複数言語からの Wikipedia リンクの変換

穂積正隆　　綱川隆司　　新谷誠　　梶博行
静岡大学大学院情報学研究科
gs14040@s.inf.shizuoka.ac.jp {tuna, araya, kaji}@inf.shizuoka.ac.jp

1　はじめに

Wikipedia は「信頼されるフリーなオンライン百科事典、それも質・量ともに史上最大の百科事典を、共同作業で創り上げることを目的とするプロジェクト」[1]である。Wikipedia では記事から他の記事へ“言語内リンク”が張られており、関連する記事を容易に参照することができる。しかし、言語内リンクの付与は記事の編集者には大きな負担となる。アンカー（言語内リンクを付与する語）を選定し、その語が複数の意味を持つ場合は記事中で使用されている意味に対応する記事を正しく選択しなければならない。このためリンクが不足している記事、あるいはリンク先記事が間違っているリンクが散見される。

テキスト中の語句に Wikipedia 記事へのリンクを自動付与する wikification の手法[1]-[3]を Wikipedia 記事に適用することは可能である。しかし、適切なアンカーの選定にはテキスト中の語句の重要性・関連性を測る方法が、適切なリンク先記事の決定には語句の表す意味を判別する方法がそれぞれ必要であり、wikification はいまだ発展途上の課題である。Tsunakawa et al. [4] は、他言語版の記事における言語内リンクを変換することにより、上記二つの課題を直接解くことなく新しい言語内リンクを付与できることを示した。

本稿では、複数言語版の Wikipedia 記事を変換元記事として利用するように Tsunakawa et al.の方法を拡張する。複数言語版の記事を用いることでより多くの言語内リンクを得ることが期待される。しかし、リンク数が多ければよいというわけではなく、重要なリンクに絞ることが望ましい。このため、変換されたリンクに順位をつける方法を提案する。

2　言語内リンクの言語間変換[4]

Wikipedia は多言語百科事典であり、同様の事柄について記述された各言語の記事は言語間リンクで対応づけられている。しかし、各言語版の記事は異なる編集者によって記述されているため、通常、対訳ではなく、言語内リンクも対応しているとは限らない。図1の例において英語記事“Market failure”は“Economic efficiency”へのリンクをもつが、“Market failure”に対応する日本語記事“市場の失敗”は“Economic efficiency”に対応する“効率性”へのリンクをもっていない。このとき、日本語記事“市場の失敗”から“効率性”へのアンカーとなり得る語を探索する。実際、“効率性”という語が含まれているので、これをアンカーとして日本語記事“効率性”へのリンクを付与する。このようにして“Market failure”中の“efficient”から“Economic efficiency”へのリンクを、“市場の失敗”中の“効率性”から“効率性”へのリンクに変換する。

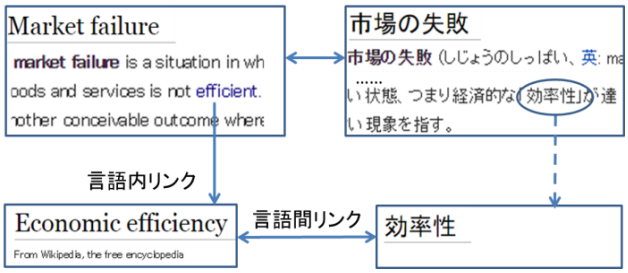

図 1：言語内リンクの言語間変換

上の例は次のような方法に一般化される。言語間リンクで結ばれた二つの言語の Wikipedia 記事pとqに対し、p中のアンカーaのリンク先記事p_dと言語間リンクで結ばれた記事q_dへのアンカーとなり得る語をqから探索し、そのような語が見つかればそれをアンカーbとするq_dへのリンクをqに付与する。ここで、アンカーとなり得る語が複数見つかることもある。その場合は次式で定義されるアンカー翻訳確率が最大の語を選択する。

$$P(b|a) = \frac{count(a,b)}{\sum_{b'} count(a,b')}$$

ここに、$count(a,b)$は、aとbが言語間リンクで結ばれた Wikipedia 記事対にアンカーとして出現し、それらのリンク先記事が言語間

[1] https://ja.wikipedia.org/wiki/Wikipedia:ウィキペディアについて

リンクで結ばれている回数である。

上記の方法において、アンカーとして選ばれた語bが多義語であることがある。しかし one sense per discourse [5] の仮説から、bは変換元記事に含まれる語（多くの場合、変換される言語内リンクのアンカーa）の訳語と同じ意味で用いられていると考えられる。したがって、変換によって得られた言語内リンクは、アンカーも含め適切であると考えられる。

3　複数言語の変換元記事を用いる拡張方法

複数の変換元言語を考え、各言語の Wikipedia 記事から 2 で述べた方法によって変換されるリンクの和集合を得られるリンクとする。したがって、変換元言語が一つの場合に比べてより多くのリンクを得ることが期待される。しかし、変換元言語の間でコンフリクトが生ずる場合があり、コンフリクトを解消することが必要である。また、多くの言語の Wikipedia 記事から変換されるリンクほど重要なリンクと考え、変換によって得られるリンクに順位を付ける。

(1)　リンク間のコンフリクトの解消

リンク先記事に対するアンカーを決定する際に用いるアンカー翻訳確率は変換元言語によって異なる。このため、同一のリンク先記事に対して、変換元言語によって異なるアンカーが得られることがある。この場合、変換元言語数が最も多いアンカーを選択することとする。変換元言語数が最大のアンカーが複数ある場合は、それらの中から無作為に一つ選択する。

(2)　リンクの順位付け

リンクの重要さを示す指標として多言語記事類似度（MLS）、変換元言語数（TN）および TN の補助指標として非アンカー言語数（NAN）を提案する。

(a) 多言語記事類似度（MLS）

Wikipedia の記事の編集方針に、関連の深いページにリンクを作成すべき、という方針がある。記事間の関連性にリンク構造の類似性を利用した研究 [6] を参考に、単言語の記事d_1とd_2の類似度$\mathrm{S}(d_1, d_2)$をアウトリンク集合$O(d_1)$と$O(d_2)$の類似度(Jaccard 係数)で定義する。

$$\mathrm{S}(d_1, d_2) = \frac{|O(d_1) \cap O(d_2)|}{|O(d_1) \cup O(d_2)|}$$

言語間リンクで結ばれた記事を一つの記事と考えて$\mathrm{S}(d_1, d_2)$を拡張したものを多言語記事類似度と呼ぶ。すなわち、多言語記事類

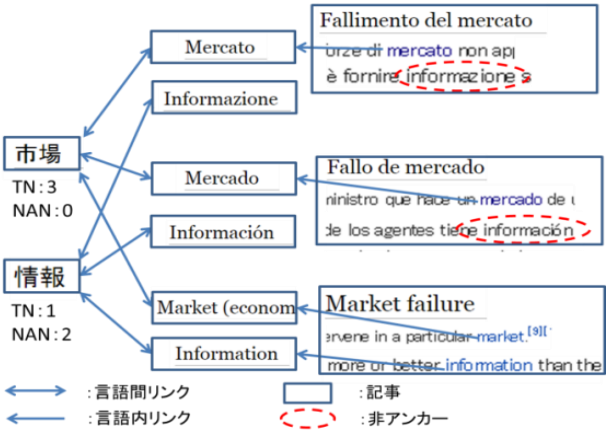

図 2 変換元言語数（TN）と非アンカー言語数（NAN）

似度$\mathrm{MLS}(d_1, d_2)$を次式で定義する。

$$\mathrm{MLS}(d_1, d_2) = \frac{\left|\left(O(d_1) \cup \cup_{d \sim d_1} O(d)\right) \cap \left(O(d_2) \cup \cup_{d \sim d_2} O(d)\right)\right|}{\left|\left(O(d_1) \cup \cup_{d \sim d_1} O(d)\right) \cup \left(O(d_2) \cup \cup_{d \sim d_2} O(d)\right)\right|}$$

ここに、$d \sim d_i$は記事dが記事d_iと言語間リンクで結ばれていることを表す。

変換先記事ごとに、変換によって得られる各リンクのリンク先記事と変換先記事の多言語記事類似度を計算し、その降順にリンクを順位付ける。

(b) 変換元言語数　(TN)

多くの言語の Wikipedia 記事から変換されたリンクほど重要なリンクと考え、変換元言語数の降順にリンクを順位付ける。図 2 に、日本語記事“市場の失敗”に対して複数言語から変換されたリンクの一例を示す。“市場”へのリンクは 3 言語から変換され、“情報”へのリンクは 1 言語から変換されているので、変換元言語数はそれぞれ 3、1 である。

(c) 非アンカー言語数（NAN）

図 2 で、“情報”へのリンクに変換された言語版以外の二つの言語版の記事は“情報”に対応する記事へのアンカーとなり得る語を含んでいるが、“情報”に対応する記事にはリンクされていない。これらの言語版では“情報”に対応する語が重要でないと判断されたものと考えることができる。このように、変換元記事において、アンカーとなり得るにもかかわらずアンカーとなっていない語、すなわち非アンカーを含む変換元記事の数を非アンカー言語数と呼ぶ。図 2 で“市場”へのリンクと“情報”へのリンクの非アンカー言語数はそれぞれ 0, 2 である。非アンカー言語数はリンクの順位を下げるいわば負の情報であり、正の情報である変換元言語数から非アンカー言語数を減じた値の降順にリンク

を順位づける。

4 評価実験

リンク増加率とアンカー正解率の測定、リンクの順位付け指標の比較評価を行う。いずれの実験においても変換先記事の言語を日本語とし、2014 年 4 月時点での純記事数が多かった 15 言語を変換元言語とした。実験に使用したのは 2014 年 10 月時点のダンプデータである。

4.1 リンク増加率とアンカー正解率測定

日本語 Wikipedia から無作為に選択した 8276 記事に対して 3 で述べた方法を適用した。変換元記事の言語は en、fr、de、es、zh、it、pl、nl、pt、ru、sv、uk、vi、war、cab の順に 1 言語ずつ増やすこととした。各段階で得られたリンクを、記事に付与されているリンクとの関係から以下に分類した。

(a) 既存リンク（完全一致）: 記事に付与されているリンクとリンク先記事もアンカーも一致したリンク

(b) 既存リンク（アンカー不一致）: 記事に付与されているリンクとリンク先記事のみが一致したリンク

(c) 新規リンク: 記事に付与されているリンクが指していないリンク先記事を持つリンク

(d) アンカーなしリンク: アンカーとなる語が記事中で発見されなかったリンク先記事

分類結果に基づいて、リンク増加率 = (c) / ((a) + (b))とアンカーなしリンク増加率 = (d) / ((a) + (b))、およびアンカー正解率 = (a) / ((a) + (b))を算出した。図 3 には、順に言語を追加していった場合のリンク増加率、アンカーなしリンク増加率、およびアンカー正解率を示した。

リンク増加率は、英語単体では 35.0%であり、fr、de、es、it は追加することで 3%以上増加率が上昇した。war、ceb の増加率の上昇はほぼ 0%であった。これらの言語は内容が短い記事が多いため、変換されるリンク数も少ないためである。最終的な増加率は 56.4%であるが、sv 以降は 1%未満の増加率となる。アンカーなしリンク増加率についても同様の傾向が見られ、ru までの増加率は概ね 10%以上だが、sv 以降の上昇率は 5%以下に減少する。このため ru を追加する 10 言語程度まで言語数を拡張することで、新規リンク、アンカーなしリンクともに大半を得ることができる。

中国語は特異であり、いずれのリンク増加率も低下する。これは、他の言語に比べ中国語は日本語と地理・文化的に近いことから類似する内容を持つ記事が多く、新しいリンクが変換されない一方、今まで変換されなかった既存リンクが多く得られた結果だと解釈できる。参考にする言語版の記述が異なるほど効果的に新たなリンクを変換できるので、類似する記事内容が多い言語はリンク追加効率が悪いといえる。

アンカー正解率は、en、fe、zh、it を追加した時点で若干低いが、大きな変動は見られず、いずれの時点においても 95%を上回った。変換元言語数が増えることで発生するコンフリクトは、最も多くの言語から変換されたリンクを採用するという提案方法で解消できるといえる。なお、本来のアンカーが“イタリア共和国”であるのに対して“イタリア”をアンカーとして選択した例なども（b）に分類されるため、アンカー正解率は過小評価されている。

4.2 リンクの順位付け指標の比較評価

3 で提案した順位付け指標を評価するため、改めて日本語版記事 358 記事を選択し、それらに対して人手によるリンクの順位付けを行った。日本語版記事は、言語内リンク数が 30 から 50 で、en、fr、de、es、zh、it、pl、nl、pt、ru、sv、uk、vi、war、cab のうち少なくとも 5 言語の対応する記事を持つことを条件とした。記事は技術史、情報、歴史、地形、国際のサブカテゴリのうち少なくとも一つに属している。3 名の評価者が、358 記事の各々について、特に重要と考えるリンクをその記事の言語内リンク数の 1/4 だけ抽出し、抽出した評価者数をリンクの重要度とした。

提案したリンクの順位付け指標の妥当性評価に NDCG(Normalized discounted cumulative gain) [7]を使用した。これは人手で定めた順位と、提案指標が定めた順位の類似度を 0〜1 の数値で表す。

$$\mathrm{NDCG} = \frac{\mathrm{DCG}(relS_1, relS_2, relS_3, \ldots, relS_n)}{\mathrm{DCG}(relD_1, relD_2, relD_3, \ldots, relD_n)}$$

ここに、$relS_i$はシステムが第 i 位に出力した記事の重要度、$relD_i$は重要度を降順に並べた時の第 i 位の重要度である。

$$\mathrm{DCG}(rel_1, rel_2, rel_3, \ldots, rel_n) = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{log_2(i)}$$

ここに、rel_iは第 i 位の記事に人手で付けられた重要度である。

表 1 に各指標に対する NDCG をまとめた。比較のため、提案指標のほかに単言語の記事類似度とキーフレーズネス[1]を指標とした場合およびランダムな順位付けに対しても NDCG を求め

た。表からわかるように、変換元言語数（TN）と非アンカー数（NAN）を組み合わせた順序付けが最良の結果となった。非アンカー数が効果的に働いていることから、負の情報はリンクの重要度を考慮する上で有効であることが分かる。記事間類似度を多言語に拡張したことで得られた効果は 0.001 ポイントの変化に留まり、多言語化の効果はほとんど見られなかった。

5 関連研究

テキストに出現する用語に Wikipedia 記事へのリンクを付与する wikification を行う手法はこれまでに多く提案されている。アンカーの選定にはキーフレーズネス[1]を利用する方法、リンク先記事の選択にはアンカーからリンクされる確率、周辺文脈の類似度[2]、周辺言語内リンクとの一貫性[3]などが用いられる。本研究のように他言語版のリンクを変換するアプローチは例がない。

6 おわりに

他の言語の Wikipedia の言語内リンクを変換することにより Wikipedia 記事に言語内リンクを付与する方法を複数の変換元言語を用いるように拡張した。その結果、アンカー正解率を低下させることなく、より多くの言語内リンクを得られることを実験により確認した。また、変換されたリンクを順位付けるための指標を提案した。特に、変換元言語数と非アンカー言語数（アンカーとなり得る語を含むがアンカーとしていない変換元記事の数）の組合せが有効であることを実験により確認した。

今後の課題として、引き続きリンクの順位付け指標の評価を行う必要がある。本稿では人手による正解データを用いた評価を行ったが、評価者 3 人が 358 記事を判定したのみで、十分な数が揃っていない。大規模な評価実験を行うための準備を整えたい。

謝辞 本研究は、一部、JSPS 科研費 15K16096 の助成を受けて行った。

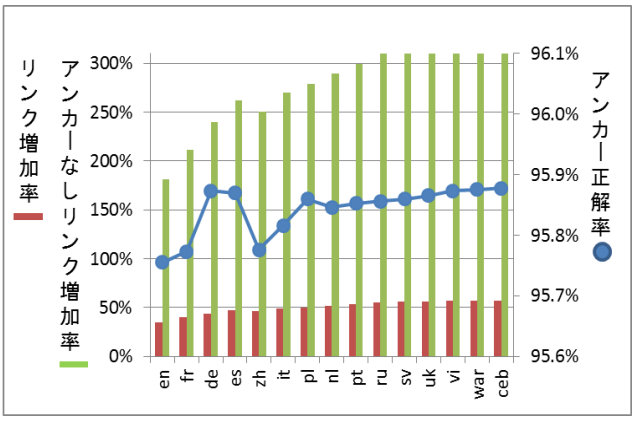

図 3：変換元言語数に対するアンカー正解率（右軸）リンク増加率、アンカーなしリンク増加率（左軸）の推移

表 1：提案指標の NDCG

指標	NDCG
多言語記事類似度(MLS)	0.745
変換元言語数（TN）	0.757
変換元言語数（TN）― 非アンカー数（NAN）	**0.763**
単言語記事類似度	0.744
キーフレーズネス	0.729
ランダム	0.656

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*, 2007, pp. 233–242.

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

[3] L. Ratinov, D. Roth, D. Downey, and M. Anderson, "Local and global algorithms for disambiguation to Wikipedia," *in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 1375–1384.

[4] T. Tsunakawa, M. Araya and H. Kaji, "Enriching Wikipedia's Intra-language Links by their Cross-language Transfer," in *Proceedings of the 25th International Conference on Computational Linguistics (Coling 2014)*, 2014, pp. 1260–1268.

[5] W. A. Gale, K. W. Church, and D. Yarowsky, "One sense per discourse," in *Proceedings of HLT '91 Workshop on Speech and Natural Language*, 1992, pp. 233–237.

[6] D.Milne and Ian H. Witten, "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," in *Proceedings of AAAI* 2008.

[7] K.Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," in *Proceedings of the ACM Transactions on Information Systems Vol. 20, No.4*, 2002 pp. 422–446.