構造を用いた英語パラグラフ検索システムにおける パラグラフ抽出機能の実現

An Implementation of a Paragraph Extraction Function for an English Paragraph Retrieval System Using Structure

芝﨑 恭史[*1], 國近 秀信[*2]
Kyoshi SHIBAZAKI[*1], Hidenobu KUNICHIKA[*2]
[*1] 九州工業大学大学院情報工学府
[*1] Kyushu Institute of Technology Graduate School of Information Engineering
[*2] 九州工業大学情報工学研究院
[*2]Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
Email: shibazaki.kyoshi674@mail.kyutech.jp

あらまし：英語初学者にとって，英語の論理展開法に則ってパラグラフライティングを行うことは困難である．論理展開法の参考にするために，Web ページ上の大量のパラグラフの中から探す方法が考えられるが，学習者自身でパラグラフの構造の適切性を判断することは難しい．そのため我々は，Web ページから大量のパラグラフを収集し，ユーザーが必要な構造と内容のパラグラフの検索を可能にするシステムの実現を目指している．本研究は，本システムの一部であるパラグラフ抽出機能の実現を目的とする．
キーワード：英語学習，パラグラフライティング，検索システム，Web からのパラグラフ収集

1. はじめに

英語初学者にとって，英語の論理展開法に則ってパラグラフライティングを行うことは困難である．論理展開法の参考にするために，Web ページ上の大量のパラグラフの中から探す方法が考えられるが，学習者自身でパラグラフの構造の適切性を判断することは難しい．先行研究(1) では，Web ページからパラグラフを収集したと仮定した上で検索機能を実現した．本研究では，Web ページからのパラグラフ自動収集機能の下位機能であるパラグラフ抽出機能の実現を目的とする．

2. 構造を用いたパラグラフ検索

本研究は，英語の論理展開法の理解が不十分なユーザーを支援対象とする．そのようなユーザーにとっては，適切な構造で書かれたパラグラフであるだけでなく，書こうとしている内容に近いパラグラフを参考にできることが望ましい．そのため，構造についてはユーザーが種類や構成要素などを指定して検索でき，内容についてはユーザーが指定した内容のパラグラフがヒットするように大量のパラグラフから検索できる必要がある．そのため，本研究では Web からのパラグラフ収集を行う．しかし，Web ページには文章以外にも多くの種類の情報が含まれており，パラグラフの質もさまざまである．よって，多くの種類の情報が含まれるページの中からパラグラフのみを収集する機能，適切な構造のパラグラフを選択する機能が必要である．本研究は，前者のパラグラフ抽出機能の実現に焦点を当てる．

3. パラグラフ検索システム

パラグラフ検索システムの概要を図 1 に示す．本システムは，パラグラフ抽出機能，パラグラフ選択機能，検索機能，パラグラフデータベース，パラグラフ展開スキーマデータベースおよび表示機能およびから成る．まずパラグラフ抽出機能により Web ページからパラグラフを集め，選択機能により構成要素を同定し適切な構造のパラグラフをデータベースへ格納する．検索時には，ユーザーが条件を入力すると，本システムはデータベースの中のパラグラフをパラグラフ展開スキーマと比較・採点し，その結果を表示する．

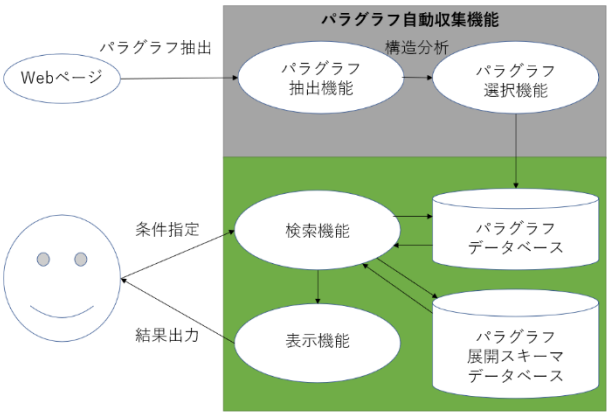

図 1 パラグラフ検索システムの概要

4. パラグラフ抽出機能

Web ページから英文パラグラフを抽出するためには，特定のサイトにアクセスし，Web ページの中から英文パラグラフに該当する部分を取り出す作業が必要となる．まず Web サイトについては，数多くのサイトの中から学習に適した内容および構造の英文

を有するサイトを対象とする必要がある．適切なサイトを自動的に判断することは困難であるため，本研究では，あらかじめ利用する Web サイトを指定しておくという方法を取る．また Web ページ内のパラグラフの特定については，サイトごとに構造が異なるため，現時点で自動的に処理することは困難である．よって本研究では，あらかじめ，各サイトのページ構造を確認し，文章の場所を特定して各サイト固有の情報として記録・利用することとした．なお，関連情報として，パラグラフの出典(URL，タイトル)，更新日時およびレベルも同時に抽出する．

本研究では，正しい英文を提示でき，幅広いレベルの学習者に対応できるようにするため，パラグラフの情報源として Breaking News English[2]，News in Levels[3]，毎日新聞(英語)[4]および alpha japantimes[5]の 4 つのサイトを使用した．続いて，サイトの固有情報について，News in Levels を用いて説明する．本サイトのページ構造を図 2 に示す．本サイトでは， タグの中にパラグラフが存在する．より具体的には，パラグラフ内の各文は，<p>タグで囲まれている．ただし，"<p>Difficult words:" 以降については，パラグラフの英文ではないため，抽出対象から除外する必要がある．

```
<main class=~ …>↵
…↵
<div id=”nContent”>↵
<p>(更新された日付、時刻)</p>↵
<p>本文</p>↵
<p>本文</p>↵
<p>本文</p>↵
<p>Difficult words: ~ …</p>↵
<p>レベル 3 でオリジナルビデオが観ることができるという説明</p>↵
</div>↵
```

図 2　News in Levels のページ構造

その他，Breaking News English については，<article>タグで囲まれた箇所の<p>タグによりパラグラフが判別可能である．ただし，同じ<p>タグにより，広告が途中に存在する場合があるため，広告か否かを判別して広告の場合は除外するよう実現した．また，毎日新聞については，Breaking News English と同様に，<article>タグでパラグラフが囲まれている．しかし，日付や著者などパラグラフ以外にも<p>タグがつけられているため，それらを抽出しないよう実現した．

5.　実行例

News in Levels の記事ページ[6]の例を図 3 に示す．この例では，赤線で囲まれた英文のみを抽出する必要がある．

図 3　News in Levels の記事ページ

図 4 に，パラグラフ抽出機能で抽出した結果を示す．ここで，図 4 の下線部は，図 3 の下線部に対応する．

ID:1↵
URL: https://www.newsinlevels.com/products/la-palma-volcano-level-1↵
タイトル: La Palma volcano – level 1 - News in Levels↵
更新日時:03-02-2022 07:00↵
レベル:1↵
パラグラフ:La Palma is small island. It belongs to Spain's Canary Island. A ↵
volcano erupts on La Palma in September 2021. Lava destroys thousands ↵
of homes. 7,000 people must leave. The area changes a lot. There are no green ↵
hills. The place is dark. Some people cannot go back home. There are many ↵
craters …↵

図 4 ：News in Levels から抽出した結果

6.　おわりに

本研究では，特定の Web ページからパラグラフを抽出する機能を実現した．パラグラフ自動収集機能を完成させるためには，収集したパラグラフの構造を分析し，適切な構造のパラグラフを選択する機能が必要である．今後は，このパラグラフ選択機能を実現する予定である．また，サイトの固有情報の一般化や，文章の難易度判定にも取り組む予定である．

参考文献

(1) 片岡拓也：“パラグラフ展開スキーマを用いたパラグラフ検索システムの改良”，2019 年度九州工業大学卒業論文（2020）
(2) Sean Banville: Breaking News English Lessons: Easy English World News Materials – ESL，https://breakingnewsenglish.com/（参照 2022.02.03）
(3) English in Levels: News in Levels，https://www.newsinlevels.com/（参照 2022.02.03）
(4) 毎日新聞社: The Mainichi - Japan Daily News，https://mainichi.jp/english/（参照 2022.02.03）
(5) ジャパンタイムズ株式会社: The Japan Times Alpha Online ― 英語学習者のための英字新聞，https://alpha.japantimes.co.jp/（参照 2022.02.03）
(6) English in Levels: La Palma volcano - level 1，https://www.newsinlevels.com/products/la-palma-volcano-level-1（参照 2022.02.03）