Towards Building a Multilingual Semantic Network: Identifying Interlingual Links in Wikipedia

Bharath Dandala
Dept. of Computer Science
University of North Texas
Denton, TX
BharathDandala@my.unt.edu

Rada Mihalcea
Dept. of Computer Science
University of North Texas
Denton, TX
rada@cs.unt.edu

Razvan Bunescu
School of EECS
Ohio University
Athens, Ohio
bunescu@ohio.edu

Abstract

Wikipedia is a Web based, freely available multilingual encyclopedia, constructed in a collaborative effort by thousands of contributors. Wikipedia articles on the same topic in different languages are connected via interlingual (or translational) links. These links serve as an excellent resource for obtaining lexical translations, or building multilingual dictionaries and semantic networks. As these links are manually built, many links are missing or simply wrong. This paper describes a supervised learning method for generating new links and detecting existing incorrect links. Since there is no dataset available to evaluate the resulting interlingual links, we create our own gold standard by sampling translational links from four language pairs using distance heuristics. We manually annotate the sampled translation links and used them to evaluate the output of our method for automatic link detection and correction.

1 Introduction

In recent years, Wikipedia has been used as a resource of world knowledge in many natural language processing applications. A diverse set of tasks such as text categorization, information extraction, information retrieval, question answering, word sense disambiguation, semantic relatedness, and named entity recognition have been shown to benefit from the semi-structured text of Wikipedia. Most approaches that use the world knowledge encoded in Wikipedia are statistical in nature and therefore their performance depends significantly on the size of Wikipedia. Currently, the English Wikipedia alone has four million articles. However, the combined Wikipedias for all other languages greatly exceed the English Wikipedia in size, yielding a combined total of more than 10 million articles in more than 280 languages.[1] The rich hyperlink structure of these Wikipedia corpora in different languages can be very useful in identifying various relationships between concepts.

Wikipedia articles on the same topic in different languages are often connected through interlingual links. These links are the small navigation links that show up in the “Languages” sidebar in most Wikipedia articles, and they connect an article with related articles in other languages. For instance, the interlingual links for the Wikipedia article about ”Football” connect it to 20 articles in 20 different languages. In the ideal case, a set of articles connected directly or indirectly via such links would all describe the same entity or concept. However, these links are produced either by polyglot editors or by automatic bots. Editors commonly make mistakes by linking articles that have conceptual drift, or by linking to a concept at a different level of granularity. For instance, if a corresponding article in one of the languages does not exist, a similar article or a more general article about the concept is sometimes linked instead. Various bots also add new interlingual links or attempt to correct existing ones. The downside of a bot is that an error in a translational link created by editors in Wikipedia for one language propagates to Wikipedias in other languages. Thus, if a bot introduces a wrong link, one may have to search for

[1] http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 30–37,
Montréal, Canada, June 7-8, 2012.

Language	Code	Articles	Redirects	Users
English	en	4,674,066	4,805,557	16,503,562
French	fr	3,298,615	789,408	1,250,266
German	de	3,034,238	678,288	1,398,424
Italian	it	2,874,747	319,179	731,750
Polish	pl	2,598,797	158,956	481,079
Spanish	es	2,587,613	504,062	2,162,925
Dutch	nl	2,530,250	226,201	446,458
Russian	ru	2,300,769	682,402	819,812
Japanese	jp	1,737,565	372,909	607,152
Chinese	cn	1,199,912	333,436	1,171,148

Table 1: Number of articles, redirects, and users for the top nine Wikipedia editions plus Chinese. The total number of articles also includes the disambiguation pages.

the underlying error in a different language version of Wikipedia.

The contributions of the research described in this paper are two-fold. First, we describe the construction of a dataset of interlingual links that are automatically sampled from Wikipedia based on a set of distance heuristics. This dataset is manually annotated in order to enable the evaluation of methods for translational link detection. Second, we describe an automatic model for correcting existing links and creating new links, with the aim of obtaining a more stable set of interlingual links. The model's parameters are estimated on the manually labeled dataset using a supervised machine learning approach.

The remaining of this paper is organized as follows: Section 2 briefly describes Wikipedia and the relevant terminology. Section 3 introduces our method of identifying a candidate set of translational links based on distance heuristics, while Section 4 introduces the methodology for building a manually annotated dataset. Section 5 describes the machine learning experiments for detecting or correcting interlingual links. Finally, we present related work in Section 6, and concluding remarks in Section 7.

2 Wikipedia

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage, and this "freedom of contribution" has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential errors are quickly corrected within the collaborative environment) of this online resource.

The basic entry in Wikipedia is an *article* (or *page*), which defines and describes an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article. Articles are organized into *categories*, which in turn are organized into category hierarchies. For instance, the article *automobile* is included in the category *vehicle*, which in turn has a parent category named *machine*, and so forth.

Each article in Wikipedia is uniquely referenced by an identifier, consisting of one or more words separated by spaces or underscores and occasionally a parenthetical explanation. For example, the article for *bar* with the meaning of *"counter for drinks"* has the unique identifier *bar (counter)*.

Wikipedia editions are available for more than 280 languages, with a number of entries varying from a few pages to three millions articles or more per language. Table 1 shows the nine largest Wikipedias (as of March 2012) and the Chinese Wikipedia, along with the number of articles and approximate number of contributors.[2]

The ten languages mentioned above are also the languages used in our experiments. Note that Chi-

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias#Grand_Total

Relation	Exists	Via
SYMMETRY		
en=Ball de=Ball	Yes	-
en=Hentriacontane it=Entriacontano	No	-
TRANSITIVITY		
en=Deletion (phonology) fr=Amuïssement	Yes	nl=Deletie (taalkunde)
en=Electroplating fr=Galvanoplastie	No	-
REDIRECTIONS		
en=Gun Dog de=Schiesshund	Yes	de=Jagdhund
en=Ball de=Ball	No	-

Table 2: Symmetry, transitivity, and redirections in Wikipedia

nese is the twelfth largest Wikipedia, but we decided to include it at the cost of not covering the tenth largest Wikipedia (Portuguese), which has close similarities with other languages already covered (e.g., French, Italian, Spanish).

Relevant for the work described in this paper are the *interlingual links*, which explicitly connect articles in different languages. For instance, the English article for *bar (unit)* is connected, among others, to the Italian article *bar (unitá di misura)* and the Polish article *bar (jednostka)*. On average, about half of the articles in a Wikipedia version include interlingual links to articles in other languages. The number of interlingual links per article varies from an average of five in the English Wikipedia, to ten in the Spanish Wikipedia, and as many as 23 in the Arabic Wikipedia.

3 Identifying Interlingual Links in Wikipedia

The interlingual links connecting Wikipedias in different languages should ideally be symmetric and transitive. The symmetry property indicates that if there is an interlingual link $A_\alpha \rightarrow A_\beta$ between two articles, one in language α and one in language β, then the reverse link $A_\alpha \leftarrow A_\beta$ should also exist in Wikipedia. According to the transitivity property, the presence of two links $A_\alpha \rightarrow A_\beta$ and $A_\beta \rightarrow A_\gamma$ indicates that the link $A_\alpha \rightarrow A_\gamma$ should also exist in Wikipedia, where α, β and γ are three different languages. While these properties are intuitive, they are not always satisfied due to Wikipedia's editorial policy that accredits editors with the responsibility of maintaining the articles. Table 2 shows actual cases in Wikipedia where these properties fail due to missing interlingual links. The table also shows examples where the editors link an article from one language to a redirect page in another language.

Link type	Total number of links	Newly added links
DL	26,836,572	-
RL	26,836,572	1,277,760
DP_2/RP_2	25,763,689	853,658
DP_3/RP_3	23,383,535	693,262
DP_4/RP_4	21,560,711	548,354

Table 3: Number of links identified in Wikipedia, as direct, symmetric, or transitional links. The number of newly added links, not known in the previous set of links, is also indicated (e.g., DP_3/RP_3 adds 693,262 new links not found by direct or symmetric links, or by direct or reverse paths of length two).

In order to generate a normalized set of interlingual links between Wikipedias, we replace all the redirect pages with the corresponding original articles, so that each concept in a language is represented by one unique article. We then identify the following four types of simple interlingual paths between articles in different languages:

DL: Direct links $A_\alpha \rightarrow A_\beta$ between two articles.

RL: Reverse links $A_\alpha \leftarrow A_\beta$ between two articles.

DP_k: Direct, simple paths of length k between two articles.

RP_k: Reverse, simple paths of length k between two articles.

Relation	Number of paths
DL	
en=Ball de=Ball	1
en=Ball it=Palla (sport)	1
en=Ball fr=Boule (solide)	0
de=Ball fr=Ballon (sport)	0
RL	
en=Ball de=Ball	1
en=Ball it=Palla(sport)	1
en=Ball fr=Boule (solide)	0
de=Ball fr=Ballon (sport)	0
DP_2	
en=Ball de=Ball	1
en=Ball it=Palla (sport)	2
en=Ball fr=Boule (solide)	1
de=Ball fr=Ballon (sport)	2
DP_3	
en=Ball de=Ball	1
en=Ball it=Palla (sport)	0
en=Ball fr=Boule (solide)	1
de=Ball fr=Ballon (sport)	1
DP_4	
en=Ball de=Ball	0
en=Ball it=Palla (sport)	0
en=Ball fr=Boule (solide)	1
de=Ball fr=Ballon (sport)	0
RP_2	
en=Ball de=Ball	1
en=Ball it=Palla (sport)	2
en=Ball fr=Boule (solide)	0
de=Ball fr=Ballon (sport)	2
RP_3	
en=Ball de=Ball	1
en=Ball it=Palla (sport)	0
en=Ball fr=Boule (solide)	0
de=Ball fr=Ballon (sport)	1
RP_4	
en=Ball de=Ball	0
en=Ball it=Palla (sport)	0
en=Ball fr=Boule (solide)	0
de=Ball fr=Ballon (sport)	0

Table 4: A subset of the direct links, reverse links, and inferred direct and reverse paths for the graph in Figure 1

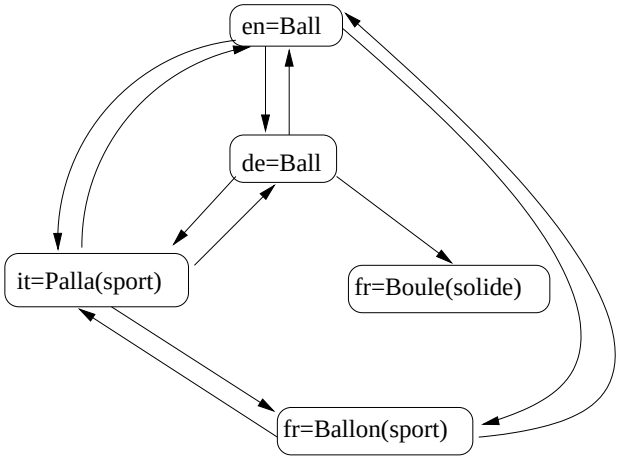

Figure 1: A small portion of the multilingual Wikipedia graph.

Figure 1 shows a small portion of the Wikipedia graph, connecting Wikipedias in four languages: English, German, Italian, and French. Correspondingly, Table 4 shows a subset of the direct links DL, reverse links RL, direct translation paths DP_k and reverse translation paths RP_k of lengths k = 2, 3, 4 for the graph in the figure.

Using these distance heuristics, we are able to extract or infer a very large number of interlingual links. Table 3 shows the number of direct links extracted from the ten Wikipedias we currently work with, as well as the number of paths that we add by enforcing the symmetry and transitivity properties.

4 Manual Evaluation of the Interlingual Links

The translation links in Wikipedia, whether added by the Wikipedia editors (direct links), or inferred by the heuristics described in the previous section, are not guaranteed for quality. In fact, previous work (de Melo and Weikum, 2010b) has shown that a large number of the links created by the Wikipedia users are incorrect, connecting articles that are not translations of each other, subsections of articles, or disambiguation pages. We have therefore decided to run a manual annotation study in order to determine the quality of the interlingual links. The resulting annotation can serve both as a gold standard for evaluating the quality of predicted links, and as supervision for a machine learning model that would automatically detect translation links.

Language pair	0	1	2	3	4
(English, German)	46	8	29	2	110
(English, Spanish)	22	19	19	13	123
(Italian, French)	30	7	19	7	132
(Spanish, Italian)	21	8	17	13	136

Table 6: Number of annotations on a scale of 0-4 for each pair of languages

From the large pool of links directly available in Wikipedia or inferred automatically through symmetry and transitivity, we sampled and then manually annotated 195 pairs of articles for each of four language pairs: (English, German), (English, Spanish), (Italian, French), and (Spanish, Italian). The four language pairs were determined based on the native or near-native knowledge available in the group of annotators in our research group. The sampling of the article pairs was done such that it covers all the potentially interesting cases obtained by combining the heuristics used to identify interlingual links. The left side of Table 5 shows the combination of heuristics used to select the article pairs. For each such combination, and for each language pair, we randomly selected 15 articles. Furthermore, we added 15 randomly selected pairs for the highest quality combination (Case 1).

For each language pair, the sampled links were annotated by one human judge, with the exception of the (English, Spanish) dataset, which was annotated by two judges so that we could measure the inter-annotator agreement. The annotators were asked to check the articles in each link and annotate the link on a scale from 0 to 4, as follows:

4: Identical concepts that are perfect translations of each other.

3: Concepts very close in meaning, which are good translations of each other, but a better translation for one of the concepts in the pair also exists. The annotators are not required to identify a better translation in Wikipedia, they only have to use their own knowledge of the language, e.g. "building" (English) may be a good translation for "tore" (Spanish), yet a better translation is known to exist.

2: Concepts that are closely related but that are not translations of each other.

1: Concepts that are remotely related and are not translations of each other.

0: Completely unrelated concepts or links between an article and a portion of another article.

To determine the quality of the annotations, we ran an inter-annotator study for the (English-Spanish) language pair. The two annotators had a Pearson correlation of 70%, which indicates good agreement. We also calculated their agreement when grouping the ratings from 0 to 4 in only two categories: 0, 1, and 2 were mapped to *no translation*, whereas 3 and 4 were mapped to *translation*. On this coarse scale, the annotators agreed 84% of the time, with a kappa value of 0.61, which once again indicate good agreement.

The annotations are summarized in the right side of Table 5. For each quality rating, the table shows the number of links annotated with that rating. Note that this is a summary over the annotations of five annotators, corresponding to the four language pairs, as well as an additional annotation for (English, Spanish).

Not surprisingly, the links that are "supported" by all the heuristics considered (Case 1) are the links with the highest quality. These are interlingual links that are present in Wikipedia and that can also be inferred through transitive path heuristics. Interestingly, links that are only guaranteed to have a direct link (DL) and no reverse link (RL) (Case 2) have a rather low quality, with only 68% of the links being considered to represent a perfect or a good translation (score of 3 or 4).

Table 6 summarizes the annotations per language pair. There appear to be some differences in the quality of interlingual links extracted or inferred for different languages, with (Spanish, Italian) being the pair with the highest quality of links (76% of the links are either perfect or good translations), while English to German seems to have the lowest quality (only 57% of the links are perfect or good). For the (English, Spanish) pair, we used the average of the two annotators' ratings, rounded up to the nearest integer.

Cases	Combinations of heuristics to extract or infer interlingual links									Link quality on a 0-4 scale				
	DL	RL	DP_2	RP_2	DP_3	RP_3	DP_4	RP_4	Samples	0	1	2	3	4
Case 1	y	y	y	y	y	y	y	y	30	6	3	6	6	129
Case 2	y	n	-	-	-	-	-	-	15	15	3	6	3	48
Case 3	n	y	-	-	-	-	-	-	15	13	3	8	4	47
Case 4	n	n	y	y	-	-	-	-	15	6	3	16	4	46
Case 5	n	n	-	-	y	y	-	-	15	13	9	12	4	28
Case 6	n	n	-	-	-	-	y	y	15	15	8	3	8	37
Case 7	n	n	n	n	-	-	-	-	15	19	8	11	5	31
Case 8	n	n	-	-	n	n	-	-	15	13	8	11	5	32
Case 9	n	n	-	-	-	-	n	n	15	25	4	11	2	33
Case 10	y	y	n	n	-	-	-	-	15	6	3	4	3	59
Case 11	y	y	-	-	n	n	-	-	15	6	2	3	0	64
Case 12	y	y	-	-	-	-	n	n	15	3	6	2	4	60

Table 5: Left side of the table: distance heuristics and number of samples based on each distance heuristic. 'y' indicates that the corresponding path should exist, 'n' indicates that the corresponding path should not exist, '-' indicates that we don't care whether the corresponding path exists or not. Right side of the table: manual annotations of the quality of links, on a scale of 0 to 4, with 4 meaning perfect translations.

5 Machine Learning Experiments

The manual annotations described above are good indicators of the quality of the interlingual links that can be extracted and inferred in Wikipedia. But such manual annotations, because of the human effort involved, do not scale up, and therefore we cannot apply them on the entire interlingual Wikipedia graph to determine the links that should be preserved or the ones that should be removed.

Instead, we experiment with training machine learning models that would automatically determine the quality of an interlingual link. As features, we use the presence or absence of direct or symmetric links, along with the number of inferred paths of length $k = 2, 3, 4$, as defined in Section 3. Table 7 shows the feature vectors for the same four pairs of articles that were used in Table 4. The feature values are computed based on the sample network of interlingual links from Figure 1. Each feature vector is assigned a numerical class, corresponding to the manual annotation provided by the human judges.

We conduct two experiments, at a fine-grained and a coarse-grained level. In both experiments, we use all the annotations for all four language pairs together (i.e., a total of 780 examples), and perform evaluations in a ten-fold cross validation scenario.

For the fine-grained experiments, we use all five numerical classes in a linear regression model.[3] We determine the correctness of the predictions on the test data by calculating the Pearson correlation with respect to the gold standard. The resulting correlation was measured at 0.461. For comparison, we also run an experiment where we only keep the presence or absence of the direct links as a feature (DL). In this case, the correlation was measured at 0.418, which is substantially below the correlation obtained when using all the features. This indicates that the interlingual links inferred through our heuristics are indeed useful.

In the coarse-grained experiments, the quality ratings 0, 1, and 2 are mapped to the *no translation* label, while ratings 3 and 4 are mapped to the *translation* label. We used the Ada Boost classifier with decision stumps as the binary classification algorithm. When using the entire feature vectors, the accuracy is measured at 73.97%, whereas the use of only the direct links results in an accuracy of 69.35%. Similar to the fine-grained linear regression experiments, these coarse-grained experiments further validate the utility of the interlingual links inferred through the transitive path heuristics.

[3] We use the Weka machine learning toolkit.

Concept pair	DL	RL	DP_2	DP_3	DP_4	RP_2	RP_3	RP_4	Class
en=Ball de=Ball	1	1	1	1	0	1	1	0	4
en=Ball it=Palla (sport)	1	1	2	0	0	2	0	0	4
en=Ball fr=Boule (solide)	0	0	1	1	1	0	0	0	1
de=Ball fr=Ballon (sport)	0	0	2	1	0	2	1	0	4

Table 7: Examples of feature vectors generated for four interlingual links, corresponding to the concept pairs listed in Table 4

6 Related Work

The multilingual nature of Wikipedia has been already exploited to solve several number of language processing tasks. A number of projects have used Wikipedia to build a multilingual semantic knowledge base by using the existing multilingual nature of Wikipedia. For instance, (Ponzetto and Strube, 2007) derived a large scale taxonomy from the existing Wikipedia. In related work, (de Melo and Weikum, 2010a) worked on a similar problem in which they combined all the existing multilingual Wikipedias to build a stable, large multilingual taxonomy.

The interlingual links have also been used for cross-lingual information retrieval (Nguyen et al., 2009) or to generate bilingual parallel corpora (Mohammadi and QasemAghaee, 2010). (Ni et al., 2011) used multilingual editions of Wikipedia to mine topics for the task of cross lingual text classification, while (Hassan and Mihalcea, 2009) used Wikipedias in different languages to measure cross-lingual semantic relatedness between concepts and texts in different languages. (Bharadwaj et al., 2010) explored the use of the multilingual links to mine dictionaries for under-resourced languages. They developed an iterative approach to construct a parallel corpus, using the interlingual links, info boxes, category pages, and abstracts, which they then be used to extract a bilingual dictionary. (Navigli and Ponzetto, 2010) explored the connections that can be drawn between Wikipedia and WordNet. While no attempts were made to complete the existing link structure of Wikipedia, the authors made use of machine translation to enrich the resource.

The two previous works most closely related to ours are the systems introduced in (Sorg and Cimiano, 2008) and (de Melo and Weikum, 2010a; de Melo and Weikum, 2010b). (Sorg and Cimiano, 2008) designed a system that predicts new interlingual links by using a classification based approach. They extract certain types of links from bilingual Wikipedias, which are then used to create a set of features for the machine learning system. In follow-up work, (Erdmann et al., 2008; Erdmann et al., 2009) used an expanded set of features, which also accounted for direct links, redirects, and links between articles in Wikipedia, to identify entries for a bilingual dictionary. In this line of work, the focus is mainly on article content analysis, as a way to detect new potential translations, rather than link analysis as done in our work.

Finally, (de Melo and Weikum, 2010b) designed a system that detects errors in the existing interlingual links in Wikipedia. They show that there are a large number of links that are imprecise or wrong, and propose the use of a weighted graph to produce a more consistent set of consistent interlingual links. Their work is focusing primarily on correcting existing links in Wikipedia, rather than inferring new links as we do.

7 Conclusions

In this paper, we explored the identification of translational links in Wikipedia. By using a set of heuristics that extract and infer links between Wikipedias in different languages, along with a machine learning algorithm that builds upon these heuristics to determine the quality of the interlingual links, we showed that we can both correct existing translational links in Wikipedia as well as discover new interlingual links. Additionally, we have also constructed a manually annotated dataset of interlingual links, covering different types of links in four pairs of languages, which can serve as a gold standard for evaluating the quality of predicted links, and as supervision for the machine learning model.

In future work, we plan to experiment with additional features to enhance the performance of the classifier. In particular, we would like to also include content-based features, such as content overlap and interlinking.

The collection of interlingual links for the ten Wikipedias considered in this work, as well as the manually annotated dataset are publicly available at http://lit.csci.unt.edu.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation IIS awards #1018613 and #1018590 and CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

G.R. Bharadwaj, N. Tandon, and V. Varma. 2010. An iterative approach to extract dictionaries from Wikipedia for under-resourced languages. Kharagpur, India.

G. de Melo and G. Weikum. 2010a. MENTA: inducing multilingual taxonomies from Wikipedia. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1099–1108, New York, NY, USA. ACM.

G. de Melo and G. Weikum. 2010b. Untangling the cross-lingual link structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 844–853, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2008. An approach for extracting bilingual terminology from Wikipedia. In *Proceedings of the 13th International Conference on Database Systems for Advanced Applications*.

M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. 2009. Improving the extraction of bilingual terminology from Wikipedia. *ACM Transactions on Multimedia Computing, Communications and Applications*, 5(4):31:1–31:17.

S. Hassan and R. Mihalcea. 2009. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suntec, Singapore.

M. Mohammadi and N. QasemAghaee. 2010. Building bilingual parallel corpora based on Wikipedia. *International Conference on Computer Engineering and Applications*, 2:264–268.

R. Navigli and S. Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

D. Nguyen, A. Overwijk, C. Hauff, D. Trieschnigg, D. Hiemstra, and F. De Jong. 2009. WikiTranslate: query translation for cross-lingual information retrieval using only Wikipedia. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, pages 58–65, Berlin, Heidelberg. Springer-Verlag.

X. Ni, J. Sun, J. Hu, and Z. Chen. 2011. Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 375–384, New York, NY, USA. ACM.

S. Ponzetto and M. Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1440–1445. AAAI Press.

P. Sorg and P. 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*.