3N-02

オンライン百科事典への日中言語間エンティティリンキング手法の提案 ー適切な用語説明ページの抽出手法ー

宋　翔[†]　周　佳良[††]　堀田 健介[†]　木村　文則[‡]　前田　亮[††]

[†]立命館大学情報理工学研究科　[††]立命館大学情報理工学部　[‡]立命館大学衣笠総合研究機構

1 はじめに

近年, インターネット環境は世界中で普及しており, 様々な言語で多くの情報が蓄積されている. また, Web 上のハイパーリンクにより, ある情報から複数の関連情報が得られる. しかし, 情報によっては利用者の母国語で提供されているとは限らず, 言語の壁によって利用者の役に立つ情報を得ることが難しい場合がある.

本論文では, ある言語で書かれた文書中のキーワードから, そのキーワードを説明する別言語の Wikipeida 記事を自動的に発見する手法を提案する. これにより, ある言語を学習している留学生などが, 分からないキーワードについて母国語による説明を容易に得ることができ, 文書の内容理解や言語学習の支援となることが期待できる.

本論文では, 文献[3]の手法により得られた複数の Wikipedia 記事の候補から, 最も適切な用語説明ページを抽出する手法を中心に述べる.

2 関連研究

Wikipedia は, 誰でも Web ブラウザから利用できる多言語オンライン百科事典である. しかし, 記事によっては, ある言語では記事があるが, 別の言語の該当記事がない場合がある. この問題を解決するため, これまでに様々な研究を行っている.

白川ら[1]は, 自然文の入力に対して関連語句を取得するための枠組みを, Wikipedia とベイズ理論を用いて構築した. 本研究でも, この研究と同様にキーワードの抽出, 特にキーワードの曖昧性解消を行っている. しかし, Wikipedia の記事同士でコサイン類似度を比較し, 元の記事との関連度がもっとも高い記事にリンクを張るという点が異なっている.

綱川ら[2]は, ある記事に対して他の言語版の記事が存在する時に, 内部リンク(Wikipedia の記事において, 他の記事へのハイパーリンク)を言語間で交換することより内部リンクを自動的に付与する方法を提案している. 評価実験において, 提案方法が既存記事の内部リンクのカバー率向上に効果があることを実証した.

3 提案手法

本章では, 日本語版 Wikipedia の記事中からキーワードに対して相応しい中国語の Wikipedia 記事を検出する手法について述べる.

3.1 提案手法の概要

本研究で提案する手法は以下のように 3 つのプロセスで構成される.

図 1 に提案手法の全体の構成を示す.

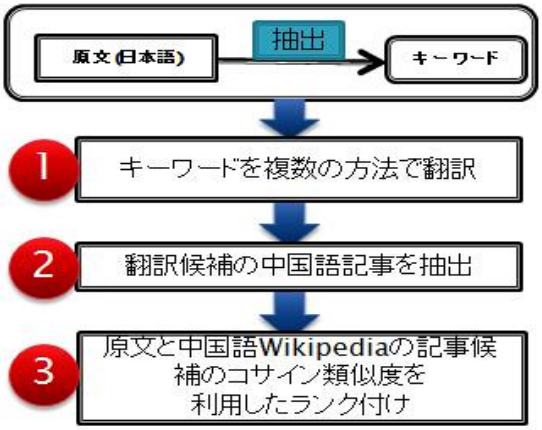
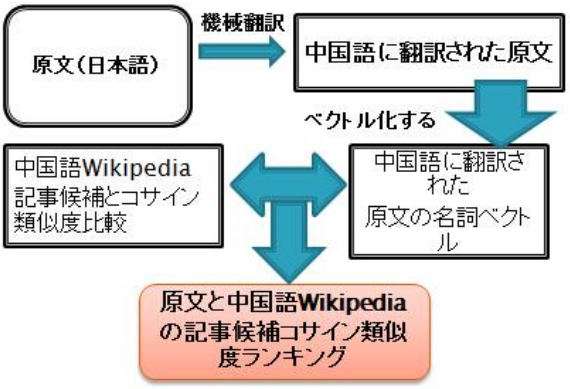


図 1：提案手法の全体の概要

まず, 日本語の文章（原文）から, キーワードを抽出する. そして, 抽出したキーワードを複数の方法で翻訳し, 翻訳候補の中国語記事を抽出する. この部分の手法については, 文献[3]で詳しく説明している. 最後に, 中国語に翻訳された原文と中国語 Wikipedia の各記事候補のコサイン類似度を比較し, ランキングを行う. 本論文では, 主にこの記事候補のランキングの部分について述べる.

A Cross-Language Entity Linking Method from Keywords in Japanese Text to Articles in Chinese Online Encyclopedia
- A Method for Extracting an Appropriate Article for a Keyword -
†Graduate School of Information Science and Engineering, Ritsumeikan University
††College of Information Science and Engineering, Ritsumeikan University
‡Kinugasa Reserch Organization, Ritsumeikan University



3.2 原文と中国語 Wikipedia の記事候補の類似度によるランキング

図 2 に提案手法におけるコサイン類似度によるランキングの手順を示す.

ランキング処理の手順は次の通りである. まず, 日本語の原文を機械翻訳器を用いて翻訳する. そして, 翻訳された原文から名詞を抽出する. 同じように中国語版 Wikipedia 全記事から名詞を抽出する. これらの名詞の出現頻度をベクトルとみなし, 原文と中国語版 Wikipedia 全記事のコサイン類似度を計算する.

4 実験

4.1 実験の概要

提案手法の有効性を評価するために実験を行った. まず, [3]の手法により, 日本語版 Wikipedia 記事 20 個（表 1）とそれぞれに対応する中国語 Wikipedia 記事候補を取得した.

表 1：日本語版 Wikipedia 記事 20 個

1 三井財閥	2 琉球征伐	3 教育ニ関スル勅語	4 麻雀
5 月岡芳年	6 鳥獣人物戯画	7 日本の地理	8 ヨーロッパ
9 アフリカ	10 三科	11 論理学	12 プラトン
13 遺伝子工学	14 制御工学	15 並列計算	16 江戸時代
17 産業革命	18 日本の漫画	19 就学率	20 資源

また, 中国語の形態素解析ツールとして ANSJ_Seg[1] を使用し, 各記事とそれぞれに対応する中国語 Wikipedia 記事候補のコサイン類似度を計算した.

[1] https://github.com/NLPchina/ansj_seg

4.2 実験結果

実験の結果としては, 抽出した 26 個のキーワードにそれぞれ対応する中国語の記事候補のコサイン類似度を計算した. 26 個の中で, 正解データがコサイン類似度ランキング一位になる記事は 21 個がある. 正解率が 80%になった.

表 2：原文に対応する中国語記事候補の数

三井財閥：21	琉球征伐：3	教育ニ関スル勅語：3	麻雀：3
月岡芳年：306	鳥獣人物戯画：4	日本の地理：20	ヨーロッパ：7
アフリカ：3	三科：2	論理学：30	プラトン：7
遺伝子工学：16	制御工学：2	並列計算：4	江戸時代：3
産業革命：3	日本の漫画：3	就学率：4	資源：14

4.3 考察

記事のタイトルの文字の数とキーワードはうまく一致する場合に正解数が多い. また, 結果が悪い例としては正解と認められているがコサイン類似度のランキングで一位になれなかった記事も含めている.

5 おわりに

本稿では, 日本語の文章中のキーワードに対して, 中国語 Wikipedia における適切な用語説明ページに自動的にリンクする手法を提案した.

今後の課題としては, 日中間言語の記事の中で意義がある言葉に新たなリンクを貼ることを目指す. また, Wikipedia だけではなく, 中国の百度百科（中国の検索エンジンである百度が 2006 年 4 月に公開したオンライン百科事典）も対象として, 任意の日本語の文章と百度百科の間の潜在的なリンクを発見することも今後の課題である.

参考文献

[1] 白川真澄, 中山浩太郎, 原隆浩, 西尾章治郎. 「Wikipedia とナイーブベイズを用いた自然文に対する関連語句取得手法」. データ工学と情報マネジメントに関するフォーラム. 2012.

[2] 綱川隆司, 新谷誠, 梶博行. 「言語版の内部リンクを利用した Wikipedia 内部リンクの自動付与」. 言語処理学会第 20 回年次大会, 発表論文集, 2014.

[3] 周佳良, 宋翔, 堀田健介, 木村文則, 前田亮. 「オンライン百科事典を対象とした日中言語間エンティティリンキング手法の提案-日本語文書中の重要語の翻訳手法-」. 情報処理学会第 77 回全国大会, 2015.