Vol.2016-NL-229 No.14
2016/12/21

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

小谷亮太† 綱川隆司† 西田昌史† 西村雅史†

概要：一般の文書から Wikipedia 記事へ自動的にリンクを貼る wikification の研究が現在注目されており，日本語においては日本語 Wikification コーパスが公開されている．Wikification にはリンクを貼るべき語句（アンカー）の選定と，選定したアンカーのリンク先記事決定の 2 つの課題が挙げられる. 前者の課題では wikification タスクをどう定義するかによってさまざまな選定方法が考えられることから，日本語 Wikification コーパスではアンカー選定の対象を固有表現に限定してリンク先記事のアノテーションを実施している．しかし，Wikipedia では記事において重要な一般名詞等がアンカーとして選択されるケースも多く，また，重要でない自明な固有名詞はアンカーとして選択されないこともあり，一般の文書に対するアンカー抽出の評価を行うには不十分である．本研究では一般名詞等を含むより広範囲なアンカー抽出の評価を行うため，日本語 Wikification コーパスに対して新たに人手でアンカー選定作業を行った．また，Wikipedia のリンクデータとこのコーパスを学習データとして用いて SVM によるアンカー抽出器を作成し，性能を評価した

Performance Evaluation of Anchor Extraction Using the Japanese Wikification Corpus

RYOTA KOTANI† TAKASHI TSUNAKAWA† MASAFUMI NISHIDA†
MASAFUMI NISHIMURA†

1. はじめに

Wikipedia は巨大なハイパーテキストであることを特徴とする Web 上の百科事典である．Wikipedia 記事に付与されたリンクにより，他の Wikipedia 記事を参照することができる．一般の文書から Wikipedia 記事を容易に参照できるようにするため, Wikipedia 記事に自動的にリンクを張る wikification の研究が盛んに行われている[1][2][3].

Wikification は，リンクを貼るべき語句（アンカー）の選定を行う第 1 ステップと，抽出されたアンカーのリンク先記事を決定する第 2 ステップから成っている[1]. 第 2 ステップは語義曖昧性解消の問題であり，様々な手法が試みられている. Milne and Witten [4]は語句が特定の記事にリンクされる頻度と語句同士の意味間の関係の強さに加え，語句が出現する文脈を考慮している．袁ら[5]は機械学習の手法である決定リストを用いた方法を提案している．これに比べると第 1 ステップに関する研究は少ない. Wikipedia でアンカーとなっている語句全てをアンカーとして採用する方法もあるが，本研究では，文書中の重要な語句や当該文書の読者が十分な知識をもっていないような事項を表す語句のみをアンカーとして抽出する方法に焦点を当てる.

一方, 日本語を対象とした wikification の研究もいくつか存在するが[6][7]，Wikipedia 記事を評価に使用しているものが多く，新聞記事など Wikipedia 記事以外での評価は少ない.

BCCWJ（現代日本語書き言葉均衡コーパス）のコアデータに対して, 関根の拡張固有表現(Version 7.1)[a]の境界情報を人手でアノテートした拡張固有表現タグ付きコーパスが存在する．Davaajav ら[7]はこのコーパス内の新聞記事 340 記事にアノテートされている拡張固有表現(ENE)に対して（時間表現，数値表現，アドレス，称号名，施設部分名は除く），Wikipedia エントリを付与した日本語 Wikification コーパスを作成し，リンク先記事決定の評価を行った．このコーパスは現在公開されているが，本研究の目的とするアンカー抽出にはそのまま使用することはできない．このコーパスは文書の内容を問わず ENE にエントリを付与しているが，本研究のアンカー抽出では文書の内容によっては単なる一般名詞や動詞・形容詞であってもアンカーとして選択され, 逆に ENE であっても文書の内容によってはアンカーとして選択されない場合があると考える.

本稿では公開されている日本語 Wikification コーパスに対して新たにアンカーとなりうる可能性がある語句（アンカー候補語句）を追加した上で，それぞれアンカーとするかどうかの判断を人手で行い，アンカー抽出のための日本語 Wikification コーパスを再構築した.

我々の以前の研究において使用した素性[8]を用いて Wikipedia のリンクデータを学習データとした場合と Wikipedia のリンクデータに今回新たに再構築した日本語 Wikification コーパスを一部追加して学習データとした場合のそれぞれに対してアンカー抽出器を学習し，日本語 Wikification コーパスを用いて評価を行ったので, その結果について報告する.

† 静岡大学
Shizuoka University

a https://sites.google.com/site/extendednamedentityhierarchy/

2. アンカー抽出のための日本語 Wikification コーパスの作成

2.1 アンカー選定方法

公開されている日本語 Wikification コーパス内でアノテートされている固有表現とは別に，新聞記事の内容によっては一般名詞や動詞・形容詞の中にもなどもアンカーとすべき語句は存在すると考えられるため，それらをアンカー候補語句に追加する．候補語句は日本語版 Wikipedia の全記事約 100 万件の中でアンカーになったことのある語句とする．しかし，Wikipedia 記事はさまざまな編集者によって作成されているため，アンカーの指定方法が Wikipedia のガイドラインに沿っていない例が存在する．このような語句は学習に悪影響を及ぼすため，Wikipedia 記事の中で 5 回以上アンカーとして出現した語句を対象とすることにした．

日本語 Wikification コーパスでは同じ対象を指す ENE が複数回出現した場合，全ての出現位置にアノテートされているので，Wikipedia のガイドラインに従い，見出しを除く本文中の最初の出現箇所のみをアンカー抽出の対象にする．

アンカー抽出対象の選択は詳細なガイドラインなどが存在しないため[9]，人によってアンカーの付け方に偏りが生じる．本研究では Wikipedia のガイドラインに準じ， 以下に示す 3 つの基準を用いてアンカー選定作業を行った．

関連度

新聞記事の主題と注目する語句の間の関連性の高さを関連度とする．特に，注目する語句が新聞記事の主題の属性になっているかどうかを考慮する．例えば自動車の記事に対して「経営」や「顧客」などは「自動車」という主題に対する属性ではないため関連度は低いとする．一方で「エンジン」や「ブレーキ」は「自動車」という主題の属性であるため関連度を高いとする．

重要度

新聞記事の本文を要約したときに注目する語句が残るかどうかを考え，より短い要約でも残る語句であればより高い重要度を持つとする．

認知度

客観的に見て，注目する語句が一般に認知されているかどうかを考える．認知されていない語句ほど認知度が低いとする．

以上 3 つの基準の中で関連度か重要度が特に高いと判断した語句，または認知度が特に低いと判断した語句についてはアンカーとして採用し，それ以外の場合は 3 つの基準をもとに総合的に判断する．

また，同じ語句でなくても同じリンク先記事を表す語句が複数出現した場合（「鈴木宗男」と「鈴木」,「彼」など），最初に出現した語句をアンカー抽出の対象とする．

2.2 アンカー選定作業の結果

アンカー選定作業の結果得られたアンカー数の内訳を表 1 に示す．

表 1 日本語 Wikification コーパス新聞記事 100 件のアンカー数とその内訳

	選定前	選定後
アンカー数	4771	2939
ENE	4771	2412
ENE以外	0	527

日本語 Wikification コーパス内の新聞記事 340 件の中から無作為に選択した 100 件を対象とし，それぞれ 3 人がアンカー選定作業を行った．このうち 2 人以上がアンカーとして選定した語句を採用する．アンカー選定作業者の 2 者間の一致率の平均は約 77%となった．

拡張固有表現としてアノテートされている語句は延べ 4771 あり，その中の 2359 がアンカーとして採用されなかった．この中には同じリンク先記事を表す語句が複数回出現した場合も含まれる．得られた合計のアンカー数は 2939 あるが，この中にはリンク先記事が存在しない(NIL)ものも複数存在する．本研究ではアンカーにすべき語句は Wikipedia 記事が存在することを前提とするのでそれらの語句を除いたものを使用する．NIL となったものは 558 あるので使用するアンカー数は 2381 となった．

3. アンカー抽出器で使用する素性

2 節で作成した日本語 Wikification コーパス 100 記事に対し，SVM（サポートベクターマシン）を用いてアンカー抽出性能を評価する. 1) Wikipedia のリンクデータを学習データとした場合と，2) Wikipedia のリンクデータに，作成した日本語 Wikification コーパスの一部を追加して学習データとした場合，のそれぞれについて教師付き学習を行った．全データを 10 に分割し（903 語句），うち 9 を教師付きの学習データ，残り 1 を評価対象として交差検証を行った．

今回使用した素性を以下に説明する.これらは先の先行研究[8]において,アンカー抽出に対する有効性を確認した素性である.

(1) **keyphraseness**

keyphraseness は候補語句が出現した Wikipedia 記事のうち, その語句がアンカーとして出現する記事の割合を表す.

$$Key(a) = \frac{|\{D_w | a \in Anchor(D_w)\}|}{|\{D_w | a \in D_w\}|}. \quad ①$$

ここに，aは候補語句，D_wは Wikipedia 記事，$Anchor(D_w)$は記事D_wに含まれるアンカーの集合とする．

(2) 候補語句の前接語・後接語

候補語句の前後の語句によって候補語句がアンカーになりやすいかどうかに影響すると考えられる．例えば，候補語句の直後が「等」や「的」である場合，候補語句はアンカーになりやすい傾向がある．このような考えに基づいて以下の 2 つの素性を検討する．

(2a) 前接語のプリアンカー確率

語xのプリアンカー確率をxの次の語がアンカーである確率として定義する．すなわち，

$$PreAnchor(x) = \frac{\left|\{D_w | \exists y \in Anchor(D_w).\ x \cdot y \in Bigram(D_w)\}\right|}{|\{D_w | x \in D_w\}|}. \quad ②$$

ここに，$Bigram(D_w)$は記事D_wに含まれるバイグラムの集合である．・（ドット）は語の連接を表す．

候補語句aの素性としてはaの前接語$pred(a)$のプリアンカー確率$PreAnchor(pred(a))$を用いる．

(2b) 後接語のポストアンカー確率

語xのポストアンカー確率をxの前の語がアンカーである確率として定義する．すなわち，

$$PostAnchor(x) = \frac{\left|\{D_w | \exists y \in Anchor(D_w).\ y \cdot x \in Bigram(D_w)\}\right|}{|\{D_w | x \in D_w\}|}. \quad ③$$

候補語句aの素性としてはaの後接語$succ(a)$のポストアンカー確率$PostAnchor(succ(a))$を用いる．

(3) 候補語句の条件付き keyphraseness

候補語句と共起する候補語句との間の関連の強さによってアンカーへのなりやすさが関係すると考えた．例えば，候補語句「BMW」は「ドイツ」や「ベンツ」などと共起する場合，アンカーになる確率が高いのではないかと思われる．アンカー抽出の研究で用いられている関連度を測る指標としては relatedness 指標[4]が存在するが，これは暫定的にリンク先記事を決定する必要があり，リンク先記事決定のタスクは本研究では対象外のため使用することができない．そのため新たにリンク元の記事情報を使用する共起候補語句を条件とする候補語句の keyphraseness を素性として提案する．すなわち，共起候補語句yをもつ候補語句xの条件付き keyphraseness を次式で定義する．

$$Pair\_cond\_key(x|y) = \frac{|\{D_w | x \in Anchor(D_w) \wedge y \in D_w\}|}{|\{D_w | x \in D_w \wedge y \in D_w\}|}. \quad ④$$

ここで，条件付き keyphraseness の条件とする共起候補語句は候補語句と関連の強いものに限定すべきである．そこで，$Pair_cond_key(x|y)$を用いるx，yの組を Wikipedia 中で共起する記事数が一定の閾値以上の組に限定し，それ以外の組に対しては値を 0 とする．

共起回数の閾値は 10 回，15 回，20 回，25 回，30 回それぞれの場合で予備実験を行った結果から共起回数は 15 回とした．

その上で，文書D中の候補語句aの条件付き keyphraseness を，D中の共起候補語句がaに与える条件付き keyphraseness の最大値として定義する．ただし，共起候補語句はアンカーであるようなaと特に関係が強いものに限定する．すなわち

$$Cond\_key(a, D) = \max_{y \in D, LLRR(a,y) \geq \theta(a,D)} Pair\_cond\_key(a|y) \quad ⑤$$

$$LLRR(x, y) = \frac{LLR(x_{anchor}, y)}{LLR(x_{nonanchor}, y)} \quad ⑥$$

$$\theta(x, D) = \left(\prod_{y \in D} LLRR(x, y)\right)^{1/n} \quad ⑦$$

ここに，$LLR(x_{anchor}, y)$はアンカーとして出現したンカーxとyの対数尤度比[10]，$LLR(x_{nonanchor}, y)$は通常のテキストとして出現したxとyの対数尤度比，nは D 中の共起候補語句の数である．

式⑦では予備実験として相加平均，相乗平均，LLRR の上位 80%それぞれの場合を比較した結果，相乗平均を用いた．

4. 評価実験

4.1 実験方法

評価実験は 1) Wikipedia のリンクデータを学習データとした場合と，2) Wikipedia のリンクデータに，作成した日本語 Wikification コーパスの一部を追加して学習データとした場合のそれぞれについて教師付き学習を行った．それぞれ 10 分割交差検定により評価を行った．

(1) 使用データ

評価実験に使用する学習データは 2016 年 3 月 10 日付 Wikipedia から無作為に抽出した 1000 記事を使用した．評価指標として accuracy，precision，recall，F 値を用いた．

(2) 使用ツール

候補語句の前後の語句を抽出するために形態素解析ソフト MeCab[b]を使用し，識別器としては機械学習には SVM（サポートベクターマシン）Libsvm[c]を使用した．

4.2 実験結果

表 2 に Wikipedia のリンクデータのみを学習データとした場合（実験①），Wikipedia のリンクデータと作成した日本語 wikification コーパスの一部を学習データとした場合

b) http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
c) http://www.okuma.nuee.nagoya-u.ac.jp/~sakaguti/wiki/index.php?LibSVM

（実験②）の結果を示す.

表 2 実験結果

	Accuracy(%)	precision(%)	recall(%)	F値
実験①	73.0	55.3	65.9	0.601
実験②	73.1	55.6	66.0	0.604

表 2 の結果より Wikipedia のリンクデータに対して作成した日本語 wikification コーパスの一部を追加し，学習データとした場合，F 値が 0.003 向上した．アンカー選択作業者間の一致率と表 2 の accuracy を比較すると，提案方法によるアンカー抽出は，アンカー付与・非付与の判定という観点では人手による精度と近い結果が得られた．また，実験①，実験②において F 値がおよそ 0.6 であり，Wikipedia 記事に対してアンカー抽出を行った場合よりも約 0.18 ポイント F 値が低くなった[8].

4.3 考察

抽出結果の中には既にアノテートされている拡張固有表現の中でリンク先記事は存在するが Wikipedia 記事中で一度もアンカーになったことがないため，3 章の素性値を計算できないものがいくつか存在した．それらの語句は分類の結果アンカーとして抽出することができなかった.

また，今回は見出しを除く本文中の語句に対してアンカー抽出を行ったが，新聞記事は Wikipedia 記事と異なり，見出しに本文中の語句が省略された形で出現することがある．図 1 の例では「危険器具使用中止呼びかけ」「厚労省」が見出しとなっておりその本文がそれ以下となる. ここで，見出しの「厚労省」は本文中の「厚生労働省」の略称である. 見出しを含めて最初の出現のみをアンカーとする場合，「厚労省」をアンカーとして採用すると，「厚生労働省」はアンカーとして採用されなくなる．本研究では Wikipedia のリンクデータを学習データとしており，Wikipedia では「厚労省」よりも「厚生労働省」のほうがアンカーへのなりやすさが高く，一般的に略称された語句よりも略称される前の語句のほうがアンカーになりやすい．よって見出しに対してもアンカー抽出を行う場合，見出しと本文は別文書とみなしアンカー抽出を行う必要がある.

危険器具使用中止呼びかけ
厚労省
厚生労働省は 23 日、日本医師会などを通じて全国の医療機関に対し、麻酔マスクの接続器具である米デュパコ製ノーマン・エルボの在庫を調べ、使用を中止するよう呼びかけを始めた。

図 1 略称が見出しで現れる例

4.4 既存の評価用コーパス

英語を対象とした wikification では AIDA CoNLL-YAGO データセットなどが使用されており[11]，このデータセットは固有表現がアノテート対象となっている.

一方で日本語を対象とした wikification では Murawaki and Mori[12]は本研究と同様に日本語 Wikification コーパスを作成して wikification の実験を行っている. この研究で作成したコーパスは BCCWJ 内のサブコーパス白書(OW)と Yahoo! Blog(OY)を元にしている．いくつかの閾値以上の NIL を除いたリンクできる語句全てをアンカーとして採用している．本研究とは使用しているデータ，アンカー選定作業の方法，素性が異なるため直接の比較は難しい.

5. おわりに

Wikipedia 以外の文書に対してアンカー抽出を行うため，公開されている日本語 Wikification コーパスを加工し，アンカー抽出のための日本語 Wikification コーパスを作成した．新聞記事に対するアンカー抽出の実験を行い，Wikipedia のリンクデータを学習データとした場合 F 値が 0.601 となり，Wikipedia のリンクデータに加え作成した日本語 Wikification コーパスの一部を学習データとした場合 0.604 となった.

Wikipedia 記事に対してアンカー抽出した場合と比較すると, Wikipedia 記事と新聞記事の差異から性能は低下した. 今回作成した日本語 Wikification コーパスはアンカー抽出の学習のために十分な量ではないため，より多くの記事に対してアンカー選定作業を行う必要がある．また，Wikipedia のリンクデータを新聞記事のアンカー抽出の性能向上へ寄与させることが今後の課題である.

謝辞 本研究は，JSPS 科研費 JP15K16096 の助成を受けたものです．また，日本語 Wikification コーパスの作成に協力してくれた方々に感謝を申し上げます.

参考文献

[1] R. Mihalcea and A. Csomai. “Wikify! Linking documents to encyclopedic knowledge.” In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp.233-242. (2007).
[2] D. Milne and l. H. Witten. “An open-source toolkit for mining Wikipedia.” *Artificial Intelligence 194*, pp.222-239. (2013).
[3] 林良彦, 山内健二, 永田昌明, 田中貴秋. “言語間の情報補完を用いた対訳文の Wikificaton.” 人工知能学会全国大会論文集 28, 1A2-3(1A2-2). (2014).
[4] David Milne and Ian H. Witten. “Learning to link with wikipedia.” In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp.509-518. (2008).
[5] 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 回年次大会発表論文集, pp.688-691. (2015).
[6] Kensuke Horita, Fuminori Kimura, and Akira Maeda. “Automatic

Keyword Extraction for Wikification of East Asian Language Documents" *International Journal of Computer Theory and Enginnering*, pp.32-35. (2016).
[7] Davaajav Jargalsaikhan, 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).
[8] 小谷亮太, 綱川隆司, 梶博行. "Wikification における SVM を用いたアンカー抽出," 言語処理学会第 22 回年次大会発表論文集, pp.1093-1096. (2016).
[9] Ling, X., Singh, S., and Weld, D. "Design challenges for entity linking." *Transactions of the Association for Computational Linguistics*, pp.315–328. (2015).
[10] Ted Dunning. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics*, 19(1):61-74. (1993).
[11] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen F¨urstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. "Robust disambiguation of named entities in text." In *Proceedings of EMNLP*, pp.782–792. (2011).
[12] Yugo Murawaki and Shinsuke Mori. "Wikification for Scriptio Continua" In *Proceedings of the 10th Edition of its Language Resources and Evaluation Conference*, pp. 1346-1351. (2016).