他言語版の内部リンクを利用した Wikipedia 内部リンクの自動付与

綱川 隆司　　新谷 誠　　梶 博行

静岡大学大学院情報学研究科

{tuna, araya, kaji}@inf.shizuoka.ac.jp

1　はじめに

Wikipedia の記事において，他の記事へのハイパーリンクは“内部リンク”と呼ばれ，記事に現れる概念を参照する上で重要な機能を持つ．内部リンクをクリックすることにより，記事中の用語に関する説明記事を参照し，記事の内容を効率よく理解することができる．内部リンクを充実させることは Wikipedia の有用性を高めるうえで重要であり，Wikipedia のガイドラインでも推奨されている[1]．

Wikipedia に新しい記事を追加する場合，同時に内部リンクを付与する必要がある．内部リンクを付与する際には，適切なアンカーを選択するとともに，リンク先の記事を正しく指定しなければならない．アンカーとして選択した語句が複数の意味を持つ場合，記事中で用いられる意味に対応する記事を正しく選んでリンクする必要がある．このため内部リンクの付与はコストのかかる作業となっている．また，既存の記事においても，内部リンクが十分に付与されているとは限らないという問題がある．

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内部リンクを自動的に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異なる言語版の二つの記事の間で，アンカーが互いに対訳であるような内部リンクが含まれるとき，それらの指す記事は同じ事柄に関するものであり，したがって言語間リンクで結ばれているという仮定に基づく．評価実験において，提案方法が既存記事の内部リンクのカバー率向上に効果があることを実証する．

[1] http://ja.wikipedia.org/wiki/Wikipedia:記事同士をつなぐ

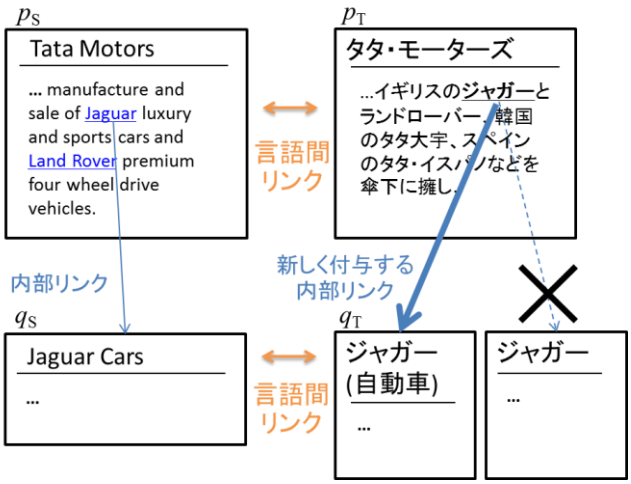

図　1　英語記事の内部リンクに基づく日本語記事への内部リンク付与

2　提案方法

2.1　基本アイデア

Wikipedia の各記事は，Wikidata を介してその記事と同じ事柄を説明する他の言語版の記事と関係付けられている．これを記事間の言語間リンクと呼ぶ．言語間リンクで結ばれた各言語版の記事はそれぞれ独立に作成されるため，その内容は対訳関係になっているとは限らない．しかし，言語間リンクで結ばれた複数言語の記事を通して，互いに訳語となっている語句の組は一つの意味で用いられると考えられる．これは，「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 [1] の仮説において，複数言語の記事全体を一つの談話とみなすことで仮定できる．

この仮定に基づき，他の言語版の記事に存在する内部リンクのリンク先記事から，新しい内部リンクの正しいリンク先が推定できる．例えば図 1 において，日本語記事 p_T 中の“ジャガー”をアンカーとする場合，リンク先記事の候補として“ジャガー (自動車)”のほか，動物の意味の“ジャガー”や“ジャガー・レーシング”等が

ある．そこで記事 p_T と言語間リンクで結ばれた英語記事 p_S に含まれる内部リンクを調べ，“ジャガー”の訳語“Jaguar”がアンカーで，そのリンク先記事が“Jaguar Cars”になっている内部リンクがあるとする．このとき，上述の仮定から記事 p_S, p_T の双方で“ジャガー”と“Jaguar”が意味する概念は同じであり，“ジャガー”のリンク先記事として“Jaguar Cars”と言語間リンクで結ばれた記事“ジャガー (自動車)”を選択することができる．

2.2 提案方法

本稿では，内部リンク l を以下のように定義する：

$$l = (a, q)$$

ただし，a はアンカー（内部リンクを付与する用語），q はリンク先記事とする．また，Wikipedia 記事 p に含まれるすべての内部リンクの集合を L(p) とする[2]．さらに，言語 T の Wikipedia 記事全体の集合を W(T) とし，現れるすべての内部リンクの集合を $L(T) = \cup_{p \in W(T)} L(p)$ とする．

言語 S の記事 p_S 中の内部リンクに基づいて，p_S と言語間リンクで結ばれた言語 T の記事 p_T に新しい内部リンクを付与する手順は以下の通りである：

記事 p_S に含まれる各内部リンク $l_S = (a_S, q_S) \in L(p_S)$ について以下の処理を行う．

(1) リンク先記事 q_S と言語間リンクで結ばれた言語 T の記事 q_T を求める．

(2) L(T) から，記事 q_T をリンク先とする内部リンクのアンカー候補集合 $A(q_T)$ を求める．すなわち：

$$A(q_T) = \{a' \mid \exists l' \in L(T).\ l' = (a', q_T)\} \cup \{\mathrm{Title}(q_T), \mathrm{RawTitle}(q_T)\}$$

ただし，$\mathrm{Title}(q_T)$ は記事 q_T のタイトル，$\mathrm{RawTitle}(q_T)$ は $\mathrm{Title}(q_T)$ のうち，末尾に分野を示す“ (…)”がある場合にその部分を取り除いた文字列とする．

(3) 記事 p_T のテキストに $A(q_T)$ の要素 a_T が出現する場合，p_T に新しい内部リンク $l_T = (a_T, q_T)$ を付与する[3]．

図 1 の例では，英語記事 p_S “Tata Motors”に存在する内部リンク (Jaguar, “Jaguar Cars”) について，リンク先記事“Jaguar Cars”と言語間リンクで結ばれた日本語記事“ジャガー (自動車)”のアンカー候補集合は {ジャガー (自動車), ジャガー, …} となる．日本語記事 p_T “タタ自動車”に文字列“ジャガー”が現れるため，これをアンカーとし，新しい内部リンク (ジャガー, “ジャガー (自動車)”) が付与される．

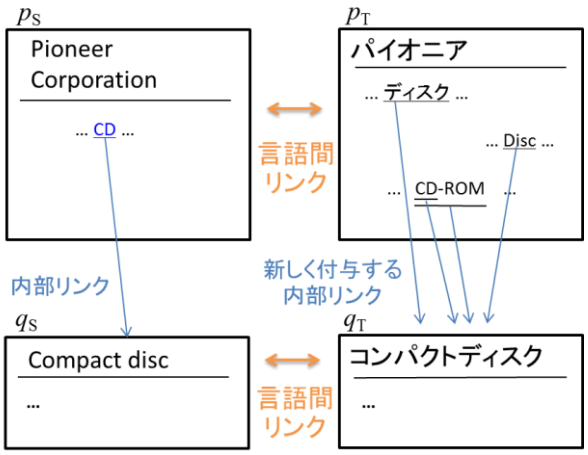

図 2 一つのリンク先記事に複数のアンカーが対応する例

2.3 アンカー選択方法の代替案

提案方法のステップ(3)において，一つのリンク先記事 q_T に対して複数のアンカー候補が記事 p_T に現れる場合がある．例えば，図 2 において，リンク先記事“コンパクトディスク”に対するアンカー候補集合 {CD, CD-ROM, ディスク, Disc, コンパクトディスク, …} のうち複数が本文中に出現している．実際の Wikipedia 記事では，一つのリンク先記事に対し，対応するすべての語句に内部リンクを付与するのではなく，代表的な語句のみに付与されることが多い．そこで，最も適した語句のみに内部リンクを付与するため，アンカー候補に優先順位をつけるための代替案を提案する．優先順位をつける基準として以下の 4 つを実験的に比較することとする．

(A) 出現順序

一般に，ある事柄について最初に出現した語句をアンカーとすることが多い．本文中で出現順序が早い順に優先順位をつける．

(B) リンク先記事タイトルとの文字列類似度

リンク先記事のタイトルはその事柄を表す代表的な表記であるため，アンカーとして選ばれやすい．アンカー候補 a_T とリンク先記事タイトル $t = \mathrm{RawTitle}(q_T)$ の編集距離を $d(a_T, t)$ とするとき，文字列類似度

$$\mathrm{sim}(a_T, t) = 1 - \frac{d(a_T, t)}{\max(a_T\text{の文字数},\ t\text{の文字数})}$$

で得られる値が大きい順に優先順位をつける．

[2] 本稿では、内部リンクの記事中の出現位置は無視する．実際の記事中には、一つの記事中にアンカー、リンク先記事がともに等しい内部リンクが複数回現れることがある．

[3] 同じアンカーが複数回現れる場合、本文中の最初の出現箇所に内部リンクを付与する．

(C) アンカー候補に対するリンク先記事の相対頻度 $p(q_T|a_T)$

Wikipedia 記事全体に出現したアンカーが a_T の内部リンクのうち，リンク先記事が q_T であったものの割合，すなわち

$$p(q_T|a_T) = \frac{\mathrm{count}((a_T,\ q_T))}{\sum_{q_T'} \mathrm{count}((a_T,\ q_T'))}$$

が大きい順に優先順位をつける．ただし，count(l) は内部リンク l が Wikipedia 記事全体で現れた回数．

(D) リンク先記事に対するアンカー候補の相対頻度 $p(a_T|q_T)$

Wikipedia 記事全体に出現したリンク先記事が q_T の内部リンクのうち，アンカーが a_T であったものの割合．すなわち

$$p(a_T|q_T) = \frac{\mathrm{count}((a_T,\ q_T))}{\sum_{a_T'} \mathrm{count}((a_T',\ q_T))}$$

が大きい順に優先順位をつける．

3 評価実験

3.1 使用データ

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）のダンプデータを用いた．言語間リンクで結ばれた英語記事と日本語記事の組は 366,358 対存在する[4]．このうち評価データとして 3,655 対を無作為に選び，残りを内部リンク集合 L(T) を得るための訓練データとして用いた．

3.2 アンカー選択方法の評価

評価用データの日本語記事 3,655 件に対して，提案方法により内部リンクを付与した．得られた内部リンクのうち，Wikipedia 記事に既に存在するものが 33,005 件あった．2.3 節で述べたアンカー選択方法の各基準を比較するため，それぞれの選択方法を適用した場合に優先順位 1 位の内部リンクが既存リンクになっているものの数を表 1 に示した．結果，基準(D)による選択方法が最も適しており，95.5% にあたる 31,525 件が既存のリンクと一致した．以下，基準(D)を用い，同じ記事中では一つのリンク先記事に対し優先順位 1 位のアンカーの内部リンクのみを付与する．

3.3 内部リンクの増加率の評価

評価用データの日本語記事に対して提案手法を適用した結果，既存のリンク先記事と一致する 33,005 件の内部リンクに加え，新たに 13,953 件の新しい内部リンクが得られ，内部リンク数は 42.3% 増加した．表 2 に日本語記事“パイオニア”に対して提案方法により得られた新しい内部リンクのアンカーとリンク先記事の例を示す．

新たに得られた内部リンクについて，選択されたアンカーの適切さを既存の内部リンクから直接評価することはできないが，3.2 節の結果から，新しい内部リンクについても高い割合で適切なアンカーが選択されていることが期待できる．また，リンク先記事の適切さについては，従来研究において対訳のアンカーからリンクされる記事について 92.0% の割合で言語間リンクが存在することが示されており[2]，提案方法で得られた新しい内部リンクについても高い割合で対象記事に対し適切なリンク先記事が指定されていることが予想される．

表 1 アンカー選択方法の比較評価の結果

優先順位基準	既存リンクとの一致数	既存リンクとの一致率
(A)	26,684	80.8%
(B)	31,267	94.7%
(C)	29,473	89.3%
(D)	**31,525**	**95.5%**

表 2 日本語記事“パイオニア”に対して提案方法によって新しく付与した内部リンク（抜粋）

アンカー	リンク先記事
CD	コンパクトディスク
欧州	ヨーロッパ
HD	高精細度テレビジョン放送
株式会社	株式会社（日本）
カラオケ	カラオケ
多国籍企業	多国籍企業
テレビ	テレビ
東京都	東京都
東証 1 部	東京証券取引所

4 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タスクはエンティティリンキングあるいは Wikification [3] と呼ばれ近年さかんに研究されており，これを Wikipedia 記事に適用することで内部リンクの自動付与が可能である．しかし，テキスト中のどの語句をアンカーとするか，またアンカーからどの記事にリンク

[4] 記事集合から年号に関する記事（“2011 年”等）を除いた．また、アンカーが年号の内部リンク、および表中に現れる内部リンクについては実験の対象外とした．

するかという二つの課題が十分解決されているとはいえない. 従来の方法では Wikipedia 全体でアンカーになったことのある全ての用語について内部リンクを付与するかかどうか検討しなければならないのに対し，提案方法では他言語版で内部リンクとして扱われたもののみを考慮するため，より適切なアンカーが選ばれると考えられる．また，従来方法ではリンク先記事は周辺に現れる語を主な手掛かりとして決定されるが，提案方法では他言語版で既にリンクされた記事という強力な手掛かりを用いることができる．

Adafre and Rijke [4] 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，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用いる方法が開発されている[5]–[7]．これらの課題はいずれも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て Wikipedia 自身の品質を向上する研究も進められている．Sorg and Cimiano [8] は, Wikipedia の言語間リンクを新たに発見するため，記事に含まれる内部リンクのリンク先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素性の一つとして用いた．Wang ら [9] はさら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素性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．

5 おわりに

本稿では，Wikipedia 記事に対して他言語版の記事を利用することで内部リンクを自動的に付与する方法を提案した．他言語版記事の内部リンクを，言語間リンクを介して対象言語の記事に変換することで，既存の内部リンクに加えて新しい内部リンクが得られ，既存記事における内部リンクのカバー率が向上することを確認した．また，内部リンクのアンカー候補が複数存在する場合，リンク先記事に対するアンカーの相対頻度に基づく優先順位付けによって，非常に高い割合で適切なアンカーを選択できた．

今後の課題として以下の点が挙げられる。

- 提案方法によって得られた新しい内部リンクの妥当性評価
- アンカー選択方法の基準の組合せによる既存リンクとの一致率の改善
- 本方法を三言語以上の組合せに拡張することによる性能向上

また，既存の不適切な内部リンクの検出を試みる．二言語間で既存の内部リンクを提案方法と同様の方法で比較することにより，一方にしか存在しない内部リンクを誤り候補として検出することができる．

参考文献

[1] 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.

[2] 綱川隆司, 梶博行, “Wikipedia 内部リンクの言語間変換,” 情報処理学会第 214 回自然言語処理研究会, 2013, No.9, pp. 1–6.

[3] 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.

[4] S. F. Adafre and M. de Rijke, “Discovering missing links in Wikipedia,” in *Proceedings of the 3rd International Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)*, 2005, pp. 90–97.

[5] I. H. W. David Milne, “An effective, low-cost measure of semantic relatedness obtained from Wikipedia links,” in *Proceedings of the Wikipedia and AI Workshop of AAAI*, 2008, pp. 25–30.

[6] A. Fogarolli, “Word sense disambiguation based on Wikipedia link structure,” in *Proceedings of 2009 IEEE International Conference on Semantic Computing*, 2009, pp. 77–82.

[7] 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.

[8] P. Sorg and P. Cimiano, “Enriching the crosslingual link structure of Wikipedia - a classification-based approach,” in *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*, 2008.

[9] Z. Wang, J. Li, and J. Tang, “Boosting cross-lingual knowledge linking via concept annotation,” in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 2733–2739.