Enriching Wikipedia's Intra-language Links by their Cross-language Transfer

Takashi Tsunakawa **Makoto Araya** **Hiroyuki Kaji**
Graduate School of Informatics, Shizuoka University
3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka 432-8011, Japan

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

Abstract

Although hyperlinks enhance the utility of Wikipedia, embedding them in articles imposes a burden on contributors. To alleviate this burden as well as enrich hyperlinks in Wikipedia articles, we propose a method for transferring intra-language links between different-language articles linked via an interlanguage link. The method avoids anchor selection and disambiguation problems by which usual wikification methods are affected, by exploiting the analogy between different language editions of Wikipedia. The effectiveness of the method was demonstrated through an experiment of transferring intra-language links from English to Japanese. It increased the number of intra-language links in Japanese articles by 40.9%, and the accuracy of anchors selected was estimated to be 96.3%.

1 Introduction

Wikipedia is a Web-based encyclopedia constructed collaboratively by many contributors and continues to enlarge and improve daily. Because of its overwhelming scale, improved quality, and multilingual nature, it has acquired a huge number of readers worldwide. One of the distinguishing features of Wikipedia is that it is a hypertext, which greatly enhances its usefulness and usability. That is, an article is linked to its related articles in the same language via intra-language links as well as to its counterpart articles in different languages via inter-language links (ILLs), and readers can navigate within millions of articles.

Editing Wikipedia articles naturally includes linking them to their related articles, which imposes an additional burden on contributors. As a result, Wikipedia articles may remain incomplete; they sometimes lack important links as well as contain incorrect links. Thus, it is desirable to automate link-related editorial tasks such as embedding links in new articles and verifying links in existing articles. Linking a plain text, usually non-Wikipedia articles, to Wikipedia articles is called wikification, and much effort has been devoted to developing a variety of wikification methods over the past decade (Mihalcea and Csomai, 2007; Milne and Witten, 2008a; Fogarolli, 2009; Ratinov et al., 2011). However, wikification methods are still immature and affected by two hard problems; anchor selection, which involves keyword extraction or term recognition, and destination-article determination, which is a kind of word sense disambiguation (WSD).

We focused on the comparability of intra-language links between different language editions of Wikipedia, and developed a method for transferring intra-language links in one language edition to another language edition. Although the method is not applicable to texts other than Wikipedia articles, it avoids the problems of anchor selection and destination-article disambiguation by using analogy with different language editions. It does not require any language resources other than Wikipedia itself. When the target language is a morphologically rich one, a morphological analyzer is also required. Although the method is applicable to any language pairs, we evaluated its effectiveness through an experiment of transferring intra-language links from English to Japanese.

2 Basic Idea

In Wikipedia, an article in one language is often linked to another article in another language via an ILL. These two articles, which describe the same entity, concept or topic, are comparable. Note that this comparability holds not only for texts in articles but also for intra-language links, each of which links an anchor or an important term within an article to another same-language article describing the entity, concept, or topic denoted by the anchor term. Figure 1 gives an example pair of ILL-linked articles; an English article “Tata Motors” and a Japanese article “タタ・モーターズ.” The former has an intra-language link from an anchor “Jaguar” to the English article “Jaguar Cars,” while the latter has an intra-language link from an anchor “ジャガー” to the Japanese article “ジャガー (自動車).” These two intra-language links are comparable: namely, the anchors are translations of one another and the destination articles are linked via an ILL.

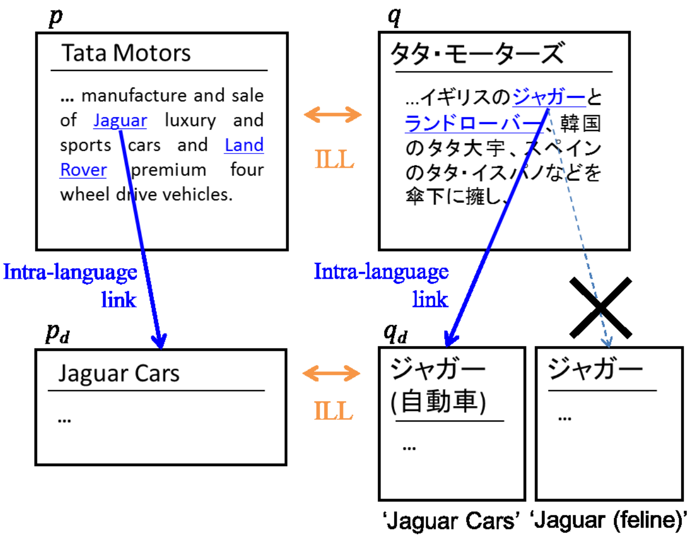

p
Tata Motors
... manufacture and sale of Jaguar luxury and sports cars and Land Rover premium four wheel drive vehicles.
q
タタ・モーターズ
...イギリスのジャガーとランドローバー、韓国のタタ大宇、スペインのタタ・イスパノなどを傘下に擁し、
ILL
Intra-language link
Intra-language link
p_d
Jaguar Cars
...
ILL
q_d
ジャガー (自動車)
...
ジャガー
...
‘Jaguar Cars’
‘Jaguar (feline)’

Figure 1. Transferring intra-language link.

The above fact inspired us to develop a method for transferring intra-language links between ILL-linked articles to enrich the intra-language links in each article. Suppose an extreme case in which an article q in one language, which is linked to its counterpart p in another language via an ILL, has no intra-language links. An intra-language link can be transferred from p to q as follows. First, following an intra-language link (p to p_d) and then the ILL (p_d to q_d), the final destination article q_d is identified as that to be linked from q. Second, the text of q is searched for possible anchors for the destination article q_d, which are learned from the entire Wikipedia beforehand. If two or more possible anchors are found, the most appropriate one will be selected according to a certain criterion. For example, suppose all intra-language links are missing from the Japanese article “タタ・モーターズ” in Figure 1. The intra-language link from the English article “Tata Motors” to “Jaguar Cars” and the ILL from “Jaguar Cars” to “ジャガー (自動車)” suggest that the Japanese article “タタ・モーターズ” should have an intra-language link to “ジャガー (自動車).” The possible anchors for “ジャガー (自動車),” which have been learned from all the Wikipedia articles linked to it, include “ジャガー (自動車),” “ジャガー,” and others. Since the text of “タタ・モーターズ” contains “ジャガー,” it is selected as the anchor for the destination article “ジャガー (自動車).”

It should be noted that our proposed method avoids the two hard problems in wikification, anchor selection and disambiguation, by exploiting the intra-language links provided by Wikipedia in another language. Resulting anchors are certainly important terms within q, since their counterparts have been selected as anchors by the author of counterpart p in another language. Even if an anchor were an ambiguous term, i.e., had two or more possible destination articles, it would be certainly linked to the appropriate one due to the “one sense per discourse” hypothesis (Gale et al., 1992). The hypothesis is extended to a pair of ILL-linked articles, p and q, as follows. A pair of corresponding anchors should be regarded as a single term and express the same sense in a discourse shared by p and q. In other words, they should be linked to articles that are linked via an ILL. Since the proposed method relies on this extended hypothesis, it will select correct destination articles for anchors in q as long as anchors in p have been linked to their correct destination articles.

It should also be noted that the proposed method first determines the destination articles then the anchors for them, while usual wikification methods first select anchors then determine their destination articles. The main reason for this is convenience of implementation; cross-language mapping of destination articles is one-to-one (or one-to-zero), while that of anchors can be one-to-many. Determining destination articles prior to anchors, however, results in an additional advantage that allows a destination article to be proposed without an anchor for it. Since the pair p and q is not parallel but just comparable, the counterpart of an anchor in p is not always found in q. This is often the case when q is incomplete, under construction, or written in a different style from that of p. In such a case, our method proposes a destination article q_d without an anchor, and q will be linked to q_d once q is enlarged to contain a term appropriate as the anchor for q_d.

3 Proposed Method

The proposed method is divided into two steps; the preprocessing step for collecting possible anchors for all Wikipedia articles in a target language as well as estimating probabilities required in the succeeding step and the main step for transferring intra-language links in a source-language article p to the target-language article q linked to p via an ILL. In this section, a triplet (p, a, p_d) denotes an intra-language link from anchor a in article p to destination article p_d and, likewise, a triplet (q, b, q_d) does. Note that although an article can have two or more intra-language links from the same anchor at different positions in the text to the same destination article, they are treated as a single link.

3.1 Preprocessing Step

Collecting Possible Anchors for Wikipedia Articles

The title of a Wikipedia article can be used as an anchor for the article. However, a title is often accompanied by a parenthesized note indicating the domain of the article to discriminate from other articles with the same title. The title "ジャガー (自動車)" of an article that describes a car named Jaguar is an example; the parenthesized note "(自動車)" discriminates the article from another article "ジャガー", which describes an animal belonging to the cat family. Such a title accompanied by a parenthesized note rarely occurs in usual texts, and the title with the parenthesized note deleted is often marked as an anchor. Accordingly, we also regard a title with a parenthesized note deleted (e.g., "ジャガー") as a possible anchor. Other terms, typically synonyms of the article title, are often used as anchors. Therefore, we collect terms that are actually used as anchors for each article from the entire Wikipedia. Finally, we threshold possible anchors by their keyphraseness to eliminate general words. The keyphraseness $\kappa(b)$ of a term b is defined as the probability that b is used as an anchor in Wikipedia articles (Mihalcea and Csomai, 2007), i.e.,

$$\kappa(b) = \frac{|\{q \mid \exists q_d. (q, b, q_d) \in L_t\}|}{\mathrm{df}(b)},$$

where L_t is a set consisting of all intra-language links in the target-language Wikipedia and $\mathrm{df}(b)$ is the number of Wikipedia articles in which b occurs.

In summary, a set of possible anchors $\mathrm{A}(q_d)$ are constructed for a target-language destination article q_d as follows:

$$\mathrm{A}(q_d) = (\{title(q_d), title'(q_d)\} \cup \{b \mid \exists q. (q, b, q_d) \in L_t\}) \cap \{b \mid \kappa(b) \geq \theta\},$$

where $title(q_d)$ and $title'(q_d)$ are q_d's title with and without the parenthesized note, respectively, and θ is the threshold for the keyphraseness.

Estimating Probabilities

The following probabilities, which will be used to select one from among possible anchors for a destination article, are estimated from the entire Wikipedia.

- The probability that the target-language anchor is b on the condition that its source-language counterpart is a, i.e.,

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

$$\mathrm{count}(a,b) = \left|\left\{((p,a),(q,b)) \,\middle|\, \begin{array}{r} \exists p_d.\exists q_d.(p,a,p_d) \in L_s \land (q,b,q_d) \in L_t \\ \land\, (p,q) \in ILL \land (p_d,q_d) \in ILL \end{array}\right\}\right|,$$

where L_s is a set consisting of all intra-language links in the source-language Wikipedia, and ILL is a set of all pairs of ILL-linked articles.

- The probability that the anchor is b on condition that the destination article is q_d, i.e.,

$$\mathrm{P}(b|q_d) = \frac{|\{q|(q,b,q_d) \in L_t\}|}{|\{q|\exists b'.(q,b',q_d) \in L_t\}|}$$

- The probability that the destination article is q_d on condition that the anchor is b, i.e.,

$$\mathrm{P}(q_d|b) = \frac{|\{q|\ (q,b,q_d) \in L_t\}|}{|\{q|\exists q_d'.(q,b,q_d') \in L_t\}|}$$

3.2 Main Step

Let p and q be source-language and target-language articles that are linked via an ILL, respectively. Intra-language links in p are transferred to q as follows:

(i) For each source-language intra-language link (p,a,p_d), do (ii) to (v).

(ii) If p_d has an ILL to an article in the target language, let q_d be the destination article of the ILL from p_d. Otherwise, output "NOT TRANSFERRED" and move to the next intra-language link.

(iii) If $\mathrm{A}(q_d)$ is empty, output the transferred intra-language link (q, NULL, q_d), which means that q should be linked to q_d but does not contain a term appropriate as the anchor, and move to the next intra-language link.

(iv) For each possible anchor $b \in \mathrm{A}(q_d)$, search the text of q for b. If found, let $\mathrm{pos}(b,q)$ denote the position of its first occurrence in the text; otherwise, let $\mathrm{pos}(b,q) = -1$.

(v) If at least one possible anchor is found, choose the most appropriate one $\hat{b}$ according to an anchor priority score $\mathrm{Score}(b)$, i.e.,

$$\hat{b} = \underset{b \text{ s.t. } b \in \mathrm{A}(q_d) \land \mathrm{pos}(b,q) \geq 0}{\mathrm{argmax}} \mathrm{Score}(b).$$

and output the transferred intra-language link $(q,\hat{b},q_d)$. Otherwise, output the transferred intra-language link (q, NULL, q_d).

We have the following five alternative anchor priority scores in step (v) above.

- Anchor translation probability: $\mathrm{Score}_1(b) = \mathrm{P}(b|a)$.
  This score favors the anchor that occurs most frequently as counterpart to the source-language anchor.

- Anchor probability: $\mathrm{Score}_2(b) = \mathrm{P}(b|q_d)$.
  This score favors the anchor by which the destination article is pointed most frequently.

- Destination article likelihood: $\mathrm{Score}_3(b) = \mathrm{P}(q_d|b)$.
  This score favors the anchor that is most likely to point the destination article.

- Spelling: $\mathrm{Score}_4(b) = 1 - \mathrm{dist}(b, title'(q_d))/\max\{\mathrm{len}(b), \mathrm{len}(title'(q_d))\}$,
  where $\mathrm{dist}(s,s')$ is the Levenshtein distance between character strings s and s' (Levenshtein, 1966), and $\mathrm{len}(s)$ is the length of character string s.

  This score favors the anchor with the highest similarity to the article's title without a parenthesized note, which is the most representative term denoting the entity, concept, or topic described in the article.

- Position: $\mathrm{Score}_5(b) = 1/\mathrm{pos}(b,q)$.
  Note that in a Wikipedia article, among two or more occurrences of an important term, the first one tends to be marked as an anchor.

4 Experiment

4.1 Experimental Settings

We conducted an experiment on transferring intra-language links from the English edition to the Japanese edition of Wikipedia.

Input Data

The English edition of Wikipedia (2013-04-03 dump), consisting of 4,241,324 articles, and the Japanese edition of Wikipedia (2013-03-28 dump), consisting of 951,411 articles[1], were used for the experiment. Intra-language links were extracted from each dump file, and ILLs were obtained from Wikidata (2013-03-28 dump). Redirect pages were resolved preliminarily, i.e., if the destination of an intra-language link or ILL was a redirect page, the destination was replaced with an article pointed by the redirect page.

Anchor	English translation	Keyphrase-ness
ベイジアン・ネットワーク	Bayesian network	1
地獄の辞典	Dictionnaire Infernal	0.810
悪魔学	demonology	0.678
パリ	Paris	0.574
オカルト	occult	0.304
悪魔	devil	0.135
ニコラス	Nicholas	0.039
対立	conflict	0.001
半分	half	7.8×10^{-5}
より	Yori (kana)	4.4×10^{-6}

Table 1. Example of keyphraseness values.

From among a total of 366,358 pairs of English and Japanese articles linked by ILLs, 3,595 pairs were randomly selected as a test set. The remaining pairs were used as training data for constructing English and Japanese intra-language link sets, L_s and L_t. The English articles in the test set contained 179,963 intra-language links in total; these were input to the algorithm of the proposed method.

Keyphraseness Threshold

Limiting possible anchors to meaningful ones and gaining many links are in a trade-off relation adjustable by the keyphraseness threshold θ. In the experiment, θ was set to 0, 0.01, 0.05, and 0.1.

Keyphraseness values of several anchors are listed in Table 1. Technical words (e.g., "ベイジアン・ネットワーク" – Bayesian network) and uncommon proper names (e.g., "地獄の辞典" – Dictionnaire Infernal) tend to have high keyphraseness values. Common words (e.g., "悪魔" – devil and "対立" – conflict) and proper names (e.g., "パリ" – Paris and "ニコラス" – Nicholas), especially identical to a general noun, have middle or low values according to their commonness. Although some functional words (e.g., "より" – from) may be included in possible anchors for the Wikipedia articles of their homographic content words (e.g., "より" – Yori (kana)), they naturally have extremely low values. By setting θ to a value slightly greater than zero, functional words could be removed from possible anchors.

Comparison of Anchor Priority Scores

To determine the most effective anchor priority score, the accuracy of anchors selected according to each score was evaluated, assuming the existing intra-language links in the original Japanese articles as gold standard. That is, anchor accuracy Acc is defined as the percentage of originally pointed destination articles for which correct anchors were selected, i.e.,

$$\mathrm{Acc} = \frac{|Result \cap GoldSTD|}{|\{(q, b, q_d) \in Result \mid \exists b'. (q, b', q_d) \in GoldSTD\}|},$$

where $Result$ is a set consisting of all transferred intra-language links and $GoldSTD$ is the gold standard intra-language link set. Table 2 lists the anchor accuracies for each anchor priority score and each θ. Anchor translation probability exhibited the best results and, therefore, we adopted anchor translation probability as the anchor priority score.

[1] Redirect pages and articles with no intra-language links were not included in these counts.

Anchor priority score	**Anchor accuracy (%)**			
	$\theta = 0$	$\theta = 0.01$	$\theta = 0.05$	$\theta = 0.1$
Anchor translation probability	**96.3**	**93.9**	**93.0**	**92.0**
Anchor probability	95.6	93.3	92.4	91.4
Destination article likelihood	90.7	90.8	91.5	91.3
Spelling	95.1	93.1	92.5	91.8
Position	88.2	87.3	87.9	87.6

Table 2. Anchor accuracy.

4.2 Experimental Results

We inputted 179,963 English intra-language links to the algorithm of the proposed method and classified them into the following five classes. Examples of each class, except class B, are given in Figure 2, which is an excerpt from the results for the pair of English article "Jacques Collin de Plancy" and Japanese article "コラン・ド・プランシー."

A. Transferred to a Japanese intra-language link in the gold standard (bold underline in Figure 2)

B. Transferred to a Japanese intra-language link whose anchor is not the same as the gold standard link to the same destination article

C. Transferred to a Japanese intra-language link not in the gold standard (double underline in Figure 2)

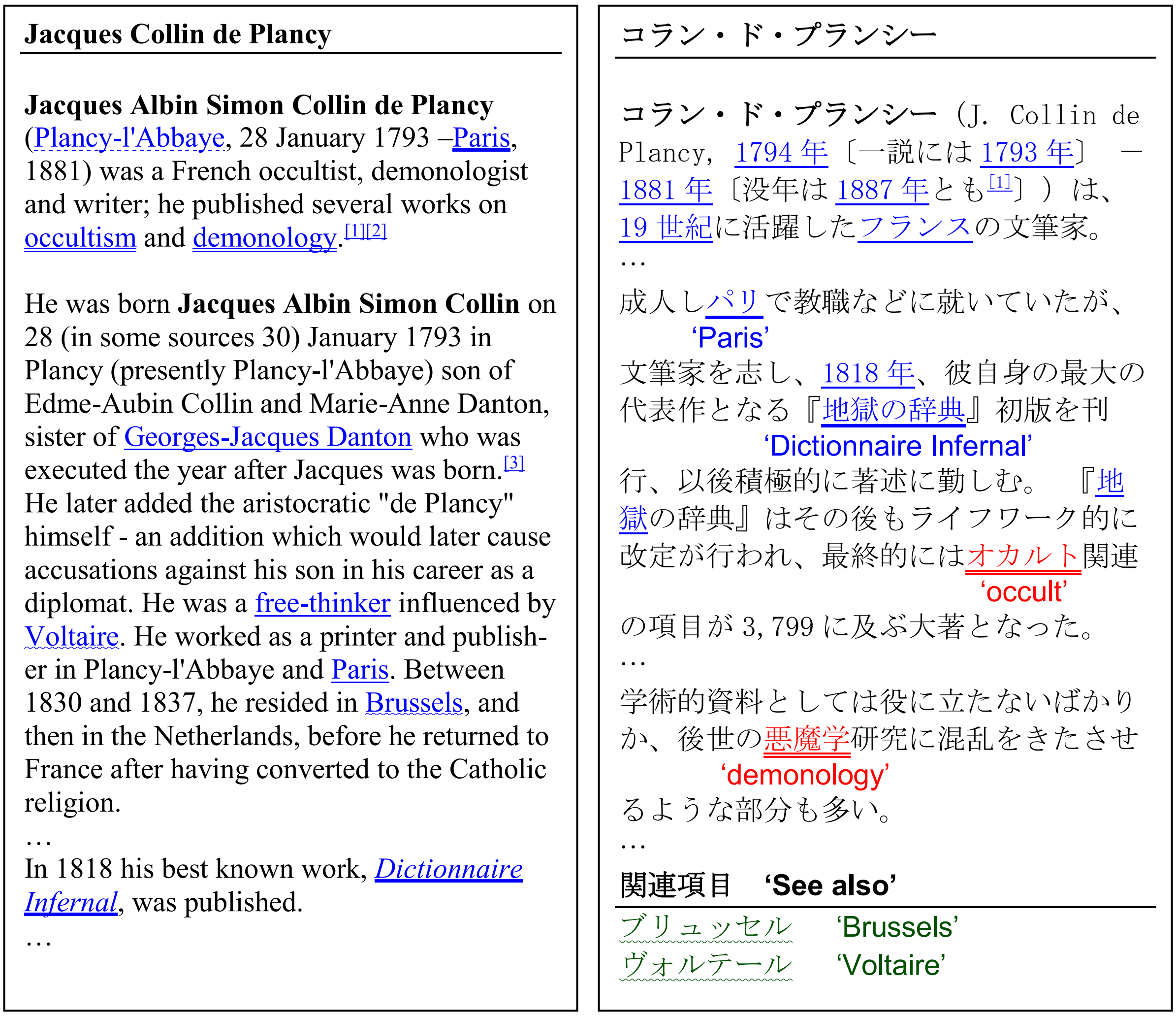
D. Transferred to a Japanese intra-language link without anchor (wavy underline in Figure 2)

E. Not transferred to a Japanese intra-language link (dashed underline in Figure 2)

Table 3 lists the numbers of English intra-language links per class. The proposed method added many new intra-language links to Wikipedia articles. Since the total number of existing Japanese intra-language links in the test-set articles was $T = 161{,}940$, the increase rate of Japanese intra-language links was $100(C + D)/T = 100(13{,}916 + 52{,}275)/161{,}940 = 40.9\%$ ($\theta = 0$). When new links without anchors were excluded, the increase rate was $100C/T = 100 \cdot 13{,}916/161{,}940 = 8.6\%$ ($\theta = 0$).

The anchor accuracy of existing links was $100A/(A + B) = 100 \cdot 31{,}770/(31{,}770 + 1{,}219) = 96.3\%$ ($\theta = 0$). Anchor accuracy of new intra-language links could not be calculated because of the unavailability of gold standard data. However, the proposed method specifies the anchor b for destination article q_d only when possible anchors for it is found in the target-language article q. The specified anchor b is likely to be the counterpart of source-language anchor a pointing to p_d that is the source-language counterpart of q_d, regardless of whether b already points to q_d or not. Thus, the anchor accuracy of new links should be similar to that of existing links.

Among the $S = 179{,}963$ input English intra-language links, $100D/S = 100 \cdot 52{,}275/179{,}963 = 29.0\%$ ($\theta = 0$) were transferred to Japanese intra-language links with the anchor unspecified. This was because different language articles contain different contents even though they are linked via an ILL. The anchor-unspecified links are put in the "関連項目" sections ("See also" sections) of target-language articles, and Wikipedia authors are expected to enlarge or revise the articles so that these anchor-unspecified links can be converted to anchor-specified links. Additionally, among the $S = 179{,}963$ input English intra-language links, $100E/S = 100 \cdot 80{,}783/179{,}963 = 44.9\%$ were not transferred to Japanese intra-language links. We assumed this was mainly due to missing Japanese articles. Note that the total number of Japanese articles is less than one-fourth that of English articles. The percentage of not-transferred links will decrease with the growing number of Japanese articles.



Jacques Collin de Plancy

Jacques Albin Simon Collin de Plancy (Plancy-l'Abbaye, 28 January 1793 –Paris, 1881) was a French occultist, demonologist and writer; he published several works on occultism and demonology.[1][2]

He was born **Jacques Albin Simon Collin** on 28 (in some sources 30) January 1793 in Plancy (presently Plancy-l'Abbaye) son of Edme-Aubin Collin and Marie-Anne Danton, sister of Georges-Jacques Danton who was executed the year after Jacques was born.[3] He later added the aristocratic "de Plancy" himself - an addition which would later cause accusations against his son in his career as a diplomat. He was a free-thinker influenced by Voltaire. He worked as a printer and publisher in Plancy-l'Abbaye and Paris. Between 1830 and 1837, he resided in Brussels, and then in the Netherlands, before he returned to France after having converted to the Catholic religion.

…

In 1818 his best known work, *Dictionnaire Infernal*, was published.

…

コラン・ド・プランシー

コラン・ド・プランシー（J. Collin de Plancy, 1794年〔一説には1793年〕 – 1881年〔没年は1887年とも[1]〕）は、19世紀に活躍したフランスの文筆家。

…

成人しパリで教職などに就いていたが、

‘Paris’

文筆家を志し、1818年、彼自身の最大の代表作となる『地獄の辞典』初版を刊

‘Dictionnaire Infernal’

行、以後積極的に著述に勤しむ。『地獄の辞典』はその後もライフワーク的に改定が行われ、最終的にはオカルト関連

‘occult’

の項目が3, 799に及ぶ大著となった。

…

学術的資料としては役に立たないばかりか、後世の悪魔学研究に混乱をきたさせ

‘demonology’

るような部分も多い。

…

関連項目 ‘See also’

ブリュッセル ‘Brussels’

ヴォルテール ‘Voltaire’

Figure 2. Example results of transferring intra-language links.

	Transfer result		**Number (percentage)**			
	Destination	**Anchor**	$\theta = 0$	$\theta = 0.01$	$\theta = 0.05$	$\theta = 0.1$
A	Existing	Correct	31,770 (17.7%)	30,951 (17.2%)	30,661 (17.0%)	30,298 (16.8%)
B		Incorrect	1,219 (0.7%)	2,025 (1.1%)	2,298 (1.3%)	2,625 (1.5%)
C	New	Found	13,916 (7.7%)	12,812 (7.1%)	11,421 (6.3%)	10,335 (5.7%)
D		Not found	52,275 (29.0%)	53,392 (29.7%)	54,800 (30.5%)	55,922 (31.1%)
E	Not transferred		80,783 (44.9%)			

Table 3. English intra-language links classified according to results.

4.3 Additional Comments on Experimental Results

Among alternative anchor priority scores, anchor translation probability seems most effective because this is a posterior probability of the target-language counterpart to the source-language anchor. Anchor probability is also useful because this is a posterior probability of the anchor for the destination. Higher accuracy with spelling score indicates that Wikipedia editors tend to use the title of the destination as an anchor. This may be caused by manually specifying the anchor and destination independently. Contrary to expectations that the first occurrence likely becomes an anchor, position score exhibited the worst results. More detailed analysis of the context in which a term tends to be selected as an anchor is necessary.

Table 2 shows that the anchor probability, unexpectedly, decreases with a rise of the keyphraseness threshold. It was caused by articles that have only one possible anchor with keyphraseness value be-low the threshold (e.g., "駅" – station). When the threshold was set high, the possible anchor set for such an article became empty and, as a results, the algorithm failed to reproduce the existing links to it.

In this experiment, we transferred English links onto Japanese articles. Since the English edition of Wikipedia is richer than Japanese, it has been assumed that an English-to-Japanese direction is more effective than the inverse. However, among the 179,963 links in English and 161,940 links in Japanese extracted from the test set of English-Japanese article pairs, only 32,989 links are paired with their counterparts and others do not have counterparts. This fact indicates that a Japanese-to-English transfer of links is also useful for enriching English articles. It also leads a low anchor recall, which is the percentage of correct links among existing links: $100A/T = 100 \cdot 31{,}770/161{,}940 = 19.6\%$ ($\theta = 0$). Combining usual wikification techniques should help improve the anchor recall.

5 Discussion

We now discuss two future directions, an extension to multiple language combination and a variation for inappropriate intra-language link detection.

The proposed method can be straightforwardly extended to three or more language combinations: Even if two source articles in two different languages are handled separately, the target article would be more enriched with the union of two transferred link sets. While this contributes to increasing the coverage of links, the reliability of links can also be improved by taking the intersection of the two transferred link sets. A more sophisticated combination of multiple source languages is a further problem.

In the experiment, existing links were used as the gold standard for evaluation, despite the fact that they are not always appropriate because they are manually created by unspecified contributors. For example, there is a biology-related article containing an anchor "translation" linking to an article "Translation" describing language translation, not to another article "Translation (biology)." Such an incorrect intra-language link may be detected using a similar method as the proposed one. In the above example, suppose the Japanese counterpart article contains an anchor "翻訳" linking to an article "翻訳 (生物学)." Two anchors "translation" and "翻訳" correspond to each other but their destination articles are not linked via an ILL. This inconsistency may be evidence for an inappropriate intra-language link. Note that which of the English and Japanese links is inappropriate cannot be easily determined. How to estimate the appropriateness of intra-language links is a problem to be solved.

6 Related Work

Wikification, which aims at linking mainly non-Wikipedia articles to Wikipedia articles, can be naturally applied to linking between Wikipedia articles. There has been much research on wikification, most of which focused on disambiguation of destination articles (Milne and Witten, 2008a; Fogarolli, 2009; Ratinov et al., 2011). Determining an appropriate destination article for an anchor term is a special case of WSD. Although a variety of ideas for WSD have been adapted to wikification, their performance is not satisfactory and there is room for further improvement. Another important issue with wikification is anchor selection, although most literature on wikification avoids the issue by selecting every term that is used as an anchor in any Wikipedia article. Anchor selection is a keyword extraction problem, which has been tackled using syntactic, statistical, and/or machine learning techniques but remains room for further improvement (Jacquemin and Bourigault, 2003). It should be added that our proposed method avoids both disambiguation and anchor selection problems by exploiting link information in another language edition of Wikipedia.

Adafre and de Rijke (2005) proposed a method for finding "missing intra-language links" in a Wikipedia article by assuming that an intra-language link represents the relatedness between concepts described by the linked articles. Their method adds intra-language links to an article by using articles with similar link structures as that of the article in question. Similar methods that use the Wikipedia's link structures as a semantic network have been proposed for entity linking (Milne and Witten, 2008b; Fogarolli, 2009; Ratinov et al., 2011). These still remain monolingual methods; the availability of other language editions cannot be assumed.

A bilingual approach to improving quality of Wikipedia articles has also been studied. Sorg and Cimiano (2008) proposed a method for finding new ILLs by using a classifier whose features include the number of ILLs between articles pointed by an article in one language and those pointed by an article in another language. Wang et al. (2013) improved the classifier by extending the intra-language links to increase the number of features. Both methods and our proposed method exploit the comparability between intra-language links in different language editions. However, while the former find new ILLs, the latter finds new intra-language links.

7 Conclusion

We proposed a method for enriching intra-language links in Wikipedia articles. It transfers intra-language links between a pair of different language articles linked by an inter-language link through the following two steps: first, determine destination articles to which the target-language article should be linked by following a source-language intra-language link and an ILL successively from each of the anchors in the source-language article; second, determine an anchor for each of the destination articles by searching the target-language article for possible anchors and selecting the most appropriate one according to the anchor translation probability criterion if two or more possible anchors are found. Unlike usual wikification methods, our method avoids anchor selection and disambiguation problems by exploiting the comparability of intra-language links between different language editions of Wikipedia.

We conducted an experiment of transferring intra-language links from the English edition to the Japanese edition to evaluate the effectiveness of our method. The method increased the number of intra-language links in Japanese articles by 40.9%, and the accuracy of anchors selected was estimated to be 96.3%. Future work includes an extension to multiple language combination and a variation for inappropriate intra-language link detection.

References

Adafre, Sisay Fissaha and Maarten de Rijke. 2005. Discovering missing links in Wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery: Issues, Approaches and Applications*, pages 90–97.

Fogarolli, Angela. 2009. Word sense disambiguation based on Wikipedia link structure. In *Proceedings of 2009 IEEE International Conference on Semantic Computing*, pages 77–82.

Gale, William A., Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of HLT '91 Workshop on Speech and Natural Language*, pages 233–237.

Jacquemin, Christian and Didier Bourigault. 2003. Term extraction and automatic indexing. In Ruslan Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics*, pages 599–615. Oxford University Press.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

Mihalcea, Rada and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pages 233–242.

Milne, David and Ian H. Witten. 2008a. Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518.

Milne, David and Ian H. Witten. 2008b. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the Wikipedia and AI Workshop of AAAI*, pages 25–30.

Ratinov, Lev, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384.

Sorg, Philipp and Philipp Cimiano. 2008. Enriching the crosslingual link structure of Wikipedia - a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence*.

Wang, Zhichun, Juanzi Li, and Jie Tang. 2013. Boosting cross-lingual knowledge linking via concept annotation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2733–2739.