事典的Web検索サイトにおける複数文書要約の応用

藤井 敦　　渡邉まり子　　石川徹也
筑波大学　図書館情報大学　筑波大学
fujii@slis.tsukuba.ac.jp

1　はじめに

World Wide Web 上の検索エンジンを事典のように使って様々な調べ物をすることが日常的になっている．Web には既存の事典に載っていない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．

筆者らは，Web の事典的な利便性を向上させる目的で検索サイト Cyclone [1][1] を構築した．Cyclone には，見出し語とその説明情報を Web 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，入力語「XML」に対する検索結果である．画面の下半分には，複数の Web ページから個別に抜粋された説明段落が，抽出元のタイトルと一緒に 3 件提示されている．説明段落は，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．

人間が編纂する辞典や事典は，1 つの見出し語に関して多面的な観点から過不足のない説明を記述している．岩波情報科学辞典 [2] 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，同じような内容を何度も読むといった無駄が生じてしまう．

本研究は，1 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．

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である．

[1]http://cyclone.slis.tsukuba.ac.jp/

図 1: 「XML」に対する Cyclone の検索結果

2　提案する要約手法

2.1　概要

本研究で行う要約は，複数文書要約の一種である．すなわち，異なる著者が書いたか，もしくは同一の著者が異なる時期に書いた文書群を入力として，冗長性がない単一の文書を生成することが目的である．単一文書の要約と比較した場合に，入力における内容の重複や発散の度合いが大きいため，情報の類似点や相違点を検出することが重要である．そこで，以下の 2 点を適切に設定しなければならない．これらは車の両輪のように密接に関連している．

- どのような単位で 2 つの情報を比較するのか
- どのような基準で 2 つの情報が類似していると判断するのか

本研究では，「定義」，「目的」，「例示」といった説明の「観点」を比較の単位として，観点が同じ情報は類似していると判断する．

本研究の特長は，説明の観点を要約手法に導入する点と，それによって玉石混淆の Web 情報から統制された説明情報を自動生成する点にある．

ただし，考慮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．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手法の提案と評価を行った．

複数文書要約に関する既存の手法は，およそ以下のような手順に分解することができる [3]．

(1) 特定
入力された文書群から，比較の単位（語，文節，文などのユニット）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．複数のユニットを箇条書きにしたり，自然言語生成によって新しい文や文章を合成する．

(5) 提示
生成された要約を目的に応じた方法で提示する．

本研究の要約手法もこの手順に基づいて実行する．ただし，各手順の実現方法は既存の手法とは異なる．以下，2.2〜2.6 節で，上記 (1)〜(5) の各手順について説明する．

2.2 特定

特定処理は，要約処理の最小単位を検出する．本研究では，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おり，説明段落を単文に分割する．

重文や複文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[2]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 [4] を適用して，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から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「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．
>
> ↓ 2 つの単文に分割

- 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．

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ために照応解析が適応可能である．しかし，現在は人手で作成した規則を用いて対処している．

[2]http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

2.3 照合

照合処理は，同じ観点に対応する単文を 1 つのグループにまとめる．ここで 2 つの異なる分類手法を併用する．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る．形態素解析によって名詞や動詞などの内容語を抽出し，Dice 係数によって語の重複度を計算する．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．見出し語「XML」に関する例を以下に示す．

(a) XML とは，拡張可能なマーク付け言語です． → 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 ... 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した．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．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．

2.4 選出

選出処理は，照合処理で構成されたグループから，グループ全体の傾向を反映する良質な文を選択する．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が高い代表文を選択する．

(1) グループ全体の傾向を反映させるために，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優先する．Cyclone の検索結果に含まれる不適切な説明や，単文の抽出誤りによって生じた文の断片を最終的な要約から排除する効果もある．

(2) 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．

(3) 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．現在は，(1), (2), (3) の順番に大きな重みを与えている．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」グループには，種々の観点に対応する単文が混在している．冗長な要約生成を避けるために,「その他」からは，一般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．

ただし，文章の一貫性を保持するためには，なるべく同じ説明段落から代表文を選出するという基準が有効であり，今後検討する必要がある．

2.5 生成

生成処理は，各グループから選出された代表文を観点名とともに箇条書きで出力する．このとき，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 2 は，見出し語「XML」に関する説明段落（上位 50 件，合計 11224 文字）を要約した結果である．この例では，397 文字で多面的な観点から XML について概観できる要約が生成された．

図 2 の要約には問題もある．例えば「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．

生成処理には改善の余地がある．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．

2.6 提示

提示処理では，ハイパーリンク機能を利用して要約を提示する．すなわち,「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．例えば「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．そこで，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．

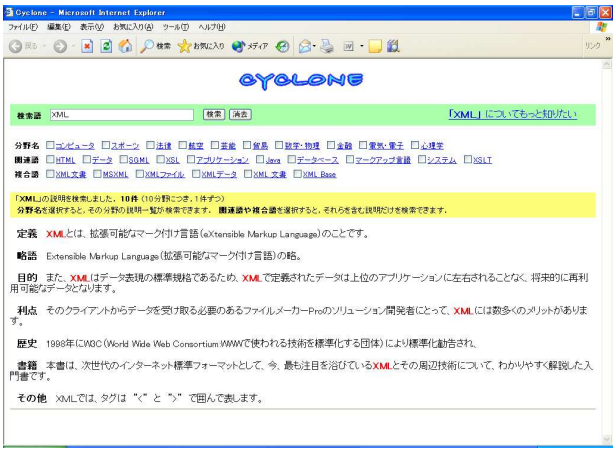

図 2: 「XML」に関する説明を要約した結果

3 評価実験

3.1 方法

要約手法の評価基準は複数ある．例えば「自動生成された要約が既存の事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，入力として Cyclone の検索結果を用いるため，既存の事典にしかない，もしくは既存の事典にはない観点が存在した場合には評価が難しくなる．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．

- 要約率

  Cyclone の検索結果をどれだけ短縮できたか．
- 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．

評価実験には，以下に示す見出し語 15 件を用いた．

> 10BASE-T，2 進数，ASCII，SQL，XML，アキュムレータ，アセンブラ，クロスケーブル，シソーラス，データウェアハウス，マクロウイルス，解像度，回転待ち時間，主記憶装置，並列処理

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．

評価の正当性と客観性を高めるために，筆者以外の大学生 2 人が個別に判定を行った．判定者は要約結果を見ずに，Cyclone の検索結果上位 50 件を吟味して単文に観点を付与した．事前に予備判定を行って必要な観点を適宜追加した結果，以下に示す 30 種類の観点が判定対象になった．

> 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，上位概念，下位概念，性質，属性，現在，予測，原因，結果，反意語，訳語，間接的説明，比較，比喩，読み，入手方法，語源，別の意味（多義語の場合），その他

表 1: 要約手法の評価実験結果

			判定者 A による網羅率 (%)				判定者 B による網羅率 (%)			
			観点数 12		観点数 28		観点数 12		観点数 28	
代表文数	要約文字数	要約率 (%)	本手法	リード法	本手法	リード法	本手法	リード法	本手法	リード法
1	616	5.97	56.62	52.84	49.49	44.84	50.00	53.61	49.49	47.56
2	998	9.61	73.43	57.23	59.26	53.70	64.50	62.96	60.75	57.37
3	1309	12.61	76.04	59.29	63.13	56.44	67.83	64.81	65.22	60.84

このうち，提案した要約手法が対応しているのは，最初の 12 種類である．

説明として適切でない単文に観点を付与しないよう，判定者には既存の辞典に掲載された説明を見せて対象の用語に関する知識を与えた．そこで，要約に不適切な単文が含まれた場合は網羅率が低下する．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，既存の辞典とは別の意味で使われている説明には「別の意味」を付与し，評価の対象外とした．

要約率と網羅率は，式 (1) で計算した．

$$\begin{aligned} \text{要約率} &= \frac{\text{自動要約の文字数}}{\text{要約しない場合にユーザが読む文字数}} \\ \text{網羅率} &= \frac{\text{要約に含まれた観点の異なり数}}{\text{判定者が付与した観点の異なり数}} \end{aligned} \quad (1)$$

式 (1) の要素には複数の解釈がある．今回は紙面の都合上，一部について検討する．要約率の分母は，Cyclone の検索結果 50 件に含まれる文字数とした．網羅率における「観点」については，要約処理が対応している 12 種類に限定した場合と，判定者に依頼した観点のうち，「別の意味」と「その他」を除く 28 種類を考慮した場合を評価した．同じ観点に対応する単文は全て等価であると見なした．また，観点名の適否は考慮しなかった．

各観点のグループから出力する代表文の件数を 1，2，3 と変化させた．ただし「その他」は複数の観点が混在しているので，常に 5 件の代表文を選出した．

提案手法の比較対象として「リード法」を用いた．これは，文書の先頭から一定文字数を抜粋する単純な要約手法である．今回の実験では，Cyclone 検索結果の上から，提案手法が出力した文字数だけ抜粋した．すなわち，要約率を揃えて，網羅率の優劣を比較した．

3.2 結果と考察

表 1 に実験結果を示す．以下，結果について考察する．

まず，選出する代表文の数を増やすことで要約率は劣化し，網羅率は向上した．要約率が 10%前後だったのに対して，網羅率を 50〜70%代に保持することができた．ユーザは 10 分の 1 程度の労力で，Cyclone の検索結果に含まれる観点を半分以上取得することができる．

代表文を 1 文ずつ選出した場合は，論文抄録と同程度か若干長めだった．代表文を 3 文ずつ選出した場合は 1309 文字であった．既存の事典でも，1 つの見出し語について長く説明する場合には，この程度の文字数を使う．

処理対象とした 12 種類の観点に限定した場合の方が，28 種類の観点を対象にした場合より網羅率が高かった．今後，処理対象の観点種を増やす必要がある．

「代表文数 1，観点数 12，判定者 B」の 1 つを除いて，提案手法はリード法よりも高い網羅率を達成し，同じ文字数で多くの情報を与えることが分かった．ただし，用語によってはリード法の網羅率が高い場合もあった．今後，原因を分析して手法を改善する必要がある．

以上の傾向は，判定者によらずにほぼ同じだった．付与された観点は判定者によって揺れがあったものの，評価結果への影響は少なかった．

岩波情報科学辞典で必須の「定義」,「例示」,「同義語」について，網羅率の内訳を分析した．その結果,「定義」と「例示」の網羅率が 60〜90%だったのに対して,「同義語」の網羅率は 50%以下だった．同義語は括弧表現で示されることが多い．しかし，括弧の用法は多様であるため，同義語だけを高精度で検出することは難しい．今後はこの問題を解決する必要がある．

4 おわりに

事典検索サイト Cyclone において，複数の説明情報を統合し，1 つの見出し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] 藤井敦, 伊藤克亘, 秋葉友良. 事典的 Web 検索サイトの構築. 言語処理学会第 9 回年次大会発表論文集, pp. 129–132, 2003.

[2] 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.

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

[4] 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.