1ZK-06

日本語学習者のための習熟度に合わせたドラマ推薦システム

劉　雨鑫†　　大寺　亮†
神戸情報大学院大学　情報技術研究科†

1. はじめに

近年のインターネットおよび動画配信サービスの普及に伴い，日本のアニメやドラマ，漫画などをきっかけに日本や日本語に興味を持つようになった外国人は多い．また，留学生を筆頭に日本語を学ぶ外国人は年々増え続けている．しかし，アメリカ国務省の「外国語習得難易度ランキング」というデータにおいて，日本語は最強難易度の「カテゴリー5＋」とされており，仮名文字や漢字，同音異義語の多い日本語は日本語を学習する外国人にとって，難しいと言われている．その一方で，興味を持つきっかけともなるドラマやアニメ，漫画は，日常会話で頻繁に使用される日本語の話し言葉表現が多く含まれており，日本語の教材としての可能性が指摘されている[1]．そこで，本研究においても，ドラマ視聴を中心とした日本語学習に着目する．

ドラマを使った日本語学習支援技術の例として，ドラマ内の字幕に対し，わからない単語を辞書引きできる機能などが既に実装されている．これは単語の学習には有効な手段と言えるが，言語学習は自身の習熟度に合わせて段階的に行うことが重要であると考えられる．そこで，本研究では，インターネットで動画配信されている日本語ドラマやアニメに関して，それらの中で実際に使用されている日本語の難易度を分析・分類し，自分の日本語習熟度に応じたコンテンツを推薦する日本語支援システムを提案する．

2. 提案手法

動画配信サービスとして，2 社のサービスを利用する．字幕は各配信サービスの字幕機能をそのまま利用する．

Japanese TV Series Recommendation System for Japanese Language Learners According to User's Language Proficiency Level
†Liu Yuxin, Ryo Ohtera, Department of Information Systems, Graduate School of Information Technology, Kobe Institute of Computing

提案手法の簡易フローチャートを図 1 に，提案するアプリの画面を図 2 に示す．

はじめに，動画内の字幕から文字抽出を行う．次に，得られた文章から日本語の難易度を解析する．種々のドラマに対して上記の処理を行い，Google 社が提供する Web ベースの表計算ソフトであるスプレッドシートを利用し，ドラマ難易度の簡易データベースを構築する．最後に，Web アプリとしてスプレッドシートと連携したドラマ推薦システムを構築する．

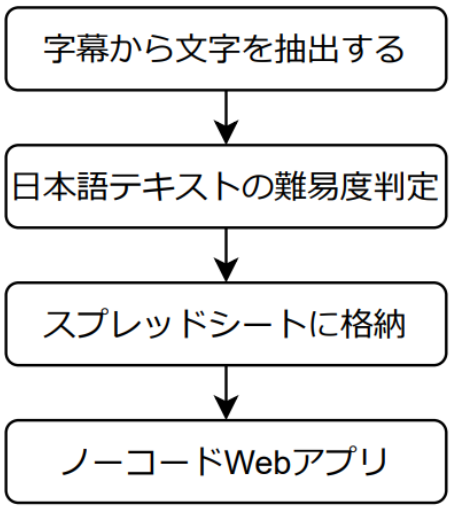

図 1　提案手法の簡易フローチャート

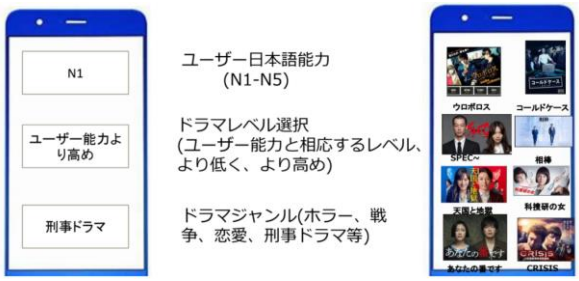

図 2　Web アプリの画面例

2.1 文字抽出部

字幕からの文字抽出は，Google 社が開発したオープンソースの OCR 文字認識ソフトである Tesseract-OCR を利用する．Tesseract-OCR は，基本的には文書画像などを処理することを想定しているため，字幕抽出のためにパラメータの調整を行った．具体的には，レイアウトパラメータと，辞書データの変更である．また，白い紙に文字が書かれている文書とは違い，字幕には背景が存在するため，影響を除去する工夫を行った．まず，OCR をかける領域を，映像の下部に絞りこむ．次に，字幕の文字は輝度値が非常に高いこと，ドラマの背景には一定以上の輝度の物体が映りこまないことに着目し，輝度値が 251 以下は背景として除去した．なお，字幕の抽出に関しては，毎フレームに対して行う必要はないため，動画の全体の長さに合わせて取得間隔を調整した．

2.2 日本語の難易度判定

日本語の難易度判定には，名古屋大学が開発した日本語テキストの難易度判定ツールである「帯」を利用した[2]．帯は難易度の規準として，小中高大の教科書から抽出したサンプルを利用しているため，難易度判定結果も小学校～大学・一般までの 13 段階で表示される．しかし，本研究は日本語学習者を対象としているため，帯の結果を，日本語学校の教員の協力のもと，日本語検定試験によるレベル分類である N1～N4 に再分類した．

帯による日本語難易度と日本語検定レベルの相関表を表 1 に示す．なお，日本語検定レベルは N5 まで存在するが，N5 レベルは，日本語の初学者であり，ドラマを見ることが難しいため除外している．

表 1 難易度対応表

帯難易度	日本語検定レベル
高校 1 年，2 年，3 年，大学 中学 1 年，2 年，3 年	N1
小学 5 年，6 年	N2
小学 3 年，4 年	N3
小学 1 年，2 年	N4

2.3 日本語難易度データベース

Google スプレッドシートに前節で得られた日本語検定による日本語難易度（N1～N4）と，ドラマ（アニメや映画も含む）のタイトル，ジャンルなどの情報をまとめる．

2.4 Glide によるアプリ作成

本研究では，ドラマ推薦システムを Web アプリとして構築する．そのために，コードを書かずにアプリを作成するノーコードツールである Glide を利用した．

3. 実験

本研究では，実験としてジャンルごとの難易度の分析を行った．動画配信サービスにおけるジャンル分けをそのまま利用し，「(a) ヒューマンドラマ」，「(b) サスペンス・ミステリ」，「(c) 犯罪」，「(d) アニメ」，「(e) ホラー」，「(f) コメディ」，「(g) ラブロマンス」，「(h) ドキュメンタリー」など 8 個のジャンルに分類し，77 作品について難易度分析を行った．結果を表 2 に示す．基本的には N2 レベルの作品が多くみられジャンルによる偏りは大きくない．ただし，「(c)犯罪」に関しては，作中の用語の難しさがしっかりと反映され，N1 レベルの割合が非常に高い結果となった．

表 2 各ジャンル難易度分析結果本数

	N1	N2	N3	N4
(a)	5	26	4	0
(b)	2	6	0	0
(c)	4	2	1	0
(d)	4	12	0	0
(e)	2	4	0	0
(f)	2	5	0	0
(g)	1	3	0	0
(h)	0	2	0	1

4. おわりに

本研究では，外国人の日本語学習者に対して，習熟度に合わせたドラマ推薦システム開発した．日本語の難易度判定はあくまでテキストレベルであったため，声の小ささや，抑揚，なまりなど発声による難易度は判定できていない．日常会話における会話の難しさを解決するためには，このような話し言葉の個性も今後加味して難易度判定する必要がある．

参考文献

[1] 岡崎正道“ドラマ・漫画による日本語教育，”岩手大学人文社会科学部紀要，vol.53, pp.39-53, 1993.

[2] S. Sato, S. Matsuyoshi and Y. Kondoh,“Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus,” Proceedings of the Sixth International Conference on Language Resources and Evaluation, pp.654-660, 2008.