Enriching the Crosslingual Link Structure of Wikipedia - A Classification-Based Approach -

Philipp Sorg and **Philipp Cimiano**
Institute AIFB, University of Karlsruhe,
D-76128 Karlsruhe, Germany
{sorg,cimiano}@aifb.uni-karlsruhe.de

Abstract

The crosslingual link structure of Wikipedia represents a valuable resource which can be exploited for crosslingual natural language processing applications. However, this requires that it has a reasonable coverage and is furthermore accurate. For the specific language pair German/English that we consider in our experiments, we show that roughly 50% of the articles are linked from German to English and only 14% from English to German. These figures clearly corroborate the need for an approach to automatically induce new cross-language links, especially in the light of such a dynamically growing resource such as Wikipedia. In this paper we present a classification-based approach with the goal of inferring new cross-language links. Our experiments show that this approach has a recall of 70% with a precision of 94% for the task of learning cross-language links on a test dataset.

Introduction

From the natural language processing perspective, a very interesting feature of Wikipedia, besides the overwhelming amount of content created daily, is the fact that information is linked across languages. This is accomplished via so called *cross-language links* mapping articles in one language to equivalent articles in another language. Obviously, such links have a natural application in cross-lingual natural language processing, e.g. in machine translation, cross-lingual information retrieval, projection of information across languages, alignment etc.

However, if natural language processing applications are expected to exploit the cross-language link structure, it should have enough coverage. A first analysis of the coverage for one language pair, i.e. German/English, shows that only a percentage of the pages are connected via such cross-language links. Thus, in this article we present a novel method for learning additional cross-language links in order to enrich Wikipedia. The method is based on a classification-based approach which classifies pairs of articles of two different languages as connected by a cross-language link or not. The features used by the classifier range from a simple calculation of the edit distance between the title of the articles over word overlap counts through to more complex link patterns as features. The results of the approach are encouraging as they show a prediction recall of 70% with a precision of 94% on the task of finding the corresponding article in another language.

Given our encouraging results, we have started processing the German Wikipedia. We will provide the additional links as download for the research community.

The remainder of this article is structured as follows: First we motivate our approach and analyse the availability of cross-language links for the language pair German/English. The core of our approach is explained and its assumptions motivated quantitatively. Then the classification-based approach and the used features are described in detail. Afterwards we present the experimental results. Before concluding we discuss related work.

Motivation

As stated above, the cross-language links in Wikipedia can be used for various cross-lingual NLP tasks. But in order to be able to perform these tasks, the cross-language link structure should be consistent and needs to have enough coverage.

In the context of this paper, we have chosen the German and English Wikipedia and computed statistics about the German/English cross-lingual link structure to get a clear picture about its consistency and coverage.

These findings motivate our approach to learning new cross-language links in Wikipedia.

Statistics about German/English Cross-Language Links

For the analyis of the German and English Wikipedia we counted the absolute number of articles in the English and German Wikipedia, the number of cross-language links between the English and German Wikipedia and classified these links into bidirectional links, links with no backlink and links with backlink to another article[1]. Articles are defined as Wikipedia pages that are not redirect[2] pages and are in the default namespace. Cross-language links ending

[1] E.g.: "3D rendering" to "3D-Computergrafik" back to "3D computer graphics"

[2] Redirect Pages are used to disambiguate different surface forms, denominations and morphological variants of a given unambiguous NE or concept to a unique form or ID.

	Articles	Cross-Language Links		
English Wikipedia	2,293,194	English → German (EN2DE)	321,498	14.0%
German Wikipedia	703,769	German → English (DE2EN)	322,900	45.9%

	EN2DE C.-L. Links		DE2EN C.-L. Links	
Bidirectional links	303,684	94.5%	303,684	94.1%
No backlink	9,753	3.0%	12,303	3.8%
Backlink to another article	7,845	2.4%	6,132	1.9%

Table 1: Statistics on the English (October 18, 2007) and German (October 09, 2007) Wikipedia Corpus.

in redirect pages were resolved to the corresponding article. All the results of the analysis are presented in Table 1.

The results show that only a small fraction (14%) of articles in the English Wikipedia is linked to articles in the German Wikipedia. The fraction of German articles linked to English articles is much bigger, but with 45.9% it is still less than half of all articles in the German Wikipedia. For some articles there may not be a corresponding article in another language due to the local context of the specific country. But as this is probably not the case for half of the German Wikipedia, there is still a big margin to learn new meaningful cross-language links.

As the fraction of bidirectional links is around 95% in the English and German Wikipedia, the consistency of cross-language links seems to be good. This motivates to use them in a bootstrapping manner to find new cross-language links.

Chain Link Hypothesis

One problem in learning new cross-language links between the German and English Wikipedia is the huge number of pages (see number of articles in Table 1). It will surely not be possible to use a classifier on all article pairs, such that a preselection of candidate articles seems appropriate.

In order to preselect a number of relevant articles, we rely on the *chain link hypothesis*. This hypothesis builds on the notion of a chain link:

Definition 1 *For two Wikipedia databases WP_α, WP_β with corresponding languages α, β, a* **chain link (CL)** *between two articles $A_\alpha \in WP_\alpha$ and $A_\beta \in WP_\beta$ is defined as the following link structure:*

$$A_\alpha \xrightarrow{pl} B_\alpha \xrightarrow{ll} B_\beta \xleftarrow{pl} A_\beta$$

with $B_\alpha \in WP_\alpha$ and $B_\beta \in WP_\beta$. Pagelinks between articles are displayed as $\xrightarrow{pl}$ and cross-language links between articles in different languages as $\xrightarrow{ll}$. The articles B_α and B_β are called **chain link intermediate articles (CLIA)**.

An example for such a chain link between a German and an English article is visualized in Figure 1. The article "Horse" ($= A_\alpha$) in the English Wikipedia is connected through the displayed chain link to the article "Hauspferd" ($= A_\beta$) in the German Wikipedia. The articles "Mammal" ($= B_\alpha$) and "Säugetiere" ($= B_\beta$) are CLIAs of this chain link that is formed by the pagelink from "Horse" to "Mammal", the cross-language link from "Mammal" to "Säugetiere" and the pagelink from "Hauspferd" to "Säugetiere".

Based on chain links we formulate the chain link hypothesis, the basic hypothesis for the selection of candidates for new cross-language links: *Every article is linked to its corresponding article in another language through at least one chain link.*

In order to empirically verify the plausibility of the above hypothesis, we have generated the RAND1000 dataset containing 1000 random articles of the German Wikipedia with existing cross-language links to the English Wikipedia. For all articles in the RAND1000 dataset, we have checked if the hypothesis is indeed fulfilled. For an article A_α in the dataset, connected to the article A_β in the English Wikipedia through a cross-language link, this means that we have to check if A_β is in the *candidate set* $\mathcal{C}(A_\alpha)$. The candidate set of an article A_α are all articles that are connected to A_α through at least one chain link.

However, we noticed that on average the number of articles in each candidate set is still to big. In case of the RAND1000 dataset the mean size of the candidate set is 153,402. This means that an approach to find a cross-language link for an article A, that considers all articles in $\mathcal{C}(A)$ as potential candidates, can be very expensive from a computational point of view.

Thus, we also consider a reduction of the number of candidates. Therefore we define the *support* of a candidate C in respect to an article A in the dataset as the number of existing chain links between A and C. For each article A, we limit the number of candidates to less than 1000 by requiring a minimal support via an appropriate threshold. For each article, we call the set of these candidates the *restricted candidate set* $\mathcal{C}'(A)$, which is restricted by definition to at most 1000 candidates. The following table contains the percentage of articles for which the chain link hypothesis is fulfilled using the full candidate set and the restricted candidate set:

	Percentage
Full candidate set	95.7 %
Restricted candidate set	86.5 %

This means that for 95.7% of pages in the RAND1000 dataset the corresponding article in the English Wikipedia is included in the full candidate set. For the restricted candidate set the hypothesis holds for 86.5% of the pages. With respect to the decrease in performance time by processing at most 1000 instead of 153,402 articles on average, this decrease in terms of best case accuracy seems a good trade-off.

Overall, the chain link hypothesis is therefore strongly supported by this evaluation on the RAND1000 dataset,

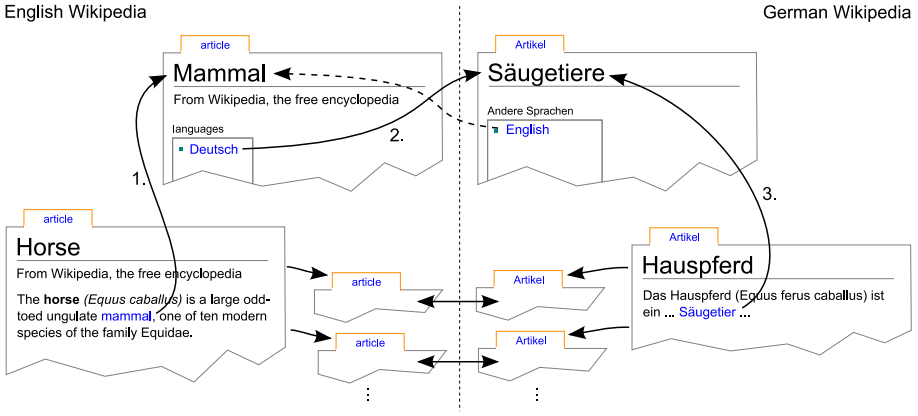

Figure 1: Visualisation of a chain link that is used to find candidate pages for new cross-language links. 1. is a pagelink in the English Wikipedia, 3. a pagelink in the German Wikipedia and 2. a cross-language link from the English to the German Wikipedia.

even after restricting the candidate set to at most 1000 candidates for each article. Based on these findings the usage of the chain link hypothesis to restrict the set of candidate articles for new cross-language links seems to be promising. The approach presented in the remainder of this paper strongly relies on the chain link hypothesis as a feature for training a classifier which is able to predict whether a pair of articles in two languages (German/English in our case) should be connected via a cross-language link or not. Having motivated our approach and the underlying hypothesis empirically, we describe the approach in more detail in the next section.

Classification-based Approach

The main idea behind our approach to learn new cross-language links is to train a classifier which is able to predict whether a pair of articles (A, B) where $A \in \text{WP}_\alpha$ and $A \in \text{WP}_\beta$ should be cross-linked. As it is not feasible to apply the articles to all pairs in two languages, for the article A we only consider the candidates $\mathcal{C}'(A) \subset \text{WP}_\beta$ as potential cross-links.

As classifier we used the popular Support Vector Machine (SVM) implementation *SVMlight* by Joachims (1999) with a linear kernel function. The classifier is trained with a number of features which we describe below in more detail. Features are defined on article-candidate pairs $(A, C) \in \text{WP}_\alpha \times \text{WP}_\beta$ with $C \in \mathcal{C}'(A)$ and are based on different information sources. Based on our chain link hypothesis, the support of C in respect to A, defined above as the number of chain links between these articles, is considered and the link structure of the CLIAs is exploited. In addition, the categories of A and C are also considered. As categories are also linked by language links it is possible to align categories across languages. Finally, we also use simple features based on the title and text of articles.

Feature Design

The features can be classified into two classes: graph-based and text-based features. The former are based on different link types in Wikipedia, i.e. pagelinks, category links and language links. The latter are based on the title and text of the Wikipedia articles.

For the definition of graph-based features, we need to define the number of inlinks of an article. Inlinks of an article $A \in \text{WP}_\alpha$ are pagelinks from another article that are targeted to A. The number of inlinks of A is therefore defined as $\text{INLINKS}(A) = |\{B \in \text{WP}_\alpha \mid B \xrightarrow{pl} A\}|$.

For the definition of text-based features we need to introduce the *Levenshtein Distance* (Levenshtein, 1966), a string metric that is based on the edit distance between two strings. The edit distance is defined as the minimal number of insert, delete and replace operations that is needed to transform one string to another. We use a version of the Levenshtein Distance that is normalized by the string lengths.

As described above, features are based on article-candidate pairs. In the following, we will refer to the article as A and to the candidate as C with $C \in \mathcal{C}'(A)$.

Graph-based Features:

Feature 1 (*Chain Link Count Feature*)
This feature is equal to the support of C with respect to A.

Feature 2 (*Normalized Chain Link Count Feature*)
This feature is the value of Feature 1 normalized by the support threshold that was used to restrict the candidate set for A.

Featureset 3 (*Chain Link Inlink Intervals*)
Given an article A and a candidate C we compute all the chain links between these and classify them into 20 intervals defined over the number of inlinks that the CLIA of the chain link has, i.e. we classify a CLIA B into a bucket according to the value $\text{INLINKS}(B)$. Thus, we yield 20 features corresponding to the 20 intervals.

The motivation behind this classification is the assumption that chain links containing CLIAs with fewer inlinks are probably more specific for a topic and therefore more important for the choice of the correct article.

By classifying the chain links into different classes using the number of inlinks of the CLIAs this assumption can be explored by the classifier.

Feature 4 (*Common Categories Feature*)
The output of this feature is the number of common categories of two articles in different languages. Common category means that both articles are member of categories that are linked through existing cross-language links.

Feature 5 (*CLIA Graph Feature*)
This feature is based on a similarity measure on graphs. Given two graphs G_α and G_β on the same set of vertices, the similarity is defined as the number of common edges of these graphs normalized by the number of vertices. For the article A and the candidate C, the graphs G_α and G_β are defined on the set of chain links between A and C as vertices. Edges in G_α between two chain links exist if the CLIAs in WP_α of these chain links are linked by a pagelink in WP_α. Analogous, edges in G_β between two chain links exist, if the CLIAs in WP_β of these chain links are linked by pagelink in the WP_β. The value of this feature is the value of the defined similarity measure between G_α and G_β.

Text-based Features:

Feature 6 (*Editing Distance Feature*)
The output of this feature is the normalized Levenshtein Distance on the titles of the candidate articles pair.

Feature 7 (*Text Overlap Feature*)
This feature computes the text overlap between the text of the candidate article pair. To remain independent of lexical resources there is no translation involved. This feature will be useful if the articles for example share many named entities.

Evaluation

The evaluation is based on the RAND1000 dataset. As described above, this dataset consists of 1000 articles of the German Wikipedia with an existing language link to an article in the English Wikipedia.

In the following we first analyse this dataset to get a lower bound for the classification experiment. Afterwards we describe the experimental setup. Finally we present further results on articles without an existing language link.

Baseline

In order to find a lower bound for recall, we define a simple method to find language links by matching the titles of articles. The recall of this method on the RAND1000 dataset is equal to the percentage of articles that are linked to English articles with identical title. The analysis of the RAND1000 dataset showed that 47.0% of the articles in this dataset are linked to English articles with identical title. The reason for this high share is the fact that many Wikipedia articles describe named entities and thus have the same title in different languages. This value defines a lower bound for recall as this method to find new language links is very simple and straightforward. Any other method should exceed the results of this baseline.

Evaluation of the RAND1000 Dataset

In the experiments we used a random 3-1-split of the RAND1000 dataset. The first part containing 750 articles was used for training the classifier. The remaining 250 articles were used for the evaluation.

In order to evaluate the correctness of our approach, we consider the TOP-k with $k \in \{1..5\}$ candidates with respect to a ranking determined on the basis of the example's (directed) distance from the SVM-induced hyperplane. The larger the distance, the higher is the classifier's certainty that it is a positive example. Hereby, we do not distinguish between positive examples, which have a positive distance to the margin and negative examples, which have a negative one. Thus, it is possible that in absence of positive examples, also negative examples appear at the top of the ranking.

TOP-k Evaluation As quality measure for the TOP-k evaluation we defined TOP-k-Accuracy as the share of articles in the test set for which the correct linked article was part of the k top ranked candidates [3].

One important problem in learning the classifier is the discrepancy between positive and negative training data. For every article in the training set there exists at most one positive example but up to 1000 negative examples. Using all this training data will most likely yield a classifier which always predicts new examples to belong to the majority class, the negative examples in our case (compare Provost (2000)). In order to avoid this, the training data has to be balanced, such that we only used a portion of the negative examples in order to train the classifier. For each article in the training set, 2, 5 and 10 negative examples were randomly selected and together with all positive examples were used to train the classifier.

To be able to measure the quality of different features we trained the classifier with different feature sets. First we used only the *Chain Link Count Feature*. In this case candidate articles with a higher number of chain links are ranked higher. The purpose of the results of this experiment is to support the hypothesis that chain links are a prominent clue for language links between articles. In another set of experiments we used the text features only as well as the graph features only, respectively. This allows to assess the influence of each of the different features. Finally, the classifier was trained with all features to find out if it is indeed worth considering all the features together.

Results of the experiments are shown in Table 2. The table shows the accuracy with respect to the top k candidates with varying sizes of negative examples considered. Overall it seems that the choice of negative/positive ratio does not have a strong impact on the results. However further experiments showed that using too many negative examples leads to learning a trivial classifier as is the case when using the chain link count feature alone for a negative/positive ratio of 10:1. A negative/positive ratio of 5:1 seems therefore reasonable and will be used in the further experiments described below. The accuracy of the prediction, when considering only the chain link features, ranges from 42.4% (TOP-

[3] TOP-k-Accur. $= \frac{|\{A_\alpha \in \text{RAND1000} \mid \exists A_\beta \in \text{TOP-}k(A_\alpha) : A_\alpha \xrightarrow{ll} A_\beta\}|}{|\text{RAND1000}|}$

Ratio -/+ data	Feature selection		TOP-k-Accuracy TOP-1	TOP-2	TOP-3	TOP-4	TOP-5
2:1	1	(*Chain Link Count Feature*)	42.4%	51.2%	60.0%	62.8%	64.8%
	6-7	(*Text features*)	68.4%	71.2%	73.6%	74.8%	75.2%
	1-5	(*Graph features*)	54.8%	64.0%	68.4%	70.8%	72.0%
	1-7	(*All features*)	71.2%	76.0%	78.8%	79.6%	80.0%
5:1	1	(*Chain Link Count Feature*)	42.4%	51.2%	60.0%	63.2%	64.8%
	6-7	(*Text features*)	68.8%	72.8%	74.4%	74.8%	75.2%
	1-5	(*Graph features*)	55.2%	62.8%	67.6%	68.8%	70.0%
	1-7	(*All features*)	74.8%	**79.2%**	**79.2%**	80.0%	80.4%
10:1	1	(*Chain Link Count Feature*)	0.0%	0.4%	0.4%	0.4%	0.4%
	6-7	(*Text features*)	68.4%	72.4%	74.4%	74.8%	75.2%
	1-5	(*Graph features*)	55.6%	62.4%	67.6%	69.2%	70.4%
	1-7	(*All features*)	**76.0%**	78.4%	78.8%	**80.4%**	**81.2%**

Table 2: Results of the evaluation on the RAND1000 dataset. The first column describes the negative/positive ratio of training examples. The second column describes the feature selection. TOP-k-Accuracy is used as quality measure.

1) to 64.8% (Top-5). Considering the TOP-1 results, we conclude that the classifier trained with the chain link features alone does not improve with respect to our baseline of 47% consisting of considering articles with the same title. The text and graph features alone yield results in terms of accuracy between 68.8% (TOP-1) and 75.2% (TOP-5) as well as 55.2% (TOP-1) and 70% (TOP-5). Both types of features thus allow to train a classifier which outperforms the naive baseline. Considering all features yields indeed the best results, leading to a prediction accuracy of between 76% (TOP-1) and 81.2% (TOP-5). Thus, we have shown that the number of chain links seems to be the weakest predictor for a cross-language link between two articles in isolation. When considering all features, the results certainly improve, showing that the number of chain links crucially contributes towards making a good decision in combination with the other features used. As we use articles from the English and German Wikipedia as test data, the text features based on text overlap and similarity are strong features with good classification results. However, even using only graph features, thus operating on a completely language-independent level, the results exceed the trivial baseline. Thus, we can assume that our method will produce reasonable results for any language pair of Wikipedia, even if they use different alphabets or if their languages are from different linguistic families. In those cases the text based features will play a negligible role.

Best Candidate Retrieval In order to automatically induce new language links, it is necessary to choose exactly one candidate for each source article and to decide whether this candidate is the corresponding article or not. To achieve these goals we define Best Candidate Retrieval as a modified TOP-1-Retrieval which selects that positive example which has the largest (positive) margin with respect to the SVM-induced hyperplane. This differs from the TOP-k retrieval introduced above in that the latter one performs a ranking on the basis of distance to the discriminating hyperplane, also considering examples on the "wrong side" of the plane. The Best Candidate Retrieval produced the following results:

Ratio -/+ data	Feature selection	Recall	Precision
10:1	All features	69.6%	93.5%

The recall of this experiment is 22.6% higher than the lower bound. Due to the preselection of candidates, the maximum recall is 86.5%. It is important to note that a recall of 69.6% means that we find 80% of the cross-language links that can be found at all given our preselection on the basis of the candidates' support.

As our aim is to learn correct links, high precision is a requirement. In this sense our approach seems very promising as new language links are learned with high precision of 93.5% and a reasonable recall. It could therefore be used to enrich the Wikipedia database with new language links.

Learning New Language Links

In order to test our approach in a "real scenario" with the aim of inducing new cross-language links instead of merely reproducing the existing ones, we have started processing the German Wikipedia, considering all those articles which do not have an existing cross-language link to the English Wikipedia. As our algorithms are still in a state of research prototype and as we do not have the computational power it was not possible for us to process all of these articles. Because of that we defined a relevance ranking on the articles based on the number of incoming pagelinks and sorted the articles according to this ranking. We processed the first 12,000 articles resulting in more than 5,000 new cross-language links according to best candidate retrieval as described above. The file with the results can be downloaded from our website [4].

The first 3,000 links were manually evaluated. As for 2,198 links the titles were identic, these links were assumed to be correct. The remaining 802 links were evaluated by 3 independent persons. They annotated them as correct links, wrong links and links between related articles. The annotator's correlation was reasonable with a Pearson's product-moment correlation coefficient between 0.80 and 0.84. As

[4] http://www.aifb.uni-karlsruhe.de/WBS/pso/learned_language_links_(German-English).tsv

overall result we got a precision of 81.9% for learning correct cross-language links. Further, the manual evaluation showed that 92.2% of the links connected at least related articles. These are very satisfactory results.

Related Work

Several authors have considered exploiting the cross-language link structure of Wikipedia for cross-lingual natural language applications. Adafre & de Rijke (2006) have for example used the language links to find similar sentences across languages. They have also considered discovering additional links in Wikipedia (Adafre & de Rijke, 2005). However, the latter approach only aimed to add additional links to articles within the same language. Based on earlier results showing that multilingual resources such as EuroWordNet can be used for cross-language Question Answering (see Ferrández & Ferrández (2006)), the same authors have shown that using Wikipedia in addition to EuroWordnet can even improve results on the cross-language Question Answering task (see Ferrández *et al.* (2007)). The reason is that Wikipedia contains more complete and up-to-date information about named entities. Other researchers have shown that the multilingual information in Wikipedia can be successfully used to improve a cross-lingual information retrieval system (see Schönhofen *et al.* (2007)). Very recently, Wentland et al. have considered the cross-lingual link structure of Wikipedia to extract multilingual contexts for named entities contained in Wikipedia. Such multilingual contexts can then be used for the disambiguation of named entities across multiple languages (Wentland *et al.*, 2008).

To the best of our knowledge, we are not aware of any approach aiming at finding new cross-language links in Wikipedia. However, such an approach would be beneficial for all of the cross-lingual applications mentioned above.

Conclusion

We have presented an approach for inducing new cross-language links for Wikipedia. Such links can be beneficial for any cross-language natural language processing task exploiting Wikipedia as source of multilingual knowledge. Our approach works for language pairs for which a number of cross-language links are already available and bootstraps on the basis of these existing links to discover new ones. No other lexical resources are needed. We have shown that our method achieves a satisfactory level of recall of around 70% and a high level of precision of around 94%. These results hold for that subset of Wikipedia pages which have been already linked across languages. To get a better estimate of the accuracy of the approach, we started to induce new cross-language links for articles in the German Wikipedia without a cross-language link to an article in the English Wikipedia and manually evaluated the first 3000 learned links. The results of this evaluation show that around 82% of the links are correct and that 92% of the links connect at least related articles. For a productive use of our methods, the algorithm needs to be optimized from a computational point of view. On a standard dual core computer using a MySQL database, the extraction of the candidates for the RAND1000 dataset and the computation of all features took 26 hours. Most expensive are the selection of candidates and the computation of graph features. The computational costs could therefore possibly be reduced by optimizing the database and by identifying the most relevant graph features. However this remains future work.

Acknowledgements

This work was funded by the Multipla project sponsored by the German Research Foundation (DFG) under grant number 38457858 and by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

Adafre, S., and de Rijke, M. 2005. Discovering missing links in wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*.

Adafre, S., and de Rijke, M. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the EACL Workshop on New Text, Wikis, Blogs and Other Dynamic Text Sources*.

Ferrández, S., and Ferrández, A. 2006. Cross-lingual question answering using inter lingual index module of eurowordnet. In *Advances in Natural Language Processing. Research in Computing Science*.

Ferrández, S.; Toral, A.; Ferrández, .; Ferrández, A.; and Muñoz, R. 2007. Applying wikipedia's multilingual knowledge to cross-lingual question answering. In *Proccedings of the 12th International Conference on Applications of Natural Language to Information Systems*.

Joachims, T. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*.

Provost, F. 2000. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI Workshop on Imbalanced Data Sets*.

Schönhofen, P.; Benczúr, A.; Bíró, I.; and Csalogány, K. 2007. Performing cross-language retrieval with wikipedia. In *Working Notes of the Cross Language Retrieval Forum (CLEF)*.

Wentland, W.; Knopp, J.; Silberer, C.; and Hartung, M. 2008. Building a multilingual corpus for named entity disambiguation, translation and transliteration. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. To appear.