム内に要約対象となる Web ページを示し，下部フレーム内に 5 種類の重要文を並列かつランダムに提示するものを用意した．被験者には，各 Web ページについて「検索語に関連して Web ページの内容をよく表す」順に順位をつけてもらい，出力される重要文が同じ若しくは同程度とユーザが判断し

(注2)：検索結果の概要はおよそ 120〜150 字であり，比較に用いられるテキストには長い文を含むものがあるため，長すぎる出力とならないよう最重要文を対象とした．

図 4　実験に用いた Web ページと重要文
Fig. 4　Page and topic sentences for experiment.

表 10　順位付けの実験結果
Table 10　Ranking results of experiments.

Text	1st	2nd	3rd	4th	5th	Total Rank
The System	79	66	32	14	9	408
Text_unit	58	59	43	32	8	473
Sentence	57	36	23	45	39	573
Non_sentence	45	39	44	35	37	580
Html_text	52	28	26	52	42	604

表 11　一対比較の結果（勝数-敗数）
Table 11　Comparison results with pairs. (win-lose)

	Text_unit	Sent.	Non_sent.	Html
The System	84-36	102-55	120-52	121-67
Text_unit	——	91-62	67-52	106-79
Sentence	62-91	——	92-98	83-93
Non_sentence	52-67	98-92	——	95-95
Html_text	79-106	93-83	95-95	——

た場合には同じ順位としてもらった．実験時間は無制限で，一つの検索語に対する 10 の Web ページについて回答してもらった．

被験者による合計 200 回答の順位の内訳を表 10 に示す．表中の Total Rank は，順位と各回答数の積を，各システムについて加算して合計した値を表す．また一対比較によって，5 種類のテキストから各々二つずつの順位を比較して順位の良い方を勝ちとし，その数を数えた結果を表 11 に示す(注3)．

結果として，HTML テキスト分割システムの出力（The System）の順位が最も良く，一対比較においても他の手法すべてに対して高い勝率をおさめた．また，引き分けの内容もそのほとんどが出力が同じことによっていた．このことから，意味の切れ目によって区切られ整形されたテキストが，Web ページの内容を必要十分に表す要約を得る上で，他のテキストに比べて適切であったといえる．

次に良い評価が得られたのがテキストユニット（Text_unit）である．この要約では，リンク部分がすべて分割されてしまうため 1 文が短くなる傾向にあり，検索語のみが一つの重要文とされてしまうなど，要約文としての情報量が足りなくなることがあった．すなわち，検索語の前後の単語が検索語が出現した文脈を知る上で必要であったことが分かる．

これら二つのテキストに比べて，文ユニットや非文ユニットの評価が大きく下がっている．この理由は，例えば文ユニットから重要文を抽出する際に，非文ユニット中にのみ検索語が含まれている場合，抽出された重要文と検索語とのかかわりが弱くなったためと考えられる．特に文ユニットは非文ユニットと HTML ソースとの一対比較で負けており，従来の重要文抽出システムによって Web ページ中の文章のみから得られる要約では，検索結果を表すための十分な要約が得ることが難しいことが分かる．

また最も評価が低かった HTML ソースは，リンクなど句点が含まれていない部分が連続する場合に，後続の文の終端である句点までが 1 文となってしまうために，1 文の内容が冗長になることが多かった．意味のつながりがないテキストを連結することは誤解や勘違いを招きやすく，検索語出現の文脈を把握する上で大きな妨げとなる．また，検索語にかかわる話題以外の出力は不要な情報であり，必要な情報を簡潔に表すという要約の目的と相反する出力になりかねない．

6. むすび

本論文では，Web ページを重要文抽出システムへの入力とするために，HTML テキストをセグメントに分割する HTML テキスト分割システムを提案した．今後の課題としては，現在はすべてのブロックレベル要素を同等に扱っているが，箇条書き等は連続するリンクタグに近い意味をもつなど，各タグの意味に応じた整形手法の改良や，対象とする Web テキストの構成（文章が多い，リンクが多い，画像が多いなど）に応じたシステム構築などが挙げられる．

HTML テキストは現在の情報化社会において最も有効な情報源であり，本システムのように重要文やキーワードを抽出するための形式を整えることは，今

(注3)：引き分けの数は全回答数 200 から勝数と敗数を引いた数である．

後の世の中の発展に役立てられるものと筆者らは考える．また実験結果から，検索における適切な要約文のもつべき性質として以下の点が挙げられる．

1. 検索語を含んでいる．
2. 検索語出現の文脈を知るための，検索語の前後の単語が必要である．
3. 検索語の前後を意味の切れ目において区切る．

HTML テキスト分割システムが出力するテキストはこれらの条件を満たす要約を与えるのに有効であり，今後システムを検索エンジンに実装して実用化を目指したい．

文　献

[1] (URL) http://www.goo.ne.jp/
[2] (URL) http://www.google.co.jp/
[3] A.L. Berger and V.O. Mittal, “OCELOT: A system for summarizing web pages,” Proc. 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.144–151, 2000.
[4] 清田陽司，黒橋禎夫，“WWW テキストの自動要約と KWIC インデックスの作成,” 情処学研報，NL137, pp.31–38, 2000.
[5] G. Salton and C. Buckey, “Term-weighting approaches in automatic text retrieval,” in Readings in Information Retrieval, pp.323–328, Morgan Kaufmann Publishers, San Francisco, CA, USA, 1997.
[6] 渡辺日出雄，“Web 文書に対する言語処理の問題点と言語処理を援助するタグセットについて,” 情処学研報，NL127, pp.95–100, 1998.
[7] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke, “Seeing the whole in parts: Text summarization for web browsing on handheld devices,” Proc. Tenth International World-Wide Web Conference, pp.652–662, 2001.
[8] S. Chakrabarti, “Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction,” Proc. Tenth International World-Wide Web Conference, pp.211–220, 2001.
[9] 亀田雅之，“段落間および文間関連度を利用した段落シフト法に基づく重要文抽出,” 情処学研報，NL121, pp.119–126, 1997.
[10] R. Barilay and M. Elhadad, “Using lexical chains for text summarization,” in Advances in Automatic Text Summarization, pp.1–12, The MIT Press, London, 1999.
[11] A. Tombros and M. Sanderson, “Advantages of query biased summaries in information retrieval,” Proc. 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.2–10, 1998.
[12] 大澤幸生，ネルスベンソン，谷内田正彦，“KeyGraph：単語共起グラフの分割・統合によるキーワード抽出,” 信学論（D-I），vol.J82-D-I, no.2, pp.391–400, Feb. 1999.
[13] 松本裕治，北内　啓，山下達雄，平野義隆，松田　寛，浅原正幸，日本語形態素解析システム「茶筌」version 2.0 使用説明書第二版，NAIST-IS-TR99012, 1999.
(URL) http://cl.aist-nara.ac.jp/lab/nlt/chasen/
[14] (URL) http://www.excite.co.jp/
[15] 砂山　渡，谷内田正彦，“観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装,” 人工知能誌，vol.17, no.1, pp.14–22, 2002.

（平成 15 年 11 月 4 日受付，16 年 8 月 14 日再受付）

砂山　渡

1995 阪大・基礎工・制御卒．1997 同大大学院博士前期課程了．1999 同大学院博士後期課程中退．同年同大学大学院助手．2003 より広島市立大学情報科学部助教授，現在に至る．博士（工学）．人間の創造活動を支援する研究に興味をもつ．

井山　晃洋

2002 阪大・基礎工・システム卒．2004 同大大学院博士前期課程了．現在，TIS（株）勤務．

谷内田正彦　（正員）

1971 阪大大学院工学研究科修士課程了．同年同大基礎工学部助手，同助教授，教授を経て 1997 より同大大学院基礎工学研究科教授，現在に至る．1967〜1968 デンマーク原子力研究所留学．1972〜1973 米イリノイ大学にて Research Associate．1980〜1981 西独ハンブルグ大学 Research Fellow．1982 米ミネソタ大学 CDC Professor．コンピュータビジョン，画像処理，人工知能，移動ロボットなどの研究を行っている．著書「ロボットビジョン」（昭晃堂），「コンピュータビジョン」（丸善，編著）など．情報処理学会，ロボット学会等各会員．工博．