毎日新聞データ集の整形とそのデータベース化について

情報科学科 平野 那月　　指導教員：山村 毅

1 はじめに

自然言語処理技術などの研究には，新聞記事のような大量の言語データが必要である．新聞記事は各出版社から紙面だけでなくフルテキストで書かれた新聞記事データ集としても提供されており，特に毎日新聞データ集は，様々な研究機関で用いられてきた実績ある新聞記事データ集である [1]．しかし，毎日新聞データ集には，タグや記号などが数多く挿入されている．また，空白文字や括弧，読み仮名，改行などに関して，様々な問題が指摘されている [2]．そのため，毎日新聞データ集を言語データとして利用するためには多くの前処理を行う必要があるため，扱いづらいといった問題点がある．そこで本研究では，毎日新聞データ集を扱いやすい形に整形し，整形した新聞記事を格納したデータベースの構築を行う．このデータベースにより，言語データとして新聞記事を用いた研究をする際の手助けを行うことを目的とする．以下，毎日新聞データ集を「データ集」，データベースを「DB」と略して記述する．

2 提案手法

2.1 毎日新聞データ集の整形

本研究では，データ集の 1994 年版～2010 年版を扱う．データ集は，下に示すように「＼タグ＼中身」の形式で書かれており，下記の例では “＼Ｔ１” は記事見出しに，“＼Ｔ２” は記事本文の段落に付されている．なお，「–略–」はその部分に該当する一部を略したものであり，実際にそのような記述はない．

毎日新聞データ集の一部

```
＼Ｔ１＼［飛べニッポン］第１部　高齢社会はこわくない／１（その１）
--略--
＼Ｔ２＼　--略-- まるで、西部開拓時代の「砦（とりで）」のように、外に身構えながら。【ワシントン・逸見義行】
＼Ｔ２＼　◇日本にも退職者の街
--略--
＼Ｔ２＼　＜３面に続く　次回からは３面に掲載＞
＼Ｔ２＼■写真説明　米アリゾナ州の砂漠の中に建設された緑と　--略--
--略--
```

データ集の整形は，1999 年版の紙面との照合調査の結果に基づいて行う．本研究では，記事見出しについては括弧の一部を含めた不要な記号の削除を行う．記事本文については，記号の種類や空白文字の有無などにより判別を行うことで，各段落に分類名を与える．この分類名とは，「本文」の他に，中見出しや小見出しなどの「見出し系」，表や補足情報などの「その他」などである．また，不要な記号，記者名，読み仮名なども削除する．

データ集は年によって微妙に書式が異なる，そのため，1999 年以外の年については，データ集を観察し，1999 年版のものと大きな違いが見られた部分に関してプログラムを書き換えることで，各年に対して適切な整形を行う．

2.2 データベース化

用いる RDBMS は SQLite である．SQLite は python の標準ライブラリであるため，sqlite3 モジュールをインポートするだけで利用できる．本研究では，データ集を整形したものと，記事に関する様々な情報を基本のテーブルとして DB に格納する．この基本のテーブルとは，

- 記事の情報 { ID, 月, 日, 番, 朝/夕刊, ページ, 掲載面 }
- 記事見出し { ID, 見出し }
- 記事本文 { ID, 段落 ID, 分類別, 中身 }
- 写真や図表の説明 { ID, 種別, 説明 }

の 4 つである．また，形態素解析の結果，依存構造解析の結果などが格納されたテーブルも，それぞれオプションとして追加できるようにする．このとき，データ集について指摘されていた問題である空白文字や改行を削除してから解析を行うことで，正しい結果が得られるようにした．

3 結果

作成した DB のうち，2.1 であげた例と同じ部分が格納されているデータについて，記事見出し，記事本文，写真や図表の説明についてのテーブルのデータを取得した結果の一部を下に示す．

構築した DB からデータを取得したときの一部

```
sqlite> select * from Headline where id=4;
id  title
4   飛べニッポン　第１部　高齢社会はこわくない１（その１）
sqlite> select * from Body where id=4;
id  paragraph_id  class    contents
--略--
4   14            本文       --略-- まるで、西部開拓時代の「砦」のように、外に身構えながら。
4   15            見出し系   日本にも退職者の街
--略--
4   19            その他     ３面に続く　次回からは３面に掲載
sqlite> select * from Chart where id=4;
id  type  caption
4   写真  米アリゾナ州の砂漠の中に建設された緑と --略--
--略--
```

データ集と DB を比較すると，記事見出しは記号の一部が，記事本文は「とりで」と書かれた読み仮名，“【】” 内に書かれた記者名，記号 “◇”，“＜” と “＞” が，DB ではそれぞれ削除されていることがわかる．また，記事本文について，DB ではそれぞれの段落に本文，見出し系などの分類名が与えられており，写真の説明については別のテーブルに格納されていることがわかる．

4 おわりに

データ集を扱いやすい形に整形したものがあれば，形態素・構文解析に先立って，タグや記号を削除したり，空白文字や括弧の問題を解決したりするなど，多くの前処理を行う必要がなくなる．また，DB にすることで，より複雑な検索をすることができ，検索や解析を行なった場合にも，より正確な結果が得られる．以上より，言語データとして新聞記事を用いた研究をする際に，データ集を整形したものが格納された DB があれば，研究の手助けをすることができると考える．

参考文献

[1] 日外アソシエーツ株式会社：“学術研究・開発研究のための言語資源コーパスのご案内”，2021-02，http://www.higashiyama.city.nagoya.jp/01_annai/.（参照 2021-11-13）

[2] 長谷川守寿：“新聞記事データ集はどれほど新聞紙面に忠実か”，言語処理学会第 17 回年次大会発表論文集，2011，p340-343.