2004－NL－159　(6)
2004／1／13

複数 Web ページの要約による用語説明の自動生成

藤井 敦[†,†††]　渡邉 まり子[††]　石川 徹也[†]

[†] 筑波大学 図書館情報学系
〒 305-8550 つくば市春日 1-2
[††] 図書館情報大学
〒 305-8550 つくば市春日 1-2
[†††] 科学技術振興事業団 CREST
E-mail: fujii@slis.tsukuba.ac.jp

筆者らが構築した検索サイト「Cyclone」は，Web から言葉や事柄に関する良質な説明段落を抽出することで，入力キーワードに関する事典的な情報を効率的に取得することができる．現状では複数の Web ページから個別に抽出された説明段落の一覧が提示される．しかし，説明段落の間には関連性がないため，キーワードに関して十分な説明を取得するためには，複数の段落を読む必要があり，その結果，同じような内容を何度も読まなければならない場合がある．本研究では，複数文書要約を応用して，複数の説明段落から過不足ない単一の説明情報を自動生成する手法を提案する．あるキーワードに関する複数の段落を「定義」や「目的」といった説明の観点に基づいて解析し，観点ごとに代表的な説明文を選択し統合することで最終的な説明情報を構築する．また，評価実験によって提案手法の有効性を示す．

Automatic Generation of Term Descriptions by Web-based Multi-Document Summarization

Atsushi Fujii[†,†††], Mariko Watanabe[††], Tetsuya Ishikawa[†]

[†]Institute of Library and Information Science
University of Tsukuba
1-2 Kasuga Tsukuba, 305-8568, Japan
[††]University of Library and Information Science
1-2 Kasuga Tsukuba, 305-8568, Japan
[†††]CREST, Japan Science and Technology Corporation
E-mail: fujii@slis.tsukuba.ac.jp

We developed a Web search system called “Cyclone”, which extracts high quality term descriptions and helps users to obtain encyclopedic knowledge efficiently. In the current implementation, multiple paragraph-style descriptions extracted from different Web pages are presented in response to a user keyword. However, to obtain sufficient information for a single keyword, a user usually has to read the similar or redundant content included in multiple paragraphs. To solve this problem, in this paper we propose a multi-document summarization method that uses multiple paragraphs to generate a single description for a keyword. Our method analyzes source description paragraphs with respect to viewpoints of explanation (e.g., definitions and purposes) and generates a summary by integrating the representative sentences for each viewpoint. We also show the effectiveness of our method by means of experiments.

1 はじめに

World Wide Web 上の検索エンジンを用いて様々な調べ物をすることが日常的な知的情報活動になっている．Web を国語辞典や百科事典のように使って知らない言葉や事柄について調べることは，そのような活動の主な例である．既存の辞典や事典には新語や専門用語が収録されていないことが多いのに対して，Web には新しい情報や専門性が高い情報が多く流通しているためである．

Web が流行りはじめた当初に比べれば検索エンジンの性能は向上し，目的の情報が簡単に見つかることも多くなった．しかし，検索要求によっては依然として何を入力すればいいのか分からない場合や，膨大な検索結果から欲しい情報をどうやって選択すればよいか分からない場合がある．また，Web には統制がないため，誤字，誤解，虚偽といった低品質の情報を排除する必要がある．

上記の問題を解決するために，筆者らは Web を事典的に利用することを目的とした検索サイト「Cyclone」[2, 3, 13, 14, 15] を構築し，継続的に評価実験や問題点の改善を行っている．Cyclone には，見出し語とその説明情報を Web ページ群から抽出してコンテンツを自動構築する機能と，コンテンツを利用するための検索機能がある．

図 1 は，新型肝炎「SARS」を入力した場合の検索結果例である．画面の下半分には，複数の Web ページから個別に抽出（抜粋）された説明段落が，抽出元のページタイトルと一緒に 3 件提示されている．説明段落は専門分野に基づいて分類され，さらに説明としての尤度に基づいて順位付けされている．また，見出し語を入力するボックスの下には，説明を絞り込むための分野名や関連語が提示されている．

そこで，提示された説明段落を順番に読んだり，分野や関連語で必要な情報に絞り込むことで，既存の検索エンジンよりも効率的に入力キーワードに関する説明情報を取得することができる．

しかし，一般的に人間が編纂する辞典や事典は，一つの見出し語に関して多面的な観点から過不足のない簡潔な説明を記述している．例えば，岩波情報科学辞典 [11] では，本質的な特徴を表す内包的定義，例示による外延的定義，同義語などの観点を必須項目とし，必要に応じて任意の観点を記述している．

それに対して，図 1 に示された複数の説明は異なる Web ページから個別に抜粋された情報であるため，相互に関連性がない．一方の説明に含まれる情報が他の説明に存在しなかったり，逆に同じような情報が複数の説明に含まれる場合がある．そこで，多面的な観点から説明情報を取得するためには，複数の説明段落を横断的に閲覧する必要があり，その結果，同じような内容の説明を何度も読むといった無駄が生じてしまう．

そこで本研究では，一つの見出し語に関する複数の説明段落を統合し，過不足ない説明情報を生成するための要約手法を提案する．具体的には複数文書要約 [5, 6, 8] に相当する処理である．その結果，携帯端末など一度に表示できる文字数が制限される環境においても利便性を向上させることができる．

また，ページをスクロールしたり何度もクリックして次のページを見ないと欲しい情報が手に入らない場合，ユーザは検索サイトの利用を中断するかもしれない．最初のページで簡潔に概要を示してユーザの興味を引くためにも，説明情報を要約して提示することは有効な手段である．

以下，2 章で Cyclone の概要について説明し，3 章と 4 章で提案する複数文書要約の手法と評価実験について説明する．

2 検索サイト Cyclone の概要

図 2 に基づいて検索サイト Cyclone の機能について説明する．事典コンテンツを構築するオフライン処理と，ユーザがコンテンツを検索するオンライン処理に分けて説明する．

オフライン処理では，まず「新語検出」によって見出し語の候補を Web から自動的に収集する．次に，各候補に対して「検索」「抽出」「組織化」を順番に実行し，説明を専門分野ごとに分類する．そこで「パイプライン（処理/油送管）」のように分野によって意味が異なる多義語の説明を区別することができる．

検索処理では，見出し語を含むページを検索する．抽出処理は，HTML タグを用いて見出し語に関する説明を段落単位に抽出する．組織化処理は，a) 特定分野への関連度，b) 説明らしい言語表現を含むかどうか，c) 説明らしい HTML レイアウトかどうか，d) ページの信頼度という 4 つの尺度を統合したスコアを計算して，その値に基づいて段落を分野に分類し，順位付けする．

最後に「関連語抽出」によって，見出し語を特徴付ける語を取得する．これらの語は，オンライン検索時にユーザの情報要求を絞り込むために利用する．

関連語抽出の基本原理は，各用語の説明段落に頻出する語を検出することである．ここで，適切な語を検出する処理と検出した語を評価する尺度が必要になる．そこで，まず段落を「茶筌」で形態素解析して，品詞情報に基づいて（複合）語を構成し，関連語の候補とする．具体的には，名詞，動詞連用形，未知語，記号の連続を語として抽出する．さらに，段落における出現頻度と抽出元の段落に対する組織化のスコアを統合して関連語をソートし，上位の関連語から優先的に提示する．すなわち，良質の説明段落によく現れる語が優先される．

以上の機能に加えて，今回は新たに「要約」機能を導入した．ここでは，一つの見出し語について（各分野ごとに）複数の説明段落を統合する．原理的に

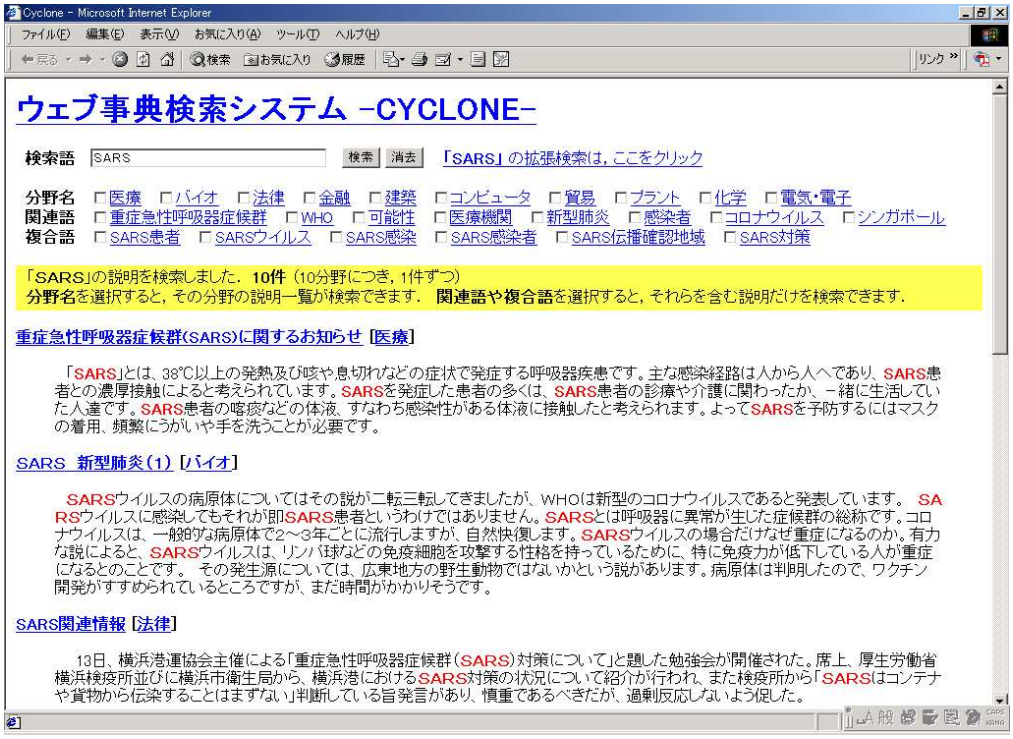

図 1: 入力語「SARS」に対する検索結果

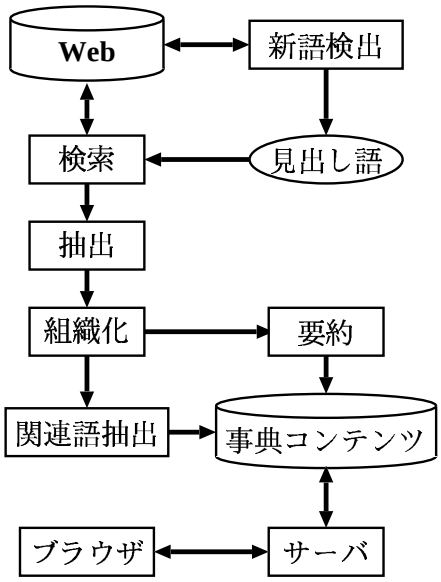

図 2: 事典的 Web 検索サイト Cyclone の概要

は，オフライン，オンラインどちらの段階でも要約機能を適用することができる．現在は，応答時間を考慮し，オフラインで事前に要約を生成している．

オンライン検索では，ユーザが入力したキーワードや質問文が見出し語になかった場合に，文字列が部分的に一致する見出し語や概念的に類似する見出し語を提示することで，ユーザを必要な情報に自然に誘導することができる．しかし，今回の焦点からは外れるため詳細は割愛する．

3 提案する要約手法

3.1 概要

単一文書を対象にした要約に比べると，複数文書の要約に関する研究は新しく，一般的なモデルが確立されていない．基本原理は，一定の基準で集められた文書群を入力として，文書間の類似点と相違点を検出し，冗長性がない単一の文書を生成する点にある．対象や目的に応じて若干の差異はあるものの，既存の手法はおよそ次のような手順に分解することができる [5].

1. 特定

   入力された各文書から，処理の最小単位（文，文節，語などのユニット）を特定する．

2. 照合

   特定処理で検出された単位で比較や照合を行い，類似するユニットを同じグループにまとめる．

3. 選出

   各グループを特徴付けるような代表ユニットを 1 つ以上選択する．また，不要なグループを適宜削除する．

4. 生成

   選択されたユニットを用いて要約を生成する．ユニットを単純に連結して箇条書きにしたり，自然言語生成技術を用いて新しい文や文章を合成する場合がある．

5. 提示

   生成された要約を目的に応じた方法で提示する．

本研究では，上記の手順に基づいて説明情報を対象にした複数文書要約のモデルを考案した．すなわち，Cyclone の組織化処理によって得られた複数の説明段落を入力とし，説明段落中のユニットを説明の「観点」に対応するグループに分類する．そして，各グループを代表するユニットを統合して最終的な説明情報を生成し，ユーザに提示する．

ただし，用意すべき観点は見出し語の種別によって変化する．例えば，専門用語と動植物では説明の観点が異なる．現行の Cyclone には，専門用語，事柄，人名，動植物など様々な種別の見出し語が約 60 万語収録されている．今回は専門用語を対象にモデルの実装と評価を行った．

既存の複数文書要約は新聞記事などを対象としているため，記事の内容やジャンルによっては観点をあらかじめ用意することが困難な場合がある．それに対して，用語説明の場合は人手によって観点をある程度列挙することが可能である．

以下の節で，上記 1〜5 の各手順について説明する．

3.2 特定

特定処理では，要約処理の最小単位を検出する．本研究では，照合処理において「観点」に対応するグループを構成するため，観点に対応する単位を説明段落から抽出する必要がある．

現在は，単文が一つの観点に対応すると仮定している．そこで，特定処理の中核は，説明段落を単文に分割することである．

文章には重文や複文が多く用いられ，これらを単文に分割することは依然として困難な問題である．本研究では「CaboCha」[12] を用いて説明段落中の各文を係り受け解析し，文構造に基づく規則 [7] を適用することで，単文抽出を行う．

ただし，単文に分割しただけでは，後方の文には主語が欠落してしまう．例えば，以下の重文を 2 つの単文に分割した場合，2 つ目の文頭には「XML とは,」を補完しなければならない．

> XML とは，eXtensible Markup Language の略であり，マークアップ言語の一種である．
>
> ↓ 単文に分割
>
> XML とは，eXtensible Markup Language の略であり，
>
> （XML とは,）マークアップ言語の一種である．

しかし，常に先頭文の主語を後続の文頭に補完すればよい訳ではない．どの要素をどのような場合に補完すればよいかを決定することは難しい．このような問題に対して，単文分割におけるゼロ主語補完の手法 [10] が適用可能である．しかし，現在は人手で作成した少数の規則を用いて対処している．

3.3 照合

特定処理によって抽出された単文は，単一の観点に対応すると仮定する．また，照合処理によって類似する単文どうしがグループにまとめられる．そこで，照合処理で構成されるグループもまた 1 つの観点に対応する．

単文を観点に基づいて分類するために，2 つの異なる戦略を併用した．まず，説明の観点に固有の定型表現を人手で作成し，定型表現を含む文を対応するグループに分類する．定型表現は「定義」における「（見出し語）とは」や「例示」における「例えば」などである．これらの表現を用いて初期分類を行う．

しかし，あらかじめ用意された定型表現を含まない（もしくは若干異なる表現が使われた）単文も存在する．そこで，次の段階では未分類の単文を既に分類された単文集合と比較し，類似度が高い単文が属するグループに分類する．類似度の計算には語の重複度を用いた．そこで，未分類の単文は，初期分類で構成されたグループの中で，最も多くの語を共有するグループに分類される．具体的には Dice 係数準拠の計算によって文の長さによる正規化を行う．しかし，全ての語を使うのではなく，形態素解析によって助詞などの機能語を削除する．見出し語「XML」に関する具体例を以下に示す．

(a) XML とは，拡張可能なマーク付け言語のことです．
→ 定義

(b) eXtensible Markup Language の略
→ 略語

(c) 1998 年 に W3C により標準化勧告され，
→ 歴史

(d) XML は Extensible Markup Language の略称です.
→ 略語

(e) この XML の標準化は，W3C で勧告された.
→ ???

この例では，初期分類によって (a)〜(d) の単文が下線を施した語や表現によって該当する観点グループに分類されたことを示している．しかし，(e) は観点グループ固有の表現を含まないため分類できなかった．そこで，語の分布に基づいて，既に分類された単文との類似度を計算する．その結果，(e) は (c) と最も類似度が高いため「歴史」に分類された．

現在，以下に示す 12 種類の観点を用意している．

定義，略語，例示，目的，同義語，書籍，製品，利点，欠点，歴史，要素，機能

また，上記の処理を行ってもいずれの観点にも分類されない単文は「その他」に分類する．

3.4 選出

選出処理では，照合処理で構成されたグループから代表的な文を 1 つ以上選択する．ここでは，グループ全体の傾向を反映しつつ，かつ良質な文を選択することが重要である．

具体的には，以下に示す種々の基準を定量化し，それらを結合したスコアに基づいて代表文を選択する．

- Cyclone の検索結果において上位の説明ほど良質である可能性が高いため，単文が抽出された元の説明段落の順位を考慮する．
- そのグループに含まれる単文に共通して現れる語を多く含む文を代表とする．その結果，説明段落や単文の抽出誤りによって生じた少数派のノイズを最終的な要約から排除する効果がある．
- 説明の文字数を考慮する．携帯端末などの利用環境によっては，表示文字数が最も強い制約になる場合がある．そのような場合には，なるべく短い文を代表文として選ぶ．

以上 3 つのスコアは互いに異なる範囲を取るため，経験的に重みを調整した上で結合している．理論的に妥当なモデルの考案は，今後の検討課題である．

「その他」には，種々の観点に対応する単文が多数混在しているか，もしくは説明文としてふさわしくないノイズが含まれる．冗長な要約になることを避けるために「その他」からは一般（「その他」以外）の観点から既に選出された単文と語の重複がなるべく少ない単文を優先的に選出する．

「その他」から複数の単文を選出する場合には，まず最初の 1 件を選出し，既に選出された単文との語の重複が少ない単文を次に選出する．この処理を再帰的に行うことで多様な単文を選出する．

3.5 生成と提示

生成・提示処理では，各グループから選出された代表文をグループ（観点）名とともに箇条書きで表示する．ここでは，選択処理におけるスコアが高い代表文から順番に提示する．図 3 と図 4 は，それぞれ，見出し語「XML」に関するコンピュータ分野の説明段落とそれらを要約した結果である．この例では，397 文字という少ない文字数で多面的な観点から見出し語について概観できる要約が生成された．

「定義」と「略語」において「eXtensible Markup Language」が重複しており，冗長性を完全に排除できていない．今後，特定処理において括弧表現なども考慮した文分割について検討する必要がある．

生成処理には工夫の余地がある．例えば，特定処理で抽出された単文の文末表現を置換することで，文字数を少なくしたり，文末らしい表現に修正することができる．しかし，単なる抜粋の範囲を逸脱して説明内容を改変することは，ページの著作権を侵害する可能性がある．研究目的として許容される範囲と Web 上で実際に運用する場合の制約について注意しなければならない．

提示処理における工夫として「定義」や「目的」などの観点名から，そのグループに属する説明文や抽出元の説明段落にリンクをはり，ユーザが選択した観点だけに絞り込むといった誘導の手法がある．「書籍」の説明は，XML に関する本の販売情報のページから抽出されたものである．この説明を手がかりにして書籍の販売情報を素早く取得することができる．

4 評価実験

4.1 方法

要約手法の評価は判定者の主観に依存する部分が大きいため工学的に評価することが困難である．新聞記事などを対象にした評価用テストコレクション [4] は存在する．また，要約手法そのものの評価ではなく，別のタスク（情報検索における適合文書の選択など）に応用した場合の性能向上によって，間接的かつ客観的に要約手法を評価する方法もある．

しかし，本研究で対象にしている用語説明に関する要約はあまり前例がないため，評価手法の問題点を洗い出すことも念頭に置いて，小規模ながら独自にテストデータを作成して評価に利用した．また，要約手法そのものに関する評価実験だけを行った．

評価の基準や尺度には複数の選択肢がある．例えば「自動生成された要約が既存の用語辞典の説明にどの程度近付いたか」という基準がある．しかし，本要約手法の入力として Cyclone の検索結果を用いるため，既存の辞典にしかない，もしくは既存の辞典にはない観点が存在した場合には評価が難しくなる．

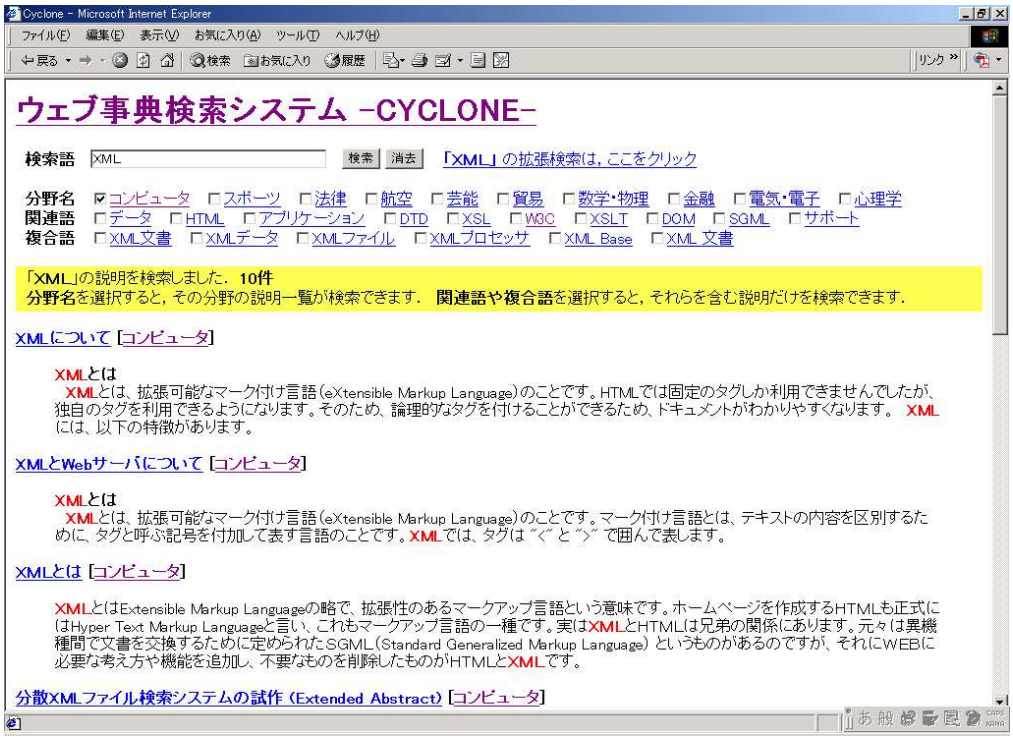

図 3: 「XML」のコンピュータ分野に関する説明

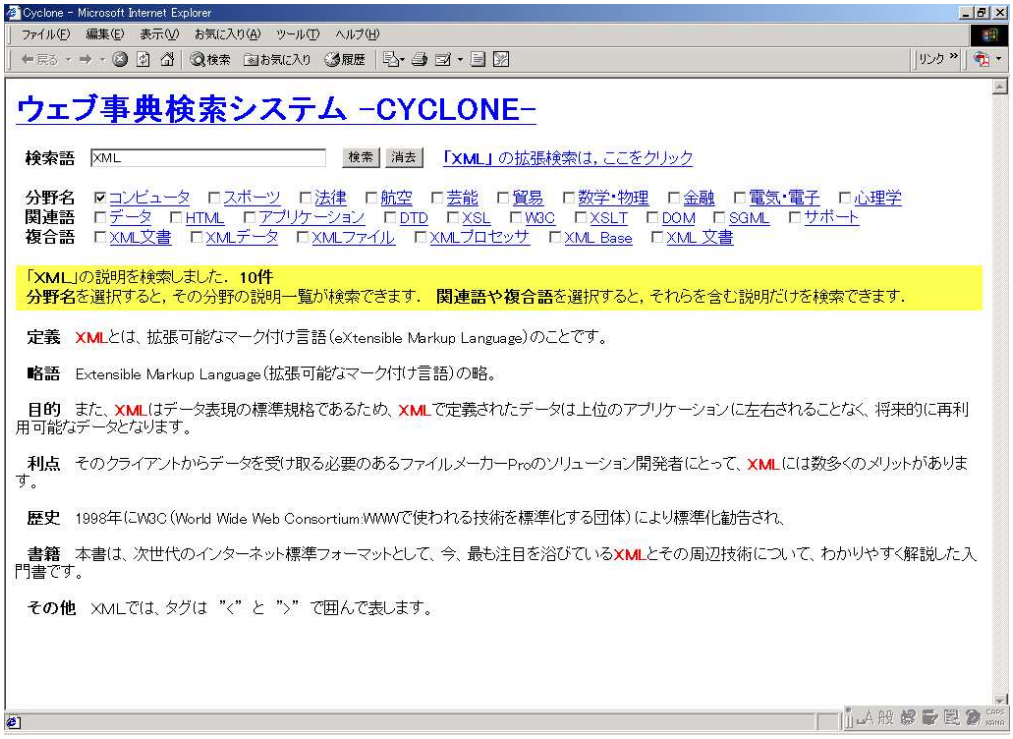

図 4: 「XML」のコンピュータ分野に関する説明を要約した結果（本文 397 文字）

そこで今回は以下の 2 種類の尺度を用いた．両者はトレードオフの関係にあり，同時に改善することが難しい尺度である．

- 要約率

  Cyclone の検索結果をどれだけ短縮することができたか．

- 網羅率

  Cyclone の検索結果に含まれる説明の観点のうち，どれだけを要約に含めることができたか．

評価実験に用いた見出し語を表 1 に示す．各見出し語について，Cyclone の検索結果のうちコンピュータ分野 50 件を要約処理の入力とした．50 件に限定したのは人手判定のコストを増やさないことが理由である．予備調査の結果，Cyclone の検索結果上位 50 件の説明段落には，既存の辞典 [9] に記述されている観点の約 85%が含まれることが分かっている．

判定者は，要約処理の結果を見ずに，Cyclone の検索結果上位 50 件だけを吟味して単文単位に観点を付与した．判定者に付与を依頼した 30 種類の観点を以下に示す．説明として適切でない単文には観点を付与しなかった．

> 定義，例示，上位概念，下位概念，要素，性質，属性，機能，目的，歴史，現在，予測，原因，結果，同義語，反意語，略語，訳語，間接的説明，比較，比喩，製品，書籍，別の意味（多義語の場合），読み，入手方法，利点，欠点，語源，その他

判定者が虚偽の説明に対して観点を付与することがないよう，既存の辞典に掲載された説明等を見せることで対象の用語に関する知識を与えた．判定者は 1 つの単文に対して 1 つ以上の観点を付与した．対象用語が多義語で既存の用語辞典以外の意味で使われている説明には「別の意味」を付与し，事実上，評価の対象外とした．

判定者に示した観点の中には，典型的な表現や語を人手で列挙できなかったために，システムがそもそも出力できないものも含まれる．これらの観点は，要約処理の結果には決して含まれないため，網羅率を下げる要因となった．

4.2 結果と考察

表 1 に要約率と網羅率を示す．各観点のグループから出力する代表文の件数（N）を 1，2，3 と変化させた．ただし「その他」からは常に代表文を 5 件選出した．要約率は式 (1) で計算した．要約率が小さいほど良い結果である．

$$\frac{\text{自動要約の文字数}}{\text{要約しない場合にユーザが読む文字数}} \quad (1)$$

表 1: 要約手法の評価実験結果

用語（見出し語）	N	要約後文字数	要約率 (%)	網羅率 (%) A	B	C
2 進数	1	605	7.8	26.7	53.3	33.3
	2	842	10.8	26.7	53.3	40.0
	3	1074	13.8	26.7	53.3	40.0
ASCII	1	551	7.2	5.9	47.1	47.1
	2	939	12.3	11.8	64.7	52.9
	3	1277	16.7	11.8	64.7	58.8
SQL	1	645	7.8	17.7	41.2	47.1
	2	1008	12.1	23.5	58.8	64.7
	3	1432	17.2	23.5	58.8	64.7
シソーラス	1	855	9.2	28.6	35.7	35.7
	2	1467	15.8	42.9	64.3	50.0
	3	1923	20.7	42.9	71.4	50.0
データウェアハウス	1	619	5.7	31.6	47.4	36.8
	2	1129	10.4	31.6	52.6	57.9
	3	1543	14.2	36.8	68.4	57.9
マクロウイルス	1	548	6.1	23.1	46.2	53.8
	2	996	11.2	23.1	53.9	61.5
	3	1307	14.6	23.1	53.9	69.2
並列処理	1	575	9.0	25.0	66.7	50.0
	2	838	13.1	33.3	75.0	50.0
	3	1093	17.1	33.3	75.0	50.0
平均	1	628	7.4	22.4	47.7	43.0
	2	1031	12.2	27.1	59.8	54.2
	3	1378	16.3	28.1	63.5	56.1

「要約しない場合にユーザが読む文字数」は，Cyclone の検索結果をユーザが上から順番に読んだ場合に，自動要約に含まれる観点を全て読むまでの文字数である．自動要約に含まれる観点を全て読めば，50 位以前でも閲覧を終了すると仮定した．観点が同じであれば，要約に含まれる文と同一である必要はない．

網羅率は式 (2) で計算した．網羅率が大きいほど良い結果である．

$$\frac{\text{要約に含まれた観点数}}{\text{人間が付与した観点数}} \quad (2)$$

なお，提案する要約手法の網羅率を以下の 2 通りの方法で評価した．

- A: 人間が付与した観点が含まれ，かつ観点名も正しかった場合のみ正解と見なした．
- B: 観点名の適否は考慮しない．

さらに，B と比較するための基準値として，以下の単純な要約手法による網羅率を計算した．

- C: Cyclone 検索結果の上位から，提案する要約手法の結果と同じ文字数を抜粋する（ただし，抜粋する点が文中の場合は文末まで出力する）．

C は単一文書の要約における「リード法」（文書の先頭から一定文字数を抜粋する単純な手法）に相当する．また，C の手法は観点名を付与することができないため，B と同様に観点名の適否は考慮しない．

表 1 の結果について考察する．まず，N の値を増やすことによって要約率は大きくなり，逆に網羅率は高くなったことが分かる．

要約率は N によって変動するものの，およそ 10%前後となり，元の情報をかなり短縮できたことが分かる．同じ量の情報を取得するために，ユーザは 10 分の 1 程度の労力を使うだけでよい．他方で，A の網羅率は平均して 20%代という低い値となった．しかし，観点名の適否を考慮しない B では，A に比べて網羅率が 2 倍以上になった．

B と C を比較すると，総じて提案手法の網羅率が高く，提案手法が同じ文字数でユーザにより多くの情報を与えることが分かった．両者の差異は小さいものの，要素技術を改善していくことで，この差は今後広がる可能性がある．

さらに，網羅率を観点ごとに分析した．紙面の制約上，Cyclone の検索結果で 6 つ以上の用語に含まれた観点を対象に，B において $N = 1$ とした場合のみを以下に示す．

> 定義（7/7），例示（4/7），同義語（1/6），性質（3/7），機能（1/7），目的（3/6），歴史（4/6），間接的説明（5/7），比較（1/7），書籍（5/6）

「例示」や「同義語」など，既存の辞典に含まれやすい観点を中心に網羅率を改善する必要がある．

以下，今後の評価実験で検討すべき事項を挙げる．

- 複数の判定者によって判定の客観性を高める．
- 致命的な誤りとそうでない誤りを区別する．
- 今回の実験では同じ観点が付与された単文を全て同等に扱った．しかし，同じ観点が付与された複数の単文に対して「どちらがより良い説明文か」という判定が必要になる．
- 本研究における要約処理の目的は，ユーザが必要な情報を効率的に取得することを補助したり，概要を簡潔に見せることでユーザの興味を引き付ける点にある．しかし，テストコレクションを用いた実験では評価できる項目に制限がある．Web 上で検索エンジンを運用し，ユーザの検索ログを解析することで要約手法の有効性を評価する必要がある [1]．

5 おわりに

Web から事典的なコンテンツを検索するサイト「Cyclone」において，複数の説明情報を統合し，1 つの用語に関して多面的な観点から概観するための要約手法を提案した．評価実験によって提案手法の有効性を示した．今後，評価実験の規模を拡張し問題点を明らかにすることで手法のさらなる改善を行う．

参考文献

[1] Peter Anick. Using terminological feedback for Web search refinement: A log-based study. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 88–95, 2003.

[2] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 488–495, 2000.

[3] Atsushi Fujii and Tetsuya Ishikawa. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 196–203, 2001.

[4] Takahiro Fukushima, Hidetsugu Nanba, and Manabu Okumura. Text summarization challenge 2: Text summarization evaluation at NTCIR workshop 3. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[5] Inderjeet Mani. *Automatic Summarization*, chapter 7, pp. 169–208. John Benjamins, 2001.

[6] Dragomir R. Radev and Kathleen R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, No. 3, pp. 469–500, 1998.

[7] 武石英二, 林良彦. 接続構造解析に基づく日本語複文の分割. 情報処理学会論文誌, Vol. 33, No. 5, pp. 652–663, 1992.

[8] 奥村学, 難波英嗣. テキスト自動要約に関する最近の話題. 自然言語処理, Vol. 9, No. 4, pp. 97–116, 2002.

[9] 藤本喜弘（編）. 第二種・シスアド情報処理用語辞典. 経林書房, 1998.

[10] 江原暉将, 金淵培. 確率モデルによるゼロ主語の補完. 自然言語処理, Vol. 3, No. 4, pp. 67–86, 1996.

[11] 長尾真. 辞典形式での専門分野の知識の体系的構築法. 人工知能学会誌, Vol. 7, No. 2, pp. 320–328, 1992.

[12] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[13] 藤井敦, 伊藤克亘, 石川徹也. WWW は百科事典として使えるか? –大規模コーパスの構築–. 情報処理学会研究報告, 2002-NL-149, pp. 7–14, 2002.

[14] 藤井敦. Web 情報を用いた事典検索サイトの構築. 情報の科学と技術, Vol. 53, No. 4, pp. 201–204, 2003.

[15] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌, Vol. J85-D-II, No. 2, pp. 300–307, 2002.