Wikipedia情報収集による可読性向上のための機械学習的要約手法の開発

富山県立大学電子・情報工学科　　指導教員：奥原浩之
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。

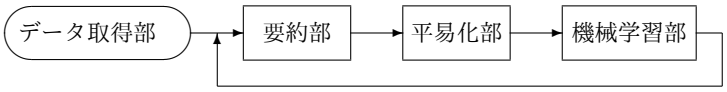
例えば、Amazon でのユーザにあった商品を提示するレコメンド機能、Google でのユーザの年齢層や地域によって検索結果を変えるパーソナライズ検索といったような試みが行われている。そしてこれらはユーザに合わせて提示するデータを動的に変える必要がある。また、ニュースの見出し電光掲示板など、限られた文字数で本文の内容を要約する技術も以前から必要性を唱えられている [1]。

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

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。

そして、以下の章ではその詳細な手法・設計について述べる。



1 システム全体のフローチャート

3 自動要約

3.1 自動要約の従来手法のまとめ

まず、自動要約は利用目的に応じて指示的要約と報知的要約の二つのタイプに分けられる。

指示的要約 (indicative)

原文を読者が読む必要があるかどうかを判断させるため材料としての要約文を生成する。

報知的要約 (informative)

原文の代わりとなる要約文を生成する。

次に自動要約を行う際に使われる諸手法を示す。まず、H.P.Luhn のテキストの重要文抽出法が有名である、この方法では、原文の中で語の頻度を高頻度、中頻度、低頻度と分けて中頻度が文中での重要語と位置付けその語が含まれる文を抽出する抽出法を示した [2]。重要語の判定では tf-idf 法が有名であり、これは出現頻度が高い単語ほど重要であるが、複数の文書で横断的に使用されている語は重要ではないとする手法である。

また、抽出法の一種で You Ouyang らが考案した重要語を重み付けするだけではなく、文中の重要語とその出現回数をペアにして重み付けするという手法も存在する [3]。しかしこれらの手法はニュースの見出しなど指示的な要約では効果を発揮できるが、文章自体の代わりとする報知的な要約では、重要文以外で捨てられる対象が文単位であるため、情報が大きく欠落する可能性がある [4]。

そのため、報知的要約では山本らが提案した文中ごと連体修飾要素や比喩表現等、文章内で重要でない節を削除する方法が提案されている [5]。また、商品レビューに対して文章内の商品の性質・意見をそれぞれ属性値・評価値と定義して 2 値分類の機械学習を行いの意見情報抽出する手法も取られている [6]。この他にも原文をオントロジや関係データベースを用いて意味関係を理解させ、再構築する生成的方法がある。これは報知的要約でも効果が期待できるが、抽出法に比べて実現方法が複雑である。

3.2 本研究での自動要約手法

本稿では要約の対象を Wikipedia の記事として、Wikipedia の記事を構造化データとして扱える DBpedia[1] を使用して記事の取得・抽出を行った。

そして、自動要約のタスクとして以下を設定した。

(1)　文章をより短縮する
(2)　文章の情報量を動的に変更する
(3)　文章の可読性（難易度）を動的に変更する

これらのタスクを達成するために、ユーザのアクティビティ（フィードバック）に合わせ文章の報知性と指示性を動的に適応させる手法を用いる。

つまり、要約文に対して報知的要素と指示的要素を期待するユーザがそれぞれいると仮定すると、機械学習を用いてそれぞれのユーザに適した文章を生成させる必要がある。ここでは要約の手法に焦点を合わせ、学習方法は x 章で述べることにする。

本稿ではまず、You Ouyang らが考案した Basic Summarization Model を元に抽出法により文章を要約する。これは Luhn が用いた方法では重要語が含まれる文章を均一に得点付けしていたのに対し、その重要語が文の何回目の登場かも判断基準に含める手法である。このモデルは以下の式によって表される。

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

式 (1) ではある一文で出現する単語 w について、w の何回目の出現順と $pos(w_i)$ と単語の位置の特徴を表す関数をかけることで文章の得点づけを行い、その文中に出現する単語それぞれに得点づけして、その合計点をその文自体の点数としている。

You Ouyang らの論文では、式 (1) の $pos(w_i)$ には四つの手法 $f(i)$ を明示的に選択するようにしていたが、ここでは一番効果が示された Binary function という手法を用いる。Binary function では w の一番目の出現では $f(i) = 1$, それ以外では $f(i) = \lambda$ を返す手法である。なお λ は小さい正の実数で今回は 0.001 とした。

この手法では単語の最初には得点を高くつけ、それ以外では、低い得点を与えることで、新しい単語が出現する文の方が重要性が高いとしている。

4 単語と文章の平易化

4.1 平易化の従来手法のまとめ

文章の平易化は、以下の三つの工程を通して行われる。

1 Syntactic Simplification

原文に対して、文圧縮と文分割を行う

2 Lexical Simplification

語句の言い換えやフレーズベース SMT(Statical Machine Translation) を実行

3 Explanation Generation

難易語に対して注釈をつける辞書引きを導入 [7]

まず、語句の言い換えでは難解な文と平易な文のパラレルコーパスを用いるのが実用的であるが、日本語のパラレルコーパスで一般利用可能なものは存在しない。そのため日本語に対しては、美野らのように、語句の難易度の基準について日本語能力試験 (JLPT) の見出し語に対して語釈文を置換単語の候補とする方法 [8] や梶原、山本のように日本語 WordNet 同義語データベースから語彙的還元を行い、述語項構造を用いる方法がある [9]。

梶原らの先行研究から平易化においては語句を置き換えるだけでなく、置き換え語の文章の校正も必要となることが理解出来る。そのため、完全な平易化には 2. の言い換えだけでなく 1,3 のような複数の手法・指標を用いる必要がある。

しかし、これらの研究ではシソーラス辞書のような膨大なデータ資源が必要である。それに対して梶原、小町が新たに考案したコーパスを用いないで単言語パラレルコーパス構築する手法 [10] により、データ資源が少ない言語でもに対しての平易化への敷居が下がったと考えられる。

4.2 文章の難易度とスコアリング方法本論では要約と平易化の複合タスクのため、前述した 1,3 の手法のように文の構成を変化させることは対象外とし、2 の言い換えのみに焦点を当てることにする。

そこで、本稿では梶原らの先行研究をもとに平易化部を作成する。

[1] ユーザレベル取得

[2] 語の難易度判定

[3] 類義語検索

[4] 類義語の難易度判定

[5] ユーザのレベルに適切な語をセット

2 平易化部のおおまかなアルゴリズム

なお本稿では、パラレルコーパスに載っていないレベル未定義語は言い換え対象外とした。まず、ユーザの理解度を参照して、要約文中の語句のレベルを日本語能力試験 (JLPT) から XX らが作成したデータから参照する。そしてその語句レベルのリストの中で、現在のユーザの文章理解度レベルと一致しないものを抽出。その後抽出した語句をユーザの（簡単・難しい・適切）のレスポンス結果に応じて対象の語句を平易化したり、難易化を行う。

そして、言い換え可能単語をすべて変換した後、山村が提案した以下の定式 [11] を用いて文章自体の難易度を数値化する。

$$D(W) = \frac{R(W) - L(W)}{L(W)} \tag{2}$$

- D : 文字削減率
- W : 単語の集合
- R :W の読みの長さの合計
- L : 文章の総文字数

式 2 では、文章中の読みの長さに着目し、文字数に対して読みが長ければ情報が圧縮され文章の難易度は高いと評価できる。言い換えると漢字が多く含まれるほど文章が難しくなるといえる。そこで、本稿では、要約文中の漢字数を総文字数で除したものでこの数式を再現している。そして、この難易度をパラメータとして次の機械学習部に $mongoDB$ を通して受け渡す。

5 可読性向上のための機械学習

5.1 ユーザごとの文章理解度把握システムの提案

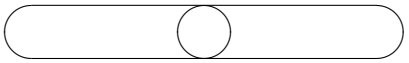
研究の被験者（以下ユーザと呼ぶ）に対して本稿の手法で開発した Web アプリケーションを使って文章に対しての不満度を評価する。

まず、ユーザの語彙力や文章の理解度を測るため、Web ページにある同一文章から作成した 6 つの要約文を提示する。なお、6 つの文章は前述したレベルごとに分けられている。そしてその文章から、ユーザ本人が一番読みやすいと感じたものをクリックしてもらう。クリックした時、内部の機械学習部に対して選んだ文章のレベルを渡す。そして、この工程を別文章で 10 回繰り返し、その結果を元に一時的なユーザ理解度レベル（以下ユーザレベル）を設定する。

次に、図 1 のように要約部に再帰して、そのユーザレベルに応じた要約文を提示してその文章に対して平易と感じたら、より難解で情報が圧縮された文章を変換する。逆に、文章が難解と感じたら文章の情報量を下げてでも平易な文章に変換する。なお変換のプロセスは 4 章の平易化部を再帰的に呼び出し実行することにした。

5.2 ユーザの文章理解度判定モデルの構築

そして、ユーザのそのアクティビティに対して実装したのが図 2 のスライド式バーである。このバーに付属するボタンを左右に動かすことで文章に対してのフィードバックを得る。このバーのボタンは初期位置はちょうど真ん中に配置されており、提示された要約文に対して、ユーザが充分満足している場合はボタンを動かさず別記事へ移動。ユーザが難解と感じ文章を平易にしたい場合は左に、その逆は右に移動させてフィードバックを直感的に観測できるように設計した。なおバーは左右それぞれ段階まで移動させることで、ユーザの文章に対する不満度を細かくできるようにした。



2 フィードバックのためのスライドバー

これらのフィードバックを元に機械学習を組み込む。まず、ユーザの記事に対してのボタンの移動量を取得して、前回のフィードバックに比べボタンの移動量が減少したら報酬を与え、移動量が増加したら報酬を与えない強化学習的手法を採択した。そして、強化学習のフォワード先として、要約の文字数（初期値は 100 文字）、語と文章の難易度、ユーザレベルの三つのパラメータとする。このパラメータの比率をどのように変化させるかは現在検討中である。

6 数値実験結果ならびに考察

6.1 ユーザによる難易度判定における不満度の評価

本稿の検証では、著者が作成したポータルサイトを用いてユーザ 20 人に対して執り行う。まず、ユーザに対して 1 章で示したシステム全体の流れを 20 回（この回数は完成後調整する）行い、スライド式バーの移動量が小さくなっていることを確認する。

6.2 機械学習学習による可読性向上の有意差検定また、上記の検定とは別に、各ユーザに対して、ペーパベースの検定を行う。各ユーザに対して、同じ記事に対しての学習前の要約文とそのユーザに最適化された要約文を提示する。そして、そのどちらが読みやすいかアンケートを取る。そのアンケートで最適化された要約文を選ぶユーザの方が多ければこのシステムの有効性は示されたこととなる。

7 まとめと今後の課題

本稿では、自然言語処理、機械学習の複合的なシステムを構築した。上に示す通り、有効性は確認できた。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] 難波英嗣, 奥村学, ”特集 テキスト要約”, 情報処理, vol. 43, NO. 12, pp. 1-8, 2002.

[2] Luhn, H. P, ”The Automatic Creation of Literature Abstracts”, IBM Journal of Research and Development, vol. 2, No. 2, pp.159-165, 1958.

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

[4] 佐藤理史, 奥村学, ”電脳文章要約術 ー計算機はいかにしてテキストを要約するかー”, 情報処理, vol 40, No. 2, pp. 1-5, 1999.

[5] 山本和英, 増山繁, 内藤照三, ”文章内構造を複合的に利用した論説文要約システム GREEN”, 自然言語処理, vol2, No. 1, pp.1225-1234, 1995.

[6] 柏木潔, 小町守, 松本裕治, ”レビュー文書からの省略された属性の推定を含めた意見抽出” 言語処理学会, 第 19 回年次大会 発表論文集, pp. 528-531, 2013.

[7] 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.

[8] 美野秀弥, 田中英輝, ”国語辞典を使った放送ニュースの名詞の平易化”, 言語処理学会, 第 16 回年次大会 発表論文集, pp. 760-763. 2010.

[9] 梶原智之, 山本和英, ”日本語の語彙平易化システムの構築”, 情報処理学会, 第 77 回全国大会, pp. 167-178, 2015.

[10] 梶原智之, 小町守, ”平易なコーパスを用いないテキスト平易化のため単言語パラレルコーパスの構築, ”情報処理学会研究報告”, vol.2016-NL-229, No. 13, pp.1-8, 2016.