決定リストの機械学習による wikification

袁 楊　綱川 隆司　梶 博行
静岡大学大学院情報学研究科
gs13050@s.inf.shizuoka.ac.jp , {tuna, kaji}@inf.shizuoka.ac.jp

1. はじめに

Web 上の百科事典 Wikipedia は記事間にリンクがはられ、関連記事を容易にたどることができる。一般のテキストから Wikipedia 記事にリンクを張る“wikification”が実現されれば、新聞、論文、その他さまざまなテキストを読む際に Wikipedia を容易に参照することができるようになる。新たに作成した Wikipedia 記事に wikification を適用することももちろん可能であり、記事作成者のリンク付与作業の負担を軽減することが期待される。

Wikification は、一般に、アンカーテキストとよばれるリンク元の語句を選定するステップと選定されたアンカーテキストのリンク先記事を決定するステップからなる[1]。字面だけでは複数の概念／エンティティを表し得るアンカーテキストも多く、リンク先記事の曖昧性解消が wikification における最大の技術課題となっている。リンク先記事の曖昧性解消は Word Sense Disambiguation（語義曖昧性解消）の一ケースと考えられ、WSD のさまざまな手法の適用が試みられている。

本稿では、決定リスト[2]を用いた WSD 手法の wikification への適用について述べる。Yarowsky によって提案された決定リストを用いた WSD では、多義語の周辺に出現する語を手がかりとする語義決定ルールをルールの確信度順に並べた決定リストをコーパスから学習する[3]。これに対し、本稿で提案する wikification では、同一テキスト中の他のアンカーテキストを手がかりとするリンク先記事決定ルールをルールの確信度順に並べた決定リストを Wikipedia のリンクデータから学習する。

決定リストを用いた WSD は、テキスト中に複数の手がかりが含まれる場合でも、最も有力な単一の手がかりに基づいて語義を決定することが特徴で、決定プロセスの透明性が高く理解しやすい方法となっている。この特徴は、他のアンカーテキストを手がかりとしてリンク先記事を決定する場合にも適していると思われる。よって、決定リストによる wikification を提案し、実験結果を報告する。

2. 関連研究

Wikification におけるリンク先記事の曖昧性解消に対する代表的なアプローチは教師ありの機械学習で、ナイーブベイズ、決定木、サポートベクターマシンなどさまざまな手法が適用されている[4][5][6]。Wikipedia のリンクデータ、すなわちアンカーテキストとリンク先記事のタグの組がトレーニングデータとして利用される。素性としては、通常の WSD と同様にアンカーテキスト周辺の語とその品詞などが用いられている。

アンカーテキストを含むテキストとリンク先記事候補の間の文脈の重なり度合を利用する方法も典型的な方法である[7]。これは、語義説明文を用いた WSD において語義説明文を Wikipedia 記事に置き換えた手法と考えることができる。性質の異なるこの手法と機械学習手法を組み合わせてリンク先記事を決定する方法も提案されている[1]。

以上は、アンカーテキストごとにリンク先記事を決定するいわば局所的な方法であるが、同一テキスト中の各アンカーテキストのリンク先記事の間の関連も考慮する大域的な方法も提案されている[5][7][8][9]。記事間の関連としては、記事間のリンクを直接利用するほか、文脈の重なり度合、参照している記事の共通性（グーグル距離）、記事の属するカテゴリ間の距離などが用いられる。

3. 基本アイデア

(1) “One sense per collocation”仮説

決定リストによる WSD は“One sense per collocation”仮説[10]に基づいている。Wikification では、リンク先記事がアンカーテキストの sense に対応するが、テキスト中の他のアンカーテキストを collocation と考えるとやはりこの仮説が有効と思われる。例えば、アンカーテキスト“Jaguar”のリンク先記事候補として“Jaguar”（動物のジャガーについての記事）、“Jaguar Cars”（スポーツカー“ジャガー”についての記事）、その他がある。“Jaguar”をリンク先とするアンカーテキスト“Jaguar”を含む Wikipedia 記事は“lion”, “tiger”など、“Jaguar”に対して動物のジャガーを連想させるアンカーテキストを含んでいる。したがって、Wikipedia のリンクデータから、「IF “lion” co-occurs with “Jaguar” THEN link “Jaguar” to “Jaguar”.」というようなリンク先決定ルールを学習することができる。

決定リストは決定ルールを確信度の降順に並べたものである。ルールの確信度としては、条件部の共起アンカーテキストとアクション部のリンクとの関連度を用いることが考えられる。すなわち、Wikipedia に含まれる大量のリンクデータから対数尤度比などの関連度を計算すればよい。

(2) “One sense per collocation”の反例への対策

上に述べた方法によれば、「IF “BMW” co-occurs with “Jaguar” THEN link “Jaguar” to “Jaguar Cars”.」というルールも上位に順位付けられるであろう。しかし、このルールは誤ったリンク先をもたらす可能性がある。実は、“Jaguar”のリンク先記事候補には前記の２つのほかに“Jaguar Racing”（F1 参加チーム“ジャガー”についての記事）が含まれる。そして、共起アンカーテキスト“BMW”をもつアンカーテキスト“Jaguar”のリンク先が“Jaguar Racing”であることも多い。要するに“One sense per collocation”仮説が成立しないのである。

このような場合を考慮すると、ルールの確信度として、競合するリンクとの相対的な関連度の大きさを用いるほうがよいかもしれない。そこで、(i)関連度によるルールの順位付け、(ii)相対的な関連度によるルールの順位付け、(iii)関連度と相対的な関連度の組合せによるルールの順位付けの 3 案の比較実験を行うこととする。

“One sense per collocation”仮説が成立しない共起アンカーテキストによるリンク先記事の決定誤りを防止する方

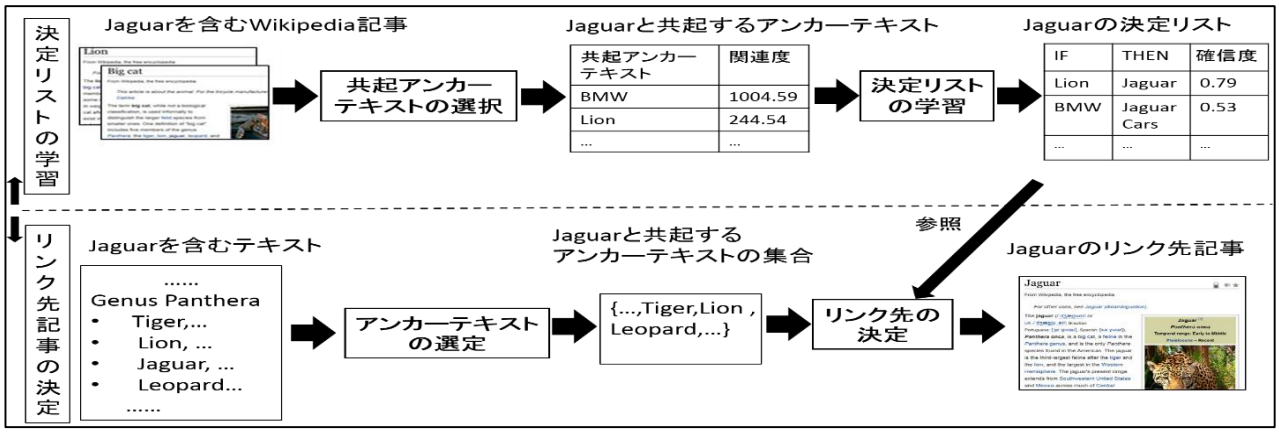

図 1　提案方法

法としては、複数の共起アンカーテキストに基づいて決定することも考えられる。そのような考え方を決定リストに組み込む方法として次の 2 案があるだろう。

(a) ルールの仕様を拡張し、条件部に共起アンカーテキストの連言（AND）を記述できるようにする。

(b) リスト中のヒットする（条件が成立する）上位 k 個のルールの多数決によってリンク先を決定する。決定リストの精神からはずれないように、k は比較的小さな値とする。

(a)案は、共起アンカーテキストの組合せが爆発的に増え、データスパースネスの問題が深刻になる。そこで、(b)案を採用し、k のいくつかの値について実験的に比較することとする。

4. 提案方法

3 章で述べた考え方にしたがって、次の 2 つのステップからなる方法を提案する。

I. Wikipedia のリンクデータをトレーニングデータとしてアンカーテキストごとに決定リストを学習する。

II. リンクをもたない入力テキストに対し、アンカーテキストを選定し、I で学習した決定リストを用いてリンク先記事を決定する。

図 1 に概要を示し、各ステップの詳細を 4.1 節と 4.2 節で述べる。

4.1. 決定リストの学習

アンカーテキストaのリンク先記事を決定するための決定リストを以下のステップで学習する。

(1)　共起アンカーテキストの選択

アンカーテキストaのリンク先決定の手がかりとする共起アンカーテキストとして、aとの共起頻度が閾値θを超えるアンカーテキストを選択する。それらをa_i $(i = 1, 2, \ldots)$とする。

(2)　共起アンカーテキストとリンクの関連度の計算

アンカーテキストaのリンク先記事候補がD_j $(j = 1, 2, \ldots)$であるとする。共起アンカーテキストa_iとaからの各リンク$l(a, D_j)$の関連度$\mathrm{Ass}\left(a_i, l(a, D_j)\right)$を計算する。関連度として相互情報量、$t$ スコア、対数尤度比のいずれかを用いる。それぞれ以下のように定義される。ただし、定義式中の$m,\ n_1,\ n_2,\ N$はアンカーテキストとリンクの共起データの分割表に示すとおりである。

	$l(a, D_j)$	$\neg l(a, D_j)$	すべて
a_i	$\boldsymbol{m}$	$n_2 - m$	$\boldsymbol{n_2}$
$\neg a_i$	$n_1 - m$	$N - n_1 - n_2 + m$	$N - n_2$
すべて	$\boldsymbol{n_1}$	$N - n_1$	$\boldsymbol{N}$

- 相互情報量MI：

$$\mathrm{MI}\left(a_i, l(a, D_j)\right) = \log_2 \frac{m/N}{(n_1/N)(n_2/N)}$$

- t スコアT：

$$\mathrm{T}\left(a_i, l(a, D_j)\right) = \frac{m - n_1 n_2/N}{\sqrt{m}}$$

- 対数尤度比LLR：

$$\begin{aligned}\mathrm{LLR}\left(a_i, l(a, D_j)\right) = -2(&\log L(m, n_1, r) \\ &+ \log L(n_2 - m, N - n_1, r) \\ &- \log L(m, n_1, r_1) \\ &- \log L(n_2 - m, N - n_1, r_2))\end{aligned}$$

$$\mathrm{L}(k, n, r) = r^k (1 - r)^{n-k}$$

$$r_1 = m/n_1,\ r_2 = (n_2 - m)/(N - n_1),\ \mathrm{r} = n_2/N$$

上の定義からもわかるように、関連度$\mathrm{Ass}\left(a_i, l(a, D_j)\right)$の計算においてアンカーテキスト$a$と共起アンカーテキスト$a_i$の記事中での距離は考慮しない。アンカーテキストの連想関係と出現位置はあまり関係がないと考えるからである。

(3)　リンク先決定ルールの生成

共起アンカーテキストa_iの各々について、a_iとの関連度が最大のリンク先を選択するルール

r_i: IF a_i co-occurs with a THEN link a to $\hat{D}(a, a_i)$.

を生成する。ここに $\hat{D}(a, a_i) = \underset{D_j}{\mathrm{argmax}}\, \mathrm{Ass}\left(a_i, l(a, D_j)\right)$である。

(4)　決定ルールの順序付け

(3)で生成したリンク先決定ルールを順序付け、最後にデフォールトルールを追加する。デフォールトルールはアンカーテキストaの最頻リンク先記事に無条件にリンクするルールである。順序付けの方法として、以下の 3 案を比較評価する。

(i) 関連度による順位付け

関連度をルールの確信度と考える。すなわち、$\mathrm{CF}_1(r_i) =$

表 1　Wikipedia 全体とテストセット T200 のアンカーテキスト*)

(a) 出現頻度 f の分布

f	f<100	100≤f<500	500≤f<1000	f≥1000	合計
Wikipedia 全体	690800 (87%)	80320 (10%)	12107 (2%)	10694 (1%)	793921
テストセット	0 (0%)	168 (84%)	24 (12%)	8 (4%)	200

*) Wikipedia 全体についても曖昧性のないアンカーテキストを除外している。

(b) リンク先記事候補数 n の分布

n	2	3	4	5	6	7	8	9	≥10	合計
Wikipedia 全体	449192 (57%)	141733 (18%)	63919 (8%)	35501 (4%)	21607 (3%)	14893 (2%)	10443 (1%)	7919 (1%)	48710 (6%)	793921
テストセット	22 (11%)	26(13%)	34 (17%)	34 (17%)	26 (13%)	21 (10%)	29 (15%)	8 (4%)	0 (0%)	200

(c) 最頻リンク先記事の確率 p の分布

p	0<p<0.4	0.4≤p<0.5	0.5≤p<0.6	0.6≤p<0.7	0.7≤p<0.8	0.8≤p<1	合計
Wikipedia 全体	82007 (10%)	46962 (6%)	181319 (23%)	126958 (16%)	89818 (11%)	266857 (34%)	793921
テストセット	19 (9%)	31 (15%)	53 (27%)	45 (23%)	26 (13%)	26 (13%)	200

$\mathrm{Ass}\left(a_i, l\left(a, \widehat{D}(a, a_i)\right)\right)$の降順に並べる。

(ii) 相対的な関連度による順位付け

他のリンク先候補の関連度と比べた相対的な関連度の大きさをルールの確信度と考える。すなわち、$\mathrm{CF}_2(r_i) = \mathrm{Ass}\left(a_i, l\left(a, \widehat{D}(a, a_i)\right)\right) / \sum_j \mathrm{Ass}\left(a_i, l(a, D_j)\right)$の降順に並べる。

(iii) 関連度と相対的な関連度の組合せによる順位付け

(i)による順位と (ii)による順位の平均をルールの順位と考えて順位付ける。$\mathrm{CF}_1(r_i)$と$\mathrm{CF}_2(r_i)$はとり得る値の範囲が異なり、直接結合することが困難であるので、それぞれによる順位を求めたうえで組み合わせる。

4.2. リンク先記事の決定

入力テキスト中のアンカーテキストの選定は、本研究の主たる目的ではないので、以下のような簡単な方法を用いる。Wikipedia において 1 回でもアンカーテキストとして選定されている語句をすべてアンカーテキストとして選定する。それらのアンカーテキストの各々について、それ以外のアンカーテキストを共起アンカーテキストと考え、以下の手順によってリンク先記事を決定する。ここに、kはリンク先決定に用いるルール数（の下限）を定めるパラメータである。

(1) ヒットするルールの集合R_Hを空にする.

(2) 決定リストの上から順にヒットするルール、すなわち条件部に書かれた共起アンカーテキストが入力テキストに含まれるルールを求めてR_Hに加える。同順位の複数のルールがヒットする時はそれらをすべて加える。これを$|\mathrm{R}_H| \geq k$になるまで繰り返す。

(3) R_Hに含まれるルールの多数決によりリンク先記事を決定する。票数が最大のリンク先記事が 1 つ得られたら終了であるが、票数が最大のリンク先候補記事が 2 つ以上の場合は(4)へ。

(4) ヒットするルールを求めてR_Hに加える処理を再開し、$|\mathrm{R}_H|$が 1 でも増加したら、(3)へ。デフォールトルールに到達したら、デフォールトルールによってリンク先記事を決定し、終了。

5. 評価実験

英語版の Wikipedia 記事をトレーニングデータ及びテストデータとして評価実験を行った。使用したのは 2014 年 02 月 03 日の Wikipedia で、記事の総数は 10,771,274 であった。アンカーテキストタイプの総数は 10,508,316 で、うち曖昧性のあるもの（複数のリンク先記事候補をもつもの）が 793,921（7.6%）、曖昧性のないものが 9,714,395（92.4%）であった。

5.1. 代替案・パラメータの比較

提案方法は、関連度の定義とルールの順位付け方法に複数の案があり、また共起アンカーテキストの共起頻度に対する閾値θとリンク先決定に用いるルール数の下限kの 2 つのパラメータをもつ。最適な組合せを決定するため、200 のアンカーテキストタイプを対象として、5 分割交差検定による比較実験を行った。

(1) アンカーテキストのテストセット T200

曖昧性のあるアンカーテキスト 200 個からなるテストセット T200 を構成した。その際、決定リストの学習のためにある程度の量のトレーニングデータが必要であるので、頻度 100 以上のアンカーテキストに限定した。また、リンク先記事候補数が 10 未満のものに限定した。リンク先記事候補が多すぎると、リンク先候補記事あたりのトレーニングデータの量が少なすぎる可能性が高いからである。

表 1 に、Wikipedia 全体と T200 それぞれの出現頻度 f、リンク先記事候補数 n、最頻リンク先記事の確率 p の分布を示す。Wikipedia 全体と比較して、T200 はリンク先記事候補数 n が大きいアンカーテキスト、最頻リンク先記事の確率 p が低いアンカーテキストの比率が高い。そのようにした理由は、代替案の間の優劣やパラメータの値による差違をはっきりさせるためである。

(2) 実験結果と検討

代替案・パラメータの各組合せによるリンク先決定の正解率を表 2 にまとめた。ベースライン（最頻リンク先記事をリンク先とする方法の正解率）は 0.61 であり、どの代替案・パラメータの組合せもこれを上回っている。実験の結果、関連度は対数尤度比が、ルールの順位付けは相対的な関連度によるのが一番よいことが明らかになった。

共起頻度の閾値θは 2 がよく、関連度あるいは相対的な関連度で順位付けるので、共起頻度の低い共起アンカーテキストを除外する必要がなかったといえる。また、リンク先決定に用いるルール数の下限は、k=1 との差が小さいものの k=3 が最適であるとの結論を得た。k=10 では明らかに正解率が低下しており、1 、2 の有力な共起アンカーテ

表 2　代替案・パラメータの比較実験の結果

関連度 Ass	共起頻度閾値θ	ルールの順位付けの方法	リンク先決定の正解率			
			k=1	k=3	k=5	k=10
MI	2	(i) 関連度	0.65	0.73	0.73	0.67
		(ii) 相対的な関連度	0.73	0.78	0.77	0.70
		(i)(ii)の組合せ	0.71	0.77	0.76	0.69
	5	(i) 関連度	0.69	0.73	0.70	0.65
		(ii) 相対的な関連度	0.73	0.75	0.73	0.67
		(i)(ii)の組合せ	0.73	0.75	0.72	0.66
	10	(i) 関連度	0.69	0.70	0.69	0.66
		(ii) 相対的な関連度	0.70	0.71	0.69	0.67
		(i)(ii)の組合せ	0.7	0.71	0.69	0.67
T	2	(i) 関連度	0.74	0.77	0.76	0.71
		(ii) 相対的な関連度	0.75	0.79	0.79	0.73
		(i)(ii)の組合せ	0.77	0.79	0.78	0.72
	5	(i) 関連度	0.73	0.74	0.73	0.69
		(ii) 相対的な関連度	0.75	0.77	0.75	0.70
		(i)(ii)の組合せ	0.75	0.76	0.74	0.70
	10	(i) 関連度	0.70	0.71	0.70	0.68
		(ii) 相対的な関連度	0.72	0.72	0.71	0.69
		(i)(ii)の組合せ	0.72	0.72	0.71	0.69
LLR	2	(i) 関連度	0.75	0.77	0.77	0.71
		(ii) 相対的な関連度	0.79	**0.80**	0.79	0.73
		(i)(ii)の組合せ	0.77	0.79	0.78	0.72
	5	(i) 関連度	0.74	0.75	0.73	0.69
		(ii) 相対的な関連度	0.76	0.77	0.75	0.70
		(i)(ii)の組合せ	0.75	0.76	0.74	0.69
	10	(i) 関連度	0.71	0.71	0.70	0.68
		(ii) 相対的な関連度	0.72	0.72	0.71	0.69
		(i)(ii)の組合せ	0.72	0.72	0.71	0.68

表 3　大量のアンカーテキストを対象とした評価実験の結果

関連度 Ass	共起頻度閾値θ	ルールの順位付けの方法	リンク先決定の正解率
			k=3
LLR	2	(i) 関連度	0.9045
		(ii) 相対的な関連度	0.9036
		(i)(ii)の組合せ	0.9044

キストに基づいてリンク先記事を決定するのがよいと言える。

5.2. 大量のアンカーテキストを対象とした評価

(1) アンカーテキストのテストセット T10000

曖昧性のあるアンカーテキスト 10000 個からなるテストセット T10000 を構成した。頻度 100 以上という条件のもとで、ランダムに選択したので、このテストセットに対する評価結果を Wikipedia の全アンカーテキストに対する評価結果と考えてよいであろう。

(2) 実験結果と検討

5.1 節の結果に従い、関連度は対数尤度比、共起頻度の閾値θ=2、リンク先決定に用いるルール数の下限 k=3 とした。ただし、ルールの順位付けについては、(ii)相対的な関連度と(iii)関連度と相対的な関連度の組合せとの正解率の差はそれほど大きくなかったので、(i)関連度を含む 3 つの案を実行した。表 3 に結果をまとめた。T10000 に対するベースラインの正解率は 0.85 であり、ルールの順位付け方法がどれであっても提案方法は有効であるといえる。

ルールの順位付け方法による正解率の差はほとんどなかったが、 T200 の場合とは逆の順になった。T200 と T10000 の違いから、次のように推察される。T200 は曖昧性が高いアンカーテキストの比率を恣意的に高めており、“One sense per collocation”が成立しないアンカーテキストも多かったので、それに強い(ii)の正解率が相対的に高かった。これはあくまでも推察であり、結果を詳細に分析する必要がある。

6. おわりに

Wikification におけるリンク先記事の曖昧性解消のため、共起アンカーテキストによるリンク先決定ルールを確信度順に並べた決定リストを学習する方法を提案した。英語 Wikipedia 記事を用いた交差検定による評価実験では正解率 90%を達成した。最頻リンク先を選択する方法によるベースラインは 85%であり、提案方法の有効性が確認された。出現頻度が小さいアンカーテキストに対する決定リストの学習が今後の課題である。

参考文献

[1] R. Mihalcea and A. Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the16th ACM Conference on Information and Knowledge Management*, pp.233-242.

[2] R. L. Rivest. 1987. Learning decision lists. *Machine Learning*, Vol.2, pp.229-246.

[3] D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp.189-196.

[4] R. Bunescu and M. Paseca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp.9-16.

[5] D. Milne and I. H. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pp.509-518.

[6] R. Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.196-203.

[7] L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp.1375–1384.

[8] D. Turdakov and P. Velikhov. 2008. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation. In *Proceedings of the Spring Young Researcher's Colloquium on Database and Information Systems*, pp.35-40.

[9] S. Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.708-716.

[10] D. Yarowsky. 1993. One sense per collocation. In *Proceedings of the Workshop on the Human Language Technology*, pp.266-271.