1A2-3

言語間の情報補完を用いた対訳文の Wikification

Improving Wikification of Bitexts by Completing Cross-lingual Information

林 良彦[*1] 山内 健二[*2] 永田 昌明[*3] 田中 貴秋[*3]
Yoshihiko Hayashi Kenji Yamauchi Masaaki Nagata Takaaki Tanaka

[*1]大阪大学大学院言語文化研究科
Graduate School of Language and Culture, Osaka University

[*2]大阪大学基礎工学部情報科学科
School of Engineering Science, Osaka University

[*3]NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

This paper proposes a method for "wikifying" bitexts, which leverages the interlingual nature of Wikipedia. The method, given a pair of sentences, first extracts named entities (NEs), and then tries to link them to corresponding Wikipedia entries in the same languages ("monolingual forward wikification"). Due to several reasons, this process may fail to link an NE properly. Our experimental results, however, show that the entire process of Wikification could be augmented by completing the results of Wikification in another language ("cross-lingual backward wikification"). This paper analyzes the errors from the forward wikification, and cases-studies the possibility of the backward wikification.

1. はじめに

テキスト中に現れる実体 (entity) への参照表現 (固有表現; 以下，NE) を認識し，それを参照知識ベースにおける該当するエントリにリンクするタスクは，"entity linking" と呼ばれ，様々なタイプのテキストに対する entity linking の関心が高まっている．最近の Text Analysis Conference (TAC) においても主要なタスクの一つとして取り上げられ，TAC2013 では，さらに言語横断のタスクも追加された[*1]．参照知識ベースとしては，Wikipedia が用いられる場合が多く，この場合はとくに，Wikification [Mihalcea and Csomai 07] と呼ばれる．

本研究では，入力テキストが対訳文となっている場合に，双方の言語のテキストに対し並行的に Wikification を行うというタスクを取り上げ，一方の言語における Wikification の不備を他方の言語の情報により補完する方式を提案し，評価実験の結果に対するケーススタディからその可能性を検討する．

2. 対訳文の Wikification

2.1 Wikification における課題

"wikify" という用語を最初に使用したと思われる [Mihalcea and Csomai 07] においては，リンキングを行う対象は NE には限られていない．このため，(1) 対象とする語句 (キーワード) の抽出，(2) 抽出された語句に対応する Wikipedia エントリの決定，の 2 つの過程が課題とされており，とくに (2) の過程は，Wikipedia を語義のインベントリとする語義曖昧性解消の過程として扱われている．

対象を NE に限定した Wikification では，上記の (1) の過程は，既存の NE 抽出器に委ねることができるが，一般には，NE 抽出器の出力と Wikipedia エントリは対応していないため，こののミスマッチに対処することが必要である．一方，上記の (2) の多義解消の過程は，対象が NE に限定されるものの，一般の語句に対する場合とほぼ同様な曖昧性解消の過程が必要なことに変わりはない．

連絡先: 林 良彦， 大阪大学大学院言語文化研究科・言語情報科学講座，mailto:hayashi@lang.osaka-u.ac.jp

∗1 http://www.nist.gov/tac/2013/KBP/EntityLinking/

2.2 対訳文に対する Wikification

後述の評価実験で使用した「Wikipedia 日英京都関連文書対訳コーパス」[*2]に現れる以下の対訳文ペアを例に対訳文に対する Wikification を説明する．下線部を付した語句が NE としてすでに NE 抽出器により抽出されていると仮定する．

- 日本語文: 大炊御門 氏忠（おおいのみかど うじただ、乾元 (日本) 元年（1302 年） - 没年 不詳）は 南北朝時代 (日本) の公卿。
- 英語文: Ujitada OINOMIKADO (1302 - year of death unknown) was a Court noble during the period of Northern and Southern Courts (Japan).

まず，日本語文における「大炊御門 氏忠」は，この NE 文字列と Wipikedia エントリのタイトルとのマッチングにより，正しく日本語 Wikipedia のエントリ http://ja.wikipedia.org/wiki/大炊御門氏忠 へリンキングできる．一方,「南北朝時代」については，Wikipedia には日本の南北朝時代だけでなく，中国やベトナムの南北朝時代に関する Wikipedia エントリが存在するため，多義解消が必要である．

英語文における "Ujitada OINOMIKADO" は，正しく NE 抽出されていたとしても，これに対する英語 Wikipedia のエントリが存在しないため，リンキングができない．また，"period of Northern and Southern Courts" は,「南北朝時代」の説明的な表現であり，NE として抽出すること自体が困難であると想定されるが，たとえ NE として抽出できたとしてもこの語句から英語 Wikipedia のエントリを直接求めることはできない．しかしながら，日本語側で「南北朝時代」が適切に日本語 Wikipedia のエントリ http://ja.wikipedia.org/wiki/南北朝時代_(日本) へとリンキングできていれば，そのエントリから言語間リンクによってリンクされている英語 Wikipedia のエントリ http://en.wikipedia.org/wiki/Nanboku-ch%C5%8D_period [*3]を求めることにより，上記の表現のリンキング先として設定できる可能性がある．

∗2 http://alaginrc.nict.go.jp/WikiCorpus/

∗3 cho の o は長音記号 (macron) 付き:ō.

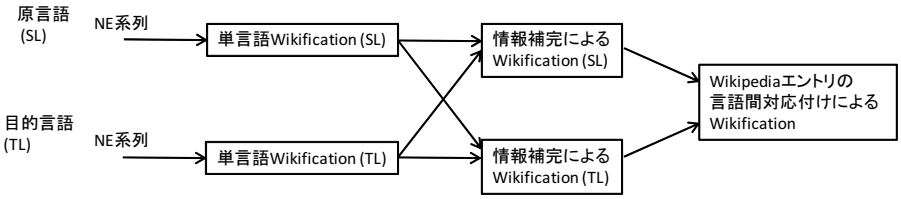

図 1: 提案方式の全体構成

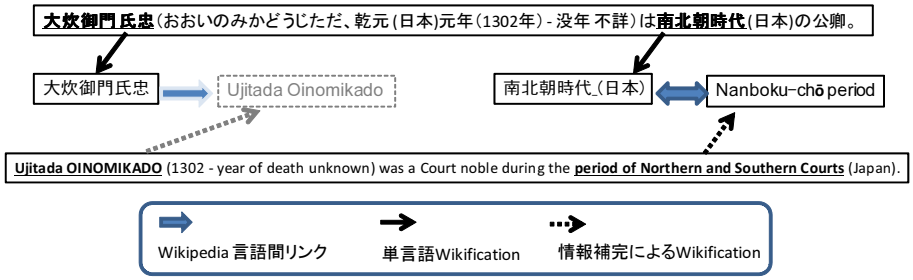

図 2: 対訳文ペア例に対する Wikification

3. 提案方式

3.1 全体構成

図 1 に提案方式の全体構成を示す．提案方式においては，入力の対訳文の双方の言語についての処理は対称的であり，特に方向性を持たないが，便宜上，一方を原言語 (SL)，他方を目的言語 (TL) と呼ぶ．先に例示した対訳文ペアに関しては，図 2 に示すようなリンキングが行えることが理想である．本稿では，紙面の関係上，「Wikipedia エントリの言語間対応付けによる Wikification」の処理過程については説明しない．

なお，実装においては，Wikipedia そのものではなく，SPARQL 言語による問い合わせが可能なエンドポイント[*4]を介して，Wikipedia を RDF により構造化したデータである DBpedia [Lehmann et al. 14] を利用した．

3.2 単言語 Wikification

単言語 Wikification は，リンキングの対象として抽出された NE 区間に対して，リンキングの候補となる Wikipedia エントリを求める候補エントリ抽出処理と，その中から最適なエントリを選出する曖昧性解消処理からなる．

3.2.1 候補エントリ抽出処理

まず，与えられた NE 文字列をタイトルに持つ Wikipedia エントリを検索する．もし，このようなエントリが得られない場合は，`dbo:wikiPageRedirects` という RDF プロパティ[*5]を利用して，リダイレクトされるエントリを検索する．

基本的には，このどちらかでエントリが検索できれば，それが適切なリンキング先である可能性が高いが，その NE が多義性を有している場合は，いわゆる「曖昧さ回避」の設定が準備されていることも多い．このため，`dbo:wikiPageDisambiguates` という RDF プロパティを利用して，「曖昧さ回避」として提示されるエントリをもリンキング対象のエントリ候補に加える．

上記の手順によってもエントリ候補が得られない場合も存在する．この原因としては，NE として抽出されている区間が適切でない (Wikipedia のエントリとマッチしない) 場合がある．そこで，まず NE 区間の拡張，次に NE 区間の縮小を行い，これらに対するエントリ候補を求める．

例えば，「奈良県斑鳩町にある法隆寺に近鉄特急で行った」という例文において，「奈良県斑鳩町」として NE 区間が抽出されたとしても，この文字列から適切なエントリ候補を求めることはできないので，「奈良県」と「斑鳩町」に NE 区間を縮小・分割して候補探索を行う．一方，「近鉄特急」は一つの NE 区間として抽出されない可能性が高いが，実際には「近鉄特急」というタイトルを持つ Wikipedia エントリが存在する．このようなケースでは，「近鉄」に対するリダイレクト情報に「近鉄特急」のエントリが含まれる可能性もあるが，リダイレクト情報の準備状況は必ずしも十分・一定ではないので，NE 区間の伸縮を試みエントリ探索を行うことが適切である．

このような，NE 区間の拡張・縮小を行うために，処理対象文に対する言語解析の結果を用いる．より具体的には，日本語では同一文節内で NE 区間の伸縮を行う．英語においては，抽出されている NE を含む名詞句部分木内で伸縮を試みる．

3.2.2 曖昧性解消処理

各候補の Wikipedia エントリに対して算出する適合度スコアにより曖昧性解消を行う．適合度のスコアリングに関しては，現時点では以下の 4 つのタイプの情報を考慮している．

- NE と Wikipedia エントリのタイトルの文字列類似度: 編集距離の一種である Jaro-Winkler の類似度 [Winkler 90] により算出する．
- 入力文と Wikipedia エントリの概要テキストの類似度: 入力文 x を文脈と考え，語義曖昧性解消で用いられる Lesk 法 [Lesk 86] と同様に，候補エントリにおける abstract テキスト y との単語オーバーラップに基づいて類似度 $textSim(x, y)$ を算出する．現時点では，Simpson 係数 ($\frac{|X \cap Y|}{Min(|X|,|Y|)}$) により類似度を算出している．なお，abstract テキストを取得するために `dbo:abstract` プロパティを利用する．
- NE タイプと DBPedia オントロジーにおけるクラスの合致: NE 抽出器が出力する NE タイプと Wikipedia エントリに付与された DBPedia オントロジーのクラスの合

*4 日本語:`http://ja.dbpedia.org/sparql`，英語:`http://dbpedia.org/sparql`

*5 ここで `dbo` とは，DBpedia オントロジーの接頭辞 `http://dbpedia.org/ontology/` の略記である．

致をチェックする．例えば，NE タイプ=LOCATION に対し，dbp:Place といったオントロジークラスが候補の Wikipedia エントリに付与されていれば，当該エントリのスコアを加点する．

- 候補抽出手段: NE 区間の拡張により抽出された候補エントリは，より適合度が高い可能性があると考え，当該エントリのスコアを加点する．

3.3 情報補完による Wikification を試みる状況

対訳文の双方の言語において，前節の単言語 Wikification を行った結果としてリンキングされた Wikipedia エントリのそれぞれについて，言語間リンク[*6]をチェックすることにより，他方の言語において対応する Wikipedia エントリを求める．これにより，例えば，図 2 の場合であれば，日本語のエントリ「南北朝時代_(日本)」に対して，英語のエントリ “Nanboku-chō period” が得られる．

情報補完による Wikification は，対象とする対訳文がパラレルであること，より具体的には，対応する NE が同様にテキスト中に分布していることを前提とし,『一方の言語 SL において Wikification されたエントリ E が存在し，かつ，それに対応する他方の言語 TL におけるエントリ E' が存在するにも関わらず，当該のエントリ E' にリンキングされた言語 TL の NE が存在しない場合，エントリ E' にリンキングすることができる NE が言語 TL に存在しないかをチェック』する．あるいは，さらにその対象を広げ，言語 TL において NE として抽出されていない表現区間を探索対象とすることも必要・有用である．単言語 Wikification が順方向 (forward) のリンキングであるとすれば，本処理は言わば，Wikipedia エントリから NE の方向への逆方向 (backward) のリンキングである．

4. 評価実験と考察

今回は，提案方式の構成要素のなかで，特に「情報補完による Wikification」の可能性を検討するための評価実験を行った．NE 抽出器の評価データに対する性能に影響を受けずに Wikification の過程の評価を行うために，人手によるアノテーションを施した評価データを準備した．

4.1 評価データ

前述の「Wikipedia 日英京都関連文書対訳コーパス」から，カテゴリ分布を考慮しつつ，1,000 件の Wikipedia エントリをランダムに選出し，さらに，これらのエントリにおける先頭の 1 文の日英対訳ペアを評価データとして抽出した．日本語・英語の各文に対しては，人手により NE 区間をアノテートし，さらに，英語については 1,000 文全てに，日本語については 202 文に対して，正解としてリンキングすべき Wikipedia エントリを付与した．なお，リンキングすべきエントリが存在しない場合には，それを示す情報 (NIL) を付与した．

なお，今回の評価では，上記で作成した人手による NE の抽出結果を用いており，NE 抽出器の出力データを用いていないため，NE タイプの情報は利用できない．また，言語解析を省略しているので，NE 区間の伸縮を行うことはできない．以下に示す評価結果は，このような条件による．

4.2 単言語 Wikification の評価

英語 (1,000 文) に関しては，全 3,898 箇所の NE に対し，2,999 箇所について正しく単言語 Wikification が行えた (正解率 (精度=再現率):76.9%)．ただし，エントリが存在しないものを NIL と判定するものも正解に含む．899 件のエラー要因を以下に分類する．

- 正解データが NIL にも関わらず Wikification 結果がある (191 件): システムによる Wikification 結果が正解と考えうる場合[*7]が 55 件含まれていたが，残り (136 件) では，本来は適切なエントリが存在しないにも関わらず，候補の中から一つをリンキング対象としてしまった．
- 正解データが存在し，Wikification 結果も存在する (152 件): やはり正解とみなしうるものが 39 件含まれていた．また，対訳文の翻訳が適切でないケース 6 件も存在した．これら以外 (107 件) は Wikification 結果が誤っている場合であり，その内訳は，多義解消に失敗しているもの (93 件)，NE 文字列による検索では適切な候補エントリを得ることが難しい (しかし，得られている候補の中からリンキング先を決定した) もの (14 件) であった．
- 正解データが存在するが，Wikification 結果が得られていない (556 件): DBpedia データの不備などの要因 (2 件) を除くと，これらは，NE 表記と Wikipedia エントリのタイトル文字列が類似しているが一致しておらず，リダイレクトにより対応されない場合 (381 件) と，類似していない場合 (173 件) に分けることができる．

一方，日本語 (202 文) に関しては，全 762 箇所の NE に対し，716 箇所について正しく Wikification できた (正解率:94.0%)．今回は京都関連のテキストを対象としたので，日本語の場合の方が結果が良い．以下では，46 件のエラーの要因を分類する．

- 正解データが NIL にも関わらず Wikification 結果がある (4 件): 英語の場合と同様，本来は適切なエントリが存在しないにも関わらず，検索された候補の中から一つをリンキング対象としたケースである．
- 正解データが存在し，Wikification 結果も存在する (18 件): 全てが多義解消の誤りであった．日本の戦国時代 (エントリのタイトル:戦国時代_(日本)) を選択すべきところを中国の戦国時代を選択してしまった場合，京都市の北区 (エントリのタイトル: 北区_(京都市)) を選択すべきところを曖昧さ回避のためのページ自体 (エントリのタイトル: 北区) を選択してしまうという，今回の対象データの特性によるものの他，そもそも NE ではなく一般語と思われる語句が対象になっていた場合も存在した．
- 正解データが存在するが，Wikification 結果が得られていない (24 件): 表記の大きな異なり (一方が漢字，他方がひらがな書きなど)，言い換え・同義語の使用，設定された NE 区間や DBpedia のエントリ不備などが含まれる．

4.3 情報補完による Wikification の可能性

定義により，一方の言語の単言語 Wikification の結果が NIL であるが，本来そのリンキング先とするべきエントリが他方の言語の単言語 Wikification の結果であるエントリから言語間リンクで結ばれている場合が「情報補完による Wikification」を行いえる状況である．このような状況の割合を形式的に調べたところ，(1) 日本語側での Wikification に寄与する可能性のある英語 NE の割合は，10.7% (416/3898) であり，(2) 英語側

*6 等価な情報を関係付ける owl:sameAs プロパティによる．

*7 評定者は DBpedia ではなく Wikipedia を用いて作業したため，両者の不整合による正解の見逃しがあったほか，正解の基準に多少の揺れがあった．より厳密な正解の基準設定は今後の課題である．

での Wikification に寄与する可能性のある日本語 NE の割合は，20.5% (746/3646) であった．もちろん，これらの中には前節で述べたような割合で誤った単言語 Wikification を行っているものも含まれているが，総合的に見れば，日本語側の単言語 Wikification の結果を用いて英語側の単言語 Wikification の結果を改善できる可能性がより高いことが分かった．

4.4 情報補完による Wikification の可能性: ケーススタディ

上記の結果を受け，特に日本語側の Wikification の結果を利用することにより，英語側の Wikification が行える可能性がある場合のケーススタディを行う．

4.4.1 NE 表記と Wikipedia エントリのタイトル文字列が類似している場合

この場合は，表記のバリエーションを考慮して DBpedia に対する検索クエリを調整することにより，単言語 Wikification により処理することも可能であるが，多義解消の負担が増加する．また，DBpedia エンドポイントへの問い合わせが増加するという問題がある．日本の情報を対象とする場合，日本語側での単言語 Wikification の成功可能性が高いことを考慮すれば，情報補完による Wikification により対応する方が良い．NE 区間として対応する語句が英語側でも抽出されていれば逆方向のリンキング処理は比較的容易であるが，そうでない場合は，英語文中のどの区間 (語句) が対応させるべき区間かを探索的に定める必要が生じる．

長音記号の有無:今回の評価データにおける英語文中において，日本語 NE に対するローマ字表記が存在する場合，長音記号を伴っていない (例: 法隆寺に対して “Horyu-ji”) が，英語の Wikipedia では原則として長音記号を伴った形でタイトルが設定されている (“Hōryū-ji”)．多くの場合は，長音記号なしの文字列での検索は，長音記号ありの文字列へリダイレクトされるが，その設定がなされていない場合は，単言語 Wikification に失敗する．しかしながら，タイトル文字列の類似度を評価することにより，逆方向のリンキング (“Hōryū-ji” のエントリから “Horyu-ji” の NE へ) が比較的容易に行える．

その他の表記上の差異:ハイフンの有無 (“Hōryūji” か，“Hōryū-ji” か)，地名などに関する接辞の有無 (“Kita-ku”，“Sakyo Ward” など)，日本人の人名における姓と名の順序の異なり (“Sokan YAMAZAKI” に対して，“Yamazaki Sōkan) などが多数観察された．これらはルールによる対処が可能であるが，言語や対象ドメインに依存するものとなる．

4.4.2 文字列が類似していない場合

このケースは，図 2 に例示した “period of Northern and Sourthern Courts” のように，翻字によらない NE 表現であることが多く，英語による説明的な表現がされる場合に出現する．これらは，そもそも NE 抽出が困難と考えられるので，逆方向のリンキングによる Wikification への期待は高い．

例えば上記の例においては，“Nanboku-chō_period” の英語エントリの本文に存在する，“The Nanboku-chō period (南北朝時代 Nanboku-chō jidai, ”South and North courts period”, also known as the Northern and Southern Courts period), ...” といった表現をとらえ，リンキング元として適切な NE 区間を求めることができるが，この処理は，名詞句のバリエーション [岡崎，辻井 08] を生成しながら，入力文中に類似の表現を求めるという探索的な処理となる．

4.5 言語並行的なリンキングの成功率

一方の言語において NE として抽出された区間に対し適切にリンキング先が得られ，かつ，対応する他方の言語の NE に対しても適切なリンキングがなされた割合を調査したところ，(1) 日本語側の NE に対し，対応する英語側の NE のリンキングに成功した割合は，全データ 1,000 文に対して，11.9% (434/3646) であり，(2) 英語側の NE に対し，対応する日本語側の NE のリンキングに成功した割合は，日本語側の正解エントリが付与された 202 文に対して，38.1% (313/822) であった．適切な Wikidpedia エントリ自体が存在しない場合は成功としてカウントしていないので，単言語 Wikification の結果と合わせて考えれば，英語側での Wikipedia エントリの不足が大きく影響している結果となった．

4.6 考察: 候補エントリの適合度スコアリング

候補の Wikipedia エントリに対する適合度スコアは，曖昧性解消における相対比較に用いるだけでなく，絶対評価にも適用できるようにする必要がある．すなわち，不適切な候補へのリンキングを行わないよう，適合スコアよる閾値処理を行うべきである．一方，情報補完による Wikification の精度が十分高いと想定できる場合には，すでに一方の言語の単言語 Wikification によって得られている結果を撤回し，逆方向のリンキングの結果に基いて修正することが有用である．提案方式は対訳文の双方に対して対称的 (バイアスがない) であるが，今回の評価データのように，対象テキストに対するトップダウン的な事前知識が得られている状況であれば，Wikipedia における情報の質量の差を考慮した処理が有効と考えられる．

5. おわりに

本論文では，対訳文に対して双方の言語で Wikification を行うというタスクにおいて，一方の言語における Wikification の不備を他方の言語の情報を用いることにより補完できる可能性を示した．今回の評価結果は人手によって作成された NE 抽出を結果対象としているが，実際の適用場面においては，NE 抽出を NE 抽出器により行う必要がある．提案方式は，対象とすべき NE 区間がその一部でも抽出されていることを要求するが，NE 抽出器の精度・性能は対象とするテキストの領域に対する学習状況に大きく依存する．このため，実際の NE 抽出器の精度・性能をある程度向上させ，NE 抽出器を用いた場合のさらなる研究課題を抽出・検討していく必要がある．

本提案手法は，言語に並行的な Wikification の結果として得られる NE の対応関係を予め句テーブルに埋め込む，または翻訳結果から動的に生成することにより，ユーザが翻訳結果を理解するための支援機能に適用できると考えられる．

参考文献

[Lehmann et al. 14] Lehmann, J. et al.: DBpedia — A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, To appear in the *Semantic Web Journal*, 2014.

[Lesk 86] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, *Proc. of SIGDOC '86*, pp.24–26, 1986.

[Mihalcea and Csomai 07] Mihalcea, P. and Csomai, A.: Wikify!: linking documents to encyclopedic knwoledge, *Proc. of CIKM 2007*, pp.233–242, 2007.

[岡崎，辻井 08] 岡崎直観, 辻井潤一: 名詞句における用語バリエーションの自動認識, 第 22 回人工知能学会全国大会, 1H1-2, 2008.

[Winkler 90] Winkler, W. E.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage, *Proc. of the Section on Survey Research Methods (American Statistical Association)*, pp.354–359, 1990.