テキスト自動抽出手法による外国人留学生向け大学受験支援アプリの開発

情報科学科　李　思遠　　指導教員：小栗　宏次

1　はじめに

日本の大学の受験を希望する外国人留学生の受験困難要因として，有用な情報を効率的に入手する方法がないことや日本語教育機関等による受験サポートの不足が挙げられる．留学生向けの適切な情報提供のためには Web に散在している入試情報を自動的に収集し，網羅的かつ正確なデータを提供することが必要である．そこで本研究では，テキスト自動抽出により入試情報を自動収集し，スマートフォンアプリによってその入試情報を提供する受験支援を実現した．

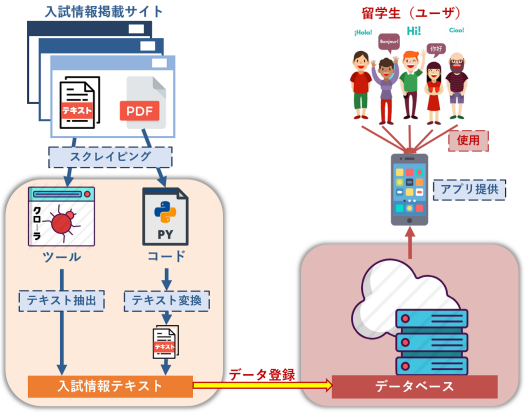

図 1　受験情報の収集方法と支援の流れ

受験情報の収集と受験支援の流れを図 1 に示す．掲載方法に応じた収集作業の後，入試情報データベースを構築する．そして受験支援アプリを通じて入試情報を提供する．

2　受験情報の自動収集

2.1　Web クローリング

Web 入試情報は HTML テキスト形式と PDF ファイル形式の 2 種類ある．HTML テキスト形式の情報収集にはスクレイピングツール Houyi を用いる．一方，PDF ファイルの情報取集にはクローラを作成し，一旦ローカルに保存し，その後 PDF ファイルのテキスト化を行う．ツールのスケジュール機能及びプログラムの周期的実行によりデータ収集を自動化する．

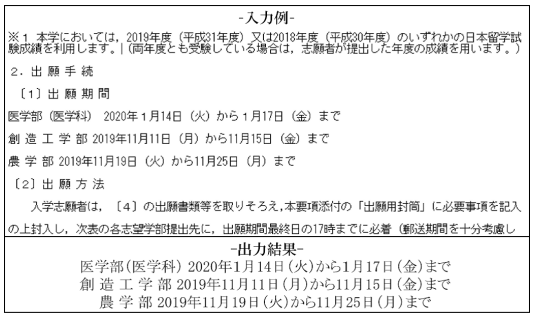
-入力例-
※１．本学においては，2019年度（平成31年度）又は2018年度（平成30年度）のいずれかの日本留学試験成績を利用します。|（両年度とも受験している場合は，志願者が提出した年度の成績を用います。）
２．出 願 手 続
〔1〕出 願 期 間
医学部（医学科） 2020年１月14日（火）から１月17日（金）まで
創 造 工 学 部 2019年11月11日（月）から11月15日（金）まで
農 学 部 2019年11月19日（火）から11月25日（月）まで
〔2〕出 願 方 法
入学志願者は，〔4〕の出願書類等を取りそろえ,本要項添付の「出願用封筒」に必要事項を記入の上封入し，次表の各志望学部提出先に，出願期間最終日の17時までに必着（郵送期間を十分考慮し
-出力結果-
医学部(医学科) 2020年1月14日(火)から1月17日(金)まで
創 造 工 学 部 2019年11月11日(月)から11月15日(金)まで
農 学 部 2019年11月19日(火)から11月25日(月)まで

図 2　正規表現によるパターンマッチング例

2.2　パターンマッチングによるテキスト抽出

正規表現は文字列を 1 つのパターン化した文字列で表現する表記法であり，パターン化した文字列にマッチする内容を抽出するのに用いる [1]．使用頻度の高いメタ文字は特殊シーケンスであり，先頭に \（バックスラッシュ）を記載する．例えば，出願期間に関する情報を取り出す正規表現式は次のようになる．

```
\W*[学部|学科]*\d*年\d*月\d*日\S*[から|まで]*
```

この正規表現式により抽出した結果の一例を図 2 に示す．入試情報を格納するデータベースには Firebase を用いる．アプリからデータの読み書きを行うため JSON ファイルでデータを構造し，入試情報データを以下の関数でやり取りを行う．

```
URLSession(configuration: delegate: delegateQueue)
JSONDecoder().decode(ResultJson.self, from: data!)
```

3　スマートフォンアプリでの情報提供の有効性

日本語学校の外国人留学生 12 人を既存の情報サイト利用（グループ A），検索エンジンと SNS 利用（グループ C），本アプリ利用（グループ C）の 3 群に分けて、入手できた情報を比較した．入手対象の大学は条件 1：英語で学位を取れる大学（4 校），条件 2：学費全額免除できる私立大学（4 校），条件 3：日本留学試験が不要な国公立大学（4 校），条件 4：書類選考だけ受験できる国公立大学（4 校），条件 5：合格率ほぼ 100 %の国立理工系大学（1 校）とした．表 1 に検証結果を示す．

表 1　各条件に対した検索結果の完成度 [%]

条件	グループ A	グループ B	グループ C
条件 1	100	100	100
条件 2	25	25	100
条件 3	0	0	100
条件 4	0	0	100
条件 5	0	0	100

この結果より，本アプリの有効性が示唆される．また，アンケート調査では本アプリが受験情報収集に有用との意見が多かったが，検証実験の対象とした学校数が少なかったためより多くの学校を対象にしたいという意見があった．

4　おわりに

本研究では，入試情報を提供するための受験支援を実現した．スクレイピングによる情報収集とパターンマッチングによるテキスト自動抽出手法で有用な情報を抽出でき，アプリで情報を提供することの有効性を示した．課題として，テキスト抽出の手法を検討し完成度の高いデータベースの構築が挙げられる．

参考文献

[1] 今岡ら．動向情報抽出における相対表現の利用効果に関する考察，知能と情報，Vol.18，No.5，pp735-744，2006