

ブラウザ上でユーザが編集可能な言語パターンマッチシステムの構築 Building a user-editable language pattern matching system in the browser

桂 辰弥¹⁾ 竹内 孔一²⁾
Tatsuya Katsura Koichi Takeuchi

1 はじめに

テキスト中の特定のフレーズや表現を見つけることは、言語および教育分野において必要となることがある。テキストデータから特定のキーワードやフレーズの出現位置や文脈を抽出するためのプログラムとしてコンコーダンサがある。テキストコーパス内の文字列や単語を検索し、検索単語を中心として、前後の文脈とともに示される KWIC 形式のように表示するコンコーダンサは既に数多く存在し [1][2]、これらを用いることで語彙や文法の理解、単語の使用法や文脈の把握に役立ち、学習者の語彙や表現力の向上に役立つ。パターンマッチングはテキストの表層で検索を行う正規表現とは異なり、情報を抽出したい文を対象に予め関係する文や文の一部に対応する文構造のパターンを用意し、そのパターンに合致する結果を取得するものである。有名なコンコーダンサの例として、Sketch Engine¹⁾がある。Sketch Engine¹⁾はクエリ言語として CQL²⁾が使用されており、正規表現を用いて複雑な条件やパターンに基づいたデータの検索や抽出が可能である。他にも StruAP[3]があり、係り受け構造を利用し、部分木パターンによる関係抽出ができるツールである。しかしこれらのコンコーダンサやパターンマッチを行えるツールの多くは商用ツールであるため、利用しにくい。言語解析者や語学学習者が文構造を考慮してテキスト中の特定フレーズや表現を抽出するようなシステムを構築することは容易ではない。そこで本研究では解析モジュールで解析した結果をユーザ自身が求める表現をあらかじめ用意された検索ブロックで組み合わせるシステムに投入し、事例を検索できるシステムの開発を行っている。先行研究 [4][5]において WEB アプリケーションとして JavaScript と Python を利用した基本システムを構築したが、システムの本格利用にはいくつかの課題が残されている。そこで本報告では検索エンジンの中心部分である Prolog データベースの実装の改良、および、大規模なテキストが扱えるためにデータベースをシステムに導入したので、この改良について報告する。

2 提案するパターンマッチシステムの概要

本章では開発したパターンマッチシステムを構築する環境と実際のシステムの処理の流れについて述べる。

2.1 提案するパターンマッチシステムの構成

本システムは図 1 のようにユーザが視覚的に操作を行うフロントエンドシステムとユーザが要求したテキストの処理バックエンドシステムに切り分けて構成している。フロントエンドシステムは JavaScript のライ

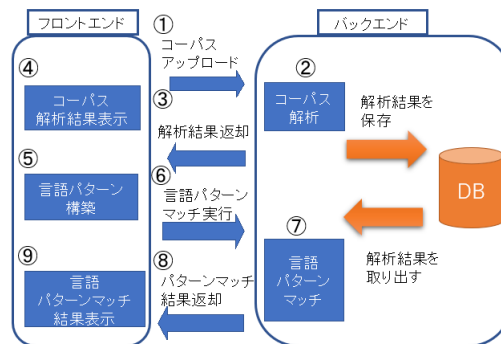


図1 システムの構成図

ブラリである React で構成されており主な機能としてはテキストファイルのアップロード、検索する言語パターン構築、解析結果の表示、検索結果の表示などがある。バックエンドシステムは Python の Web フレームワークである Django とデータベースシステムである Elasticsearch³⁾で構成されており、主な機能としてはテキストファイルの解析、検索クエリの言語パターンマッチ実行などがある。詳しい処理の流れについては以降の節で述べる。

2.2 バックエンドの処理の流れについて

バックエンドの処理の流れとしてテキスト解析、言語パターンマッチ実行の処理についてそれぞれ説明する。

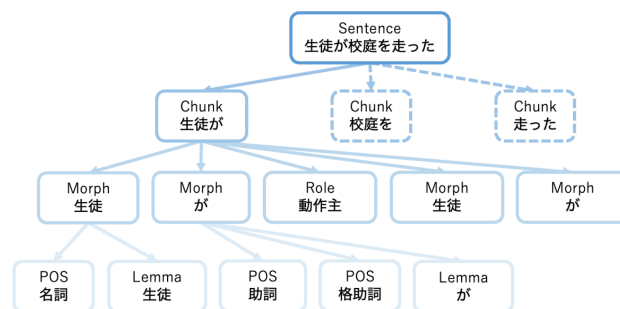


図2 文を解析した木構造の例

表1 Prolog の述語一覧

述語	第1引数	第2引数	第3引数
chunk(, 0,)	文番号	0 固定	文節 ID
morph(, ,)	文番号	文節 ID	形態素 ID
main(, ,)	文番号	文節 ID	主形態素
part(, ,)	文番号	文節 ID	副形態素
role(, ,)	文番号	文節 ID	意味役割
semantic(, ,)	文番号	文節 ID	概念
surf(, ,)	文番号	ノード ID	表層
surfBF(, ,)	文番号	形態素 ID	基本形
sloc(, ,)	文番号	文節/形態素 ID	文中出現位置
pos(, ,)	文番号	形態素 ID	品詞
dep(, ,)	文番号	文節 ID	係り受け文節 ID

1) 岡山大学大学院環境生命自然科学研究科 Graduate School of Environmental, Life, Natural Science and Technology, Okayama University

2) 岡山大学学術研究院 Academic Research Assembly, Okayama University

1) <https://www.sketchengine.eu/>

2) <https://www.sketchengine.eu/documentation/corpus-querying/>

3) <https://www.elastic.co/jp/elasticsearch/>

テキスト解析の処理はユーザがテキストファイルのアップロードを行うことで実行される。送られたテキストファイルの文に ASA⁴⁾を用いて形態素解析と項構造を適応させる。解析した結果の文の木構造を図2に示す。その後 Prolog 述語に変換を行い、データベースに保存する。データベースに保存する際には、1文ごとに対応した解析データを保存している。変換する Prolog 述語は以下の表1のように定義している。ASAで解析したデータがJSON形式であるため、Elasticsearchをデータベースとして活用することで、JSON形式の大規模な解析データの柔軟な取り扱いや高速な検索、リアルタイムな更新、スケーラビリティ、高度な分析など、データベースとしての利点を最大限に生かすことができる。言語パターンマッチの処理はフロントエンドからユーザが構築した言語パターンが送られると、データベースから各文に対応する Prolog データを取得し、1文ずつ Prolog データと検索クエリでパターンマッチを実行し、マッチ解の生成を行っている。パターンマッチを実行する Prolog 処理系として SWI-Prolog の Python モジュールである pyswip を使用している。Prolog 処理系として採用した SWI-Prolog⁵⁾はCで書かれた高機能の Prolog 処理系であり、SWI-Prolog の強力な論理プログラミング機能と Python のデータ処理能力を組み合わせることで、高速でより拡張性の高い環境が提供される。

2.3 フロントエンドの表示機能について

次にフロントエンドでの解析結果、検索クエリとなる言語パターンの構築、検索結果の表示について説明する。

テキストの解析結果

文をクリックすることで、アップロードされたテキストの解析結果を確認できます。

	ASA	PROLOG	PROLOG木構造
生徒が校庭を走った			
有川浩が図書館戦争を書いた			
彼は本を買ったけど僕はその本を売った			
又吉直樹が火花を書いた			
尾田栄一郎がワンピースを書いた			
僕は日記を書いた			
泥棒に財布を盗まれる			
昨日友達と喧嘩した			

図3 解析結果の例(ASA)

フロントエンドはテキスト解析の処理終了後、解析結果をデータベースから取得して表示できる。以下の図3,4は解析結果の表示例である。1文ごとに対応した解析データを保存するようにデータベースに保存しているので、解析データの表示の際には文をクリックするたび

4) https://github.com/Takeuchi-Lab-LM/python_asa/

5) <https://www.swi-prolog.org/>

テキストの解析結果

文をクリックすることで、アップロードされたテキストの解析結果を確認できます。

	ASA	PROLOG	PF
生徒が校庭を走った			
有川浩が図書館戦争を書いた			
彼は本を買ったけど僕はその本を売った			
又吉直樹が火花を書いた			
尾田栄一郎がワンピースを書いた			
僕は日記を書いた			
盗まれる			
昨日友達と喧嘩した			

図4 解析結果の例(Prolog)

にその文に対応する解析データを取得するように実装している。

```

を格と動詞 (SENTENCE_ID, Wo_sloc, (Verb_slock):-
chunk(SENTENCE_ID,0,(Wo_chunk_id))
and
part(SENTENCE_ID,(Wo_chunk_id),を)
and
sloc(SENTENCE_ID,(Wo_chunk_id),Wo_sloc)
and
chunk(SENTENCE_ID,0,(Verb_chunk_id))
and
morph(SENTENCE_ID,Verb_chunk_id,Verb_morph_id)
and
pos(SENTENCE_ID,Verb_morph_id,(動詞))
and
sloc(SENTENCE_ID,(Verb_morph_id),Verb_slock)

```

図5 「ヲ格と動詞」の言語パターン

言語パターンを構築する際には Blockly⁶⁾を用いて生成できる。前節で示した Prolog 述語のブロックをユーザが自ら組み合わせることで、複雑な検索クエリを構築することができる。図5はその例である。パターンマッチ実行後、バックエンドから検索結果を受信し、結果の表示の際には KWIC, テーブル, 強調の3つの表示形式を用いることができる。以下の図6,7,8は図5の検索結果の表示例である。引数 Wo_Slock, Verb_Slock は文中での出現位置を示しており、検索パターンの引数に _slock を含めると表示形式で強調する要素を選択できるようになる。図6は「ヲ格」、図8は「動詞」の要素を強調して表示し、視覚的にわかりやすくなっている。

3 動作評価実験

システムの動作評価実験を行い、パターンマッチシステムの処理性能の向上の確認を行う。

3.1 実験内容

表2に示すテキストファイルを用意し、テキスト解析とパターンマッチを行い、これらの処理時間を先行研究のシステムと提案するパターンマッチシステムでそれぞれ

6) <https://developers.google.com/blockly?hl=ja>

表示形式 KWIC	キーワード Wo_sloc	
検索結果の表示形式を指定します。		
-----	キーワード Wo_sloc	-----
生徒が	校庭を	走った
有川浩が	図書館戦争を	書いた
彼は	本を	買ったけど僕はその本を売った
彼は	本を	買ったけど僕はその本を売った
彼は本を買ったけど僕はその	本を	売った
彼は本を買ったけど僕はその	本を	売った
私は誰かに	パソコンを	盗られました
私は誰かに	パソコンを	盗られました

図 6 KWIC 表示

表示形式 テーブル		
検索結果の表示形式を指定します。		
	SENTENCE_ID	Wo_sloc
		Verb_sloc
0	詳細	0
1	詳細	1
2	詳細	2
3	詳細	2
4	詳細	2
5	詳細	2
6	詳細	3
7	詳細	3

図 7 テーブル表示

表示形式 強調	キーワード Verb_sloc	
検索結果の表示形式を指定します。		
キーワード Verb_sloc		
生徒が校庭を走った		
有川浩が図書館戦争を書いた		
彼は本を買ったけど僕はその本を売った		
彼は本を買ったけど僕はその本を売った		
彼は本を買ったけど僕はその本を売った		
彼は本を買ったけど僕はその本を売った		
私は誰かにパソコンを盗られました		
私は誰かにパソコンを盗られました		
先生に褒められました		

図 8 強調表示

表 2 ファイルサイズ(バイト)

文の数	ファイルサイズ
1	46
10	440
100	5,342
1000	53,248
5000	262,199
10000	524,399

れ計測した。具体的にはフロントエンドからバックエンドに送信し、バックエンドからデータが返ってくるまでを処理時間として計測する。これらの処理時間は Chrome のデベロッパーツールを用いて計測を行う。検索クエリは表 5 の「ヲ格と動詞」の言語パターンを用いる。

3.2 実験結果

表 3 テキスト解析の処理時間(秒)

文の数	先行研究のシステム	提案するシステム
1	0.083	0.144
10	0.414	0.453
100	3.50	3.56
1000	35.3	36.4
5000	174	192
10000	計測不能	768

表 4 パターンマッチの処理時間(秒)

文の数	先行研究のシステム	提案するシステム
1	0.110	0.242
10	0.440	0.495
100	8.15	1.65
1000	960	30.9
5000	計測不能	78.4
10000	計測不能	160

パターンマッチの動作評価実験の結果をそれぞれ表 3, 4 に示す。テキスト解析, パターンマッチ実行はともに文の数が増えるにつれ, 処理時間も増加していることが読み取れる。

テキスト解析の処理時間については先行研究のシステムは 10000 文のテキストファイル进行处理する際には動作が止まってしまったが, 提案するシステムでは 10000 文でも動作が確認できた。提案するパターンマッチシステムは先行研究のシステムに比べ, 少し大きくなっているが微差である。

パターンマッチシステムの処理時間については先行研究のシステムは 5000 文のテキストファイル进行处理する際には動作が止まってしまったが, 提案するシステムでは 10000 文でも動作が確認できた。提案するシステムは 10000 文のテキスト解析, パターンマッチ実行がともに動作を確認できた。また 100 文以上のパターンマッチの処理時間が向上しており, 全体的にシステムの処理性能の向上が確認できた。

ただし 10000 文の処理を行った際にはブラウザの挙動が重くなっており, ユーザが操作可能であるとは言い難い。

4 考察

テキスト解析の処理時間は先行研究のシステムに比べ, 少し遅くなったが, これはデータベースを 1 文ごとに対応した解析データを保存するように改良を行ったためであり, 10000 文を解析を行った際には 10000 個の解析データを保存する必要があるため, バックエンドでの Django とデータベースシステムとのやりとりの時間が増加してしたためである。今回の実装では 1 文ずつ Prolog データベースに追加し, パターンマッチを生

成したが、これは全文の Prolog データベースの処理を行うファイルの生成に時間がかかるため、1 文ずつ生成するように変更した。このことにより処理速度が向上したが、これは全文の Prolog データを記載した大規模なファイルを生成する必要があったためである。

しかしこの実装には問題点として、例えば 10000 文のパターンマッチを行う際には 10000 回処理を行う Prolog ファイルの生成を行う必要があり、バックエンドシステムへの負荷が大きい、そのためファイル生成を行わずにパターンマッチを実行する実装方法を検討する必要がある。Elasticsearch は、1 度に取得することができるドキュメントの最大件数は、デフォルトでは 10,000 件であるためであり、10000 文以上のテキストファイルの解析は現状のシステムでは不可能である。また提案するパターンマッチシステムは 10000 文まで解析やパターンマッチが可能となったが、さらに解析後のブラウザの挙動がかなり重くなっており、ユーザが利用可能とは言い難い。今後さらに処理性能の向上させるには、パフォーマンスの問題やネットワークの制約などに留意して実装する必要がある。

現在の実装では、柔軟なパターンマッチを行うためには、ユーザが Prolog に精通している必要がある。より簡単な操作を実現する方法としては表層に“*”のような曖昧性を含んだ正規表現を持つクエリと Prolog のパターンマッチの実装が考えられる。今回 Prolog 処理系として導入した SWI-Prolog は正規表現に対応可能な regex パッケージが存在するため、今後の課題として正規表現マッチのシステムの導入が考えられる。

5 まとめ

本研究では、先行研究のシステムをベースに開発を進め、検索エンジンの中心部分である Prolog デー

タベースの実装の改良として、Prolog 処理系として SWI-Prolog を導入し、1 文ずつ Prolog データと検索クエリでパターンマッチを実行し、マッチ解を生成するように実装した。また、大規模なテキストが扱えるためにデータベースとして Elasticsearch をシステムに導入した。実装した機能の処理性能を確認するための動作評価実験を行い、10000 文のテキストの動作することが確認できた。また今後の課題としてさらなる処理性能の向上と正規表現マッチの追加が考えられる。

謝辞

本研究の一部は、日本学術振興会科学研究費補助金(助成番号 22K00530)の助成を受けています。

参考文献

- [1] 中條清美, 西垣知佳子, アントニ・ローレンス. フリーウェア WebParaNews オンライン・コンコーダンサーの英語授業における活用. 日本大学生産工学部研究報告 B (文系), Vol. 47, pp. 49–63, 2014.
- [2] 小木曾智信, 中村壮範. 通時コーパス用『中納言』: Webベースの古典語コンコーダンサー. 第2回コーパス日本語学ワークショップ, 2012.
- [3] Kohsuke Yanai, Misa Sato, Toshihiko Yanase, Kenzo Kurotsuchi, and Yoshiki Niwa Yuta Koreeda. StruAp: A tool for Bundling Linguistic Trees through Structure-based Abstract Pattern. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 31–36, 2017.
- [4] 岡田魁人. Blockly を利用した言語パターンマッチシステム ASAGAO の構築. 修士論文, 岡山大学大学院自然科学研究科, 2022.
- [5] 小笠原崇. 意味役割付与とテキストに対する Prolog パターンマッチシステムの構築. 修士論文, 岡山大学大学院自然科学研究科, 2022.