E-005

Wikification における前接語・後接語を用いたアンカー抽出
Anchor Extraction for Wikification Using Preceding and Succeeding Words

小谷亮太† 綱川隆司† 西田昌史† 西村雅史†
Ryota Kotani Takashi Tsunakawa Masafumi Nishida Masafumi Nishimura

1 はじめに

Wikipedia は Web 上の百科事典で､巨大なハイパーテキストであることが特徴である。記事に付与されたリンクにより、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、リンク元の文字列であるアンカーを抽出する第 1 ステップと、抽出し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[4]。これに比べると第 1 ステップの研究は少ない。Wikipedia でアンカーとなっている語句を全て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。

本稿では語句の前接語と後接語を用いた素性と、共起語の情報を用いた素性を提案し、既存研究との比較実験を行った結果を報告する。

2 関連研究

初期の wikification のためのアンカー抽出の研究[1]では対象の語句が出現した Wikipedia 記事のうち､実際にアンカーとなっている記事の割合である keyphraseness だけを用いてアンカー抽出を行っている。Milne and Witten[2]は全ての候補語句に対してリンク先記事を決定した後、その結果を利用したアンカー抽出を行っており、keyphraseness のほかにリンク先記事関連度やリンク先記事のカテゴリの深さなどを素性として使用している。

我々の以前の研究[3]では keyphraseness のみを用いた場合に比べ、語句の前接語と後接語を用いた素性と共起語の情報を用いた素性を追加した場合に、keyphraseness のみを用いた場合に比べて精度向上が見られることを報告した。本研究では keyphraseness に加えて Milne and Witten[2]において使用されている素性との比較評価を行った。

3 提案方法

一般的にアンカー抽出はアンカー候補語句の抽出とアンカーにすべき語句の選定から成っており、それぞれを以下に説明する。

・アンカー候補語句の抽出

キーワード抽出の研究においてはよくヒューリスティックな方法が用いられており、例えばストップワードを除く方法[5]、特定の品詞のみを対象とする方法[6]、N グラムを用いる方法[1][2][7]などがある。

本研究では語句の出現回数が少ないものや、品詞が動詞や副詞であっても文書によってはアンカーとして選択されることがあると考えたので、Wikipedia 記事中で一度でもアンカーになったことのある語句を候補語句とした。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適切でない例が存在する。また、極端な頻出語をアンカーとしている例も存在する。いずれの場合もその語句がアンカーになる割合が低いことから、keyphraseness の値が 0.005 未満の語句は除外した。

・アンカーにすべき語句の選定

既存研究では、候補語句にスコア付けを行い、ランク付けを行い､全ての候補語句のうちの上位 6%をアンカーとして選定する方法[1]や決定木を用いてアンカーを選定する方法[2]が存在する。

本研究では SVM(サポートベクタマシーン)を用いて実験を行った。

次に今回提案した素性を以下に説明する。

(1) 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 つの素性を提案する。

(1a) 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、

$$PreAnchor(x) = \frac{|\{D_w | \exists y \in Anchor(D_w), x \cdot y \in Bigram(D_w)\}|}{|\{D_w | x \in D_w\}|} \quad (1.1)$$

ここに、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、$Bigram(D_w)$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。

(1b) 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、

$$PostAnchor(x) = \frac{|\{D_w | \exists y \in Anchor(D_w), y \cdot x \in Bigram(D_w)\}|}{|\{D_w | x \in D_w\}|} \quad (1.2)$$

候補語句aの素性としてはaの後接語$succ(a)$のポストア

† 静岡大学大学院総合科学技術研究科情報学専攻

ンカー確率$PostAnchor(succ(a))$を用いる。

(2) 候補語句の条件付き keyphraseness

候補語句と共起する他の候補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「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。このような考えに基づき、共起候補語句を条件とする候補語句のリンク確率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の条件付きリンク確率を次式で定義する。

$$Pair\_cond\_link\_prob(x|y) = \frac{|\{D_w|x \in Anchor(D_w), y \in D_w\}|}{|\{D_w|x \in D_w, y \in D_w\}|} \quad (2.1)$$

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、$Pair_cond_link_prob(x|y)$を計算するx, yを Wikipedia 中で共起する記事数がある閾値以上の組に限定する（本実験では閾値を 15 とした）。

その上で、文書D中の候補語句aの条件付きリンク確率をD中の共起候補語句がaに与える条件付きリンク確率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、

$$Cond\_link\_prob(a, D) = \max_{y \in D, LLRR(a,y) \geq \underset{y' \in D}{\text{avg}}\, LLRR(a,y')} Pair\_cond\_link\_prob(a|y) \quad (2.2)$$

ここに、$LLRR(x, y)$はアンカーであるxとyの対数尤度比[8]とアンカーでないxとyの対数尤度比の比である。すなわち、

$$LLRR(x, y) = \frac{LLR(x_{anchor}, y)}{LLR(x_{nonanchor}, y)} \quad (2.3)$$

4 評価実験

4.1 実験方法

(1) 使用データ

評価実験に使用するデータとして 2016 年 3 月 10 日付 Wikipedia 記事中の 3000 候補語句を選択し、訓練データに 2100 語句、テストデータに 900 語句を使用した。ここで、[2]との比較実験を行うにはあらかじめリンク先記事を決定する必要があるので最頻リンク先がリンクされる確率が 90%以上の語句を採用し、最頻リンク先をリンク先記事と仮定した。評価は precision、recall、F 値を計算して行った。

(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 を使用し、機械学習には SVM（サポートベクターマシン）Libsvm を使用した。

(3) 素性の組合せ

以下の 4 通りの素性の組合せでアンカー抽出器を学習させ、実験を行った。

(i)keyphraseness

(ii)keyphraseness + Milne and Witten [2]

(iii)keyphraseness + 前接語のプリアンカー確率+後接語のポストアンカー確率 + 条件付き keyphraseness

(iv)keyphraseness + Milne and Witten [2] + 前接語プリアンカー確率 + 後接語のポストアンカー確率 + 条件付 keyphraseness

4.2 実験結果

上記の(i)~(iv)の結果を表 1 に示す。

表 1 アンカー抽出の結果

	precision(%)	recall(%)	F値
(i)	72.8	71.5	0.721
(ii)	76.1	75.9	0.760
(iii)	77.2	74.1	0.757
(iv)	78.2	77.1	0.776

表 1 から、既存研究に提案した 3 つの素性(前接語のプリアンカー確率・後接語のポストアンカー確率・条件付き keyphraseness)を追加した場合最もよい結果が得られた。本研究ではリンク先記事決定のタスクを行っていないので厳密なアンカー抽出の精度を比較評価することは難しいが、提案した素性がアンカー抽出のタスクにおいて有効であることは分かった。

5 おわりに

SVM を使ったアンカー抽出のための新しい素性を提案し、既存研究との比較評価を行った。提案した素性を加えることにより F 値を 0.016 向上させることができた。今後の課題は、本方法を訓練データが利用できない一般のニュース記事や新聞記事に適用することである。

謝辞：本研究は、JSPS 科研費 15K16096 の助成を受けたものです。

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] 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 23 回年次大会発表論文集, pp.1093-1096(2016).

[4] D. Roth, H. Ji and M. Chang and T. Cassidy. “Wikification and beyond: The challenges of entity and concept grounding”, In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials* (2013).

[5] Z. Liu, P. Li, Y Zheng, and M. Sun, “Clustering to find exemplar terms for Keyphrase extraction”, In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp.257-266 (2009).

[6] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts”, In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404-411 (2004).

[7] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, “KEA: Practical automatic Keyphrase extraction”, In *Processing of the 4th ACM Conference on Digital Libraries*, ACM Press, pp.254-255 (1999).

[8] Ted Dunning. “Accurate methods for the statistics of surprise and coincidence”, *Computational Linguistics*, 19(1):61-74 (1993).