英語パラグラフの構成要素同定機能の改良
～順序のディスコースマーカー～

Improving a Function for Identifying English Paragraph Components
~Discourse Markers for Sequencing~

林 楓[*1], 國近 秀信[*2]
Kaede HAYASHI[*1], Hidenobu Kunichika[*2]
[*1] 九州工業大学大学院　情報工学府
[*1] Graduate School of Information Engineering, Kyushu Institute of Technology
[*2] 九州工業大学大学院　情報工学研究院
[*2]Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
Email: Hayashi.kaede995@mail.kyutech.jp

あらまし：英語の論理展開法に沿った説得力のある英語パラグラフを書くための方法として，適切な構造で書かれたパラグラフを参照する方法が考えられる．その方法を支援するため，我々は，Web から英語パラグラフを入手し，その種類と構造の適切性を判断した上でユーザへ提供するパラグラフ検索システムの実現を目指している．本システムの実現のためには，パラグラフ中の文章を解析し，構成要素とその役割を同定する構成要素同定機能が必要である．これまでに構成要素同定機能を実現したものの，いくつかの問題点があった．本研究では，構成要素同定機能の改良として，順序のディスコースマーカーの同定機能の改良を行う．
キーワード：英語学習，パラグラフライティング，パラグラフ検索，論理展開

1. はじめに

一般に英語初学者は，英語文章の構造を十分には理解できておらず，また，構造を意識して英語パラグラフを書いた経験が不足しているため，英語の論理展開法に沿った説得力のある英語パラグラフを書くことが難しい．この問題を解決するため，英語初学者の書きたい内容に近く，論理展開法に則った適切なパラグラフ例を参考とする方法がある．しかし，参考書等から得られるような構造が適切なパラグラフは多くなく，また Web 上から得たパラグラフは構造の適切さが保証されない．

これを解決するため，我々は Web 上から入手したパラグラフについて，その種類と適切な構造であるか否かを判断して，適したパラグラフを提供する検索システムの実現を目指している．我々はこれまでに，パラグラフの構成要素の同定機能を実現した．本研究では，パラグラフの構成要素のうち適合率が低かった順序のディスコースマーカー“Word for enumeration”の同定機能を改良することを目的とする．

2. パラグラフの構成要素同定機能

パラグラフは，主題文やその支持文など複数の要素で構成される．また，パラグラフには目的に応じて多様な書き方があり，それぞれに対して異なる要素で構成される．

本研究では，パラグラフの種類別の構造を表すパラグラフ展開スキーマの構成要素を同定する．スキーマの例を図 1 に示す．構成要素には，Introductory sentence，Topic sentence，Concluding sentence など，文の位置からある程度は役割を推定できるものがある．また，そこからパラグラフの種類を判定することで，Item(reason)や，Item(effect) など細かな特定の役割が同定できる．さらに，順序のディスコースマーカーである Word for enumeration など 1 文では同定がしづらいものは，パラグラフ全体の解析を行い，語彙同士の前後関係やパラグラフ全体で出現する語彙関係に着目して同定する．また，これに加えて，品詞の情報を含む構成要素ごとのキーワードを利用する．

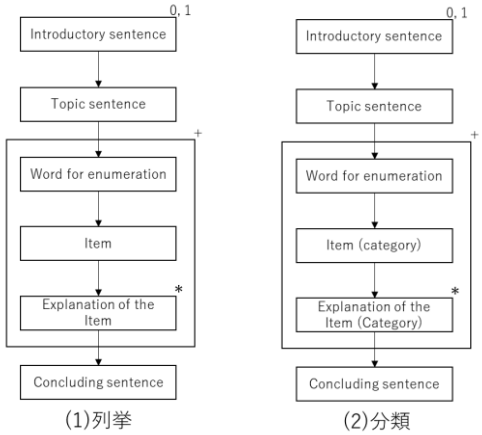

図 1 パラグラフ展開スキーマの例

3. パラグラフ検索システム

本システムは，図 2 のように，パラグラフ検索機能，パラグラフ表示機能，パラグラフデータベース，パラグラフ展開スキーマによって構成される．本システムのユーザは，参照したいパラグラフの種類や

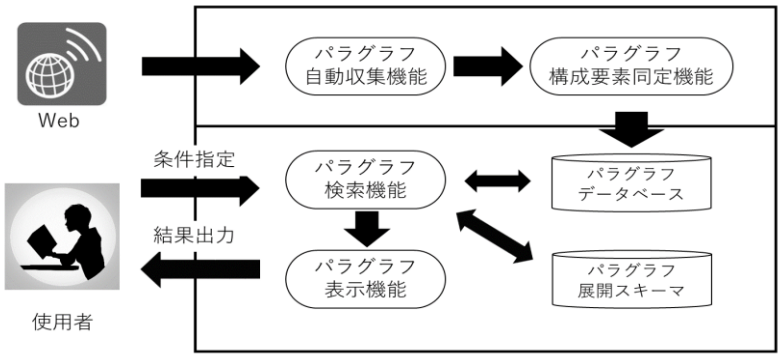

図 2 パラグラフ検索システムの構成

The differences between Christianity and Islam are often emphasized, and certainly there are great differences between these two religions. However, Christianity and Islam also have important similarities. <Word_for_enumeration> First </Word_for_enumeration>, Islam and Christianity are large and influential, with many millions of followers. <Word_for_enumeration> Second </Word_for_enumeration>, both religions teach that there is one God. <Word_for_enumeration> Third </Word_for_enumeration>, both religions teach that it is the duty of believers to take care of the poor and those who need help. For example, it is the duty of every Muslim (follower of Islam) to give alms (gifts of money) to help the poor. In addition, Christians have built institutions to help those who need help, including the poor and the sick. <Word_for_enumeration> Fourth </Word_for_enumeration>, people of both religions try to follow the teachings written down in their holy books, the Bible for Christians, and the Koran for Muslims. These teachings tell how followers should live their daily lives and carry on their relationships with God, their family, other people, the church, and their government. These are just a few of the similarities between Christianity and Islam.

図 3 Word for enumeration についての実行結果

内容に関するキーワードを条件として検索を行う．本システムはデータベースの中から，ユーザが望む構造および内容のパラグラフを提示する．

4. Word for enumeration の同定機能の改良

先行研究では，パラグラフ展開スキーマの 31 種類の構成要素全てについて同定機能を実現した．本研究では，先行研究においてプログラム上の不備が見つかった Word for enumeration の同定について，プログラムの見直しを行う．

Word for enumeration とは「数え上げのための言葉」であり，順序を示す“first”や“second”などが該当する．ここで，パラグラフ文中においてこれらの語が出現すれば必ず Word for enumeration と決められるのではなく，パラグラフ全体を通して Word for enumeration とされる単語が複数回，順番通りに出現した場合に同定される．

先行研究[(1)]では，first, secondly, next, finally などの頻出語句を定めるとともに，本要素の同定条件を次のように定めていた．

(I)パラグラフ中の頻出語句が数詞の順番に出現すること．
(II)頻出語句が同じ品詞で出現すること．
(III)next 等，数詞以外の語句が来た場合は，以降の文中に数詞は出現しない．
(IV)形容詞の頻出語句については，その語が形容する名詞も考慮する．

しかし，先行研究におけるプログラム上では，数詞の順番のチェックが適切ではない場合があること，および，構文解析器を CoreNLP[(2)]へ変更したことによりタグ付け条件の変更が必要であることがわかった．

これらの問題を解消するため，次のように変更を行った．前者については，最低限の条件を満たしたことを条件分岐で確認したうえで，さらに，順番通りに出現したもののみをタグ付けするよう条件文を追加した．後者については，構文解析器の出力結果で得られた構文によってもタグ付けが行えるように条件に加える形で変更を行った．変更後の本プログラムを実行した実行例を図 3 に示す．このパラグラフでは，First, Second, Third, Fourth が出現しており，それらに<Word for enumeration>とタグ付けされていることがわかる．

5. おわりに

本研究では，順序のディスコースマーカーである Word for enumeration というパラグラフ構成要素を対象とし，同定機能の改良を行った．順序のディスコースマーカーにはいくつかのパタンが存在するものの，本研究ではそれらを区別せずに統一的に同定する方針で実現した．今後は，それらのパタンを考慮し，同定条件を細かく設定する予定である．また，本機能の評価を行う予定である．

参考文献

(1) 坂本洵一郎:“構造を用いた英語エッセイ検索システム ～パラグラフの構成要素同定機能の評価～”, 2018 年度九州工業大学修士論文 （2019）
(2) Stanford NLP Group: CoreNLP, https://stanfordnlp.github.io/CoreNLP（参照 2023/1/29）