3N-01

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 —日本語文章中の重要語の翻訳手法—

周 佳良† 宋 翔†† 堀田 健介†† 木村 文則‡ 前田 亮†

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

1 はじめに

在日中国人留学生は日本の記事を読むことが多いが，記事中で分からない単語があるとその記事を理解することが困難となる．そのような単語を調べる際には，母国語で書かれた説明を見つける必要があるが，対応する説明を見つけることは容易ではない．Wikipedia のようなオンライン百科事典には世界各国の言語の記事があるが，必ずしも母国語で書かれた対応する記事へとリンクなどの対応付けが行われているとは限らない．

このような問題を解決するため，自動的に異なる言語間での潜在的なリンクを見つける取り組みが必要となる．その方法として，ある日本語の記事中で理解できない単語をアンカーとして抽出して，意味を説明する中国語の記事と元の記事をリンクさせることが挙げられる．このような仕組みは留学生の勉強の役に立ち，オンライン百科事典の有用性を更に高める事ができると考えられる．

本論文では，在日中国人留学生が勉強するためのオンライン百科事典を学習ツールとして，知識発見を支援するシステムを提案する．具体的には，ある日本語の記事中で理解できない単語に対して，対応する中国語の記事を自動的にリンクさせる手法を提案する．本論文では重要語の翻訳処理およびそれに対応する中国語 Wikipedia 記事の取得について述べる．

2 関連研究

Text Analysis Conference (TAC)[1]では，Cross-lingual Entity Linking (CLEL)というタスクが行われている．このタスクでは，ある記事に出現した人名や地名などのエンティティが，別の言語の記事に出現しているかどうかを探すことが目的である．本論文では，地名や人名に限定しない点で，このタスクとは異なっている．

NTCIR-9 と NTCIR-10 で行われた，Cross-lingual Linking Discovery (CLLD)というタスクでは，Wikipedia のある記事に対して対応する別言語の記事を探すことが目的であり，本論文の目的と類似している．このタスクに参加した佐藤[2]は，Wikipedia 自身を辞書として使い，他の外部的な辞書を利用しない手法を提案している．この手法では，Wikipedia から最も適切な言語横断リンクを発見することを提案している．

3 提案手法

我々は，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を行う．ここでは，日本語の Wikipedia の記事中の重要語を抽出し，それに対応する中国語の Wikipedia 記事の抽出を行う．提案システムは以下の順に処理を行う（図 1）．1. 日本語の Wikipedia の記事（原文）中から重要語を抽出する．2. 重要語を複数の手法で翻訳し，全ての訳語候補を取得する．3. 取得した全ての訳語候補に対応する中国語の Wikipedia 記事を取得する．4. 取得した各中国語の Wikipedia 記事を，中国語に翻訳した原文と比較し，最も類似度の高いものを最終的に提示する．4 の処理について参考文献[3]で詳しく述べる．

本論文では，上記の 2, 3 の処理について焦点を当て，その詳細について述べる．

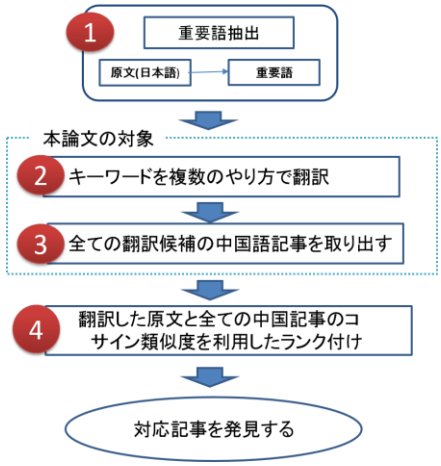

図 1：提案手法の全体の概要

3.1 複数の翻訳手法

本論文では得られた重要語を 2 種の機械翻訳（Google 機械翻訳，Bing 機械翻訳）と文字コード変換の方法で翻訳する（図 2）．その後，それぞれの翻訳方法で得られた訳語を（誤訳も含めて）全てを訳語候補とする．このような手法を用いる原因は得られた重要語を単一手法で翻訳すると，適切な訳語が得られない可能性があることを防ぐためである．

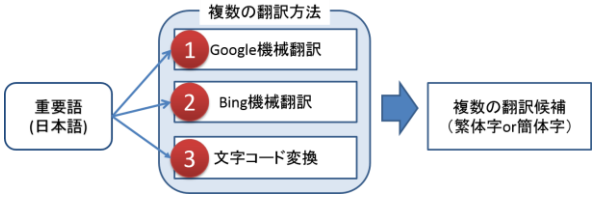

図 2：重要語の訳語候補の取得方法

A Cross-lingual Entity Linking Approach for Linking Japanese Keywords to Articles in Chinese Online Encyclopedia – An Entity Translation Method for Japanese Texts

†College of Information Science and Engineering, Ritsumeikan University
††Graduate School of Information Science and Engineering, Ritsumeikan University
‡Kinugasa Research Organization, Ritsumeikan University

3.2　対応候補の中国語記事の取得

この処理では，前節の処理で得られた全ての訳語候補に対応する中国語版 Wikipedia の記事を取得する．本手法では，全ての訳語候補と中国語版 Wikipedia 記事の見出し語を比較し，部分一致すればその見出し語の記事を対応候補記事として取得する．翻訳候補に多数の意味がある場合，完全一致の抽出手法を使うと，翻訳候補の「曖昧性の解消のページ」だけが抽出されてしまう．一つの理由としては，複合語の全てを完全に翻訳することができない場合でも，その複合語の個別の単語であればうまく翻訳できることも多いからである．

この処理の段階では，不要な記事も大量に取得してしまうことになるが，次の処理（本論文では詳細は触れない）において取得された対応候補記事のランキングを行うため，適切であると思われる記事に絞り込むことができる．それゆえ，この処理においては適切な対応候補記事の取得漏れが起きないことが重要であるため，上記のような手法を採用している．

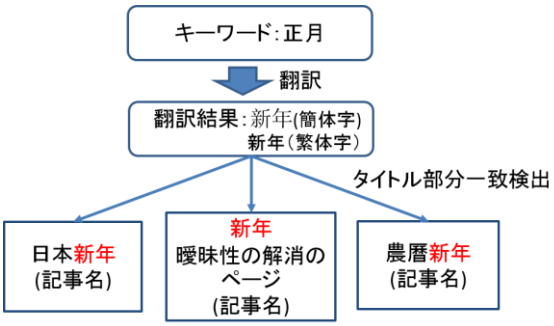

図 3：部分一致の例

4　評価実験

提案した手法の評価を行うために，日本 Wikipedia 記事中にある重要語に対応する中国語記事を取得する実験を行った．使用した日本語記事は中国人留学生が興味を持った 20 記事であり，記事ごとに 55 個の重要語（表 1）を人手により抽出し，その重要語に対する中国語の対応記事をシステムにより取得した．抽出した重要語は，文章中で重要であると思われる単語かつ，中国人留学生が理解し難い単語を選んだ．また，システムが取得した対応記事は，1 つの重要語に対して複数となることもありうる．本実験では，20 記事中の 55 個の重要語に対して，システムにより候補記事から，1350 記事を取得した．取得された記事の平均値は 26 件であり，その中で「合戦」という重要語は最大件数 187 件の対応記事を取得した．取得した中国語の対応記事が元の日本語の重要語に適しているかどうかは，筆者が取得した記事を実際に読んで判断を行った．

表 1 ：日本語版 Wikipedia 記事中の重要語の例

記事タイトル	重要語
三井財閥	財閥解体，中央集権，持株会社
琉球征伐	平田増宗，三司官，江戸城
麻雀	陳魚門，三人麻雀，脱衣麻雀
月岡芳年	幽霊図，号，合戦

訳語と記事候補の結果を以下 4 つのパターンがある．表 2 に示す．

表 2：取得した訳語と記事候補の結果

パターン	正解の訳語 有無	正解の記事候補 有無
①	○	○
②	○	×
③	×	○
④	×	×

55 の重要語のうち何語に対して，システムにより適切な中国語の候補記事が得られたかどうかを集計した結果を表 3 に示す．この集計結果は，取得結果が上記のいずれかのパターンになるかについて分類を行っている．

表 3：候補記事取得有無の実験結果

翻訳方法	①	②	③	④
Bing 機械翻訳	0.56 (31/55)	0.0 (0/55)	0.05 (3/55)	0.38 (21/55)
Google 機械翻訳	0.65 (36/55)	0.0 (0/55)	0.07 (4/55)	0.27 (15/55)
文字コード変換	0.44 (24/55)	0.0 (0/55)	0.0 (0/55)	0.56 (31/55)
上記三手法の結合結果	0.9 (50/55)	0.0 (0/55)	0.03 (1/55)	0.07 (4/55)

5　おわりに

本論文では日本語版 Wikipedia 記事中の重要語に対応する中国語版 Wikipedia 記事候補を取得する手法を提案した．

しかし，提案手法について今後改善すべきの課題が発見された．以下で今後の改善点について記す．

機械翻訳本来は文章を翻訳するものであり，単語を翻訳するのは本来の使い方ではない．今後，単語前後の文脈を考慮することにより曖昧性を解消することを検討する必要がある．対応候補の中国語記事を取得する段階では，部分一致による手法の精度が高いが，不要な記事も大量に抽出してしまうことになる．不要な記事を減らすため，ほかの手法を検討する必要がある．

参考文献

[1] Heng, J., Nothman, J., and Hachey, B.: Overview of TAC-KBP2014 entity discovery and linking tasks. *Proc. of TAC2014*, 2014.

[2] Sato, T.: Osaka Kyoiku University at NTCIR-10 CrossLink-2 link filtering by title tag of corpus as a dictionary. *Proc. of NTCIR-10*, pp. 47-50, 2013.

[3] 宋翔，周佳良，堀田健介，木村文則，前田亮：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− 名詞における日中間文章のコサイン類似度計算− ． 情報処理学会第 77 回大会，2015