情報処理学会論文誌 Vol.59 No.2 306–314 (Feb. 2018)

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

小谷 亮太[1,a)] 綱川 隆司[1,b)] 西田 昌史[1,c)] 西村 雅史[1,d)]

受付日 2017年5月9日, 採録日 2017年11月7日

概要：本稿では，日本語文書中の語句に，Wikipedia 記事へのリンクを付与する wikification タスクにおいて，リンク付与に値する重要な語句等を選択するアンカー抽出器について検討を行う．本研究では Wikipedia におけるリンクのガイドラインに準じたアンカー抽出基準をベースに，文書に適度にリンクを付与して Wikipedia 記事と結び付けることにより，文書の理解の可能性を高めることをねらいとする．日本語におけるアンカー抽出に有効と考えられる素性として，アンカーの前接語・後接語との関係をとらえた素性，および共起するアンカーの条件付き keyphraseness 素性の利用を提案する．また，一般的な日本語文書に対するアンカー抽出器の性能評価を行うため，日本語 Wikification コーパスに対して本研究で定めたアンカー抽出基準に従ってアンカー抽出作業を行い，評価用コーパスを構築した．評価実験により，提案した素性を既存手法に加えることで性能が改善することが示された．また，評価用コーパスを用いた実験では，正解率においてアンカー抽出作業者の 2 者間一致率の平均と同程度の性能が得られていることを確認した．

キーワード：Wikification，エンティティ・リンキング，アンカー抽出

Building Anchor Extractor and Evaluation Corpus for Japanese Wikification

Ryota Kotani[1,a)] Takashi Tsunakawa[1,b)] Masafumi Nishida[1,c)] Masafumi Nishimura[1,d)]

Received: May 9, 2017, Accepted: November 7, 2017

Abstract: This paper proposes a novel method for anchor extraction, which is to identify words/phrases worthy linking to the corresponding Wikipedia articles, in Japanese wikification. The aim of this study is to enhance understandability of a document by connecting the document and Wikipedia articles with appropriate links annotated on the basis of anchor extraction criteria that we defined according to the Wikipedia linking guideline. We build a Japanese anchor extractor with using two proposed new features: one is preceding/succeeding words of the target anchor and the other is keyphraseness conditioned with cooccurring anchors. In addition, we build an evaluation corpus for anchor extractor in Japanese text, which is manually annotated to the Japanese Wikification Corpus, according to our anchor extraction criteria. The experimental result shows that the proposed features are effective for improving the performance by introducing them into the existing method. We confirm that the proposed method achieves anchor detection accuracy similar to the mean agreement rate of two human annotators of the evaluation corpus.

Keywords: Wikification, entity linking, anchor extraction

1. はじめに

テキスト中に現れる固有表現等の語句の各出現に対して，それを説明する知識ベースの項目を対応付けるタスクはエンティティ・リンキングと呼ばれ，語句に対応付ける知識

[1] 静岡大学
Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
[a)] kotani.ryota.15@shizuoka.ac.jp
[b)] tuna@inf.shizuoka.ac.jp
[c)] nishida@inf.shizuoka.ac.jp
[d)] nisimura@inf.shizuoka.ac.jp

ベースを Wikipedia とする場合，特に wikification と呼ばれる．Web 上のテキストに wikification を適用することで，閲覧者はテキスト中の語句に関する知識をワンクリックで参照して補完できることから，テキスト理解の可能性を高めることができる．また，wikification は Wikipedia 記事の編集時に適切なリンクを付与するためのツールとしても有用と考えられるほかに，知識獲得や情報検索・情報抽出といった応用への基盤技術となりうる側面を持っている．

wikification は，主としてアンカー抽出とリンク先決定の 2 つのサブタスクからなる．アンカー抽出は，文書中でリンクを付与すべき語句の箇所（アンカー[*1]）を同定することである．ある語句にリンクを付与するかどうかの基準はある程度共通性があるものの，wikification の応用目的によって異なりうる．一方，リンク先決定は各アンカーに対してそれらを説明する Wikipedia 記事の有無を判定し，その記事があればそれを特定することである．これは曖昧性を解消する問題の一種であり，従来の語義曖昧性解消手法をはじめとする様々な手法が試みられている．リンク先決定タスクは wikification の応用目的にかかわらず同一の課題を解く問題であるのに対し，アンカー抽出はそれぞれの目的に応じて問題設定を変える必要がある．たとえば，テキスト中の単純な名詞句をアンカーとしたい場合は，アンカー抽出は浅い構文解析によって実現できる．また，アンカーを固有表現に限る場合は，テキスト中の固有表現を特定する固有表現認識タスクとなる．

本研究は，Wikipedia 記事のようにテキストに適度なリンクを付与することでテキスト理解の可能性を高めるため，Wikipedia と同様に，リンクの有用性を基準としたアンカー抽出を行うことを目的とする．Wikipedia では，リンク付与のガイドライン[*2]が定められており，記事の内容に関連する語句や重要な名称のみをアンカーとし，それ以外の一般名詞等の単語や，独立した記事を作るに値しないもの[*3]を表す固有表現はアンカーとしない．これにより，ほとんどクリックされることのないリンクを排除し，重要な語句のみをアンカーとして強調表示することで可読性を向上させることができる．

この基準によるアンカー抽出は，Wikipedia 記事のリンクを訓練例として用いる教師あり学習によって実現する方法が提案されている．教師あり学習に用いる素性としては，語句のアンカーへのなりやすさを示す keyphraseness [1] あるいはリンク確率 [2], [3] や，文書との関連性 [2] 等が用いられている．

ここでは，リンクの有用性を示す新しい素性を 2 つ提案し，その有効性を示すための評価実験を実施する．1 つはアンカーの前後に出現する単語の分布に着目した素性であり，もう 1 つはアンカー抽出において有効である keyphraseness 指標について，周辺の共起語の条件を付けることで関連性を考慮した素性である．

また，Wikipedia 記事のようなテキストだけでなく，他の日本語テキストにおけるアンカー抽出性能の評価を行うため，新たに日本語 Wikification コーパス [4] から評価用コーパスを構築した．本コーパスは新聞記事からなっており，Wikipedia のガイドラインを参考にアンカーを人手で抽出した．Wikipedia 記事を訓練例とするため，百科事典的テキストと新聞記事のテキストではアンカー抽出の基準が異なる可能性がある．そこで，本研究では Wikipedia 記事を主要な学習データとし，少量の新聞記事を学習データに加えた場合の分野適応的効果についても評価を行った．

2. 関連研究

Mihalcea ら [1] は，Wikipedia の記事を編集する際に必要となる文書中の重要な語句を特定し，それぞれに対応する他の記事へのリンクを付与するタスクを wikification と定義した．一方で，文書中の固有表現等（メンション）を同定してそれらが意味する知識ベース中の項目と対応付けるタスクはエンティティ・リンキングと呼ばれ [5]，wikification を一般化したタスクと解される．エンティティ・リンキングにおいては同定するメンションをテキスト中の固有表現とする場合や，知識ベースに存在する項目に対応する語句のすべてとする場合がある．また，知識ベースにない固有表現も同定し，知識ベースにないことを示すタグを付け，さらに同一のエンティティを表すメンションをクラスタリングすることもある．これらのタスク設定は，wikification およびエンティティ・リンキングの応用目的によって異なる部分があるが，共通しているのは，リンクを付与すべき語句（アンカーまたはメンション）の抽出と，リンク先決定の 2 つのタスクがあることである．以下では，本研究において取り組むアンカー抽出についての関連研究について述べる．

初期の wikification のためのアンカー抽出の研究 [1] では，アンカー抽出のための指標として tf-idf，χ 二乗検定，および，語句が出現した記事数のうち，その語句がアンカーとして出現した記事の割合 keyphraseness の 3 つを比較し，keyphraseness の有効性を示した．以降，keyphraseness またはその類型であるリンク確率 [2], [3] は基本的な指標として用いられている．

Milne ら [2] は，先にすべての語句に対してリンク先を決定してから，その結果を利用してアンカー抽出を行った．これにより，リンク先決定において重要な素性である commonness（アンカーがある記事を指す確率）等を利用できる．また，アンカー抽出すべきかどうかの基準の 1 つである記事の内容との関連性を，他のアンカーのリンク先記事から求めることができる．

*1 アンカーテキストまたはメンション（言及）とも呼ばれる．
*2 https://ja.wikipedia.org/wiki/Wikipedia:記事どうしをつなぐ
*3 たとえば，有名でない人物，書籍，楽曲等．

アンカー抽出は，固有表現認識と同様に，単語単位の系列ラベリング問題ともとらえられるため，CRF（条件付き確率場）に基づく手法も用いられている [6]．Wikipedia 記事の各単語に対してアンカーかどうかを示す BIO 方式のタグを付け，Wikipedia 記事を用いた学習によりテキスト中の各単語のタグを予測することでアンカーを抽出した．このモデルにおいても tf-idf やリンク確率のほか，前後 n 単語の出現分布や品詞情報等の素性が用いられる．また，単語の表層表現だけでなく単語や文字の分散表現 [7] も入力に用いられている．

アンカー抽出における素性には，keyphraseness 等の言語に依存しないものと，英語において先頭が大文字であることといった言語に依存するものがある．日本語の Wikification あるいはエンティティ・リンキングに関する研究はまだ多くはなく，アンカー抽出まで考慮に入れているものはほとんどみられない．日本語の学術文献中の専門用語に対する wikification [8]，日英対訳文の wikification [9]，およびソーシャルメディアからの話題抽出のためのエンティティリンキング [10] では，英語版 Wikipedia を用いた言語横断的な方法をとっている．前者 2 つはアンカーとして抽出する対象を専門用語や固有表現に限定している．中村ら [10] は，短い文書を対象とする wikification システム TAGME [3] のアンカー抽出手法をベースに，アンカーテキストの前後の単語・文字の統計情報を用いている．本研究においてもアンカーの前後の統計情報を学習のための素性として導入する．

3. アンカー抽出

3.1 アンカー抽出基準

wikification およびエンティティ・リンキングにおいて，アンカーの抽出方法はその応用に応じて異なる．本研究では，Wikipedia と同様のアンカー抽出結果を得ることを目標とする．ただし，Wikipedia におけるリンク付与のガイドラインは記事執筆のための案内であるため，本研究で用いる評価用コーパスを作成する際のアンカー抽出基準をより明確化するために，アンカー抽出の基準として以下の重要度，関連度，無名度を定めた．

重要度 文書の内容を表すのに不可欠な語句ほどアンカーになりやすい．当該語句の重要度はその不可欠さの度合いとする．いい換えると，文書を要約した際に残すべき語句は重要であり，より短い要約であっても残すべき語句は，より重要度が高いとする．

関連度 文書の内容と直接関連のある語句はアンカーにすべきである．当該語句の関連度は，文書の主題と当該語句の関連性の高さとする．当該語句が主題の 1 つの属性を表すものであるときは特に関連度が高いとする[*4]．

無名度 認知度が低い語句に対するリンクはクリックされやすく，有用性が高い．当該語句の無名度は，その語句を知っている人の割合の少なさの程度とする．ただし，本研究では，当該語句を説明する記事が存在しない場合はアンカーとしない[*5]．

これらの基準は相対的なものであり，本研究では直接定量化をせず，既存の Wikipedia 記事のアンカーを教師データとして用い，上記の基準を間接的に表す素性を導入してアンカー抽出器を構築する．

また，1 つの文書中に同一の語句が複数回出現する場合[*6]や，同一の記事がリンク先となる異なる複数の語句が出現する場合は，いずれも最初に出現する語句のみをアンカーとして抽出する．本研究におけるアンカー抽出結果の評価の際は，Wikipedia 記事中のアンカーの出現位置にかかわらず，各語句がアンカーになっているか否かでアンカー抽出結果の正誤を判定する．

3.2 アンカー抽出方法

本研究では，日本語の入力文書 d からアンカー候補語句リストを作成し，それぞれのアンカー候補語句 a をアンカーとして抽出すべきかどうかを SVM（サポートベクタマシン）を用いた教師あり学習によって構築したアンカー抽出器で判定する．

アンカー候補語句は，日本語版 Wikipedia の少なくとも 1 つの記事中において，アンカーとなったことのある文字列[*7]とする．入力文書を前から順に探索し，その文字から始まるアンカー候補語句があるときは，その中で最も長い候補語句を抽出し，抽出した候補語句の次の文字から再び探索を継続する．また，抽出した各アンカー候補語句について最初の出現位置を記録しておく．これにより，入力文書に現れるアンカー候補語句集合を得る．

入力文書から得た各アンカー候補語句を，アンカーかどうか判定する SVM による二値分類器に入力し，アンカーと

[*4] たとえば，文書の主題が “自動車” であるとき，“エンジン” や “運転” といった語句は “自動車” のある属性を示すものであり，“自動車” と関係のある “経営” や “顧客” といった語句より関連度が高いとする．

[*5] 通常の wikification では，対応する記事が存在しないことを判定し，リンク先として NIL を割り当てる．記事が存在しない理由は主に，(1) 作成されるべき記事であるがまだ書かれていない，(2) 独立した記事を作成する価値がない，のいずれかである．文書へのハイパーリンク付与という観点では，前者を表す語句はアンカーとして抽出すべきであるが，後者はリンクを作成する意味がない．本研究の提案方法では両者を区別できないため，記事が存在しないことが明らかな語句はアンカーとして抽出しないこととした．

[*6] 同一の語句が 1 つの文書中で異なる意味で用いられている場合には両方ともアンカーとして抽出すべきであるが，例外的なケースであるため本研究では考慮しないこととした．

[*7] ただし，数字，年を表す表記（“2016 年” 等），および漢字以外の 1 文字からなる文字列を除く．また，存在しない記事へのリンク（赤リンク）のアンカーも除く．

判定されたものを最終的なアンカー抽出結果とする．SVMによる二値分類器は，Wikipedia 記事および 4 章で述べる評価用コーパスのアンカーを訓練データおよびテストデータに分け，3.3 節に述べる素性を用いて学習する．

3.3 アンカー抽出に用いる素性

入力文書 d 中の各アンカー候補語句 a について，素性 i の素性値を $f_i(a,d)$ で表す．

3.3.1 keyphraseness

アンカー候補語句 a の keyphraseness $\mathrm{key}(a)$ は，a が出現する記事のうちアンカーとして出現する記事の割合であり，式 (1) で定義され，この値を素性値として用いる．

$$f_{\mathrm{key}}(a,d) = \mathrm{key}(a) = \frac{\mathrm{af}(a)}{\mathrm{df}(a)}. \tag{1}$$

ただし，$\mathrm{df}(a)$ は a が文字列として出現する Wikipedia 記事数，$\mathrm{af}(a)$ は a がアンカーとして出現する記事数とする．一般的には，Wikipedia 記事において語句 a の重要度が高いほどアンカーになりやすいため，keyphraseness は a の重要度をある程度反映していると考えられる．また，使用頻度の高い，よく知られている名詞はアンカーになりにくく，この点では a の無名度も反映している．

3.3.2 アンカー候補語句の前接語・後接語素性

ある語句がアンカーになりやすいかどうかは，その前後の語句にも依存していると考えられる．たとえば，ある語句の直後に “等” という語がくる場合，その語句は何らかの具体例を示していて，他の語が直後にくる場合よりもアンカーになりやすい傾向がみられる．この考えに基づき，前接語および後接語から求められる確率値を素性として導入する．

ある語 x のプリアンカー確率 $\mathrm{Pr}_{\mathrm{pre}}(x)$ を，x が現れる文書のうち x の直後の語句がアンカーになっている文書の割合として定義する．すなわち，

$$\mathrm{Pr}_{\mathrm{pre}}(x) = \frac{\mathrm{df}_{\mathrm{preanchor}}(x)}{\mathrm{df}(x)}, \tag{2}$$

ただし，$\mathrm{df}_{\mathrm{preanchor}}(x)$ は，アンカーの直前に x が現れる記事の数とする．この値が大きいほど，x はアンカーの直前の語として特徴的に出現しやすい語であることを示す．

アンカー候補語句 a の前接語に関する素性値は，文書 d 中の a の最初の出現箇所[*8]の直前の語 $\mathrm{pred}_d(a)$ のプリアンカー確率とする．すなわち，

$$f_{\mathrm{preanchor}}(a,d) = \mathrm{Pr}_{\mathrm{pre}}(\mathrm{pred}_d(a)). \tag{3}$$

図 1 にアンカー候補語句の前接語の例を示す．点線または実線の下線を付した語句がアンカー候補語句であり，そのうち実線のものが実際にアンカーとしてリンクが付与されている．太字で示した語がアンカー候補語句の前接語となる．アンカー候補語句 “類型” の前接語 “の” のプリアンカー確率は 0.38 であり，したがって “類型” の前接語に関する素性値は 0.38 である．同様に “文系” の前接語 “、” のプリアンカー確率は 0.70 であり，素性値は 0.70 とする．この素性においては後者のほうがアンカーになりやすいと判断される．

情報学部の類型には、文系的な分野と理系的な分野の双方を総合的に扱っているもの…がある。（Wikipedia 記事 “情報学部” より）

図 1 Wikipedia 記事とアンカー候補語句の例
Fig. 1 Example of a Wikipedia article and anchor candidates.

また，ある語 x のポストアンカー確率 $\mathrm{Pr}_{\mathrm{post}}(x)$ を x が現れる文書のうち x の直前の語句がアンカーになっている文書の割合と定義し，アンカー候補語句 a の後接語に関する素性値は，文書 d 中の a の最初の出現箇所の直後の語 $\mathrm{succ}_d(a)$ のポストアンカー確率とする．すなわち，

$$\mathrm{Pr}_{\mathrm{post}}(x) = \frac{\mathrm{df}_{\mathrm{postanchor}}(x)}{\mathrm{df}(x)}, \tag{4}$$

$$f_{\mathrm{postanchor}}(a,d) = \mathrm{Pr}_{\mathrm{post}}(\mathrm{succ}_d(a)), \tag{5}$$

ただし，$\mathrm{df}_{\mathrm{postanchor}}(x)$ は，アンカーの直後に x が現れる記事の数とする．

これらの素性は，テキスト中において重要な語句が現れるときに特徴的なコロケーションの出現パターンを反映しており，その語句の重要性を部分的に示していると考えられる．

3.3.3 条件付き keyphraseness

keyphraseness 素性はアンカー候補語句のみに依存しており，また，前接語・後接語素性は直近の出現パターンに依存している．いずれも，テキストの内容とアンカー候補語句との関連性をとらえていない．従来手法においては関連性をとらえるための素性として，リンク先記事のアウトリンク・インリンクの類似性を正規化 Google 距離により求める手法 [11], [12] があるが，当該語句のリンク先記事を先に推定する必要が生じる．本研究では，同一テキスト内の他のアンカーとの条件付き keyphraseness の値を導入し，アンカー候補語句とテキストの内容の関連度を考慮した素性を提案する．

たとえば，自動車のメーカを表す語 “BMW” は，メーカの国籍 “ドイツ” や競合する自動車メーカ “ベンツ” といった語と共起するとき，偶然出現した場合や単に例示として出現する場合よりもアンカーになりやすいと思われる．条件付き keyphraseness は，アンカー候補語句が特定の語と共起したときのアンカーへのなりやすさをとらえる．アンカー候補語句 x がアンカー候補語句 y と共起するときの条件付き keyphraseness を次式で定義する．

[*8] Wikipedia 記事を訓練データとして用いる場合，アンカーとしての出現箇所とする．後接語についても同様．

$$\mathrm{ckey}(x|y) = \begin{cases} \dfrac{\mathrm{af}(x,y)}{\mathrm{df}(x,y)} & (\mathrm{df}(x,y) \geq t) \\ 0 & (\text{otherwise}) \end{cases}, \quad (6)$$

ただし，$\mathrm{df}(x,y)$ は x と y がともに現れる Wikipedia 記事の数，$\mathrm{af}(x,y)$ はそれらの記事のうち，x がアンカーとして現れる記事の数，t は共起回数の閾値とする．共起回数の閾値は予備実験の結果から $t = 15$ とした．

アンカー候補語句 x の keyphraseness に比べ，特定のアンカー候補語句 y と共起した場合の条件付き keyphraseness が高いということは，y の共起によってそれだけ x がアンカーである確率が高くなったことを示していると考えられる．したがって，共起するアンカー候補語句からそれぞれ求められる条件付き keyphraseness の中で最大の値 $\max_{y \in V_d} \mathrm{ckey}(a|y)$ を用いることとする．ただし，V_d は文書 d に出現するアンカー候補語句の集合とする．

しかし，共起するアンカー候補語句の中には，x の意味にかかわらず同程度に共起するものが存在する場合があり，x が多義性を持つ場合でアンカーになりにくい意味で用いられたときに，その共起語から求められた条件付き keyphraseness を採用すると悪影響を及ぼす可能性がある．そこで，最大値を求めるために用いる共起語は x がアンカーであるかどうかの識別能力が高いものが望ましい．すなわち，x がアンカーとして出現した場合（以下，x_a とする）との関連度が高く，アンカーとしてではなく単に語句として出現した場合（以下，x_n とする）との関連度が高くない共起語のみを計算対象としたい．

そこで，これを求めるための関連度指標として対数尤度比 [13] を用いる．2 つの語 w_1, w_2 間の対数尤度比[*9]を $\mathrm{LLR}(w_1, w_2)$ とするとき，テキスト中の語句 x と，その共起語 y の間の対数尤度比のアンカー・非アンカー比 $\mathrm{LLRR}(x,y)$ を以下の式で定義する．

$$\mathrm{LLRR}(x,y) = \frac{\mathrm{LLR}(x_\mathrm{a}, y)}{\mathrm{LLR}(x_\mathrm{n}, y)}. \quad (7)$$

この値がある閾値以上の共起語 y について条件付き keyphraseness を求め，その最大値を条件付き keyphraseness を用いた素性値 f_ckey とする．すなわち，

$$f_\mathrm{ckey}(a,d) = \max_{\{y \in V_d | \mathrm{LLRR}(a,y) \geq \theta(a,d)\}} \mathrm{ckey}(a|y), \quad (8)$$

また，閾値 $\theta(a,d)$ はすべての共起語についてのアンカー・非アンカー比の相乗平均とし，次の式で求める．

$$\theta(a,d) = \left(\prod_{y \in V_d} \mathrm{LLRR}(a,y) \right)^{\frac{1}{|V_d|}}. \quad (9)$$

図 1 の例文においては，アンカー候補語句 “総合”（リンク先記事は “総合科学”）が単に現れる場合に比べ，“文系” や “理系” といった関連語が共起する場合の方がアンカーになっている頻度が高いことが予想され，そうであれば条件付き keyphraseness の素性値は “総合” の keyphraseness の値よりも高くなり，共起語による効果が現れることが期待される．逆に，“類型” や “双方” といったアンカー候補語句は，同一記事内にこれらに関連するアンカー候補語句が現れない限り，条件付き keyphraseness の素性値はそれぞれの keyphraseness の値と同程度の値にとどまることが期待される．

[*9] ここでの対数尤度比は，x の出現を x_a と x_n に分けて求める．

4. アンカー抽出性能評価用コーパス

アンカー抽出の評価には，内部リンクが付与されている Wikipedia 記事が用いられることが多い．この評価は，wikification の対象が Wikipedia 記事であるときは妥当であるが，Wikipedia 記事は百科事典的で，かつ独特の記述方法を持つテキストであり，一般のテキストを対象としたときの評価方法としては必ずしも十分ではない．wikification のうちリンク先決定に関する評価には AIDA CoNLL-YAGO データセット [14] や日本語 Wikification コーパス [4] 等を用いることができるが，アンカー抽出対象がすべての固有名詞である等，3.1 節で述べたような基準に基づくアンカー抽出評価には適用できない．そこで，本研究では日本語 Wikification コーパスをベースとして 3.1 節で述べた基準に従い新たにアンカーを人手で抽出し，これに基づく性能評価実験を行った．

日本語 Wikification コーパスは，BCCWJ（現代日本語書き言葉均衡コーパス）[*10]のコアデータに対して，関根の拡張固有表現階層-7.1.0- [*11] [15] が付与された拡張固有表現タグ付きコーパス（東京工業大学）[*12]内の新聞記事 340 記事にタグ付けされている拡張固有表現の一部に対し，それぞれの拡張固有表現が指す Wikipedia 記事をリンク先としてラベル付けしたものである．以下，日本語 Wikification コーパスに対するアンカー抽出の手順を記す．

(1) 新聞記事の，見出しを除く部分について，拡張固有表現およびそれ以外のアンカー候補語句を次のように特定する．
  (a) 新聞記事中（見出しを除く）の拡張固有表現のうち，対応する Wikipedia 記事が存在しない（リンク先が NIL となっている）もの，時間表現，数値表現，アドレス，称号名，施設部分名のいずれかに分類されるものを除くすべてをアンカー候補語句とする．ただし，1 つの新聞記事内に同一のリンク先を指す拡張固有表現が複数回出現する場合，見出しを除く最初の出現のみをアンカー候補語句として残す．

[*10] http://pj.ninjal.ac.jp/corpus_center/bccwj/index.html
[*11] https://sites.google.com/site/extendednamedentityhierarchy/
[*12] http://www.gsk.or.jp/catalog/gsk2014-a/

表 1 評価用コーパスへのアンカー抽出数（リンク先 NIL 除く）
Table 1 Numbers of annotated anchors that are not linked to NIL.

	日本語 Wikification コーパス	評価用コーパス
拡張固有表現	3,698	2,337
拡張固有表現以外	—	867
計	3,698	3,204

(b) 新聞記事中で拡張固有表現としてタグ付けされている箇所以外で，日本語版 Wikipedia から作成したアンカー候補語句リストにある語句が出現する場合，前方から文字列探索して最長一致となる箇所をすべてアンカー候補語句とする．ただし，以下の場合を除く．
- アンカー候補語句の指すリンク先が，(1a) で得られたアンカー候補語句の指すリンク先のいずれかと一致する場合．
- 他のアンカー候補語句の出現と重なる場合（前方のアンカー候補語句を優先する）．

(2) 各アンカー候補語句について，3.1 節で述べた基準に従い，下記のいずれかの条件を満たすと判断したアンカー候補語句で，かつリンク先の Wikipedia 記事が存在するものをアンカーとして抽出する．
- 重要度，関連度，無名度の少なくとも 1 つが特に高い．
- 重要度，関連度，無名度のすべてがある程度高い．

(3) 抽出したアンカーのうち，同一新聞記事内で同じリンク先を指すものが複数ある場合，最初の出現のみをアンカーとして残す．

日本語 Wikification コーパスのうち無作為に選択した新聞記事 100 件を評価用コーパスとした．各新聞記事に対し，延べ 5 名の評価者がそれぞれ独立に上記の手順に従い，各アンカー候補語句をアンカーとして抽出するかどうか判定した．最終的に 3 名以上がアンカーとして抽出すべきとしたアンカー候補語句を評価用コーパスにおける抽出すべきアンカーとした．アンカー抽出作業の一貫性の程度を調べるため，評価者のすべての 2 名の組合せにおいて各アンカー候補語句をアンカーとして抽出するかどうかの判断が一致した割合を求め，それらを平均した 2 者間アンカー一致率平均を求めたところ，約 73.3%であった．この数値は人間がこの基準によって行うアンカー抽出性能の上界に近いものと考えられる．また，Fleiss の kappa 係数は 0.509 であり，中程度の一致率であった．

日本語 Wikification コーパスおよび評価用コーパス中のアンカー数を**表 1** に示す．日本語 Wikification コーパスのアンカー数は延べ数であり，同一のリンク先を示す拡張固有表現が複数回出現する場合もすべて数えている．また，リンク先記事が存在しないものは含まれていない．日本語 Wikification コーパス中でアノテートされた拡張固有表現のうち，1,361 個は評価用コーパスにおいてアンカーとして採用されなかった．その多くは本研究において定めた，同一のリンク先を指す語句は最初の出現のみアンカーとして抽出するというルールにより除外されたものであるが，一部にはアンカー抽出の基準を満たさないと判定されたものがある．逆に，拡張固有表現以外の語句として 867 個のアンカーが得られている．一例として，家庭訪問に関するある新聞記事に現れるアンカー候補語句と評価者の抽出判定結果を**表 2** に示す．この記事の主要なテーマである “家庭訪問” や，“共働き”，“不就学” といった事柄は一般には固有表現とはみなせないが，本研究のアンカー抽出基準においてはこれらも含めるべきであると考えられる．

表 2 アンカー抽出作業結果の例
Table 2 An annotation example of anchors.

アンカー候補語句	拡張固有表現	抽出判定者の割合
家庭訪問		5/5
学校 5 日制	✓	4/5
共働き		3/5
廃止		2/5
埼玉県	✓	5/5
小学校		2/5
不就学		4/5

評価用コーパスは，アンカー抽出器の訓練データまたはテストデータとして用いるときは，手順 (1b) までで得た各アンカー候補語句を訓練例とし，アンカーとして抽出されたものを正例，それ以外を負例として扱った．

5. 評価実験

本稿で提案したアンカー抽出のための素性の有効性を示すための評価実験を行った．アンカー抽出器の訓練・テストデータに日本語版 Wikipedia を用いた場合の実験，および，日本語版 Wikipedia と評価用コーパスの両方を用いた場合の実験をそれぞれ実施した．

5.1 実験設定

本実験では 2016 年 3 月 10 日時点の日本語版 Wikipedia ダンプデータを用いた．Wikipedia 記事は大別してトピックページ，曖昧さ回避ページ，リダイレクトページ，カテゴリページの 4 つに分類され，これらのうちある概念について説明するトピックページである 868,689 記事のみを用いた．

本研究におけるアンカー抽出の対象は通常のテキストであり，アンカーが列挙された形式のものは対象としない．また，Wikipedia 記事中には Wikipedia 外のサイトへのリンクもあるが，これらは本研究の抽出対象ではない．これらのことから，本実験においては Wikipedia 記事から表，Infobox[*13]，関連文献および脚注はあらかじめ除外した．

[*13] 記事の右上に配置され，記事の主題についての要約情報を項目ごとに整理して提供する規定フォーマット．

表 3 実験に用いたデータ
Table 3 Data used for the experiments.

	学習データ	テストデータ	交差検定
実験 1	Wikipedia 記事	Wikipedia 記事	10 分割
実験 2	Wikipedia 記事	評価用コーパス	—
実験 3	Wikipedia 記事＋評価用コーパス	評価用コーパス	10 分割（評価用コーパス）

本提案方法におけるアンカー抽出は形態素解析によらず，Wikipedia から得たアンカー候補語句との文字列マッチングにより行う．ただし，アンカー候補語句の前接語・後接語素性（3.3.2 項）を求めるときの前接語・後接語を得るために形態素解析ソフト MeCab [*14] を使用した．アンカー抽出器のための SVM のライブラリとして Libsvm [*15] を利用し，ガウシアンカーネルを用いた．

抽出した記事の各アンカー候補語句をそれぞれ訓練例とし，実際にアンカーになっているものを正例，アンカーにはなっていないものを負例として扱ってアンカー抽出器を学習した．抽出した Wikipedia 記事と評価用コーパスから，**表 3** に示す 3 種類の実験を実施した．実験 1 では，Wikipedia 記事を対象とする wikification における提案方法の効果を示す．実験 2 および 3 では，新聞記事を対象とする wikification の性能評価を行う．Wikipedia 記事のみを用いた学習では評価用コーパスである新聞記事とドメインが異なるために十分な性能が得られないことが予想されるため，実験 3 において，少量の評価用コーパスを学習データに加えたときに分野適応的効果がどの程度得られるか検証を行った．実験 1 と実験 3 については学習データとテストデータに含まれるコーパスについて 10 分割交差検定を行った．

アンカー抽出の評価指標として，個々のアンカー判定に対する適合率，再現率，F 値および正解率を用いた．

5.2 Wikipedia 記事におけるアンカー抽出（実験 1）

提案した素性を用いて Wikipedia 記事に対するアンカー抽出実験を行った．用いたデータは表 3 の実験 1 のとおりである．**表 4** に実験 1 の評価結果を示す．ここで，M&W [12] は Milne ら [12] で用いられている素性のうち，リンク先決定後に得られるものを除く下記の素性を示す．

文脈との関連度 アンカー候補語句のリンク先記事と，共起するアンカー候補語句のリンク先記事の関連度

一般性 Wikipedia のカテゴリ階層における，リンク先記事が属するカテゴリの深さのレベル

アンカー候補語句の出現位置 文書内におけるアンカー候補語句の最初の出現位置，最後の出現位置，および，それらの間の距離

[*14] http://taku910.github.io/mecab/
[*15] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

表 4 Wikipedia 記事におけるアンカー抽出実験（実験 1）結果
Table 4 Experimental results of anchor extraction on Wikipedia articles (Experiment 1).

適用素性	適合率	再現率	F 値
keyphraseness のみ	0.728	0.715	0.721
keyphraseness + M&W [12]	0.761	0.759	0.760
すべて	0.782	0.771	0.776

表 5 Wikipedia 記事 “自然言語” のアンカー抽出結果
Table 5 Anchor extraction results for a Wikipedia article “Natural Language (in Japanese)”.

	実際のアンカー	非アンカー
抽出	人工言語，自然言語処理，音声，...	心理学
非抽出	人間，記号，文化，文字，...	

表 6 評価用コーパスを対象としたアンカー抽出実験（実験 2・3）結果
Table 6 Evaluation results of anchor extraction on our evaluation corpus (Experiment 2 & 3).

	正解率	適合率	再現率	F 値
実験 2	0.743	0.627	0.637	0.632
実験 3	0.745	0.630	0.638	0.636

表 4 の結果から，提案素性および Milne ら [12] の素性をすべて用いたときに最も高い性能が得られ，提案素性を用いない場合と比べ F 値で 0.016 高い値が得られた．

表 5 に，Wikipedia 記事 “自然言語” に対してアンカー抽出を行った例を示す．提案方法によって抽出できなかった “人間”，“記号” 等のアンカーは，いずれも比較的出現頻度が多い一般語句である．このような語句は keyphraseness の値が低い傾向があるためアンカーとして抽出されにくく，検討の余地が残っている．

5.3 評価用コーパスを対象としたアンカー抽出実験（実験 2・3）

表 6 に，すべての提案素性を用いたときの評価用コーパスを対象としたアンカー抽出実験の評価結果を示す．Wikipedia 記事を対象とする場合と比較して，F 値で 0.14 程度の低下がみられ，これは Wikipedia 記事と評価用コーパスの性質の違いに起因すると考えられる．評価用コーパスの一部を学習データに加えることで期待される分野適応的効果については，F 値の改善は 0.004 にとどまった．

また，正解率という観点では，評価用コーパスにおけるアンカー抽出作業者の 2 者間一致率の平均が 73.3%であり，これと比べるとアンカーとして抽出するかどうかの判定においては人手による評価と近い一致率が得られているといえる．

6. まとめ

本研究では，日本語テキストを対象とした wikification におけるアンカー抽出器を構築し，アンカーの前接語・後接語および共起アンカーの条件付き keyphraseness の 2 つの素性を提案してその有効性を示した．また，一般の日本語テキストにおけるアンカー抽出性能の評価を行うため，日本語 Wikification コーパスをもとにアンカー抽出評価用コーパスを作成し，提案方法によるアンカー抽出器の評価を実施した．Wikipedia 記事を対象とした実験ではアンカー抽出性能の改善がみられ，一般の日本語文に対しても，アンカー抽出作業者の 2 者間一致率の平均と同程度の正解率が得られた．

今回作成した評価用コーパスはアンカー抽出の学習に十分な量ではなく，一部を学習データとして用いた実験では分野適応的な効果は確認できなかった．一般のテキストに対する評価用コーパスとしては，より多くのアンカー抽出作業を行い，その上で提案手法および分野適応の効果を示す必要がある．今回評価に用いた新聞記事と Wikipedia 記事との差異をより詳細に調査し，新聞記事やその他の一般のテキストの性質を反映した素性を開発することも今後の課題である．また，本稿はリンク先決定前の評価を行っているため，抽出したアンカーの差異によるリンク先決定精度への寄与や wikification 全体の性能評価について今後行っていきたい．

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, R. and Csomai, A.: Wikify!: Linking Documents to Encyclopedic Knowledge, *Proc. 16th ACM Conference on Conference on Information and Knowledge Management* (*CIKM*), pp.233–242 (2007).

[2] Milne, D. and Witten, I.H.: Learning to Link with Wikipedia, *Proc. 17th ACM Conference on Information and Knowledge Management* (*CIKM*), pp.509–518 (2008).

[3] Ferragina, P. and Scaiella, U.: TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities), *Proc. 19th ACM International Conference on Information and Knowledge Management* (*CIKM*), pp.1625–1628 (2010).

[4] Jargalsaikhan, D.，岡崎直観，松田耕司，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 回年次大会発表論文集，pp.793–796 (2016).

[5] Hachey, B., Radford, W., Nothman, J., Honnibal, M. and Curran, J.R.: Evaluating Entity Linking with Wikipedia, *Artificial Intelligence*, Vol.194, pp.130–150 (2013).

[6] Gardner, J. and Xiong, L.: Automatic Link Detection: A Sequence Labeling Approach, *Proc. 18th ACM Conference on Information and Knowledge Management* (*CIKM*), pp.1701–1704 (2009).

[7] Ling, W., Tsvetkov, Y., Amir, S., Fermandez, R., Dyer, C., Black, A.W., Trancoso, I. and Lin, C.-C.: Not All Contexts Are Created Equal: Better Word Representations with Variable Attention, *Proc. 2015 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.1367–1372 (2015).

[8] 古川竜也，相良 毅，相澤彰子：言語横断エンティティリンキングのための語義曖昧性解消，情報知識学会誌，Vol.24, No.2, pp.172–177 (2014).

[9] 林 良彦，山内健二，永田昌明：言語間の情報補完を用いた対訳文の Wikification，2014 年度人工知能学会全国大会論文集，Vol.28, No.1A2-3, pp.1–4 (2014).

[10] 中村達哉，白川真澄，原 隆浩，西尾章治郎：ソーシャルメディアからの言語横断的な話題抽出に向けたエンティティリンキング手法，データ工学と情報マネジメントに関するフォーラム（DEIM 2015）(2015).

[11] Milne, D. and Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links, *Proc. AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pp.25–30 (2008).

[12] Milne, D. and Witten, I.H.: An open-source toolkit for mining Wikipedia, *Artificial Intelligence*, Vol.194, pp.222–239 (2013).

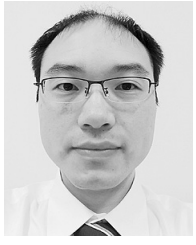
[13] Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol.19, No.1, pp.61–74 (1993).

[14] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S. and Weikum, G.: Robust Disambiguation of Named Entities in Text, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.782–792 (2011).

[15] Sekine, S.: Extended Named Entity Ontology with Attribute Information, *Proc. 6th International Conference on Language Resources and Evaluation* (*LREC*), pp.52–57 (2008).

小谷 亮太

2015 年静岡大学情報学部情報科学科卒業．2017 年同大学院総合科学技術研究科情報学専攻修士課程修了．修士（情報学）．



綱川 隆司 （正会員）

2005 年東京大学大学院情報理工学系研究科コンピュータ科学専攻修士課程修了．2008 年同博士後期課程単位取得退学．東京大学特任研究員，静岡大学学術研究員を経て，2011 年より静岡大学情報学部助教．自然言語処理に関する研究に従事．言語処理学会会員．博士（情報理工学）．

西田 昌史 （正会員）

1999 年龍谷大学大学院修士課程修了．2002 年同博士後期課程修了．千葉大学助手，同助教，同志社大学准教授，名古屋大学特任准教授を経て，2015 年より静岡大学情報学部准教授．音声情報処理，行動信号処理，福祉情報工学に関する研究に従事．電子情報通信学会，日本音響学会，人工知能学会各会員．博士（工学）．

西村 雅史 （正会員）

1983 年大阪大学大学院基礎工学研究科博士前期課程修了．同年日本アイ・ビー・エム（株）入社．同社東京基礎研究所にて，音声言語情報処理の研究に従事．2014 年より静岡大学大学院総合科学技術研究科教授．1998 年情報処理学会山下記念研究賞，1999 年日本音響学会技術開発賞受賞．IEEE，電子情報通信学会，日本音響学会，人工知能学会各会員．博士（工学）．