Wikification における SVM を用いたアンカー抽出

小谷亮太　綱川隆司　梶博行
静岡大学大学院総合科学技術研究科情報学専攻
gs15018@s.inf.shizuoka.ac.jp ，｛tuna，kaji｝@inf.shizuoka.ac.jp

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。リンク元の文字列をアンカーと言う。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。

Wikification は、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。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し、曖昧性解消に焦点を当てた研究が多い。しかし、そのような方法ではアンカーだらけの文書となるので、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をアンカーとして抽出する方法が必要である。

本稿は、機械学習、具体的には SVM を用いたアンカー抽出に関し、アンカー抽出に有効ないくつかの素性を提案する。Wikipedia のリンクデータを訓練データとしてアンカー抽出器を学習し、交差検定によって評価した結果を報告する。

2　関連研究

Wikification のためのアンカー抽出の研究では、通常、Wikipedia 記事においてアンカーとされている語句を収集し、それらをアンカー候補語句とする。入力文書内のアンカー候補語句からアンカーを選定するための素性として、Milne and Witten はリンク先記事関連度、ベクトル類似度、リンク確率の 3 つを評価し、リンク先記事関連度とリンク確率は有効であるが、ベクトル類似度は有効ではないと報告している[2]。なお、リンク先記事関連度とベクトル類似度を求めるには候補語句のリンク先記事を（暫定的に）決定することが必要であるのに対し、リンク確率はリンク先記事を決定せずに計算することができる。アンカー抽出と類似の問題として専門用語の抽出や重要語句の抽出が挙げられる。専門用語抽出に関しては、名詞と一部の特殊な形容詞を単名詞として扱い、それら単名詞の出現頻度と連接頻度を用いる方法[3]、重要語句抽出に関しては、検索エンジンのクエリログ内に現れている語句は重要であるという考えに基づく方法[4]や、候補語句の出現位置を利用する方法 [5]などが提案されている。

3　提案方法

3.1 概要

アンカー抽出は、アンカー候補語句をアンカーと非アンカーに分類する問題ととらえることができるので、図 1 に示すように、機械学習によりアンカー候補語句の分類器を学習させる。以下、図 1 の各要素について説明する。

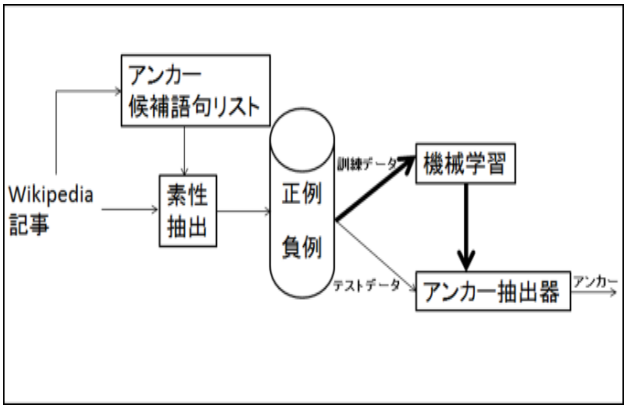

図 1 提案方法の概要

(1)　アンカー候補語句リストの作成

基本的に、Wikipedia 記事のなかで実際にアンカーとなっている全ての語句を候補語句とする。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適切でない例が存在する。また、極端な一般語をアンカーとしている例も存在する。これらを候補語句とするとノイズになるので、提案方法ではリンク確率が一定の閾値以上の語句に限定する。リンク確率とは、当該語句が出現する Wikipedia 記事のうち当該語句がアンカーとなっている記事の割合である（3.2 節の(1)で定義式

を示す)。なお、4 章の実験ではリンク確率に対する閾値を 0.005 とした。

(2) 素性抽出

Wikipedia 記事中に出現するアンカー候補語句の各々に対して 3.2 節で述べる素性の値を求める。実際にアンカーとなっている候補語句は正例、アンカーになっていない候補語句は負例である。候補語句の素性値に正例／負例のラベルを付けたデータの集合を出力する。

(3) 機械学習

(2)の出力の一部をテスト用に除外し、残りを訓練データとして用い、アンカー抽出器を学習する。

(4) アンカー抽出器

(2)の出力のうちテスト用にとっておいたデータの各々をアンカー、非アンカーのどちらかに分類する。分類結果の正誤はデータにつけられた正例／負例のラベルと比較することによって判定することができる。

3.2 素性

アンカー候補語句を分類するための素性として以下の 4 つを用いる。このうち、(1)(2)はさまざまな研究において使用されている。(3)(4)が本研究で新たに提案する素性である。なお、Milne and Witten が有効な素性であると報告している素性のうちリンク先記事関連度は用いない。アンカー候補語句に対してリンク先記事を決定するコストが大きいことがその理由である。以下の(1)から(4)において、アンカー候補語句をa、アンカー候補語句が出現した記事をDとする。

(1) 候補語句のリンク確率

Wikipedia 中で候補語句がアンカーとなっている確率で、次式で定義される。

$$Link\_prob(a) = \frac{|\{D_w | a \in Anchor(D_w)\}|}{|\{D_w | a \in D_w\}|}$$

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である。

定義式からわかるように、この素性は、候補語句が出現した文書によらず、候補語句のみによって値が決まる。

(2) 候補語句の正規化頻度

記事の中で何度も繰り返し出現する語句は読者にとって印象に残りやすく、その記事の中で重要な語句であると考えられる。したがって、記事中の候補語句出現頻度を素性とすることが考えられるが、記事のサイズの影響を受けないよう、次式のように正規化した出現頻度を用いる。

$$Norm\_freq(a, D) = \frac{f_D(a)}{\sum_{b \in D} f_D(b)}$$

ここに、$f_D(a)$は文書D内の候補語句aの出現頻度を表す。

(3) 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が違うのではないかと考えられる。例えば、候補語句の直後が「等」である場合、候補語句はアンカーになりやすい、あるいは直後が「は」である場合、候補語句は文中の主語でありアンカーになりやすいという傾向がある。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。

(3a) 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、

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

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。

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

(3b) 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、

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

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

(4) 候補語句の条件付きリンク確率

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が違うのではないかと考えられる。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。

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

ここで、条件付きリンク確率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 中の共起回数がある閾値以上の組に限定する（4 章の実験では閾値を 15 とした）。ただし、共起回数はxとyが共に出現する文書の数とし、文書中のxやyの出現回数は考慮しない。

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、

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

ここに、$LLRR(x, y)$はアンカーであるxとyの対数尤度比[6]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、

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

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 2015 年 2 月 21 日付 Wikipedia30 記事中の候補語句を選択した。各記事に含まれる候補語句は正例より負例が圧倒的に多い。そこで、正例はすべて採用し、負例は正例と同数のものをランダムに選択した。その結果、正例、負例ともに 5200 語句となった。それらを 1040 の正例と 1040 の負例からなる 5 つのデータセットに分割した。

(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[1]を使用し、機械学習には SVM（サポートベクターマシン）Libsvm[2]を使用した。

(3) 素性の組合せ

以下の 5 通りの組合せでアンカー抽出器を学習させ、それぞれによるアンカー抽出を実行した。

(i) リンク確率
(ii) リンク確率＋正規化頻度
(iii) リンク確率＋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
(iv) リンク確率＋条件付きリンク確率
(v) リンク確率＋正規化頻度＋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＋条件付きリンク確率

4.2 実験結果

アンカー抽出結果の例として Wikipedia 記事「自然言語」に対する結果を表 1 に示す。素性の組合せは 4.1 節で説明した(v)である。

また、5 通りの素性の組合せのそれぞれについて適合率(precision)、再現率(recall)、F 値を表 2 に示す。ベースラインとして、全てのアンカー候補語句を採用した場合の適合率、再現率（100%）、F 値も示す。

表 1 Wikipedia 記事「自然言語」に対する
アンカー抽出結果

抽出されたアンカー(TP)	抽出されなかったアンカー(FN)
人工言語	人間
自然言語処理	記号
形式言語	文化
プログラミング言語	文字
言語学	体系
書き言葉	
話し言葉	誤って抽出したアンカー(FP)
計算機	心理学
音声	

4.3 結果の検討

表 1 において抽出されなかったアンカーはいずれも比較的出現回数が多い一般語句である。そのような語句の抽出に関して検討の余地があるといえる。

表 2 からわかるように、すべての素性の組合せがベースラインを上回り、従来研究で有効な素性とされているリンク確率に今回提案した素性を加えることによりさらに精度が向上した。中でも、前接語のプリアンカー確率と後接語のポストアンカー確率の効果が大きいといえる。

[1]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2]http://www.okuma.nuee.nagoya-u.ac.jp/~sakaguti/wiki/index.php?LibSVM

表 2 アンカー抽出の適合率／再現率／F 値

素性の組み合わせ	*precision*	*recall*	*F値*
Link_prob(a)	75.6%	74.2%	74.9%
Link_prob(a)＋Norm_freq(a,D)	77.4%	73.6%	75.5%
Link_prob(a)＋PreAnchor(pred(a))PostAnchor(succ(a))	79.3%	76.4%	77.9%
Link_prob(a)＋Cond_link_prob(a,D)	79.1%	76.0%	77.5%
Link_prob(a) +Norm_freq(a,D)+ PreAnchor(pred(a))+ PostAnchor(succ(a))+ Cond_link_prob(a,D)	80.5%	76.9%	78.7%
全ての候補語句をアンカーとした場合	50.0%	100.0%	66.7%

5 おわりに

SVM を使ったアンカー抽出のための 2 つの新しい素性―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と候補語句の条件付きリンク確率―を提案した。従来研究で有効とされているリンク確率にこれらの新しい素性を加えることにより F 値を 3.8%向上させることができた。今後の課題は、今回の実験では除外したリンク先記事関連度の比較評価を行うことと、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、一部、JSPS 科研費 15K16096 の助成を受けて行った。

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] David Milne and lan H. Witten. An open-source toolkit for mining Wikipedia. *Artificial Intelligence 194*, pp.222-239 (2013).

[3] 中川裕志, 森辰則, 湯本紘彰. 出現頻度と連接頻度に基づく専門用語抽出. 自然言語処理, 10(1):313-320 (2003).

[4] Wen-Tau Yih, Josua Goodman, and Vitor R. Carvalh. Finding advertising keywords on web pages. In *Proceedings of the 15th International Conference on World Wide Web*, pp.213-222 (2006).

[5] Thuy Dung Nguyen and Min-Yen Kan. Keyphrase extraction in scientific publications. In *Proceedings of the International Conference on Asian Digital Libraries*, pp.317-326 (2007).

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