Wikipedia Missing Link Discovery: A Comparative Study

Omer Sunercan and **Aysenur Birturk**
Department of Computer Engineering, METU
Inonu Bulvarı, 06531, Ankara, Turkey
e157891, birturk@metu.edu.tr

Abstract

In this paper, we describe our work on discovering missing links in Wikipedia articles. This task is important for both readers and authors of Wikipedia. The readers will benefit from the increased article quality with better navigation support. On the other hand, the system can be employed to support the authors during editing. This study combines the strengths of different approaches previously applied for the task, and adds its own techniques to reach satisfactory results. Because of the subjectivity in the nature of the task; automatic evaluation is hard to apply. Comparing approaches seems to be the best method to evaluate new techniques, and we offer a semi-automatized method for evaluation of the results. The recall is calculated automatically using existing links in Wikipedia. The precision is calculated according to manual evaluations of human assessors. Comparative results for different techniques are presented, showing the success of our improvements. We employ Turkish Wikipedia, we are the first to study on it, to examine whether a small instance is scalable enough for such purposes.

Introduction

Wikipedia, the online encyclopedia, involves a large amount of up-to-date, reliable knowledge created collaboratively by authors from all over the world. It improves the traditional encyclopedia concept with abilities of information technologies and provides practical features like linking, categorizing, inserting info-boxes etc. for enabling users to access knowledge faster. These features also exhibits machine processable hypertext structures which are rich semantic resources forming a relevancy based network between articles. These features also offer low processing cost than would be the case for processing whole textual content. As a result, Wikipedia increasingly attracts the attention of various kinds of research areas.

Wikipedia's reliability of content is one of its most valuable properties. The main source of this reliability is the auto-control system depending on the large number of authors collaboratively creating and checking the content. There are guidelines instructing authors about the principles of high quality content. Moreover, social mechanisms across the authors like assigned responsibilities for quality assurance are established.

Although obtrusive errors like vandalism and absence of references can be usually detected by these mechanisms, small mistakes that are encountered more are often overlooked. Missing links in articles are example of such mistakes. Another problem is the large amount of articles that are relatively new and have not reached a mature state with respect to the size and quality of content. Since the authors usually do not add all links while entering or updating the content, some articles with missing links remain for a long time since an author has not revisited them. For example, textual part of the article "Kaynak Tanımlama Çerçevesi" *(eng: Resource Description Framework - RDF)* in Turkish Wikipedia is relatively long with about 1500 words. But it contains only a single link by January 2010, although it was created in March 2007. As a result, automatic discovery of missing links is important to improve the quality of articles and to help authors during editing.

The guidelines concerning the creation of links between articles suggests adding links if they are relevant to the context of the article. Also, links should increase the understandability by providing necessary navigation support. Technical terms, names of people and places should be selected as links instead of ordinary words in the language unless they are important for the context. On the other hand, irrelevant or insignificant links decrease the readability of the article. Additionally, again for readability purposes, if a concept is linked once in the article, it should not be linked again unless necessary.

In our study, our aim was to detect missing links according to the principles of these guidelines. The challenge for such a system is that it should be aware of the context and should recommend relevant links that do not harm the semantic consistency of the article (Adafre and de Rijke 2005). Two general, unsupervised approaches are taken as the solutions to this problem. First approach is trying to match the terms in the text to an article title in Wikipedia, and then filtering out unsuitable ones (Mihalcea and Csomai 2007). The second approach is selecting links from a set of relevant existing links, as in Adafre and de Rijke's (2005) study. We combine the strong features of these approaches and our improvements to reach the best results.

The rest of the paper is organized as follows. The next section surveys the related work on the area. A detailed description of our implementation follows. The following part gives experimental results and discussions of our evaluations. In the last section conclusions are summarized and the future work is indicated.

Related Work

Link discovery on Wikipedia has been studied in two different ways. First one is discovering the links for an article that does not contain any link. This task suits to need for suggesting links on a newly created Wikipedia article or suggesting Wikipedia links for non-Wikipedia articles. The other is the discovery of the missing links in existing articles of Wikipedia. The second one is more challenging because it is more subjective and involves offering a small set of new links to be added to a manually linked article.

Mihalcea and Csomai (2007) focus on discovering links in their system called Wikify. All article titles are traversed for identifying matching link candidates in the text. For each candidate a keyword extraction technique measuring its keyphraseness is applied to determine whether to identify it as a link discovery. They compare three different unsupervised keyword extraction techniques. After link discovery, a word sense disambiguation technique is applied to link to the correct article from the disambiguation pages.

Another study (Milne and Witten 2008) of the link discovery task uses a supervised technique. The authors applied a machine-learning technique by training a classifier in properties like the keyphraseness mentioned above, relatedness, generality and location. They also applied a similar approach for link disambiguation.

The missing link discovery problem has been studied only by Adafre and de Rijke (2005). They introduce the LTRank algorithm to cluster articles according to their relevance. The algorithm firstly queries the articles that link to the article to be clustered. A second query is applied to these articles' titles to determine the most relevant articles. As a result, for each article, a set of related articles are assigned which are similar to the articles linking to them. In the link discovery phase, all of the links in the relevant article set are traversed to find the matches with the text. All matches in this phase are accepted as relevant valid link discoveries, since they come from a related article.

Sorg and Cimiano (2008) studied enriching the cross-lingual link structure of Wikipedia for exploiting it to solve further cross-lingual natural language processing tasks. Schonhofen (2006) employs the titles and existing categories of articles to recommend new categories.

Key term and named entity detection is another similar application field and studies utilizing Wikipedia have been suggested. The algorithm LRT_{wiki}(Jakob, Müller, and Gurevych 2009) is an improves an older method by including domain specific knowledge obtained from Wikipedia. A novel approach for key term extraction is given in (Grineva, Grinev, and Lizorkin 2009), which exploits a graph-based solution and constructs the graph according to relatedness of articles. (Adafre, Jijkoun, and de Rijke 2007) proposes a method that, for a given concept, extracts related text snippets from related Wikipedia articles. Bunescu and Paşca (Bunescu and Paşca 2006) apply a supervised machine learning technique for detection and disambiguation of named entities using the knowledge in Wikipedia. (Toral and Munoz 2006) describes a mostly language independent method that uses Wikipedia to create and maintain gazetteers for named entity recognition.

Data Preprocessing

We use Lucene for indexing the articles. We used the XML dump of Turkish Wikipedia from April 7th 2008. We established two indexes. One is to access article content where each article corresponds to a Lucene document. This document contains titles, links and categories of articles. The redirect pages of Wikipedia are handled as different titles of the article. Therefore, the actual article is accessed when a redirect page is referred to. Non-article pages like user pages, image pages etc. are excluded from the index. Also, we observed that Turkish Wikipedia contains many links to date articles that contain lists of events that occurred on a named date. They do not contain any topic and should not be linked. Our opinion was also confirmed by the policies in English Wikipedia, so we decided to exclude date links from the index and our evaluations. Second index is an inverted index for fast calculation of values like term and link frequencies. In this index, a Lucene document is created for each article title and holds the title, article names containing the title in their text and the article names containing the title as a link.

Details of the two different approaches we have investigated are given below.

Discovering Links from Related Articles

The first method we have applied is searching links from a set of related articles. The relatedness of links is ensured by the assumption that related articles contain related links, similar to the assumption used in the study of Adafre and Rijke (2005). We do not use the LTRank algorithm and instead investigate different sources of related articles. Then, article relatedness is ensured with checking the link overlap. After determining the set of related articles; links in the related articles are matched in the text of the article for which links will be discovered (we will call it the target article). The steps can be summarized as follows:

1. Collect candidates for the related articles.
2. Calculate the score for each article according to the link overlap measure.
3. Select the best scoring articles as related articles.
4. Traverse the links in the related articles and find out the matches in the text to be the discoveries.

Details of these steps will be explained below.

Collecting Related Article Candidates

We aimed to experiment and compare different kinds of sources to determine candidates for related pages. The common point for the five different approaches we investigated

is employment of links and categories to determine relatedness. The details of these approaches are given below.

Articles in the Same Category Category structure of Wikipedia is another important semantic resource. Articles are hierarchically classified within the categories. Some of them bring entities of the same type together (e.g. 'Writers') and some of them are created according to the thematic relatedness of articles (e.g. 'Writing'). This method employs these relations and collects all articles from the target article categories.

Articles Linked by the Target Article This approach uses links as indicators for selecting related articles and collect all articles linked from the target article as candidates.

Articles Linking to the Target Article As opposite to the previous approach, articles that contain a link to the target article are collected by this method.

Index Search for Common Links This method accepts having common links as another indicator of relatedness. An index search is applied and a constant number of best scoring articles is selected. All of the links in the target article and the title of the target article are used for the search. The target article's name is boosted by a factor of 4 to support articles that directly link to the target article.

Index Search for Link Term Occurrence in the Text The previous approach searches in terms of shared links. This approach is different in one respect only; the search is applied over the text of the articles. This is because articles that contain the link terms inside text might also be related to the target article.

Selecting Related Articles

After collecting the candidate articles, each article is evaluated according to the number of links it shares with the target article. Adafre and de Rijke (2006) use this approach to detect similar sentences in cross-lingual articles. Therefore, it should be even more applicable to our problem, since the article scope contains many more links to obtain a better result. Another reason for preferring this measure is that it lies parallel to our aim; if we are looking for semantic relatedness in order to find similar links, the best measure should be similarity of sets of links. Similarity is measured using the Jaccard similarity over links. Calculation of the score for a candidate article ($score_c$) can be formulated as:

$$score_c = \frac{shared_{tc}}{n_t + n_c - shared_{tc}} \quad (1)$$

where n_t and n_c are the number of links in the target and candidate articles consecutively, and $shared_{tc}$ represents the number of shared links of the target and candidate articles.

After calculation of all scores, a constant number of best scoring articles are selected to be used in the link discovery procedure.

Locating the Discoveries

In this step, the links in the selected related articles are used as the candidates for a link discovery. The links in the category pages of the target article are also used as candidates, because category pages links to a list of articles that are related to the target article.

Firstly, the text is tokenized to its words and these tokens are iterated to find matches. Since links point to article titles, they frequently occur as multiple words. So, not only single token matches are searched, but matches of n-grams are also examined. If there are multiple matches of varying token numbers, then the longest n-gram match is preferred, since it is more probable to be the referent.

Discovering Links from Article Titles

In this section, we will describe the second method we applied. This method firstly searches crudely for all possible matching links, by accepting all of the article titles as candidates. At the end of this step, a large number of mostly irrelevant discoveries are found.

Next, the irrelevant discoveries are detected and removed. Mihalcea and Csomai (2007) approach to this step as a keyword extraction problem. For each link, a score of relatedness is calculated. For the calculation, three different keyword extraction techniques were examined: TF*IDF (Salton and Buckley 1987), Chi-square independence test (Manning and Schütze 1999) and keyphraseness. The first two are well known IR techniques. The third one is a specific measure for the link discovery problem. Keyphraseness is calculated according to the frequency of use of a word as a link in the collection. They report that the keyphraseness is the best performing technique they applied. Since it is specifically related to the linking domain we prefer to call this measure as *linkness*. The linkness value for a term (single or multiword phrase) is its probability of appearance as a link in any article. For a term T, linkness(T) it is calculated as:

$$linkness(T) = \frac{count(T_{link})}{count(T_{text})} \quad (2)$$

where $count(T_{link})$ represents the number of articles in which the term occurs as a link and $count(T_{link})$ is the number of articles in which the term occurs. Since, this measure is not effective for situations with a low number of occurrences, they do not apply filtering if the $count(T_{text})$ is less than five. The candidates with linkess value over a threshold are selected as discoveries.

Contextual Linkness Filter

We have observed that this linkness calculation is independent of the context of the article. Therefore, if the term is an important concept for the article but is not frequently linked in other articles, it loses its chance to be discovered. For example the term "yeşil" *(eng: green)* is not mostly used as link, but when it is used in the article "Gökkuşağı" *(eng: Rainbow)*, it should be marked as a link. On the other hand, a term which is a key concept that is linked in most of the articles, may be irrelevant for a specific article. For example the word "tür" *(eng: kind, species, type)* is mostly used as

link because of being frequent in articles about organisms. But for the use of it in the article "Çankırı", which is a city in Turkey, it has the meaning of "type" and should not be linked.

To overcome this problem, we have modified the linkness filter to include contextual information extracted from the first sentences of articles. These sentences are mostly definition sentences about the concept in the article. They usually contain important links to critical concepts about the domain of the article. Also, different names identifying the article are given inside three inverted commas (''') to mark it as stressed. For example, the first sentence of the article "Alfa Centauri" is:

'''Alfa Centauri''', [[Güneş]]'e en yakın [[yıldız sistemi]].

(eng: '''Alpha Centauri''' is the closest [[star system]] to the [[Sun]].)

The sentence contains links to related concepts "Sun" and "star system". Only the title "Alpha Centauri" is stressed for this sample. We have observed that the links and stressed terms in the first sentence carry satisfying and focused contextual information for the article. Therefore, we have employed these terms to improve the linkness measure. To achieve this, we have changed the query which is done to find the values in the linkness formula. Instead of searching for every occurrence of the terms in whole Wikipedia, we have searched only for those occurrences together with at least one of the words extracted from the first sentence. For example, the Lucene query for the documents containing the term "yıldız" *(eng:star)* is;

internal_link:"yıldız"

for the linkness filter, in order to add the context information the query is transformed to:

(text:"Alfa Centauri" OR text:"Güneş" OR text:"yıldız sistemi") AND internal_link:"yıldız"

By calculating both $count(T_{link})$ and $count(T_{text})$ values in this way the term is evaluated as more related to the context of the article. This technique also resembles human behavior where inexperienced authors use it as an important reference in checking similar pages to see whether they contain such links.

Filtering the Discoveries from Related Articles

Two methods we have examined have different characteristics by their approach to the relatedness verification of candidate link discoveries. One of them selects candidates from related articles and the other employs a filtering approach. The former one is a convincing solution but since it does not evaluate the candidates specifically, this general solution seems very close to allow some irrelevant discoveries. As a result, it can not guarantee high accuracy. On the other hand, the latter approach specifically evaluates each discovery candidate. Although it seems to be more accurate for the decision of relevancy, since this method accepts all matching article titles as a candidate, there are much more candidates to evaluate compared to the former approach. Therefore, it is also prone to allow some irrelevant candidates to be accepted by weaknesses of the filtering approach. As a result, we decided to apply a combination of both methods to mutually eliminate weaknesses of each other. Firstly, candidate discoveries are selected from related articles. Then, filtering is performed to detect irrelevant ones. By this way, both the number of candidates is controlled and all candidates are specifically checked to reach a balanced level of completeness and accuracy of the results.

Evaluations

The general approach in automatic evaluation of link discovery problem is, consequently: removing all links from the test articles, running the discovery procedure, and measuring the recall and precision by comparing the discoveries with pre-existing links. This approach firstly diverges from the missing link discovery problem since the missing links are not considered. Human judgment is needed to resolve these issues. Additionally, since our system aims missing link discovery, it exploits from the existing links in the articles. Therefore, it would not be feasible to evaluate by removing existing links. We suggest a different automated approach for measuring the recall of the system. The precision is measured by the evaluations of human assessors. Details of evaluations are given below.

For measuring recall, naturally, the set of the missing links expected to be discovered should be determined. As mentioned above using existing links is not feasible. On the other hand, it is very hard to manually determine this set because there are plenty of candidate terms in the articles. Extracting all of them is not realizable and might heavily vary between people. As a result, we preferred a different method which indirectly measures the recall. First, a single randomly selected link is removed from an article. Then, link discovery is applied to this article, and discovery of the removed link is expected. Totally, for 2000 random articles, 2000 different link discoveries are identified as expected discoveries. Consequently, the ratio of articles meeting this expectation is accepted as recall value.

Adding a new link is not an objective task even for humans, since decisions about the suitableness of the link may vary from person to person. Therefore, automatic precision evaluation is not realistic. Each article is evaluated by three human assessors who are experienced Internet (and Wikipedia) users with graduate level education. For our evaluations, we randomly selected 40 articles. All different link discovery approaches introduced above are executed and all discoveries suggested by them are collected. The assessors were given these suggestions and the text as discovered links highlighted. Participants used the Sandbox facility of Wikipedia which allows users to preview the text in Wikipedia format as an actual Wikipedia page. The assessors were informed of the general principles of linking in Wikipedia and asked to evaluate according to these principles, their insight to see the suggestions as a link and comparison with the linking approaches in similar articles involving the English counterparts.

Method	Recall	Precision	F-Measure
R.A. / No filtering	89.3	66.8	76.4
R.A. / Linkness	82.3	**83.1**	82.7
R.A. / Cont. Linkness	84.4	82.5	**83.5**
A.T. / Linkness	91.4	68.9	78.6
A.T. / Cont. Linkness	**93.6**	70.6	80.5

Table 1: Evaluation results for missing link discovery methods with different filters *(R.A.: Discovery from related articles, A.T.: Discovery from the article titles)*

Method	Recall	Precision	F-Measure
Same category	35.4	86.5	50.2
Linked by target art.	33.0	80.8	46.9
Linked to target art.	28.1	80.4	41.6
Common Links	78.5	**87.0**	82.5
Common Terms	80.4	84.0	82.1
All combined	**84.4**	82.5	**83.5**

Table 2: Evaluation results for different techniques we have applied for related article selection

Results and Discussions

The results of the different methods are given in Table 1. The best results according to F-Measure are gained applying the contextual linkness filter to discoveries from related articles. For related article discovery, all approaches used to collect related articles are combined to obtain the best results, as will be explained in next section.

Filtering brings a serious increase in the achievement of the related article method. Especially precision increases and comes to a balanced point through the relatively small decrease in recall. The contribution of filtering shows that not all links brought from related articles are actually related to the target article. This might also be interpreted as an indicator of a need for more precise selection of related articles.

It is seen that discovery from article titles with filtering approach does not give balanced results by recall and previsions values as the first approach. This result points to the fact that linkness filtering is not competent enough to remove the majority of irrelevant suggestions when applied as single expedient. A more aggressive filter might be developed to improve the method. Another disadvantage of the method is its poor operating performance, which is caused by attempting to match each of about 130,000 titles with each token of text.

The contextual linkness filter also brings a considerable improvement compared to the linkness filter. The increase in recall values is remarkable. This increase shows that elimination of some relevant links by the linkness filter is prevented "yeşil" *(eng: green)* example. There is a higher improvement in the discovery from article title method. This can be explained by the implicit filtering by selecting

Evaluation of Related Article Retrieval Techniques

Table 2 gives a comparison of techniques applied in order to select related articles. The best achieving method of the previous section is used for this experiment. The results can be separated into two main groups which are obviously parallel to the characteristic features of the techniques. The first three techniques explore related articles with a direct link or category relationship. For this group, the results seem satisfactory in terms of precision but the recall of these techniques is very low. On the other hand, techniques in the second group apply an index search on whole articles based on the link overlap with the target article. For this group, while precision is preserved, the recall increases to the same level. The difference of recall for the two groups can be associated with the number of articles they return. Techniques in the first group evaluate a small set of articles because direct relationship constraints, although the ones in the second group make the selection from a broader range of articles. On the other hand, for second group, preservation of the precision in spite of increase in article number indicates the success of related article scoring and filtering applied.

An appreciable result is the precision obtained by the articles collected from same categories. This can be interpreted as a clue to expose the fact that category relationship might be a more important semantic information that reveals the relatedness of articles compared to the link relationship.

Another interesting result is the parallelness of the behaviour of second and third approaches. These two techniques exploit, consecutively, the outgoing links and the incoming links of the target article. The similarity of results might indicate incoming and outgoing link sets of articles highly intersect. Therefore, it may point to the article clusters in Wikipedia article space, which are classified by the density of interconnections by means of links.

The results show that variety and number of articles considered increase the success. Therefore, as a most comprehensive alternative, we experimented with combining all these related article retrieval techniques and the system achieved the best results.

Evaluation of the Discovery Amount

In this section, we will discuss the link gain obtained by discovery of missing links. The 2,000 articles used in the recall experiments initially contained a total of 34,508 links, which corresponds to 17.3 links per article. Table 3 lists the discovery amounts for each method. The first column gives the number of total discoveries by method. The second one contains the normalized number of discoveries, obtained by multiplying the discoveries with the precision of the method, since the number of correct suggestions is more meaningful. The last column gives the increase ratio of normalised link numbers by applying the discovery. The results show that a considerable enrichment of the links might be gained by applying link discovery. The most remarkable finding from the results is the success of the methods with high recall values despite their low precision and thus the F-measure values. According to these results, the system might be employed as a suggestion system where the suggestion will be accepted or ignored by the user. This usage might allow omitting the precision loss of 15% and preferring high number of discov-

Method	#Discoveries	#Discoveries Normalized	Increase of links (%)
R.A. / No filtering	13,432	8,973	26.0
R.A. / Linkness	7,156	5,946	17.2
R.A. / Contextual Linkness	7,177	5,921	17.2
A.T. / Linkness	**14,173**	**9,765**	**28.3**
A.T. / Contextual Linkness	13,240	9,347	27.1

Table 3: Evaluation results for different techniques we have applied for related article selection

eries. On the other hand, other methods might be improved by configuring some constant and threshold values.

Conclusions and Future Work

In this paper, we explained our work on missing link discovery task on Turkish Wikipedia, using a method which combines different approaches. We suggest a semi-automated approach for evaluation, where this was completely manual in previous studies. The comparison of results show that our improvements bring a considerable contribution to the field.

To improve our system, better ways for implementing contextual linkness filtering can be explored. For example, the contextual information carried by the categories or infoboxes might be considered. Also, we have examined that first sentences of some articles do not provide satisfactory contextual information because of lack of necessary links. As a result, for such articles, the number of query results is less than five and the linkness filtering is performed. Therefore, a broader section like the first paragraphs might be preferred instead of the first sentences.

Satisfactory results of our system prompt us to develop a suitable GUI application to serve Wikipedia users. This application might be thought of as a suggestion system for the reader and editor of an article. The user can see the suggestions both in the article text and as a list from which the user can manipulate the results to filter out unnecessary ones. Also statistics can be more easily collected about our suggestions to measure precision more precisely.

Wikipedia allows the cross-lingual matching of articles which results to a rich multilingual resource. Exploiting the cross-lingual link structure of Wikipedia seems as to be a promising approach for various kind of multilingual tasks. It might also be benefited for our task by considering the existing links in other language versions of the articles.

Another future study might aim at determination of the best location for inserting links when there are more than one occurrences of the same term. Wikipedia guidelines suggest linking as early as possible but also considering the local (e.g. sentence or paragraph scope) relevancy of the term. Therefore, the system should select the most leading location which is contextually suitable for inserting the link.

References

Adafre, S. F., and de Rijke, M. 2005. Discovering missing links in wikipedia. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, 90–97. New York, NY, USA: ACM.

Adafre, S. F., and de Rijke, M. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* 62–69.

Adafre, S. F.; Jijkoun, V.; and de Rijke, M. 2007. Fact discovery in wikipedia. In *2007 IEEE/WIC/ACM International Conference on Web Intelligence*.

Bunescu, R. C., and Paşca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics.

Grineva, M.; Grinev, M.; and Lizorkin, D. 2009. Extracting key terms from noisy and multi-theme documents. In *18th International World Wide Web Conference (WWW2009)*.

Jakob, N.; Müller, M.-C.; and Gurevych, I. 2009. Lrtwiki: Enriching the likelihood ratio test with encyclopedic information for the extraction of relevant terms. In *Proceedings of the WikiAI 09 - IJCAI Workshop: User Contributed Knowledge and Artificial Intelligence: An Evolving Synergy*.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*. New York, NY, USA: ACM.

Milne, D., and Witten, I. H. 2008. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, 509–518. New York, NY, USA: ACM.

Salton, G., and Buckley, C. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.

Schonhofen, P. 2006. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 456–462. Washington, DC, USA: IEEE Computer Society.

Sorg, P., and Cimiano, P. 2008. Enriching the crosslingual link structure of wikipedia - a classification-based approach -. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (WikiAI'08)*.

Toral, A., and Munoz, R. 2006. A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. *EACL 2006*.