Vol.2013-NL-214 No.9
2013/11/15

Wikipedia 内部リンクの言語間変換

綱川 隆司[†1]　　梶 博行[†1]

Wikipedia の最大の特徴の一つはハイパーテキストであることだが，リンクを付与する労力の低減が望まれる．二つの異なる言語の記事が言語間リンクで結ばれていて，各記事に同じ事柄に対応する内部リンクが存在するとき，それらが指す記事の間にも言語間リンクがあると考えられる．本稿では，この仮定を用いることで，Wikipedia 記事に内部リンクを自動的に付与するために，他の言語版の記事に存在する内部リンクを変換する方法を提案する．この方法の有効性を示すため，対訳のアンカー文字列対に対してリンク先記事間に言語間リンクが存在する割合が非常に高いことを示した．また，言語間リンクが存在しない場合について考察した．

Transferring Wikipedia Intralanguage Links across Languages

TAKASHI TSUNAKAWA[†1]　HIROYUKI KAJI[†1]

Although one of the most important features of Wikipedia is that it has hypertext, reducing costs of annotating links is desired. If two Wikipedia articles in different languages are connected by an interlanguage link and they have intralanguage links corresponding to the same entity, it seems likely that linked articles are also connected by an interlanguage link. On the basis of the above assumption, this paper proposes a method for automatically annotating intralanguage links in a Wikipedia article by transferring intralanguage links of an article in another language. We showed that the anchor text pairs in two languages were almost always linked to articles connected by interlanguage links. In addition, we discuss the cases that the anchor text pairs were linked to irrelevant articles.

1. はじめに

Wikipedia の記事はハイパーテキストで記述されており，記事中のテキストにある豊富な内部リンクによって，読者は未知の語句や関連項目に素早くアクセスすることができる．また，内部リンクで結ばれた二つの記事には関連性があることから，リンク構造を利用した語句の意味的関連性の計測 [1] やパラレルコーパスの構築 [2] といった応用が幅広く展開されている．

記事中のリンクはテキスト編集時に人手で付与されているが，リンク先記事の指定も行わなければならないため，労力がかかる．近年，テキストに対して Wikipedia 記事へのハイパーリンクを自動的に付与するエンティティリンキングあるいは Wikification [3] とよばれるタスクの研究が進められており, Wikipedia 記事に適用することでリンクを自動的に付与することができると考えられる．しかし，この方法ではテキストのどの部分をアンカー文字列とするか，およびアンカー文字列に対してどの記事にリンクするかの二つの課題を解決する必要がある．

本稿では，ある記事に対して他の言語版の記事が存在するときに，内部リンクを言語間で変換することにより新たな内部リンクを付与する方法を提案する．本方法は，言語間リンクで結ばれた二つの記事について，アンカー文字列が互いに対訳であるような内部リンクがそれぞれの記事に含まれるとき，それらの指す先の記事は同じ事柄に関するものであるという仮説に基づいている． この仮説は，one sense per discourse [4][5] を Wikipedia 記事に拡大適用することから導ける．

本方法の有効性を示すため，既存の Wikipedia に含まれる内部リンクを正解として利用した正答率の予測を行う．また，既存の Wikipedia において本方法を用いた内部リンクの変換が不適切な結果になる場合について考察する．

以下, 2 節で Wikipedia の内部リンクと言語間リンクについて説明し，3 節で本研究の着眼点と提案方法について述べる. 4 節で既存の Wikipedia を用いた正答率の推定と結果についての議論を行う．5 節で内部リンクおよび言語間リンク付与に関する関連研究について述べ，6 節にて全体をまとめる．

2. Wikipedia の内部リンクと言語間リンク

Wikipedia 記事の内部リンクは同一言語の他の記事へのリンクである．内部リンクは，アンカーとするテキストの一部分（以下, “ アンカー文字列 ”と呼ぶ）と，リンク先記事からなる．リンク先記事は任意に指定できる．したがって Wikipedia 全体では，一つのアンカー文字列に対してリンク先記事が複数存在し，内部リンクを付与するときには記事によってどのリンク先が適切かを判断する必要がある．まだ存在しない記事への内部リンクを作成することも可能である. Wikipedia の内部リンク作成のガイドラインによれば，たとえ対応する記事が存在しても単なる単語にはリンクすべきでなく，内容に関連するリンクだけを作成することとされている．

Wikipedia は 200 を超える言語で記述されており, 一つの事柄に対応する各言語版の記事は，言語間リンクによって

†1 静岡大学
Shizuoka University

結ばれている．言語間リンクで結ばれた二つの記事のタイトルは多くの場合対訳関係になっており，以下本稿では，ある記事と言語間リンクで結ばれた記事のタイトルを，その記事の“対訳タイトル”と呼ぶ．

3. 内部リンクの言語間変換

3.1 着眼点

[4] は，一つの談話内で一つの語句が複数の意味で使われることはほとんどないという one sense per discourse 仮説を提唱した．Wikipedia の記事については，様々な側面を記述した長い記事も存在し，そのような場合は複数の談話からなると認められる．しかし，それらの談話は同一の事柄に関することであるから，以下が成り立つと仮定できる．

仮定 1 一つの Wikipedia 記事に含まれる全ての談話を通じて，一つの語句はほぼ一つの意味で用いられる．

また，言語間リンクで結ばれた記事対は同一の事柄について記述していることから，以下が成り立つと仮定する．

仮定 2 言語間リンクで結ばれた記事対は共通の談話を含む．

両仮定から，言語間リンクで結ばれた記事対中の共通の談話に現れる対訳のアンカー文字列対は一つの意味で用いられるはずであり，したがってそれらのリンク先記事同士も言語間リンクで結ばれていることになる．例えば，図 1 のように言語間リンクで結ばれた英語の記事“Seychelles”と日本語の記事“セーシェル共和国”について，それぞれに首都に関するアンカー文字列“Victoria”と“ヴィクトリア”が存在する．これらのアンカー文字列はともに多くの人名，地名に用いられており非常に曖昧な語であるが，リンク先記事はそれぞれセーシェルの首都に関する記事

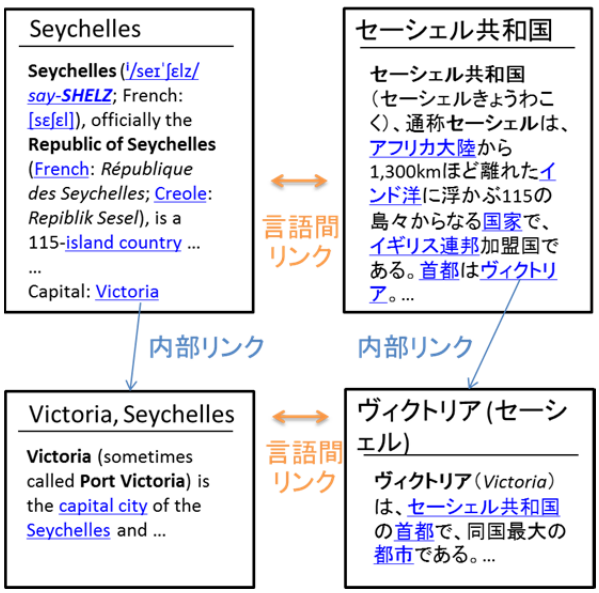

図 1 言語間リンクで結ばれた記事中の内部リンク

Figure 1 Intralanguage links in articles connected by interlanguage links.

“Victoria, Seychelles”と“ヴィクトリア (セーシェル)”であり，言語間リンクで結ばれている．

この関係を用いることで，一方の言語の記事に内部リンクがあり，もう一方の言語の記事にはない場合に，アンカー文字列の訳とリンク先記事の言語間リンク先から内部リンクを付与することができる．

3.2 提案方法

言語間リンクで結ばれた二つの異なる言語の記事を a_1, a_2 とする．a_1 に含まれる各内部リンクを a_2 上に以下の手順で変換する．

1. a_1 の内部リンクのリンク先記事 a_3 と言語間リンクで結ばれた記事 a_4 について，a_4 をリンク先とするアンカー文字列の集合を S とする．
2. a_2 のテキストのうちアンカー文字列でない部分の中で，S の要素のいずれかが最初に出現する部分をアンカー文字列とし，a_4 をリンク先記事とする内部リンクを付与する．

図 2 は，記事“Tata Motors”に含まれる内部リンク“Jaguar”を変換する例を示している．内部リンク“Jaguar”のリンク先記事“Jaguar Cars”に対応する日本語の記事“ジャガー (自動車)”について，この記事を指すアンカー文字列の集合は S = {ジャガー (自動車), ジャガー} となる．記事“タ

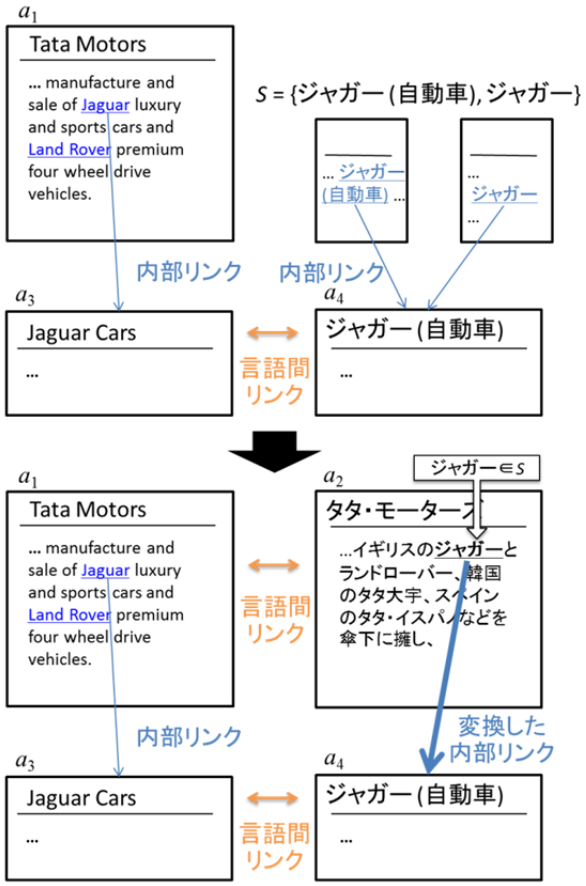

図 2 内部リンクの言語間変換

Figure 2 Transferring intralanguage links across languages.

タ自動車”の中で S に含まれる文字列を探し，その部分をアンカー文字列として記事“ジャガー (自動車) ”を指す内部リンクを付与する．

3.1 節で述べたように，記事 a_1 の内部リンクを記事 a_2 に対して変換する場合のリンク先記事は a_3 と言語間リンクで結ばれた a_4 であり，a_4 に対するアンカー文字列は既存の Wikipedia から得ることができる．また，この方法で付与した内部リンクは記事 a_1 で適切に付与されたものを変換したに過ぎないため，内部リンク作成のガイドラインも満たしているといえる．

4. 正答率の推定

4.1 推定方法

提案方法により言語間変換した内部リンクが適切であるといえるのは，変換前の内部リンクと変換後の内部リンクのリンク先記事の間に言語間リンクが存在するときである．例えば，図 1 の例は二つの内部リンクの先に言語間リンクが存在する例である．一方で，図 3 は英語記事のアンカー文字列“Hamamatsu”が記事“Hamamatsu Station”を指す一方で，日本語記事のアンカー文字列“浜松”が記事“浜松市”を指す例であり，このような場合は提案方法によって変換した内部リンクは適切でないことになる．Wikipedia において対訳タイトル対をそれぞれアンカー文字列として持つ言語間リンクのついた記事対を抽出し，その中でアンカー文字列のリンク先記事に言語間リンクがある割合を提案方法の正答率として推定した．

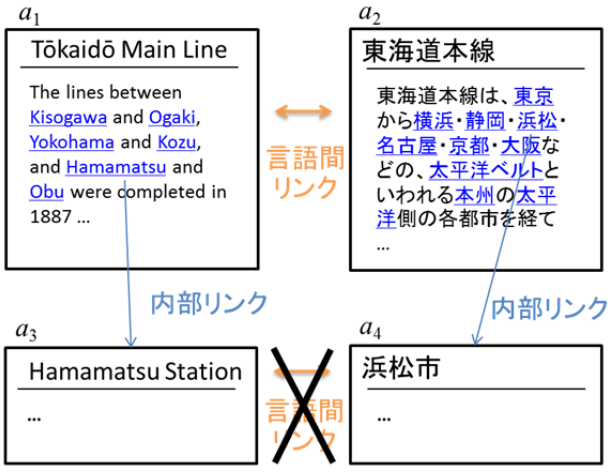

図 3　内部リンクの言語間変換が失敗する例

Figure 3　A failed case of transferring intralanguage links.

4.2 推定結果

(1)　全体の正答率

英語 Wikipedia（2013 年 4 月 3 日時点）と日本語 Wikipedia（2013 年 3 月 28 日時点）および言語間リンクのための Wikidata（2013 年 3 月 28 日時点）のダンプデータを推定に用いた．英語記事と日本語記事の対訳タイトル対は 368,849 対存在する．これらをアンカー文字列とする内部リンクが，言語間リンクのついた記事中に対で現れた回数は延べ 9,386,824 回であり，このうち 92.0%にあたる 8,637,309 回はリンク先記事間に言語間リンクが存在した．このことから，提案方法において，アンカー文字列の集合 S を記事 a_4 のタイトルのみに限定した場合の正答率は 92.0%であると推定できる．

(2)　アンカー文字列対別の正答率

アンカー文字列対のうち，少なくとも 10 記事対に出現するものは 91,016 対存在した．これらのそれぞれについて，出現する記事対数に対し，内部リンクのリンク先記事に言語間リンクがある数の割合をアンカー文字列対ごとの正答率として求めた．

正答率が 100%，すなわち，必ずリンク先記事が言語間リンクで結ばれているアンカー文字列対は 74,632 個あり，82.0%を占めている．図 4 はある正答率以上のアンカー文字列対の数を示したグラフであり，正答率 95%以上は 80,198 個（88.1%），正答率 90%以上は 83,548 個（91.8%）存在した．一方で，正答率が 50%未満のものは 2,291 個（2.5%），0%のものは 600 個（0.7%）存在した．正答率の平均は 96.2%であった．これらのことから，多くの Wikipedia 記事タイトルについて，ある記事にタイトルと同じ文字列が現れた場合，提案方法を用いてその部分をリンク文字列とする内部リンクを作成した場合，90%以上の高い確率で適切な内部リンクとなると考えられる．

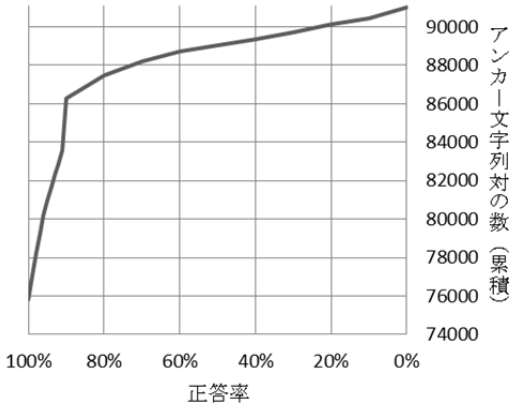

図 4　ある正答率以上のアンカー文字列対の累積数

Figure 4　Accumurative numbers of anchor text pairs above the accurate rate.

(3)　リンク先記事数別の正答率

アンカー文字列のリンクする可能性のある記事の数毎に集計したものが表 1 である．英語・日本語ともそれぞれ一つの記事にしかリンクしないアンカー文字列対（49,982 個）は必ず正答率が 100%となる．残りの 40,124 個についての正答率の平均は 92.0%であり，例えば英語・日本語とも 5〜10 個の記事にリンクする可能性のあるアンカー文字列対であっても平均 75.0%の正答率が得られた．

表 1　アンカー文字列対のリンク先記事数別の
正答率平均(%)（縦軸：英語記事数、横軸：日本語記事数）
Table 1　Average accuracy rates (%) of anchor text pairs by distinct numbers of linked articles.
(Rows: # of English articles, Columns: # of Japanese articles)

	1	2	3	4	5-10	11-99	100-
1	100	95.2	95.6	94.8	95.2	91.9	100
2	98.0	91.6	91.4	92.5	89.0	93.9	45.5
3	97.4	91.7	85.5	86.7	89.6	81.6	83.3
4	96.7	90.3	84.1	82.4	84.4	83.7	n/a
5-10	95.4	86.3	81.9	79.3	75.0	74.9	n/a
11-99	95.1	84.4	77.1	71.5	70.3	69.4	69.9
100-	90.8	86.5	84.0	80.1	74.9	71.2	73.7

4.3 正答率の低いアンカー文字列の分類

表 2 で，リンク先記事間に言語間リンクがない場合について，アンカー文字列対の例を挙げながら整理した．また，このようなアンカー文字列対を含む記事対には以下のような特徴がみられた．

- 非常に長い記事
  一つの記事が，その事柄に関する様々な側面から説明されることで非常に長くなる場合がある．このとき，一つの記事で同じアンカー文字列が複数の記事にリンクされる場合がある．例えば，“日本”という記事中には“内閣”というアンカー文字列が複数回現れ，“内閣 (日本)”および“内閣”という異なる記事にリンクされている．これは仮定 1 が成立しない場合といえる．このような場合，トピックの遷移の検出を行い二つの言語版の記事をトピック毎に対応づけることで解決できる可能性がある．
- リスト・表形式中の内部リンク
  リストや表の中に現れる内部リンクのアンカー文字列は，その位置に関係する事柄と関係づけられた記事へリンクされる．例えば，表 1 においてアンカー文字列 Jalisco が記事 Estadio Jalisco（ハリスコ州にあるスタジアム名）にリンクされるのは，メキシコのサッカーリーグに所属するチームを列挙した表中であり，「スタジアム」の列に現れるためである．これは一つの記事中で同じ語句が複数の意味で用いられる例外的な場合と考えられる．

5. 関連研究

[6] は，記事と記事を結ぶ内部リンクを概念同士の関係性を示すものとみなすことで，まだリンクされていない「欠けたリンク」の発見を行った．ある記事に欠けたリンクを付与するため，その記事と似たリンク構造を持つ関連記事を探し，関連記事に含まれるリンクを加えていく．また，一般のテキストに Wikipedia 記事等の知識ベースへのリンクを付与するエンティティリンキングにおいても，Wikipedia のリンク構造をセマンティックネットワークとして利用する研究が進められている [7][8][9]．これらの課題は単一言語上で解決されており，他の言語の情報が利用可能な状況を前提としていない．

Wikipedia の言語間リンクを利用した自然言語処理タスクへの応用に関する研究も進められている．[2] は言語間リンクで結ばれた二つの Wikipedia 記事からパラレルコーパスを構築するために内部リンクの共通性を利用し，リンクされていない固有表現も内部リンクとして拡張することで抽出できる対訳文を増やしたが，拡張した内部リンクの曖昧性解消については考慮していない．[10] は，同じ事柄を指す二つの言語の Wikipedia 記事の間に新たに言語間リンクを付与するため，それぞれの記事に含まれる内部リンクの指す先の記事間に存在する言語間リンクの数を分類器学習のための特徴の一つとして用いた．[11] は更に内部リンクを拡張することで言語間リンクの分類学習器の特徴数を増加させている．これらの研究で用いられた記事間の内部リンクと言語間リンクの連鎖的関係は本研究のものと非常に近く，本研究はこの関係を内部リンクの発見に用いている．

6. おわりに

本稿では，言語間リンクで結ばれる二つの記事のテキストに対して，一方の内部リンクを他方に変換する方法を提案した．既存の Wikipedia 記事対を用いて，本方法の正答率を予測したところ，アンカー文字列とリンク先記事のタイトルが同一の場合において 91.7%であることを確認した．また，リンク先記事が複数あるアンカー文字列に限った場合でも，平均で 90%を超える確率で適切なリンク先が割り当てられることを確認した．

今後，アンカー文字列とリンク先記事のタイトルが異なる場合を含めた正答率の推定を行う．アンカー文字列とリンク先記事タイトルは単なる同義語でなく，アンカー文字列がリンク先記事の一部の性質を示すに過ぎないことがある．このような場合，アンカー文字列は高い曖昧性を持つことから，アンカー文字列がリンク先記事を表すのに適切かどうか判定することが必要になる．また，本方法の有効性を直接示すため，既存の Wikipedia データにおいてアンカー文字列でない部分に内部リンクを付与した場合の評価や，新しい記事を作成する際に必要な内部リンク作成コストの低減を実際に示すことが今後の課題として挙げられる．

参考文献

1) Milne, D.: Computing semantic relatedness using Wikipedia link structure, In Proceedings of the New Zealand Computer Science Research Student Conference (2007).

2) Adafre, S. F. and de Rijke, M.: Finding similar sentences across multiple languages in Wikipedia, In Proceedings of EACL 2006 Workshop on New Text, Wikis and Blogs and Other Dynamic Text Sources, pp. 62-69 (2006).
3) Mihalcea, R. and Csomai, A.: Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 233-242 (2007).
4) Gale, W. A., Church, K. W., and Yarowsky, D.: One sense per discourse. In Proceedings of HLT '91 Workshop on Speech and Natural Language, pp. 233-237 (1992).
5) Carpuat, M.: One translation per discourse. In Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pp. 19-27 (2009).
6) Adafre, S. F. and de Rijke, M.: Discovering missing links in Wikipedia. In Proceedings of the 3rd International Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005), pp. 90-97 (2005).
7) Milne, D. and Witten, I. H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceedings of the Wikipedia and AI Workshop of AAAI, pp. 25-30 (2008).
8) Fogarolli, A.: Word sense disambiguation based on Wikipedia link structure. In Proceedings of 2009 IEEE International Conference on Semantic Computing, pp. 77-82 (2009).
9) Ratinov, L., Roth, D., Downey, D., and Anderson, M.: Local and global algorithms for disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1375-1384 (2011).
10) Sorg, P. and Cimiano, P.: Enriching the crosslingual link structure of Wikipedia - a classification-based approach. In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (2008).
11) Wang, Z., Li, J., and Tang, J.: Boosting cross-lingual knowledge linking via concept annotation. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence, pp. 2733-2739 (2013)

表 2　アンカー文字列対毎の正答率と特徴
Table 2　Example anchor text pairs with accuracy rates and the descriptions.

分類	アンカー文字列対（括弧内数字はリンク先記事の数）	正答率(%)	言語間リンクのないリンク先記事対の例	特徴
地名	Yamanashi (6), 山梨 (36)	100	―	Estadio Jalisco や Goa trance のように、その地名の名がついた別の事柄を表すために地名のみのアンカー文字列が用いられる場合がある。
	Jalisco (16), ハリスコ州 (2)	99.2	Estadio Jalisco, ハリスコ州	
	Goa (65), ゴア州 (2)	98.4	Goa trance, ゴア州	
	Kawasaki (29), 川崎 (61)	54.3	“Kawasaki, Kanagawa”, 川崎駅	
人名	Damon lindelof (1), デイモン・リンデロフ (1)	100	―	人名のうち姓名がはっきりしているものの大半はリンク先記事が一意に定まる。一方で姓または名のみの項目は曖昧性が高く、同じ形で地名等の他の固有名詞にも使われる場合、正答率は低くなる傾向がある。
	John George I (4), ヨハン・ゲオルク 1 世 (4)	100	―	
	Adam (227), アダム (21)	82.0	Adam (Bible), アダムとイヴ	
	Philip II (41), フィリップ 2 世 (13)	68.1	Philip the Bold, フィリップ 2 世 (フランス王)	
その他の固有名詞	Fist of the North Star (9), 北斗の拳 (10)	68.2	Fist of the North Star (1986 film), 北斗の拳 (対戦型格闘ゲーム)	芸術・娯楽作品は、異なる作者による同一タイトルが存在したり、同じ作品が様々なメディアで展開されたりすることでそれぞれに対応する記事が作成される。かつ、それらが一つの記事中で同時に現れることがあるため、言語間リンクのみの対応付けでは不十分な場合がある。この他、地名と同様に、その名称を一部とする別の事柄を表す場合がある。
	Godzilla (27), ゴジラ (18)	30.9	Godzilla (2014 film), ゴジラ (1954 年の映画)	
	Waseda University (2), 早稲田大学 (13)	82.2	Waseda University, 早稲田大学ラグビー蹴球部	
年号と年次イベント	1980 (2419), 1980 年 (88)	70.5	1980 World Series, 1980 年	Wikipedia にはある年に関する事項を並べた年号の記事が存在する．Wikipedia の編集ポリシーでは，記事の主題に密接に関係する年号や日付に対して内部リンクを作成する場合があるが，そのリンク先が単なる年号の記事である場合と年に関係するイベントの場合がある．また，イベント名と各年のイベントに対してそれぞれ記事が存在する場合も，言語によってリンク先が異なる場合がある．
	FIFA World Club Cup (14), FIFA クラブワールドカップ (9)	88.7	2006 FIFA World Club Cup, FIFA クラブワールドカップ	
	Tour de France (115), ツール・ド・フランス (103)	23.8	1972 Tour de France, ツール・ド・フランス 1975	
曖昧な用語	Skin (47), 皮膚 (3)	98.6	Human skin, 皮膚	一般名詞や専門用語の記事には、言語によって記事の整理のされ方に差異がある場合がある。例えば“Skin”の場合、英語のみ人間の皮膚についての記事が別に存在する。また、分野によって専門用語の意味が異なる場合、一方で曖昧性を解消したページにリンクしないことで問題が生じる。
	Translation (20), 翻訳 (7)	90.3	Translation (genetics), 翻訳	
国名等の略称	GER (295), GER(223)	39.7	German motorcycle Grand Prix, ドイツグランプリ	国名等の略称がアンカー文字列として使われる場合、それに関係した記事へリンクされるため、高い曖昧性が生じる．また，記事内のテーブル中に用いられた場合，アルファベットが表す国名等とテーブル中の場所の両方に関係する記事へリンクされるため、言語間リンクのみの対応では不十分である．また，最後の例のようにリンク先記事の一方が存在しない場合がある．これは一方を他方に直訳した場合によく生じる。
	AUS (284), AUS (132)	16.9	Kent Music Report, ARIA チャート	
	DEN (110), DEN (22)	5.8	2006 Grand Prix of Denver, 2006 年のグランプリ・オブ・デンバー *(※存在しないページ)*	