

多言語 Wikipedia を用いた伝統文化の差異情報抽出の提案

藤原 裕也[†] 灘本 明代[‡]

甲南大学理工学部情報システム工学科[†]

甲南大学知能情報学部[‡]

1. はじめに

近年インターネットの普及により、インターネット上で自由に他国の人との交流が可能になっている。これにより他国の情報が容易に取得することが可能になった。しかしながら、若者の海外留学及び海外旅行が激減している等、若者の海外への興味が以前より薄くなっている[1]。その理由の一つとして、自国との文化や習慣の違いなどが上げられる。そこで我々は日本の若者に海外へ興味を持ってもらうために、日本及び海外の伝統文化に注目し、これらの文化の紹介の仕方の違いを抽出し提示する手法の提案を行う。

本研究では、はじめの一步として、日本の伝統文化に注目し多言語 Wikipedia を用いて、日本と海外における日本の伝統文化の紹介の差異情報を抽出し提示するシステムを提案する。具体的には、ユーザの入力した日本の伝統文化の日本語 Wikipedia と英語 Wikipedia を比較しその差分情報を提示する。この時、情報の粒度の違いから、複数の日本語 Wikipedia の記事と英語 Wikipedia の記事を比較対象とする。本論文では、この比較対象の範囲を Wikipedia のリンク構造を用いて決定する手法の提案及び多言語間の差分情報抽出手法の提案を行う。

以下、第2章では提案システムの概要を、3章では比較対象 Wikipedia の記事の抽出手法を4章では多言語記事の差分抽出について述べ5章でまとめについて述べる。

2. システム概要

以下に処理の流れを示す。

- ① ユーザは調べたい日本の伝統文化をクエリとして入力する。

- ② 入力された伝統文化のタイトルの日本語と英語の Wikipedia の記事を各々取得する。
- ③ 比較対象となる日本語の記事群を②で取得した日本語の記事のリンク構造を解析し取得する。
- ④ ②で取得した英語の記事と③で取得した日本語の記事群を比較し差分情報を取得する。
- ⑤ 取得した差分情報をユーザに提示する。

3. 比較対象 Wikipedia の記事の抽出

日本語と英語の Wikipedia の記事を比較する時、言語や文化の違いから情報の粒度が異なり、対応する記事が複数にまたがる場合がある。特に日本の伝統文化の場合、英語の Wikipedia では1記事であるのに対し、日本語の Wikipedia では詳細に書かれており複数の記事になっている場合がある。例えば、「和歌」の場合、英語の Wikipedia では和歌の形式の1つである長歌や短歌の説明が和歌という記事1つに書いてあるのに対し、日本語の Wikipedia では和歌の記事だけでなく長歌、短歌の記事が各々存在し複数ページにまたがっている。そこで、我々は日本語の Wikipedia のリンク構造を解析する事により、比較対象の記事を抽出する。以下に抽出手順を示す。

- ① ユーザの入力したクエリと同じタイトルを持つ日本語の記事からのリンクグラフを作成する。ここで、ユーザの入力したクエリと同じタイトルを持つ記事のノードを基準ノードと呼ぶ。
- ② 基準ノードと双方向にリンクされているノードは、基準ノードの記事に深く関連すると考え、双方リンクされているノードを残し、その他のリンク（インリンク、アウトリンク）のみのノードをリンクグラフから削除する（図1参照）。

Extracting Difference Information of Japanese Traditional Culture Article from Multilingual Wikipedia

[†]Yuya FUJIWARA, and [‡]Akiyo NADAMOTO

[†]Faculty of Science and Engineering information system engineering department

[‡]Dept. of Intelligene and Informatics

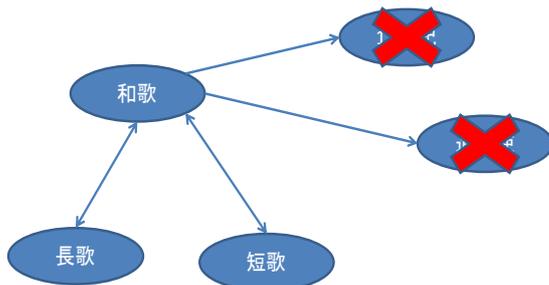


図 1: リンクグラフ

- ③ 基準ノードとリンクグラフ内のその他のノードの類似度を以下の式のコサイン相関値を用いて求める。

$$\text{式 } \text{Cos}(x, y) = \frac{\sum x_i y_j}{\sqrt{\sum x_i^2 \sum y_j^2}}$$

ここで x_i は日本語の記事の名詞の出現頻度とし y_j は英語の記事の名詞の出現頻度とする。

4. 多言語 Wikipedia の差分抽出

4.1 多言語 Wikipedia の比較

言語にかかわらず Wikipedia の記事は目次構造に基づいて段落に分かれている。つまりは、Wikipedia の段落は意味的に分かれている可能性が高いと考えられる。そこで、多言語 Wikipedia を比較してその差分情報を抽出する為に、我々はこの Wikipedia の目次構造に注目し、目次構造に基づくコンテンツの比較を行う (図 2 参照)。類似している段落の中から、その差分情報を抽出することを行う。



図 2: 目次構造とコンテンツ

ここでは、日英 Wikipedia 各々の記事の段落

毎にテキストの形態素解析を行う。ここで、日本語の形態素解析には Mecab[2] を、英語の形態素解析には Tree tagger[3] を使用する。各々の単語の品詞を取得したいために、英語版 Wikipedia においても tagger を使用する。

次に、英語を日本語に GENE95 辞書[4] を使用して翻訳する。また、GENE95 辞書に載っていない単語は Google Ajax api と Microsoft api の翻訳を使用する。

翻訳時に単語の多義性が問題になる。今回はこの多義性には考慮せず、今後の課題とする。

次に日本語版 Wikipedia の記事と英語版 Wikipedia の日本語翻訳の記事の名詞の出現頻度ベクトルを作成する。

そして、コサイン相関値を用いて、各々の段落における類似度を求め、ある閾値以上の段落を差分情報を抽出する対象段落ペアとする。

4.2 差分情報の抽出

4.1 において抽出した類似段落のペアから、差分情報を抽出する。我々は、類似段落の中でのアンカー文字列は、一般的に重要である単語である可能性が高いと考え、このアンカー文字列に注目する。類似段落ペアの各々のアンカー文字列を比較し、一方にあって、もう一方にないアンカー文字を差分情報とする。

5. まとめ

本発表では、日本の伝統文化を対象とし、日英の Wikipedia 上での差分情報を取得する手法の提案を行った。今回は翻訳時の単語の多義性に考慮していないが、今後この問題を解決する予定である。

参考文献

- [1] <http://headlines.yahoo.co.jp/hl/?a=20100911-00000848-yom-soci>
- [2] <http://mecab.sourceforge.net/>
- [3] <http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>
- [4] <http://www.namazu.org/~tsuchiya/sdic/data/gene.html>