Eliminating Incorrect Cross-Language Links in Wikipedia

Nacéra Bennacer Seghouani, Francesca Bugiotti, Jorge Galicia, Mariana Patricio, Gianluca Quercini

Eliminating Incorrect Cross-Language Links in Wikipedia

Nacéra Bennacer, Francesca Bugiotti, Jorge Galicia, Mariana Patricio, and Gianluca Quercini

LRI, CentraleSupélec, Paris-Saclay University Gif-sur-Yvette, 91190, France
nacera.bennacer@lri.fr, francesca.bugiotti@lri.fr,
jorge.galicia@student.ecp.fr,
mariana.patricio@student.ecp.fr,gianluca.quercini@lri.fr

Abstract. Many Wikipedia articles that cover the same topic in different language editions are interconnected via cross-language links that enable the understanding of topics in multiple languages, as well as cross-language information retrieval applications. However, cross-language links are added manually by the users of Wikipedia and, as such, are often incorrect. In this paper, we propose an approach to automatically eliminate incorrect cross-language links based on the observation that groups of articles that are pairwise connected through cross-language links form independent connected components. For each *incoherent* component (i.e., one that contains two or more articles from the same language edition), our approach assigns a *correctness score* to its crosslinks and removes those with the lowest score to make the component coherent. The results of our evaluation on a snapshot of Wikipedia in 8 languages indicates that our approach shows quantitative promise.

Keywords: Wikipedia, cross-language links, multi-language information retrieval

1 Introduction

Many Wikipedia articles that cover the same topic in different language editions are interconnected via cross-language links that enable the understanding of topics in multiple languages, as well as cross-language information retrieval applications [1,5,11]. Typically, the crosslinks are manually added by the users of Wikipedia and, as such, likely to be incorrect, meaning that they might connect articles that do not cover the same topic.

In this paper we describe an algorithm for the automatic elimination of incorrect crosslinks in Wikipedia. The existing literature is scarce and mostly focuses on the problem of determining missing crosslinks [3,7,8,10]. As noted by de Melo and Weikum [6], groups of articles that are pairwise connected through crosslinks (such as the ones titled `Decision Theory`, `Teoría de la decisión`, `Teoria della decisione` and "决策论") form independent connected components, if we model Wikipedia as a graph where the nodes correspond to the articles and

the edges are the cross-links. Under the hypothesis that all crosslinks are correct, all articles that belong to the same connected component cover the same topic; also, any connected component never contains two or more articles from the same language edition, because crosslinks connect articles in different languages. On the other hand, if two or more distinct articles within the same connected component come from the same language edition, at least one crosslink in the component is incorrect and the component is termed *incoherent*.

In order to detect the incorrect cross-language links, our approach looks specifically for *incoherent* connected components and iteratively removes cross-language links to turn them into coherent components. In order to determine the cross-language links of an incoherent component to eliminate, the approach assigns a *correctness score* to each crosslink and starts removing those that have the lowest score. The main contribution of this paper is the use of metrics derived from the topology of the Wikipedia graph to compute the correctness score of each crosslink; the contribution of each metric is thoroughly evaluated on a large sample of more than 1,124 crosslinks.

The remainder of this paper is organized as follows. After reviewing the related scientific literature (Section 2) and introducing the terminology (Section 3), we describe our approach in Section 4 and the the experiments in Section 5, followed by concluding statements in Section 6.

2 Related Work

While many approaches exist to finding missing crosslinks [3,7,8,10], considerably less research investigated the problem of determining the incorrect ones.

de Melo and Weikum observe that the articles that are pairwise connected through a crosslink form a connected component [6]. The aim of their approach is to obtain coherent components; to this extent, cross-links between pairs of articles that are asserted to be distinct are removed, with the constraint that the removals are minimized in order not to change the input graph too much. Their approach requires the solution of a linear program; depending on the size of the program, the solution may not be found. Rinser and colleagues point out that the stricter the definition of connectivity, the less incoherent the connected components [9]. Thus, weakly connected components are often incoherent and, as such, discarded, while any incoherent strongly connected component is split into bi-directional connected components and biconnected components. While the objective of this approach is to obtain coherent components, which can still contain incorrect crosslinks, it does not guarantee that the crosslinks that are eliminated are actually those incorrect. A more extreme approach consists in discarding all incoherent connected components [2].

Bolikowski presents an interesting study on the topology of crosslinks in Wikipedia [4]. His findings suggest that Wikipedia consists of near-complete subgraphs; some of them are connected through crosslinks, which is a sign of the presence of incorrect crosslinks. This study does not propose any approach to correct the crosslinks.

3 Terminology

Each article in Wikipedia belongs to an edition in a specific language and is characterized by a *title* (e.g., `Hot chocolate`), that is unique within its language edition, and *links* (also known as *intra-language links*) to other related articles in the same language (e.g., `Milk`, `Sugar`). Some links, known as *cross-language links* or *crosslinks*, connect articles that belong to two different language editions and cover the same topic (e.g., the English article titled `Hot chocolate` and the French article titled `Chocolat chaud`). A *redirect page* is one used to automatically link to an article whose title (e.g. `Hot chocolate`) is a synonym (or, alias) of the title of the redirect page (e.g., `Hot cocoa`). A *disambiguation page* has an ambiguous title (e.g., `Flash`) and presents a list of links to articles whose titles are its possible meanings, or interpretations (e.g., `Flash (photography)`, `Adobe Flash`).

In this paper, we model Wikipedia as a graph $\mathcal{W} = (PA, IL \cup CL)$; each node $p_\alpha \in PA$ corresponds to a Wikipedia page in language α that is either an article, a disambiguation or a redirect page; an edge is either a intra-language link $(p_\alpha, q_\alpha) \in IL$ or a crosslink $(p_\alpha, p_\beta) \in CL$ between two pages in languages α and β. Henceforth, the terms node and Wikipedia page will be used interchangeably. The *crosslink graph* $\mathcal{C} = (PA, CL)$, obtained from $\mathcal{W}$ by only keeping the crosslinks, is made of *connected components* such that two nodes belong to the same connected component if they are connected by a path of crosslinks. In other words, two nodes that belong to the same connected component correspond to two articles that cover the same topic, unless one or more crosslinks in the component are incorrect. More precisely, a sign of the presence of incorrect crosslinks is that a connected component has two or more nodes from the same language edition, in which case the component is considered as *incoherent*.

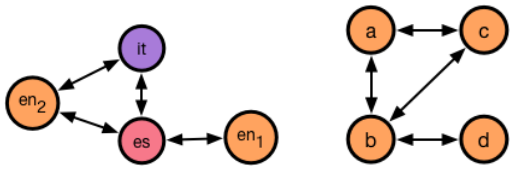
4 Approach

Our approach works through two main steps, that we term the candidate generation and the elimination step. The *candidate generation step* consists in identifying the set of incoherent connected components in the crosslink graph $\mathcal{C}$; this is done with a DFS visit on $\mathcal{C}$ that only selects the connected components of $\mathcal{C}$ that contain two or more nodes from the same language edition. The rationale for this step is to reduce the search space by focusing solely on the connected components that contain incorrect crosslinks with certainty. In the *elimination step*, the approach iterates over all candidate components with the intent of removing the incorrect crosslinks that make the component incoherent. More precisely, the approach iteratively eliminates crosslinks from a component C until C is split into two or more coherent connected components. The approach assigns a *correctness score* γ to each crosslink in C that measures the likelihood of the crosslink being correct; the links with lowest score are the first to be eliminated.

In the remainder of this section, we detail more the correctness score and the elimination algorithm.

Correctness Score. Let C be an incoherent connected component in the crosslink graph $\mathcal{C}$ with nodes $v_1, \ldots, v_n$. Our approach assigns a correctness score γ to each crosslink in C that measures the likelihood of the crosslink being correct; ideally, correct (resp., incorrect) crosslinks receive high (resp., low) values of γ. The idea is to sort the crosslinks in C by decreasing values of γ and eliminate those with the lowest values so as to turn C into a coherent component.

We observe that the graph topology of C is a good indicator as to the correctness of a crosslink. Since crosslinks in Wikipedia are added manually by different users, the likelihood that two articles from different language editions have both an incorrect crosslink to the same article is low. More precisely, if an article in the Spanish Wikipedia (es) has an incorrect crosslink to an article in the English Wikipedia (en_1), it is unlikely that the corresponding article in the Italian Wikipedia (it) links to the same article (Figure 1a); the probability of two different users doing the same mistake is low. Stated otherwise, the incorrect crosslinks are often incident with nodes that are loosely coupled to the other nodes of the component; the idea is to penalize these crosslinks by computing the



(a) (b)

Fig. 1: Explanatory examples

correctness score γ from the graph topology of the component C. We identified four topology metrics that we detail below.

Bidirectionality. Since each Wikipedia language edition is maintained independently of the others, the fact that a node v_i has a crosslink to a node v_j does not necessarily imply that v_j links back to v_i. Therefore, the fact that v_i and v_j are connected by a *mutual*, or *bidirectional*, crosslink is a strong indication as to the correctness of that crosslink. Our analysis on the crosslink graph $\mathcal{C}$ supports this intuition. Given a crosslink l, its bidirectionality score $\beta(l)$ is 1 if l is bidirectional, 0 otherwise.

Alternative Paths. Bidirectionality alone is not enough to identify incorrect crosslinks because the crosslinks of an incoherent connected component might all be bidirectional, in which case they would all have the same score. We observe that a clue of the correctness of a crosslink l between two nodes v_i and v_j in the connected component C is the number of alternative paths that lead from v_i to v_j; the higher the number of alternative paths between v_i and v_j, the higher the probability that the two nodes are strongly related (in this case, the

crosslink between them is correct). Figure shows an example, where there are three alternative paths between a and b, namely (a, b), (b, c, a) and (a, c, b). The *alternative path score* of a crosslink $l = (v_i, v_j)$ is defined in Equation 1

$$\alpha(l = (v_i, v_j)) = \frac{p(v_i, v_j)}{\max_{(v_k, v_m) \in C} p(v_k, v_m)} \quad (1)$$

where $p(v, w)$ is the number of paths between nodes v and w.

Minimal removal. We came across many examples were a connected component was incoherent because of a single incorrect crosslink. In this case, the removal of that crosslink would split the incoherent component into two or more coherent components and thus solve the problem. The *minimal removal score* $\zeta(l)$ of a crosslink l is 1 if the removal of l splits C into coherent components.

Chain links. Sorg and Cimiano observed that articles that are connected with a crosslink are often also connected by at least one chain link [10]. A chain link between two nodes v_i and v_j is a path in the Wikipedia graph $\mathcal{W}$ composed of both crosslinks and intra-language links such that: $v_i \xrightarrow{\text{intra}} w_i \xleftrightarrow{\text{cross}} w_j \xleftarrow{intra} v_j$ where *intra* (resp., *cross*) indicates an intra-language link (resp., crosslink). As an example, consider the case where v_i and v_j correspond to the articles that describe `Paris` in the English and French Wikipedia respectively and w_i and w_j are the articles that describe the Eiffel Tower. Intuitively, two nodes v_i and v_j that are connected by a crosslink and that link to many articles that are also connected via a crosslink are highly likely to cover the same topic. Equation 2 defines the chain link score of a crosslink l that connects two nodes v_i and v_j:

$$\xi(l) = \frac{cl(v_i, v_j)}{\max_{(v_k, v_m) \in C} cl(v_k, v_m)} \quad (2)$$

were $cl(v, w)$ is the number of chain-links between nodes v and w.

The correctness score. The correctness score $\gamma(l)$ of a crosslink l is obtained from a weighted average of the scores presented above, as indicated in the Equation 3.

$$\gamma(l) = w_1 \cdot \beta(l) + w_2 \cdot \alpha(l) + w_3 \cdot \zeta(l) + w_4 \cdot \xi(l) \quad (3)$$

The values of the weights w_i are such that $\sum w_i = 1$ and are discussed in Section 5.

The Elimination Algorithm. The elimination algorithm first determines the candidate incoherent components and then iterates over them to make them coherent. Each crosslink of a component C is assigned the correctness score γ and crosslinks are sorted by increasing score, meaning that the first link in S is the one with the lowest score. Finally, crosslinks are removed from C, starting from the ones with lowest score, until C is coherent.

5 Evaluation

For the evaluation of our approach, we adopted the following methodology. We downloaded eight Wikipedia language editions — English (en), German (de), French (fr), Italian (it), Spanish (es), Greek (el), Dutch (nl), Chinese (zh) — as of December 2016 and stored them as a graph $\mathcal{W}$ in Neo4j. The graph has 28,539,306 nodes that correspond to either articles, redirect or disambiguation pages, 346,165,183 intra-language links and 24,033,912 cross-language links. We computed the crosslink graph $\mathcal{C}$ and sampled 400 incoherent components where the incorrect crosslinks were manually identified to form a ground truth. We tuned the weights of the correctness score by using a subset of these incoherent components and run the approach on another subset to verify whether the crosslinks eliminated by the approach were actually those marked incorrect in the ground truth. Finally, we trained four classifiers and we compared the results. All the experiments were carried out on a computer running Windows 8 with an Intel core i7 processor, 8GB memory and a 512GB SSD hard drive. All the steps of the evaluation are detailed in the remainder of this section.

5.1 Results

In order to tune the four weights of the correctness score that set the importance of the corresponding topology metrics, we run the approach on a training set (240 connected components with a total of 683 annotated incorrect crosslinks and 7,653 correct crosslinks) and measured its ability of eliminating incorrect crosslinks by computing precision (P), recall (R) and f-measure (F), defined as follows:

$$P = \frac{|TP|}{|TP| + |FP|} \quad R = \frac{|TP|}{|TP| + |FN|} \quad F = \frac{2 \times P \times R}{P + R}$$

where TP is the set of links that are marked as incorrect by the approach that are actually so (true positives); FP is the set of links that are marked as incorrect by the approach that are actually correct (false positives); FN is the set of links that are marked as correct (or left undetermined) by the approach that are actually incorrect (false negatives). We run our approach on the test set (160 connected components with a total of 399 incorrect crosslinks and 4,207 correct crosslinks) with weights $w_1 = 0.4$ and $w_2 = 0.6$ and we obtain 0.80 for precision, recall and f-measure. Figure 2 shows that the accuracy is higher when considering small components (that account for the majority of the components in the crosslink graph). However, the approach can obtain a good precision (higher than 0.8) even on medium-sized components with 14 nodes, while the recall seems to be more sensitive to the variation of the size.

As for the time performance, the candidate generation step is the most expensive, as it takes 10 hours and 42 minutes to visit the crosslink graph and obtain all the incoherent components. The time to complete the elimination step depends on the metrics that are used to compute the correctness score. When the

chain links are not used, the elimination of the crosslinks in a given connected component takes 10 to 15 seconds on average; when considering the chain links the average time increases dramatically by 1 to 2 minutes, depending on the size of the component.

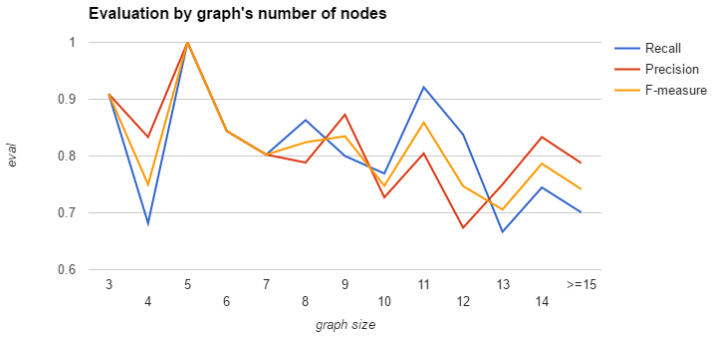


Fig. 2: Results by connected component size

Comparison. Given a cross-link, described by a set of numeric and nominal features, we can train a classifier to label the cross-link as either correct or incorrect (two classes). We used a re-sampling mechanism with no replacement to obtain ten different balanced training sets containing 916 cross-links, equally distributed across the two classes. Each cross-link (u, v) in both training and test set is described by a set of 6 features: two nominal features that indicate whether the node u (respectively, node v) is an article, a redirect or a disambiguation page; a nominal feature that indicates whether the cross-link (u, v) is bidirectional; a nominal feature that indicates whether the removal of (u, v) splits its connected component into two coherent components; the alternative path score of (u, v), as computed in Equation 1; the chain link score of (u, v), as computed in Equation 2. We trained four classifiers — SVM, Naive Bayes, Random Forests and OneR — on the ten training sets, we evaluated them on the test set and we averaged precision, recall and f-measure over the ten evaluations. As shown in Table 1, SVM is the best classifier in terms of precision (0.62), recall (0.89) and f-measure (0.73). Our approach achieves a much better precision (0.80) with a high recall (0.80) that results in the best f-measure (0.80). Among the classifiers, the results of SVM are consistent across all the training sets, while OneR has a lot of variability that depends on the sole feature that it selects to classify the cross-links; the precision ranges from 0.28 to 0.62, while the recall remains relatively stable.

SVM			Naive Bayes			R. Forests			OneR		
P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
0.62	0.89	**0.73**	0.38	**0.90**	0.53	0.45	0.89	0.60	0.39	0.86	0.54

Table 1: Result of the classification algorithms.

6 Concluding Remarks

In this paper, we presented an approach to identify and eliminate incorrect crosslinks from Wikipedia. Crosslinks are eliminated from incoherent components (those that contain two or more articles from the same language edition) starting from the links that have the lowest correctness score, which measures the likelihood of a link being correct. Our evaluation shows that the approach has quantitative promise (especially compared against classification algorithms). Future research will include the exploration of topology metrics, the elimination of incorrect crosslinks from coherent components and the parallelization of the algorithm.

References

1. S. F. Adafre and M. de Rijke. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 62–69, 2006.
2. E. Adar, M. Skinner, and D. S. Weld. Information Arbitrage across Multi-lingual Wikipedia. In *WSDM*, pages 94–103. ACM, 2009.
3. N. Bennacer, M. J. Vioulès, M. A. López, and G. Quercini. A Multilingual Approach to Discover Cross-Language Links in Wikipedia. In *WISE*, pages 539–553, 2015.
4. Ł. Bolikowski. Scale-free Topology of the Interlanguage Links in Wikipedia. *arXiv preprint arXiv:0904.0564*, 2009.
5. G. de Melo and G. Weikum. MENTA: Inducing Multilingual Taxonomies from Wikipedia. In *CIKM*, pages 1099–1108. ACM, 2010.
6. G. De Melo and G. Weikum. Untangling the Cross-lingual Link Structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 844–853, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
7. C. E. M. Moreira and V. P. Moreira. Finding Missing Cross-Language Links in Wikipedia. *JIDM*, 4(3):251–265.
8. A. Penta, G. Quercini, C. Reynaud, and N. Shadbolt. Discovering Cross-language Links in Wikipedia through Semantic Relatedness. In *ECAI*, pages 642–647, 2012.
9. D. Rinser, D. Lange, and F. Naumann. Cross-lingual entity matching and infobox alignment in wikipedia. *Information Systems*, 38(6):887–907, 2013.
10. 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 Artifical Intelligence*, pages 49–54, 2008.
11. P. Sorg and P. Cimiano. Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data Knowl. Eng.*, 74:26–45, 2012.