Wikipedia情報集収による可読性向上のための機械学習的要約手法の開発

指導教員　西田泰伸

電子・情報工学科　小野田　成晃

もくじ

1．研究背景・対象

研究背景

- ・情報爆発により、ウェブからコアな情報を得ることが難しくなっている
- ・多くの記事が外国人や子ども等の語彙レベルの人に配慮していない

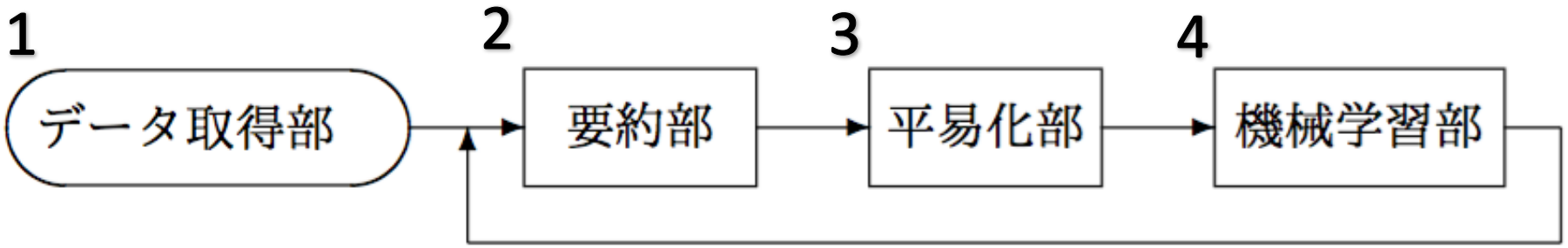
研究目的

- ・日本語学習者や子どもなどに対しても理解しやすい文書を提示したい
- ・ユーザのアクティビティを反映したシステム開発を行う

→ユーザの性質に合わせた文章を自動生成することを目標とする

2. 開発するシステムの概要

システム全体のフロー



1　システム全体のフロー図

1.文章の取得・抽出
2.要約・文章処理
3.単語レベル設定・シソーラス適用
4.ユーザによるアクティビティ（満足度・不満度）の学習

以上4つの部にわけられる複合システムである

各システム構築のための基盤技術

データ取得部

- ・**DBPedia**： Wikipediaの情報を格納した**Linked Open Data**、SQLライクな言語(**Sparql**)で問い合わせできる

平易化部

- ・**日本語ワードネット（WN）**: 日本語の語同士の関係を格納した木構造データベース
- ・**日本語教育語彙表**：李らが開発した、日本語検定出現単語から単語のレベル付けした辞書

文章の難易度の定義

定義

要約率（元文からの何%が残っているか）が低いと文字数は減って文章がシンプルになるので平易になる　厳密には逆要約率

文字難易度：高いほど難しい（level.1-7)

- *日本語教育語彙表をもとにレベル付

アクション

難解化→要約率UP、文字難易度UP

平易化→要約率DOWN、文字難易度DOWN

3. 要約部

自動要約部における方針

- 従来：自動要約では指示的・報知的という目的に応じた手法は存在した
- 動的な要約手法もあるもののuser-queryを対象として行われている（Dragonmir R. Radevら 2000）
- しかし、ユーザ個々の性格や理解度を想定した文章要約事例は少ない（例：品川らによるユーザプロファイルによるhtmlページ自動生成など）

- 本研究：自動要約・語彙的換言と強化学習を組み合わせて、ユーザに最適な要約文を提示する

自動要約の手法

・本研究では不自然な文章になりにくい抽出的要約を行う

・要約モデルとして、**You Ouyang**らが考案した**Basic Summarization Model**を導入する

2. 平易化部

平易化（語彙的換言）

平易化の手順としては、

1. ユーザレベル（語彙の理解度）取得＊以後ユーザレベルと呼ぶ
2. 日本語教育語彙表を用いて、要約文中の語の難易度判定
3. ユーザレベルから変換候補の単語をマップに格納
4. WNデータベースから変換候補の単語の類語抽出（難易度を判定も）
5. ユーザレベルに応じて変換候補のなかから適切なレベルの単語を換言

2. 機械学習部

アクティビティ反映システム

最もシンプルなアクティビィティ反映法

下の図のように要約率と語彙レベルをパタン化しておく

レベル	1	2	3	4	5	6	7
要約率	10	25	40	55	70	85	100

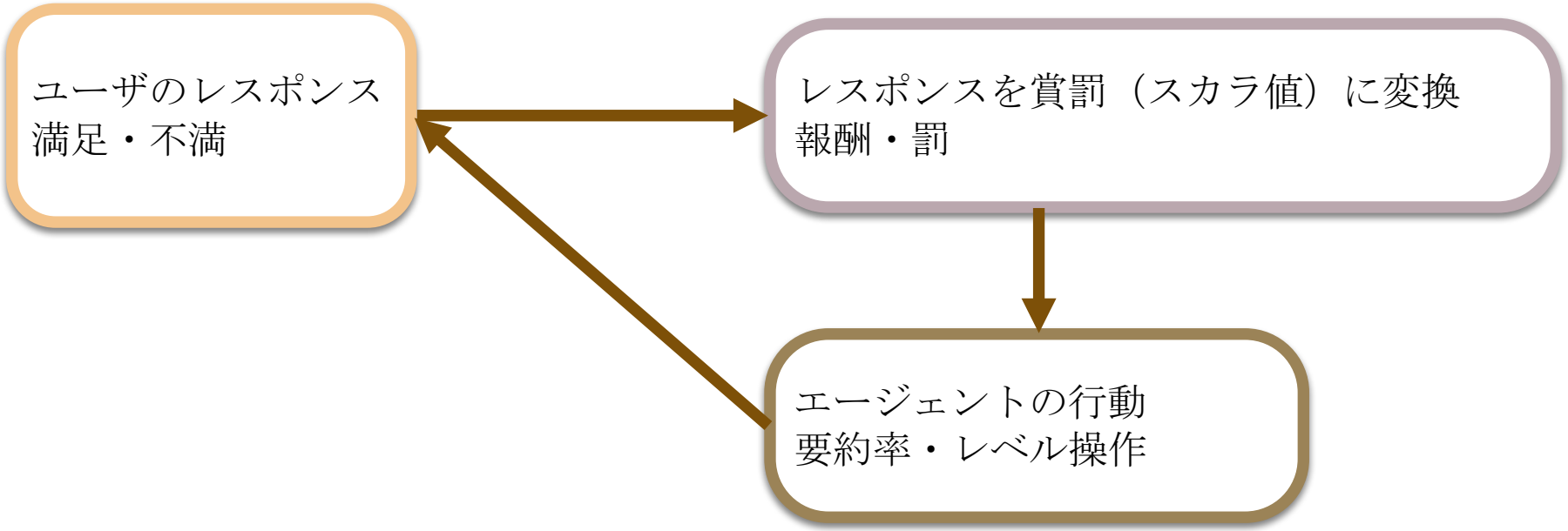
→実装がシンプルだがユーザの細かな要望に対応できない

- 例：要約率は下げなくて良いので、語彙レベルだけ下げてほしい等

強化学習の環境設定

状態(State): 文字の要約率、ユーザレベル(ユーザの語彙レベル)

行動(Action):



具体的な学習法

強化学習の中で代表的な手法であるQ学習(Q-learning)を用いる

- Q学習では報酬としての行動価値観数

$$Q(s_t, a_t) = R(t)$$

- 学習モデル

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma argmax Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right]$$

$r(t)$、$a(t)$、$s(t)$はそれぞれ施行回数tにおける報酬、行動、状態

αは学習率、γは割引率

具体的な学習法

前提

ε-greedy algorithmにより最初はランダムな行動を取りやすくし、行動の幅を広げておく

環境(s)

- 要約率(10~100)：レベル(1~7)
- アクション数（要約率,レベル）の組み合わせ4通り

目的

- Q(s,a)：91^2 *4 = 行8281:列4の33124パタンの行列
- このQ関数の行列から最適な戦略を探索（行列内は各報酬r(t)を格納）

研究の課題と考察

・生成文章について

・要約部で生成した要約文は自然な文章になっている。
・平易化部でWNを用いて換言するときに、おかしな変換候補が選択される
例：「現在」の株式市場→「プレゼント」の株式市場　等
・名詞→名詞というふうに同じ品詞にしか変換しない機構を組み込むが上のような例は修正できない。
・解決策として、別の変換用辞書を用いるか、全く別の変換手法をとる必要がある（梶原らが用いたパラレルコーパスの手法等）

・強化学習について

・概ねユーザの要望を学習しているが生成される文章が不自然なので、正しく計測できてはいない
→現在のパラメータがユーザレベル・要約率のみとなっているので増やすことで改善が見込まれる

ご清聴ありがとうございました