様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 １８ 日現在

機関番号：１３８０１
研究種目：若手研究(B)
研究期間：2015～2018
課題番号：１５Ｋ１６０９６

研究課題名（和文）Wikipediaの多言語性を利用したwikificationの高精度・高機能化

研究課題名（英文）Developing high-performance and high-functional wikification using multilinguality of Wikipedia

研究代表者

綱川　隆司（TSUNAKAWA, Takashi）

静岡大学・情報学部・助教

研究者番号：３０６１１２１４

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：テキスト中に現れる重要な固有表現等の語句にWikipedia記事を対応付けるwikificationタスクについて、他の言語版のWikipedia記事に含まれるリンクを日本語に変換する方法を提案し実証実験を行った。実験の結果、変換したリンク数の増加は確認されたもののリンク先の決定精度向上までには至らなかった。また、サブタスクであるアンカー抽出およびリンク先決定のそれぞれについて、新たな特徴量および共起語の抽象性に基づいた方法を提案し、それぞれ性能向上を確認した。

研究成果の学術的意義や社会的意義
テキストにWikipedia記事へのリンクを付与することにより、テキストの理解可能性を高めるのみならず、情報抽出等の自然言語処理タスクへの応用も期待される。本研究は、従来のWikipediaのリンク言語間変換の手法を発展させてwikificationに応用し、全言語のWikipediaのリンクデータを用いてwikificationの性能向上を図る方法を検討したものであり、リンク数を増加させることで新たなリンクの提案を一定の精度で行うことができる可能性を示した。

研究成果の概要（英文）：We proposed a method for converting links in Wikipedia articles of other languages into Japanese for wikification task, which links important phrases in text to Wikipedia articles. The experimental results did not show improvement of performance for link destination determination, although the number of converted links increased. In addition, for each of subtasks, anchor detection and link destination determination, we proposed methods based on new features and abstractness of co-occurring words, and confirmed improvement of performance.

研究分野：自然言語処理

キーワード：Wikification　エンティティリンキング　Wikipedia　多言語　情報組織化

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

## １．研究開始当初の背景

(1) 近年、テキスト中に出現する重要な語句を特定し、その語句が表すものに対応する Wikipedia 記事へのハイパーリンクを自動的に付与する、wikification と呼ばれる課題が提起され、研究が進んでいる。wikification は、Wikipedia 記事を参照する際の手間を大きく軽減することにとどまらず、テキスト中の重要な語句が表す概念を同定する技術であるので自然言語理解の発展に寄与することが期待される。

(2) wikification タスクは大きく分けて、テキスト中の重要な語句の特定（アンカー抽出）と各語句に対応するリンク先の Wikipedia 記事の決定（リンク先決定）の 2 つのステップからなる。アンカー抽出はテキストからのキーワード抽出や固有表現認識に用いられる手法が応用されているが、アンカーとするかどうかの基準は入力テキストの性質によって異なりうるため、既存のアンカー抽出に関する研究は Wikipedia 記事そのものを対象としたものがほとんどであり、かつ日本語のテキストを対象としたものは非常に少ない。

(3) 一方、リンク先決定は語義曖昧性解消と呼ばれるタスクの特殊なケースとみなすことができる。語義曖昧性解消の手法として、テキスト中のアンカーと共起する語の分布を用いるものが知られており、wikification では共起するアンカーを用いる方法が一般的である。Wikipedia 記事はアンカーのリンク先が定められているので、このリンクデータをトレーニングデータとして用いた教師あり機械学習が最も有望なアプローチである。

## ２．研究の目的

(1) wikification におけるリンク先決定のためのトレーニングデータ不足の問題を改善するため、従来研究のように対象言語の Wikipedia のリンクデータのみを利用するのではなく、全ての言語版の Wikipedia に含まれるリンクデータを利用する方法を提案した。図 1 に示すように、他の言語の記事に含まれるリンクを対象言語のリンクに変換することにより、共起アンカーの量を飛躍的に増加させることができる。ただし、他言語の記事と言語間リンクで結ばれていない記事も多く、そのようなリンクを対象言語のリンクに変換することは自明ではないため、これを解決するのが一つの技術課題となる。

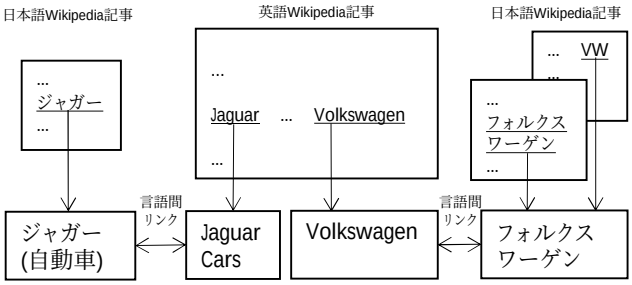


図 1　トレーニングデータの言語間変換

英語のWikipedia記事からトレーニングデータ「“Volkswagen”と共起する“Jaguar”のリンク先記事＝“Jaguar Cars”」が得られ、リンクの言語間変換により「“フォルクスワーゲン”と共起する“ジャガー”のリンク先記事＝“ジャガー (自動車)”」および「“VW”と共起する“ジャガー”のリンク先記事＝“ジャガー (自動車)”」に変換される。

(2) また、他言語の記事の利用により、アンカー抽出によりリンクを付与すべきと判定した語句に対応する記事がテキストと同じ言語版の Wikipedia に存在しない場合、英語版等他の言語版の Wikipedia に対応する記事がある場合はその記事にリンクを張る言語横断 wikification（図 2）も同様の方法を用いて実現できると考えられる。

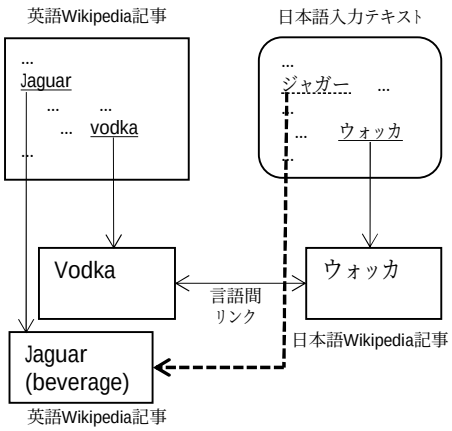


図 2　言語横断wikification

日本語Wikipediaには、飲み物のジャガーについて説明する記事が存在しないので、“ジャガー”から英語記事“Jaguar (beverage)” にリンクを張る。

(3) さらに、一般にアンカー抽出およびリンク先決定の精度を高めるため、先行研究において有効とされる手掛かりをもとに、新たにいくつかの特徴量を導入して精度の向上を図る。

## ３．研究の方法

(1) リンク先決定のためのトレーニングデータである“リンクと共起アンカーの対”を日本語版と英語を含む 15 の言語版の Wikipedia から抽出し、図 1 に示す方法により日本語以外のデータから日本語へ変換することで抽出可能なアンカーがどの程度増加するか調べ、かつその内適切と考えられるアンカーの割合を調査する。

(2) アンカー抽出性能を向上させるため、以下の 2 つの特徴量を提案し、それらの効果を示す。

① アンカー候補語句の前接語・後接語：ある語句がアンカーとして扱われやすいかどうかは、その前後の語句にも依存していると仮定し、アンカー候補語句の直前の語（前接語）および直後の語（後接語）を考慮に入れた出現確率を特徴量として導入する

② 条件付き keyphraseness：先行研究において有効性が知られている keyphraseness はアン

カー候補語句のみで求められる値であるが、同一テキスト内に共起した他のアンカーを考慮した keyphraseness を導入することで、共起関係を考慮した特徴量となる

また、wikification の性能評価に用いることのできる日本語のデータとして“日本語 Wikification コーパス”(引用①)が挙げられるが、テキスト中の固有表現に対するリンク先決定の評価を対象としたものであり、本研究ではこれをアンカー抽出評価にも用いられるよう拡張し、提案方法の評価に用いる。

(3) リンク先決定の性能を向上させるための新たな特徴量として、手掛かりとして重要な共起語を抽象的なもの（抽象的特徴）とそれ以外のもの（非抽象的特徴）に分けて考え、最初に抽象的特徴を用いてリンク先候補をクラスタリングし、抽象的特徴によってリンク先候補として適切なクラスタを特定してから、非抽象的特徴のみを用いてクラスタ内の適切なリンク先を決定する 2 段階のリンク先決定方法を提案し、その効果を示す。

＜引用文献＞

① Davaajav Jargalsaikham、岡崎 直観、松田 耕史、乾 健太郎、日本語 Wikification コーパスの構築に向けて、言語処理学会第 22 回年次大会、2016

４．研究成果

(1) リンク先決定のためのトレーニングデータを言語間変換することで新たに得られるデータ数、およびそれらが適切である割合（正解率）を人手評価セットを用いて評価した結果を表 1 に示す。

表 1　リンクの言語間変換により得られたリンクの総数と正解率

| 変換元言語（言語数） | | en (1) | +fr (2) | +de (3) | +it (4) | +zh (5) | +es (6) | +ru (7) | +マイナーな 7 言語 (14) |
|---|---|---|---|---|---|---|---|---|---|
| 既存リンク | 総数 | 1776 | 2034 | 2340 | 2451 | 2550 | 2614 | 2684 | 2684 |
| | 正解率 | 98.9% | 98.9% | 98.8% | 98.9% | 98.9% | 99.0% | 98.8% | 98.8% |
| 新規リンク | 総数 | 904 | 1193 | 1480 | 1711 | 1916 | 2052 | 2165 | 2165 |
| | 正解率 | 83.7% | 83.1% | 82.8% | 81.4% | 81.5% | 81.5% | 81.0% | 81.0% |

表 1 の“既存リンク”は変換した結果が既存のリンクのアンカーと一致したもの、“新規リンク”は既存のものと一致せず新たに得られたと判断されるものを示し、正解率はリンクの変換結果が正当であると判定されたものの割合を示す。変換元言語として用いる言語数を増やすに伴い既存リンク・新規リンクのいずれもリンク数の増加に寄与することが示された。一方で変換したリンクの正解率は新規リンクについて低下傾向にあり、単に言語数を増やすのみでは品質向上につながらないことが示された。

また、英語から日本語に変換したリンクを用いてトレーニングデータを追加した実験を行った結果、リンク先決定の正解率は追加前が 92.1%、追加後が 92.2%となり有意差はみられず、リンク先決定性能の改善には変換したリンクの品質改善やその他の特徴量への反映等が必要と考えられる。

(2) 日本語 wikification におけるアンカー抽出性能の評価を行うため、日本語 Wikification コーパスをベースに、重要度、関連度、無名度の 3 つのアンカー抽出基準を定めてアンカーを人手で付与した。この結果、日本語 Wikification コーパスにおいて対象となっている拡張固有表現 3698 個のうち 2337 個を抽出すべきアンカーと判定し、また拡張固有表現以外に新たに 867 個のアンカーを付与して評価用コーパスを作成した。

アンカー抽出の提案方法の効果を示すため、Wikipedia 記事をテストデータとして用いた実験では表 2 のように性能向上が確認された。また、評価用コーパスをテストデータとして用いた実験では正解率 0.745、F 値が 0.636 という結果が得られ、Wikipedia 記事をテストデータとして用いた場合よりも低い F 値が得られたが、評価用コーパス作成時の作業者 2 者間一致率 0.733 と近い正解率が得られており、人手による評価と近い一致率が得られているといえる結果が示された。

表 2　アンカー抽出実験結果

| | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| ベースライン（従来方法） | 0.761 | 0.759 | 0.760 |
| 提案方法 | 0.782 | 0.771 | 0.776 |

(3) リンク先決定の性能改善に向けた抽象的特徴によるリンク先候補のクラスタリングの効果、および決定リストにおいて非抽象的特徴のみを用いることによる効果を示すため、Wikipedia 記事および日本語 Wikification コーパスを用いたリンク先決定評価実験を行った。実験結果を表 3 に示す。

表 3　抽象的特徴を用いたリンク先決定の評価実験結果

| 抽象的特徴によるクラスタリング | 非抽象的特徴のみの利用 | Wikipedia 記事における正解率 | 日本語 Wikification コーパスにおける正解率 |
|---|---|---|---|
| 非適用 | 非適用 | 91.15% | 82.36% |
| 非適用 | 適用 | 93.40% | 85.03% |
| 適用 | 適用 | 91.24% | 67.44% |

いずれの評価データにおいても、クラスタリングを用いた場合は当初の方法と比べ効果がほぼ見られないか性能が低下した一方で、クラスタリングを行わずに非抽象的特徴のみを用いた場合に最も高い性能が得られた。手掛かりとしての重要な共起語の抽象性に着目し、抽象的な共起語を用いないことにより性能が向上することが示された。

(4) 本研究ではリンク先決定方法に従来手法で有効であった決定リストを用いる方法を採用しているが、近年は深層学習を用いた手法が最高精度を達成しており、wikification の対象言語以外の言語のリンクデータの情報が有用であることも報告されてきている。当初の方向性自体は有効であると考えられるものの、最新の研究を行うには手法の大きな見直しが必要であることが判明し、研究期間内では実証までには至らなかった。今後引き続き手法の検討を進めるとともにより多くの種類の情報を取り入れることを検討していく。

## ５．主な発表論文等

〔雑誌論文〕（計１件）


① 小谷 亮太、綱川 隆司、西田 昌史、西村 雅史、日本語 Wikification におけるアンカー抽出器および評価用コーパスの構築、情報処理学会論文誌、査読有、Vol.59、No.2、2018、pp.306-314


〔学会発表〕（計７件）


① Takashi Tsunakawa, Recent Activities of Speech and Language Research Group at Shizuoka University, The 18th China-Japan Natural Language Processing Joint Research Promotion Conference (CJNLP 2018), 2018
② 村上 凌悠、綱川 隆司、西田 昌史、西村 雅史、リンク先決定における特徴の抽象性を利用した wikification の精度向上、言語処理学会第 24 回年次大会、2018
③ 村上 凌悠、綱川 隆司、西田 昌史、西村 雅史、リンク先の抽象的特徴を利用した wikification の精度向上、第 15 回情報学ワークショップ(WiNF2017)、2017
④ 小谷 亮太、綱川 隆司、西田 昌史、西村 雅史、日本語 Wikification コーパスを用いたアンカー抽出性能評価に関する検討、情報処理学会第 229 回自然言語処理研究会、2016
⑤ Takashi Tsunakawa, Ryota Kotani, Ryosuke Murakami, Masafumi Nishida, and Masafumi Nishimura, Enriching Wikipedia link data for wikification, The 16th China-Japan Natural Language Processing Joint Research Promotion Conference (CJNLP 2016), 2016
⑥ 村上 凌悠、綱川 隆司、西田 昌史、西村 雅史、英語 Wikipedia リンクデータの利用による日本語 wikification、第 15 回情報科学技術フォーラム、2016
⑦ 小谷 亮太、綱川 隆司、西田 昌史、西村 雅史、Wikification における前接語・後接語を用いたアンカー抽出、第 15 回情報科学技術フォーラム、2016


〔その他〕

静岡大学情報学部 西村研/綱川研/西田研(NIST-Lab) ホームページ
http://lab.inf.shizuoka.ac.jp/nisimura/
静岡大学情報学部 綱川研究室 ホームページ
https://wwp.shizuoka.ac.jp/tsunakawa/

## ６．研究組織

(1)研究分担者
なし

(2)研究協力者
研究協力者氏名：梶 博行
ローマ字氏名：(KAJI, hiroyuki)

研究協力者氏名：小谷 亮太
ローマ字氏名：(KOTANI, ryota)

研究協力者氏名：村上 凌悠

ローマ字氏名：(MURAKAMI, ryosuke)

研究協力者氏名：西田 昌史
ローマ字氏名：(NISHIDA, masafumi)

研究協力者氏名：西村 雅史
ローマ字氏名：(NISHIMURA, masafumi)