Élimination des liens inter-langues erronés dans Wikipédia

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

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

Résumé. Un lien inter-langue dans Wikipédia est un lien qui mène d'un article appartenant à une édition linguistique à un autre article décrivant le même concept dans une autre langue. Ces liens sont ajoutés manuellement par les utilisateurs de Wikipédia et ainsi ils sont susceptibles d'être erronés. Dans ce papier, nous proposons une approche pour l'élimination automatique des liens inter-langues. Le principe de base est que la présence d'un lien erroné est révélée par l'existence d'un chemin de liens inter-langues reliant deux articles appartenant à une même édition linguistique. Notre approche élimine des liens inter-langues, à partir de ceux qui ont un faible *score de correction*, jusqu'à ce qu'il n'y ait plus de chemins entre deux articles d'une même édition linguistique. Les résultats de notre évaluation sur un sous-graphe de Wikipédia consistant en 8 langues montre que l'approche est prometteuse.

1 Introduction

Les liens inter-langues (LILs) de Wikipédia permettent de naviguer facilement entre ses différentes éditions linguistiques et sont également exploités dans des applications de recherche d'information multilingue [de Melo et Weikum (2010); Sorg et Cimiano (2012)]. Cependant, les LILs sont ajoutés manuellement par les utilisateurs de Wikipédia et ainsi ils sont susceptibles d'être erronés (c.a.d. ils relient des articles qui ne décrivent pas un même concept).

Dans ce papier, nous proposons une approche pour l'élimination automatique des LILs. Les travaux existants s'attaquent principalement au problème de la détection des LILs manquants [Bennacer et al. (2015); Moreira et Moreira; Penta et al. (2012); Sorg et Cimiano (2008)]. Comme de Melo et Weikum l'ont fait remarquer, l'existence d'un chemin de LILs entre deux articles appartenant à une même édition linguistique (et, donc, décrivant deux concepts différents) révèle la présence d'un LIL erroné [De Melo et Weikum (2010)]. Autrement dit, si deux articles provenant d'une même édition linguistique appartiennent à une même composante connexe, au moins un LIL de cette composante est erroné et la composante est dite *incohérente*.

Notre approche attribue un *score de correction* aux liens d'une composante incohérente et élimine des liens de façon itérative, à partir de ceux qui ont un score faible (c.a.d., susceptibles

d'être erronés), jusqu'à diviser la composante en deux ou plusieurs composantes cohérentes. La contribution principale de ce papier est l'exploration de métriques obtenues de la topologie du graphe Wikipédia afin de calculer la probabilité qu'un LIL soit erroné.

La présentation du papier est organisée comme suit. Dans la section 2 nous présentons un aperçu de l'état de l'art. Nous introduisons, ensuite, dans la section 3, les notations et la terminologie permettant de décrire notre approche que nous présentons dans la section 4. Nous poursuivons, dans la section 5, par les expérimentations et les évaluations menées sur un sous-graphe de Wikipédia consistant en 8 éditions linguistiques. Enfin, nous concluons et présentons nos perspectives.

2 Aperçu de l'état de l'art

Contrairement à l'identification des LILs manquants, qui a fait l'objet de nombreuses recherches [Bennacer et al. (2015); Moreira et Moreira; Penta et al. (2012); Sorg et Cimiano (2008)], peu de travaux ont porté sur l'élimination des LILs erronés.

De Melo et Weikum définissent un ensemble de critères (appelés *assertions*) pour identifier les LILs qui sont susceptibles d'être erronés dans une composante incohérente [De Melo et Weikum (2010)]. Ces critères, qui, contrairement à notre approche, ne prennent pas en compte la topologie du graphe Wikipédia, sont utilisés par un programme linéaire qui divise la composante incohérente en deux ou plusieurs composantes cohérentes tout en minimisant le nombre de liens éliminés.

L'approche proposée par Rinser et ses collègues divise des composantes incohérentes faiblement connexes en plusieurs composantes cohérentes fortement connexes [Rinser et al. (2013)]. Leur évaluation ne montre pas si les liens éliminés sont effectivement ceux erronés.

Enfin, Bolikowski présente une étude intéressante qui montre que le graphe induit par les LILs de Wikipédia consiste en sous-graphes presque complets et la présence de liens entre ces sous-graphes est souvent un signe d'incohérence [Bolikowski (2009)]. Ce papier ne propose pas d'approche automatisée pour l'élimination des LILs erronés.

3 Terminologie

Nous modélisons Wikipédia comme un graphe orienté $\mathcal{W} = (PA, IL \cup CL)$: chaque nœud $p_\alpha \in PA$ correspond à un article Wikipédia (identifié par un *titre*) dans une langue α ; un arc est soit un lien interne $(p_\alpha, q_\alpha) \in IL$ entre deux articles de la même édition linguistique, soit un LIL $(p_\alpha, p_\beta) \in CL$. Nous notons que les termes *nœud* et *article* sont synonymes dans ce contexte. Le *graphe des liens inter-langues* $\mathcal{C} = (PA, CL)$, obtenu de $\mathcal{W}$ en éliminant tous les liens internes, consiste en plusieurs *composantes connexes* ; une composante est dite *incohérente* si elle contient deux articles d'une même édition linguistique.

4 Notre approche

Notre approche identifie d'abord l'ensemble des composantes incohérentes en faisant une visite DFS du graphe $\mathcal{C}$. Ensuite, chaque composante incohérente est divisée en deux ou plusieurs composantes cohérentes en éliminant des LILs de façon itérative. Pour ce faire, chaque

lien d'une composante reçoit un *score de correction* γ qui mesure la probabilité qu'il soit correct ; l'élimination commence par les liens qui ont les scores plus faibles (susceptibles d'être erronés).

Pour le calcul du score de correction γ, nous utilisons la topologie du graphe $\mathcal{C}$. Dès lors que les LILs sont ajoutés manuellement par des utilisateurs différents, la probabilité que deux articles appartenant à deux éditions linguistiques différentes aient un LIL erroné vers un même article est faible. Dans l'exemple de la figure 1, le lien entre es (article de l'édition espagnole) et en_1 (article de l'édition anglaise) est erroné (le lien correct porte vers l'article en_2) ; il est fort improbable que it (l'article correspondant à es dans l'édition italienne) ait lui aussi un lien incorrect vers en_1. En d'autres termes, les LILs erronés sont souvent incidents à des nœuds qui sont périphériques dans leurs composantes, comme c'est le cas du nœud en_1.

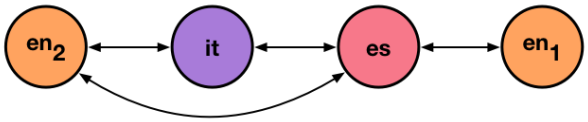

FIG. 1: Un LIL erroné est souvent incident à un nœud qui périphérique.

Nous définissons ci-dessous quatre métriques basées sur la topologie de $\mathcal{C}$.

Bidirectionnalité. Nous observons que la présence d'un LIL bidirectionnel entre deux nœuds v_i et v_j est souvent signe que le lien est correct. Pour cette raison, nous attribuons un *score de bidirectionnalité* à un LIL l qui vaut 1 si l est bidirectionnel, 0 sinon.

Chemins alternatifs. Un LIL entre deux nœuds v_i et v_j a une forte probabilité d'être correct s'il y a plusieurs chemins qui mènent de v_i à v_j, et vice-versa. Le *score des chemins alternatifs* $\alpha(l)$ d'un LIL $l = (v_i, v_j)$ est défini comme suit :

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

où $p(v, w)$ est le nombre de chemins qui mènent de v à w, et vice-versa.

Élimination minimale. Dans le sous-graphe de Wikipédia que nous considérons, nous avons trouvé plusieurs composantes connexes incohérentes à cause d'un seul LIL erroné. Dans ces cas, l'élimination de ce LIL divise la composante en deux composantes cohérentes, ce qui résout le problème. Le score d'élimination minimale $\zeta(l)$ d'un LIL l est 1 si l'élimination de l divise sa composante en deux composantes cohérentes, 0 sinon.

Chaînes de liens. Sorg et Cimiano ont fait remarquer que deux nœuds v_i et v_j reliés par un LIL sont aussi reliés par au moins une *chaîne de liens* : $v_i \xrightarrow{\text{interne}} w_i \xleftrightarrow{\text{inter-langue}} w_j \xleftarrow{interne} v_j$ [Sorg et Cimiano (2008)]. Intuitivement, deux articles v_i (par ex., *Paris* dans la Wikipédia anglaise) et v_j (par ex., *Paris* dans la Wikipédia française) qui décrivent un même concept ont des liens internes vers des articles (par ex., *Eiffel Tower* et *Tour Eiffel*) qui eux-mêmes décrivent

un même concept. Donc, plus il y a de chaînes de liens entre v_i et v_j, plus la probabilité que le LIL entre v_i et v_j soit correct est forte. Le *score des chaînes de liens* $\xi(l)$ d'un LIL $l = (v_i, v_j)$ est défini comme suit :

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

où $cl(v, w)$ est le nombre de chaînes de liens entre v et w.

Le score de correction. Le *score de correction* $\gamma(l)$ d'un LIL l est obtenu en calculant une moyenne pondérée des scores présentés ci-dessus $\gamma(l) = w_1 \cdot \beta(l) + w_2 \cdot \alpha(l) + w_3 \cdot \zeta(l) + w_4 \cdot \xi(l)$. Les valeurs des poids w_i ($\sum w_i = 1$) sont discutés dans la Section 5.

5 Expérimentations et évaluations des résultats

Nous avons évalué notre approche sur un sous-graphe de Wikipédia (stocké dans une base de données Neo4j) consistant en huit éditions linguistiques — anglaise, allemande, française, italienne, espagnole, grecque, néerlandaise, chinoise — qui datent de Décembre 2016. Le graphe a 28 539 306 nœuds, 346 165 183 liens internes et 24 033 912 LILs. Nous avons calculé le graphe des liens inter-langues $\mathcal{C}$ et sélectionné 400 composantes incohérentes où les LILs erronés ont été identifiés par les auteurs de ce papier. Les expérimentations ont été effectuées sur un ordinateur équipé de Windows 8, d'un processeur Intel Core i7, 8GB de mémoire et un disque SSD de 512 GB.

Résultats. Afin de régler les quatre poids du score de correction, nous avons appliqué notre approche sur un ensemble d'entraînement (240 composantes, respectivement 683 et 7 653 LILs erronés et corrects) et nous avons mesuré sa capacité d'éliminer des LILs erronés en calculant la précision ($P = |VP|/(|VP|+|FP|)$), le rappel ($R = |VP|/(|VP|+|FN|)$) et la F-mesure (F, moyenne harmonique de P et R). VP est l'ensemble des liens qui sont correctement considérés erronés par notre approche (vrais positifs) ; FP (liens incorrectement considérés erronés) et FN (liens incorrectement considérés corrects) sont respectivement les faux positifs et négatifs. A l'issue de la phase d'entraînement, les valeurs des poids qui donnent les meilleurs résultats sont les suivantes : $w_1 = 0.4$, $w_2 = 0.6$, $w_3 = 0$ et $w_4 = 0.1$. Nous remarquons que la métrique "Élimination minimale" entraîne une augmentation du rappel mais a un impact fortement négatif sur la précision, d'où la décision de mettre $w_3 = 0$. Nous avons appliqué notre méthode avec ces valeurs sur un ensemble de test (160 composantes, respectivement 399 et 4 207 LILs erronés et corrects) et nous avons obtenu P = 0.78, R = 0.83 et F = 0.80. La figure 2 montre que l'approche est plus efficace sur des composantes de petite taille qui constituent la majorité dans $\mathcal{C}$.

En ce qui concerne les performances de l'approche, le calcul du graphe des LILs nécessite de 10 heures (visite DFS du graphe Wikipédia) ; le temps nécessaire pour compléter l'élimination des LILs varie entre 10 et 15 secondes par composante quand l'approche n'utilise pas les chaînes de liens (sinon, il faut compter un temps variable de 1 à 2 minutes).

Comparaison. Le problème de l'élimination d'un lien erroné peut être décliné comme un problème de classification binaire. Plus précisément, nous décrivons un LIL (u, v) avec quatre

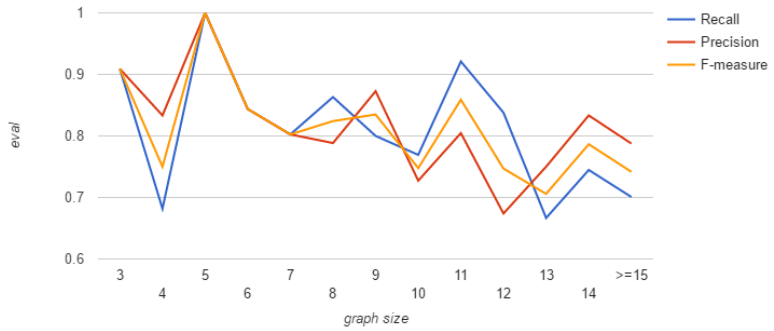

FIG. 2: Résultats en fonction de la taille des composantes.

attributs qui reprennent les métriques présentées en Section 4. Nous utilisons un mécanisme de réechantillonage sans remplacement pour obtenir dix ensembles d'entraînement équilibrés consistant en 916 LILs, distribués uniformément sur les deux classes (erroné, correct).

Nous avons entraîné quatre classificateurs — machine à vecteurs de support à noyau linéaire (*SVM*), Naive Bayes, Forêt d'arbres décisionnels (*R. Forests*) et OneR (classificateur à base de règles) — sur les dix ensembles d'entraînement et nous les avons évalués sur l'ensemble de test précédemment créé. Les résultats dans le tableau 1 montrent que SVM est le meilleur classificateur en termes de précision (0.62), rappel (0.89) et F-mesure (0.73). Nous notons que les mêmes résultats sont obtenus sans considérer l'attribut "Élimination minimale". Les résultats de tous les classificateurs, sauf OneR, sont comparables sur tous les ensembles d'entraînement.

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

TAB. 1: Résultats des classificateurs.

6 Conclusions et perspectives

Dans ce papier nous avons présenté une approche pour l'élimination des liens inter-langues (LILs) erronés dans Wikipédia. La contribution principale de cette approche est l'exploration de métriques basées sur la topologie du graphe Wikipédia. Les résultats de notre évaluation sur un sous-graphe de Wikipédia consistant en 8 langues montre que l'approche est prometteuse. Nos travaux actuels portent sur l'étude de la topologie des composantes cohérentes qui peuvent contenir des LILs erronés et que, à notre connaissance, aucune approche considère. Nous souhaitons également intégrer des heuristiques qui exploitent d'autres éléments de Wikipédia tels

que les catégories et les pages de redirection et d'homonymie. Une expérimentation sur une grande base de données et une comparaison avec toutes les approches existantes fera l'objet de nos travaux futurs, ainsi qu'une implémentation parallèle de notre approche.

Références

Bennacer, N., M. J. Vioulès, M. A. López, et G. Quercini (2015). A Multilingual Approach to Discover Cross-Language Links in Wikipedia. In *WISE*, pp. 539–553.

Bolikowski, Ł. (2009). Scale-free Topology of the Interlanguage Links in Wikipedia. *arXiv preprint arXiv :0904.0564*.

de Melo, G. et G. Weikum (2010). MENTA : Inducing Multilingual Taxonomies from Wikipedia. In *CIKM*, pp. 1099–1108. ACM.

De Melo, G. et G. Weikum (2010). Untangling the Cross-lingual Link Structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 844–853. Association for Computational Linguistics.

Moreira, C. E. M. et V. P. Moreira. Finding Missing Cross-Language Links in Wikipedia. *JIDM 4*(3), 251–265.

Penta, A., G. Quercini, C. Reynaud, et N. Shadbolt (2012). Discovering Cross-language Links in Wikipedia through Semantic Relatedness. In *ECAI*, pp. 642–647.

Rinser, D., D. Lange, et F. Naumann (2013). Cross-lingual entity matching and infobox alignment in wikipedia. *Information Systems 38*(6), 887–907.

Sorg, P. et P. Cimiano (2008). Enriching the crosslingual link structure of wikipedia-a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, pp. 49–54.

Sorg, P. et P. Cimiano (2012). Exploiting Wikipedia for Cross-lingual and Multilingual Information Retrieval. *Data Knowl. Eng. 74*, 26–45.

Summary

Many Wikipedia articles that cover the same topic in different language editions are interconnected via cross-language links 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. The rationale is that the presence of an incorrect cross-language link is revealed by the existence of a path of cross-links between two articles of the same language edition. Our approach removes cross-languages links, starting from those having a low *correctness score*, until there is no path between two articles of the same language edition. The results of our evaluation on a snapshot of Wikipedia in 8 languages indicates that our approach shows quantitative promise.