1–4　Wikipedia情報収集による機械学習的要約手法の開発

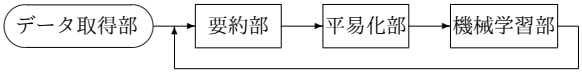
西田研究室
1415015 小野田 成晃

1. はじめに

WWW の発達により，Web 上の情報の量・種類ともに膨大となっている．そのため，Web 上でコアな情報を検索・収集することがより困難になることが危惧される．そこで，計算機を用いてユーザに適切な情報を提示することの重要性が増している．

そこで本稿では，日本語版 Wikipedia を対象として記事を自動要約して，ユーザの文章の理解度に合わせ，シソーラスを用いて要約文を平易化を行う．そしてその要約文に対してのユーザのレスポンスを学習データとして機械学習をして，最終的にユーザに適した文章を動的に提示する．これにより自然言語処理と機械学習的手法により，要約文の可読性を動的に調整する基盤技術を開発する．

2. システム全体の流れ



1　システム全体のフロー図

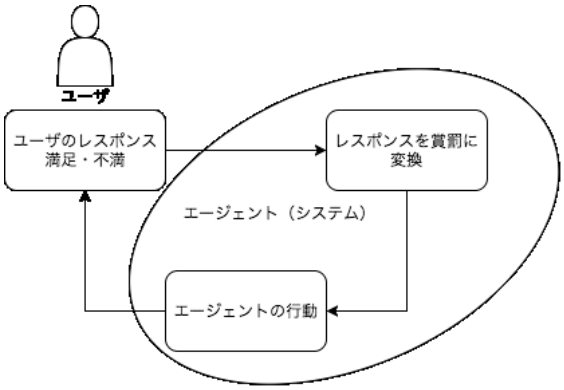
本システムでは，対象となる文章の取得・抽出，要約・文章処理，単語レベル設定・シソーラス適用，ユーザによるアクティビィティの学習と四つの部に分けられる複合システムである．そこでにおいてそれぞれデータ取得部，要約部，平易化部，機械学習部と位置づける．また，四つの部の実行フローを以下図 1 に示す．図 1 では機械学習部を実行した後，その結果を要約部以前にフィードバックする．

3. ユーザの文章理解度判定モデルの構築

まず，要約部では You らの Basic Summarization Model[1] という次式の要約モデルを用いて要約を行う．このモデルは従来は文章中の単語の出現頻度を表していたが，文章中の単語の位置も考慮したモデルとなっており，要約コンペティションでも高い精度をだしている．

$$score'(s) = \sum_i [\log freq(w_i) \cdot pos(w_i)]/|s| \qquad (1)$$

続いて，平易化部では梶原の研究を参考に難解語に対して，語彙的換言を行う．そして，この 2 つの部から生成された文章を提示し，ユーザのレスポンスを待つ．



2　強化学習の環境とエージェントのイメージ

最後に機械学習部では，図 2 のような報酬駆動システムを導入する．具体的には，ユーザのアクティビティを強化学習法を用いて学習する．そして，ユーザのアクティビティに応じてエージェントに賞罰が送られ，その結果から文章の要約率とユーザの語彙理解度レベル（ユーザレベル）を決定し再度要約文を生成する．そしてユーザが満足するまでこの工程を繰り返すことで最適な文章を探索する．

4. おわりに

本稿では，一連の分野を複合的に組み合わせることでの有効性を確認することに焦点をあてたが，システムの四つの部それぞれに対して，より細かなチューニングを行うことで，さらなる可読性の向上は期待できる．

また，今回は日本語を対象として行ったがシステム処理の内，言語特有の部分を切り離して設計することにより多言語の対応も可能になるだろう．また，機械学習で推定するパラメータに要約の手法を取り入れれば，文章の種類やドメインに応じた要約もできる可能性がある．以上より今後さらなる改良を検討したい．

参考文献

[1] You Ouyang, Wenjie Li, Qin Lu, Renxian Zhang, ”A Study on Position Information in Document Summarization”, Colin 2010: Poster Volume, pp.919-927, 2010.

[2] 梶原智之, ”語彙的換言を用いたテキスト平易化”, 第七回 NLP 東京 D の会.