E-004

英語 Wikipedia リンクデータの利用による日本語 wikification
Wikification of Japanese Articles Using Link Data in English Wikipedia

村上 凌悠† Ryosuke Murakami　綱川 隆司† Takashi Tsunakawa　西田 昌史† Masafumi Nishida　西村 雅史† Masafumi Nishimura

1. はじめに

Wikipedia は様々な分野をカバーする大規模百科事典である。記事中にはある語句からその語句を説明する Wikipedia 記事へのリンクがあり、より効率的に記事内容を理解できる。リンクを自動的に張る wikification[1]の実現により、記事の品質を高めることが期待される。

Wikification を実現するためには、リンクが張られる語句の曖昧性解消を行ってリンク先となる記事を決めるリンク先決定が必要であり、このためには機械学習を用いた手法が多く用いられる[2][3]。

機械学習の素性には、アンカー(リンクが張られる語句)、そのアンカーと共起するアンカーがあり、トレーニングデータはそれらの組からなる。しかし、アンカーによっては出現する記事が少ない場合や記事中の共起アンカーの数が少ない場合があるので、対象となる言語版のトレーニングデータだけでは十分でないものが存在する。したがって本研究では他言語版である英語のリンクデータをトレーニングデータとして日本語 wikification の学習に追加し、その影響を分析する。日本語のリンクデータのみで学習した場合と、日本語のリンクデータに英語のリンクデータを追加して学習した場合の比較実験を行う。

2. 提案方法

曖昧な語は、周囲に現れる関連のある語によって選別できる。例えば“ジャガー”は動物のジャガーと自動車のジャガーの意味があるが、周辺に“BMW”や“ポルシェ”など自動車関係の語句が出現すれば“ジャガー”は自動車に関係しているという判断ができる。アンカーのリンク先決定にも同様の考え方を用いる。

2.1 決定リストによる学習

各アンカーに対し、アンカーとリンク先の組を正解データとし、そのアンカーと共起するアンカーとの関連を学習する決定リストに基づく機械学習の手法が提案されている。本研究では袁ら[3]の方法に基づき各アンカーの決定リストを学習する。決定リストは以下のようなルールのリストである。

「IF “BMW” co-occurs with “ジャガー” THEN link “ジャガー” to “ジャガー (自動車)”」

上記のルールは、アンカー“BMW”がアンカー“ジャガー”と共起すれば、アンカー“ジャガー”は記事“ジャガー (自動車)”にリンクすることを意味する。このようなルールをトレーニングデータから求めた確信度順に並べる。確信度はクロス集計表を基にアンカー、リンクの組と共起アンカーが従属な場合と独立な場合との対数尤度比で求める。アンカーのリンク先決定時には、決定リストの上位のルールから順に該当する共起アンカーが存在するかを調べ、共起アンカーが存在する上位 3 つのルールの多数決で決定する。

なお、共起アンカーとの共起頻度は、1 つの記事に対して同じ共起アンカーが複数回出現しても 1 回とカウントする。

2.2 英語リンクデータの利用

英語版 Wikipedia リンクデータを日本語におけるリンク先決定に用いるために、英語版 Wikipedia リンクデータを翻訳する。このため同じ概念を表す英語と日本語の Wikipedia 記事間に張られた言語間リンクを用いる。図 1 は、アンカー“Jaguar”、そのリンク先“Jaguar Cars”、および共起するアンカー“Volkswagen”からなるトレーニングデータの翻訳の例を示している。リンク先“Jaguar Cars”においては言語間リンクで直接対応する“ジャガー (自動車)”を翻訳後のリンク先とする。アンカーである“Jaguar”と“Volkswagen”の翻訳は、そのリンク先記事の言語間リンクをたどり、その対応する記事にリンクを張るアンカーを翻訳語として用いる。アンカー“Volkswagen”の場合だと“VW”，“フォルクスワーゲン”が翻訳語となる。このように 1 つのアンカーに対して複数の翻訳語が得られる場合がある。

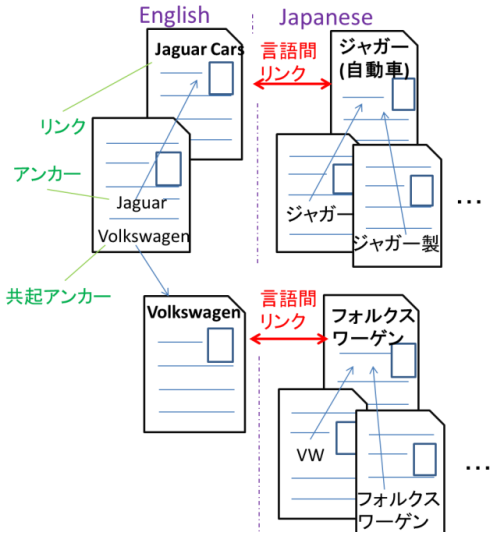

図 1 英語リンクデータの翻訳法

2.3 適切な翻訳語の抽出

上記のように 1 つの英語アンカーに対して複数の日本語アンカーが得られるが、それらの中には適切でないものが存在する。例えば、アンカーが指示語になっている場合(例 アンカー“それ ”→リンク先記事“円卓会議 (ポーランド)”)、特殊な文脈でのみリンクを意味する場合 (例 アンカー“不 二”→リンク先記事 ”富士山 ”)、単純に誤っている場合(例 アンカー“LED”→ リンク先記事“バーコ

† 静岡大学大学院総合科学技術研究科情報学専攻

ード”)が考えられる。これらの適切でない翻訳語に関するルールの確信度を小さくするため、対数尤度比の計算に用いる頻度を、アンカーaがリンク先Dを指す確率$TP(a,D)$で重み付けする。

$$\mathrm{TP}(a,D)=\frac{count(a,D)}{\sum_{D'} count(a,D')}$$

ここで、aは翻訳後のアンカーを表し、Dは翻訳に用いるリンク先記事、D'はaからリンクされる任意の記事とする。$count(a,D)$はアンカー aとリンク先 Dの組が存在する記事数である。

3. 評価実験

日本語リンクデータのみで学習した場合と日本語リンクデータに英語リンクデータを追加して学習した場合の比較実験を行った。

3.1 実験設定

2016 年 2 月 3 日時点の日本語版 Wikipedia、2016 年 2 月 4 日時点の英語 Wikipedia 記事を使用し、曖昧性解消ページを除いた標準ページからリンクデータを抽出した。リンク先記事がリダイレクトページの場合はそのリダイレクト先をリンク先記事とした。リンク先候補が 2 以上の曖昧なアンカーを 500 個ランダムに選択し、それらのアンカーを評価対象とした。

評価は以下のように行った。Wikipedia 記事全体を 10 分割し、分割した記事の部分集合のそれぞれについて、評価対象のアンカーを含む記事をテストセットとし、残り 10 分の 9 の記事集合をトレーニングデータとして学習を行う 10 分割交差検定により、正解率を評価した。

Wikipedia 記事中のリンクは、実際に指定されているリンク先記事とは異なる記事にリンクを張っていても不適切とは言えないものが存在するが、本実験では実際に指定されたリンク先記事のみを正解であるとみなす。

3.2 実験結果

日本語のみの場合の正解率は 92.0%、英語を追加した場合の正解率は 92.1%となり、正解率はわずかに向上した。また交差検定の各検定過程において評価に用いた正解リンクデータの総数は 52036 で、日本語のみで正解だったものが英語を追加して不正解になったリンクの総数は 748、日本語のみで不正解だったものが英語を追加して正解になったリンクの総数は 796 あった。このことから、改善した場合と悪化した場合の双方があることがわかった。

4. 現時点の課題

以下に実験と学習における現時点の課題を示す。

1)言語間リンクの偏りの影響

複数のリンク先候補のうち、一方にだけ言語間リンクが存在する場合、そのリンク先候補だけ非常に多くの学習データが用いられ、決定リスト中のルールの大半がそのリンク先に関するルールになり、別のリンク先候補の決定に影響を与える場合がある。

例えば、アンカー“メトロポリタン美術館”の場合リンク先候補に“メトロポリタン美術館 (みんなのうた)”、“メトロポリタン美術館”が存在するが、前者は言語間リンクが存在しない。表 1 アンカー“メトロポリタン美術館”に対する決定リストのリンク先候補別のルール数に日本語の決定リストと英語を追加した場合の決定リストそれぞれのリンク先のルールの数を示す。英語リンクデータを追加することで言語間リンクが存在するリンク先候補”メトロポリタン美術館”のルールの数が大幅に増加した。また、上位 2405 位まで全て“メトロポリタン美術館”へのリンクのルールとなった。このことで、“メトロポリタン美術館 (みんなのうた)”にリンクすべきアンカーに対しても“メトロポリタン美術館”にリンクする可能性が高まり、リンク先決定に悪影響が現れたと考えられる。

表 1 アンカー“メトロポリタン美術館”に対する決定リストのリンク先候補別のルール数

リンク先候補	ルール数 (Ja)	ルール数 (Ja+En)
メトロポリタン美術館 (みんなのうた)	2963	2963
メトロポリタン美術館	26105	218792

2)分野、地域などの共通性の影響

複数のリンク先候補に分野、地域などの共通点があれば、現れる共起アンカーも共通のものが多くみられるようになり、共起アンカーによるリンク先決定が難しくなる。例えばアンカー“姫路”はリンク先候補として“姫路城”や“姫路駅”、“姫路市”などがあるが、どれも姫路という地名に関係しているので共起アンカーとして同じ語が現れやすくなり、リンク先決定が難しくなる。英語のリンクデータを追加しても、不正解となった例にこのようなものがなお多く見られたので、単純に英語のリンクデータを追加するだけではこの問題への効果が現れなかったと考えられる。

5. おわりに

決定リストにおける日本語の wikification に英語のリンクデータを翻訳して追加する方法を提案し、その影響を分析した。追加することで正解となったものもあれば、不正解となったものもあり、全体としては正解率がわずかに向上した。なお、評価には実際に指定されたリンク先記事のみを正解としたが、そのリンク先記事以外にもリンクが張られても間違いではないような記事が存在するので、より正確な評価のためには人手評価も重要であると考えられる。

謝辞

本研究は JSPS 科研費 JP15K16096 の助成を受けたものです。

参考文献

[1] Mihalcea Rada, Andress Csomai, “Wikify!: linking documents to encyclopedic knowledge”, In Proc. of CIKM 2007, pp. 233-242(2007).

[2] D. Milne and I. H. Witten, “Learning to Link with Wikipedia” In Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp.509-518(2008).

[3] 袁楊, 綱川隆司, 梶博行, ”決定リストの機械学習による wikification” ,言語処理学会第 21 回発表論文集, pp. 688-691(2015).