Finding Missing Cross-Language Links in Wikipedia

Carlos Eduardo M. Moreira, Viviane P. Moreira

Instituto de Informática-UFRGS, Brazil
{cemmoreira,viviane}@inf.ufrgs.br

Abstract. Wikipedia is a public encyclopedia composed of millions of articles written daily by volunteer authors from different regions of the world. The articles contain links called cross-language links which relate corresponding articles across different languages. This feature is extremely useful for applications that work with automatic translation and multilingual information retrieval as it allows the assembly of comparable corpora. Thus, it is important to have a mechanism that automatically creates such links. This has been motivating the development of techniques to identify missing cross-language links. In this article, we present CLLFinder, an approach for finding missing cross-language links. The approach makes use of the links between categories and of the transitivity between existing cross-language links, as well as textual features extracted from the articles. Experiments using one million articles from the English and Portuguese Wikipedias attest the viability of CLLFinder. The results show that our approach has a recall of 96% and a precision of 98%, outperforming the baseline system, even though we employ simpler and fewer features.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: classification, cross-language links, similarity functions, wikipedia

1. INTRODUCTION

Wikipedia is a large public and collaborative encyclopedia composed of millions of articles. These articles are created and modified on a daily basis by a community of volunteer authors and editors. These articles are written in several languages, thus Wikipedia has become a valuable repository of multilingual information.

Conceptually, a Wikipedia article is represented by a page that has information about the entity it describes. A *Cross-language Link* (CLL) (also known as interlanguage link) is a very interesting feature, which allows navigating to corresponding versions of the article written in other languages. Figure 1(a) shows examples of CLLs (in the red rectangles) linking the English and Portuguese versions of the article about *Parque Farroupilha*.

Although the primary use for CLLs was to help users navigate through different versions of an article, CLLs have also been used with many other goals. Nguyen et al. [2011], Oh et al. [2008], and Erdmann et al. [2009] use CLLs between articles to create a bilingual dictionary, while Adafre and de Rijke [2006] use the CLLs to find similarities between sentences in different languages. Several studies rely on CLLs to use Wikipedia as a comparable corpus[1] from which to derive translation schemes [Adafre and de Rijke 2006; Potthast et al. 2008; Sorg and Cimiano 2008a].

[1] A comparable corpus is a collection of texts in two or more languages in which texts describe the same topic.

This work has been partially funded by CNPq project 478979/2012-6.

Table I. Statistics on the English and Portuguese Wikipedias (May, 2011)

		Cross-language Links		
Wikipedia	No. Articles	Direction	Number of CLLs	% CLLs vs No. of Articles
English	3,632,660	EN $\xrightarrow{CLL}$ PT	447,372	12.3%
Portuguese	681,499	PT $\xrightarrow{CLL}$ EN	449,305	65.9%

Table II. Estimating the number of Missing Cross-language links.

		Cross-language Links					
Wikipedia	Articles	Direction	Existing CLLs	Possible CLLs	% Existing CLLs	% Missing CLLs	Missing CLLs
English	3,632,660	EN $\xrightarrow{CLL}$ PT	447,372	613,349	72.93%	27.07%	166,033
Portuguese	681,499	PT $\xrightarrow{CLL}$ EN	449,305	613,349	73.25%	26.75%	164,070

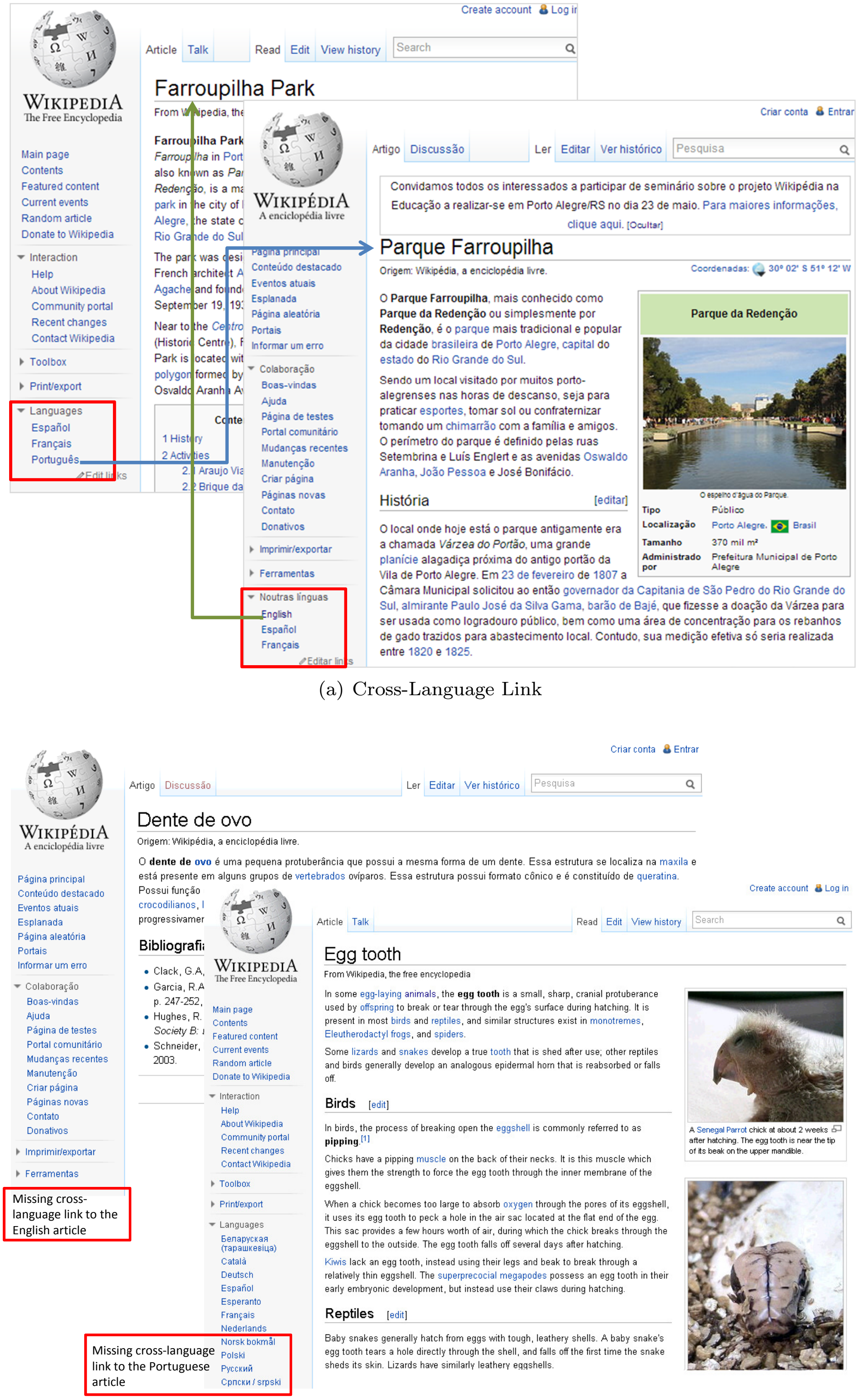
Navigating through the several languages for a particular article, allows finding versions which are more complete. And this can, in turn, be used to enrich the versions with less information [Adar et al. 2009; Bouma et al. 2009; Rinser et al. 2013]. For example, information on tourism and Brazilian culture will likely be more complete in the Portuguese Wikipedia than in other Wikipedias.

CLLs are typically added by the authors of the articles. When the author of the article does not link it to its other versions, we have a *missing CLL*. Figure 1(b) shows an example of a missing CLL. The Portuguese version of the article does not link to and is not linked by the version of the article in English. A missing CLL prevents users from navigating across languages and does not allow applications exploit the full power of Wikipedia's multilingualism. In order to enrich the multilingual capabilities of Wikipedia, the automatic discovery of missing CLLs is a highly desirable feature.

We performed a statistical analysis of the CLLs between articles from the English and Portuguese Wikipedias. As shown in Table I, it turns out that only a small fraction (12.3%) of the articles in the English Wikipedia are connected via a CLL to articles in Portuguese. However, the fraction of the articles in Portuguese that are mapped to corresponding articles in English is much larger, representing 65.9%. This is due to the different sizes between these two Wikipedias, *i.e.* the Portuguese Wikipedia is approximately 18% of the English Wikipedia.

By counting the number of articles and cross-language links in both Wikipedias, one can estimate the number of missing CLLs that can be found. This estimate is shown in Table II. The maximum number of CLLs that could be established between the Wikipedias in English and Portuguese is 681,499 (i.e., the size of the Portuguese Wikipedia, which is the smallest of the two). However, not all Portuguese articles will have a corresponding article in English since some articles describe an entity from the local context, which can be country-specific. Let us assume that about 10% of the articles fit into this category. Even with this estimate, about 27% of the Portuguese articles would still have a missing CLL to their English counterparts. In absolute numbers, this amounts to over 160K missing CLLs for this language pair. These figures justify the need for a method that is able to perform the discovery of such missing CLLs in an automatic and effective manner. Furthermore, given the large amount of data, the method needs to perform at an acceptable cost.

It is worth pointing out that the number of CLLs from a language α to a language β and the number of CLLs from a language β to a language α do not match. This can be seen in the figures shown in Table II. In some cases, the links are unidirectional, and in other cases, they may even point to a different article (e.g. article a_1 in language α has a CLL to article b_1 in language β which has a CLL to article a_1 in α). Because the definition of article equivalence is vague, some of the existing CLLs are incorrect. This problem was addressed by de Melo and Weikum [2010] and Rinser et al. [2013] and is outside the scope of this work. We are only concerned in finding new links in order to enrich Wikipedia's multilingual structure.



Farroupilha Park

From Wikipedia, the

Convidamos todos os interessados a participar de seminário sobre o projeto Wikipédia na Educação a realizar-se em Porto Alegre/RS no dia 23 de maio. Para maiores informações, clique aqui. [Ocultar]

Parque Farroupilha

Origem: Wikipédia, a enciclopédia livre.

Coordenadas: 30° 02' S 51° 12' W

O **Parque Farroupilha**, mais conhecido como **Parque da Redenção** ou simplesmente por **Redenção**, é o parque mais tradicional e popular da cidade brasileira de Porto Alegre, capital do estado do Rio Grande do Sul.

Sendo um local visitado por muitos porto-alegrenses nas horas de descanso, seja para praticar esportes, tomar sol ou confraternizar tomando um chimarrão com a família e amigos. O perímetro do parque é definido pelas ruas Setembrina e Luís Englert e as avenidas Oswaldo Aranha, João Pessoa e José Bonifácio.

História [editar]

O local onde hoje está o parque antigamente era a chamada *Várzea do Portão*, uma grande planície alagadiça próxima do antigo portão da Vila de Porto Alegre. Em 23 de fevereiro de 1807 a Câmara Municipal solicitou ao então governador da Capitania de São Pedro do Rio Grande do Sul, almirante Paulo José da Silva Gama, barão de Bajé, que fizesse a doação da Várzea para ser usada como logradouro público, bem como uma área de concentração para os rebanhos de gado trazidos para abastecimento local. Contudo, sua medição efetiva só seria realizada entre 1820 e 1825.

Parque da Redenção

O espelho d'água do Parque.

Tipo	Público
Localização	Porto Alegre, 🇧🇷 Brasil
Tamanho	370 mil m²
Administrado por	Prefeitura Municipal de Porto Alegre

(a) Cross-Language Link

Missing cross-language link to the English article

Missing cross-language link to the Portuguese article

Dente de ovo

Origem: Wikipédia, a enciclopédia livre.

O **dente de ovo** é uma pequena protuberância que possui a mesma forma de um dente. Essa estrutura se localiza na maxila e está presente em alguns grupos de vertebrados ovíparos. Essa estrutura possui formato cônico e é constituído de queratina.

Bibliografi

Egg tooth

From Wikipedia, the free encyclopedia

In some egg-laying animals, the **egg tooth** is a small, sharp, cranial protuberance used by offspring to break or tear through the egg's surface during hatching. It is present in most birds and reptiles, and similar structures exist in monotremes, Eleutherodactyl frogs, and spiders.

Some lizards and snakes develop a true tooth that is shed after use; other reptiles and birds generally develop an analogous epidermal horn that is reabsorbed or falls off.

Birds [edit]

In birds, the process of breaking open the eggshell is commonly referred to as **pipping**.[1]

Chicks have a pipping muscle on the back of their necks. It is this muscle which gives them the strength to force the egg tooth through the inner membrane of the eggshell.

When a chick becomes too large to absorb oxygen through the pores of its eggshell, it uses its egg tooth to peck a hole in the air sac located at the flat end of the egg. This sac provides a few hours worth of air, during which the chick breaks through the eggshell to the outside. The egg tooth falls off several days after hatching.

Kiwis lack an egg tooth, instead using their legs and beak to break through a relatively thin eggshell. The superprecocial megapodes possess an egg tooth in their early embryonic development, but instead use their claws during hatching.

Reptiles [edit]

Baby snakes generally hatch from eggs with tough, leathery shells. A baby snake's egg tooth tears a hole directly through the shell, and falls off the first time the snake sheds its skin. Lizards have similarly leathery eggshells.

A Senegal Parrot chick at about 2 weeks after hatching. The egg tooth is near the tip of its beak on the upper mandible.

(b) Missing Cross-Language Link

Fig. 1. Example of Matching Wikipedia Articles with (a) and without Cross-Language Links (b)

In this article, we propose a method called Cross-language Link Finder (CLLFinder) which identifies missing CLLs between articles in two given languages. There are two main issues involved in such identification, namely, *candidate selection* and *similarity computation.* Candidate selection is necessary because of the large number of articles existing for many languages, which makes exhaustive pairwise comparisons unfeasible. For example, finding missing CLLs for 160K articles for the Portuguese-English language pair would require over half a trillion comparisons, if we were to compare exhaustively each article in Portuguese to all articles in English which do not have a CLL. Thus the first step is to reduce the search space. We perform such reduction by combining our proposed *CategoryLink* and the *Chain Link Hypothesis* proposed by Sorg and Cimiano [2008b]. The second issue concerns how to compute the similarity between an article in the source language and all the candidates in the target language. In order to do that, we identify features that reflect the similarity between articles. Besides using well-known features such as edit distance on the titles of the articles, our main feature is *Cross-language Link Transitivity* (CLLTransitivity). This feature takes into account the transitivity of CLLs considering other languages as pivots. The selected features are submitted to a classifier.

We performed experiments in which we seek for missing CLLs in the Portuguese-English language pair. The results show that by combining our proposed *CategoryLink* to the Chain Link Hypothesis, we were able to increase the recall in the candidate set. Furthermore, CLLTransitivity has shown a good discriminative power, helping to identify equivalent articles. Compared to the work by Sorg and Cimiano [2008b], we achieve higher precision and recall while using fewer and simpler features.

This article is organized as follows: Section 2 covers the existing literature on discovering missing CLLs. In Section 3, we introduce CLLFinder detailing its phases and algorithms. The experimental analysis is reported in Section 4. Finally, Section 5 concludes the article.

2. RELATED WORK

The first work to address the problem of finding missing CLLs was that of Sorg and Cimiano [2008b]. They relied on the *Chain Link Hypothesis*, which states that an article in a language and its equivalent in another language are connected by a path of links which includes links in the same language as well as CLLs. To narrow down the candidate set, they keep only the 1K articles with the highest number of chain links. Then, they train a classifier which will predict whether a pair of articles match. The classifier is based on seven features, of which five are based on graph structure and two are based on the texts of the articles. They performed experiments that aim at finding missing CLLs between the German and English Wikipedias. The results show a precision of 93.5% and a recall of 69.6%.

Oh et al. [2008] proposed a method for finding missing CLLs between the Wikipedias in Japanese and English. The method creates a feature vector $V(a)$ in language α for article a considering the title and a morphological analysis on the text of the article (to identify nouns and noun phrases). This vector is compared to the vectors of $V(b)$ in language β and a similarity scored is assigned. If a $\langle V(a), V(b)\rangle$ has a similarity score higher than a threshold then $v(b)$ will be part of the set of candidates for $V(a)$. The candidates are then submitted to a classifier which employs 14 features. The authors performed an evaluation in which the method achieved 93.4% precision and 79.7% recall. It is worth pointing out that this method relies on some features which are specific to the English-Japanese language pair (e.g. in many cases, the first sentence of a Japanese article contains its English translation between brackets). Furthermore, performing a morphological analysis in every article is very costly.

More recently, Penta et al. [2012] proposed WikiCL, an algorithm for finding missing CLLs. The first step is to pre-process all articles to classify them as (i) an article that describes a non-geographic named entity, (ii) an article that describes a geographic named entity, or (iii) an article that does not describe a named entity. This classification is based on heuristics, and the idea is that an article that describes a non-geographic named entity should be matched to another article in the same category. Articles which describe geographic named entities usually have longitude/latitude information. When this is the case, the candidates are selected based on these geographic coordinates. In order to identify which articles match, a semantic relatedness measure is calculated. In their experiments, articles in English are matched to articles in Italian, German, and French. The precision on the restricted dataset was between 89% and 94%, while the recall ranged between 89% and 93%. The authors compared their approach to the one by Sorg and Cimiano [2008b] and found that WikiCL has higher recall but lower precision. It is worth noticing that precision is the most important metric for evaluating the creation of CLLs.

Unlike the approach by Oh et al. [2008], CLLFinder is designed to be language independent. Compared to existing methods, our proposed approach uses fewer and simpler features. Nevertheless, according to our experiments (detailed in Section 4), CLLFinder achieves higher precision and recall.

3. FINDING MISSING CROSS-LANGUAGE LINKS

Our proposed approach is composed of three modules: (i) selecting the set of candidates, (ii) computing similarity evidences to identify the matching article among the candidates, and (iii) submitting the dataset to the classifier. Figure 2 shows how these modules relate, as well as their inputs and outputs.

The first module aims at reducing the number of candidates in the target language β that will be compared to the source article in language α. This step is necessary since it is not viable to compare all pairs of Wikipedia articles across two languages in search for a CLL. According to the numbers in Table I, the English Wikipedia has over 3,6 million articles, so comparing each article in Portuguese against all articles in English would not be feasible. Thus, let WP_α be the set of Wikipedia articles

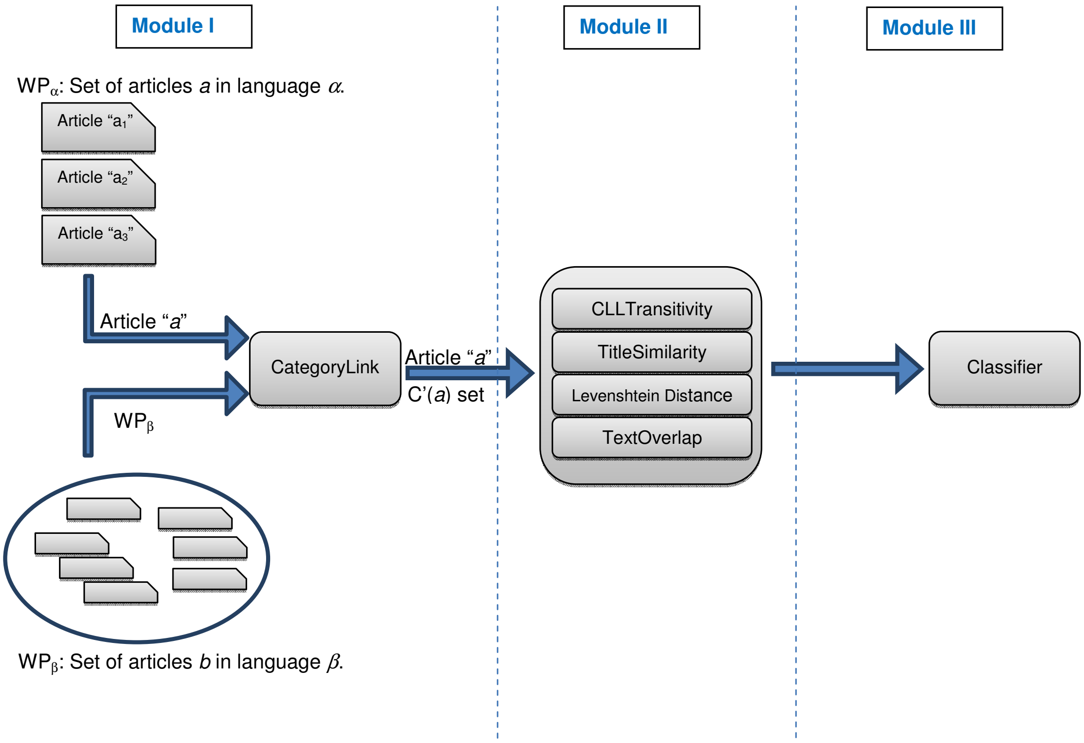


Fig. 2. Architecture of CLLFinder

in language α and let WP_β be the set of Wikipedia articles in language β. For each article $a \in WP_\alpha$, this module generates the restricted set of candidates $C'(a)|C'(a) \subset WP_\beta$.

The second module is responsible for analyzing features which reflect the similarity between a pair of articles. This way, the similarity for each pair of articles $\langle a, b\rangle$ where $a \in WP_\alpha$ and $b \in C'(a)$ is computed. At this step, the following similarity features are employed: CLLTransitivity, TitleSimilarity, Levenshtein Distance, and TextOverlap.

In the last module, the similarity coefficients calculated by Module II are submitted to a classifier which will identify the corresponding pairs.

3.1 Generating the Set of Candidates

Restricting the set of candidates is a common problem for methods that seek missing CLLs. The work by Sorg and Cimiano [2008b], for example, introduces the *Chain Link Hypothesis*. A *Chain Link* is defined as follows: for two versions of Wikipedia WP_α and WP_β, there is a chain link between two articles $a_\alpha \in WP_\alpha$ and $b_\beta \in WP_\beta$ if $a_\alpha \xrightarrow{pl} b_\alpha \xrightarrow{CLL} b_\beta \xleftarrow{pl} a_\beta$, where *pl* are pagelinks within the same language. If there is chain link between a_α and b_β, then b_β can be part of the set of candidates for a_α. The assumption is that an article is linked to its corresponding article in another language through at least one chain link.

A Wikipedia article can be assigned to one or more *categories*. For example, the article for the actor *Tom Cruise* belongs to the categories: *1962 births*, *20th-century American actors*, *21st-century American actors*, *Actors from New York*, among many others. Wikipedia categories form a large directed graph (and not a tree), since a given category may have more than a parent. In this work, we developed a method called CategoryLink which considers the categories of the article in the source language and the categories of the corresponding article in the target language. This mechanism is depicted in Figure 3.

The goal of CategoryLink is to find candidates for article $a \mid a \in WP_\alpha$, whose title is *Sustentabilidade Ambiental*. Thus, the entire set of categories of a, denoted by $CAT(a)$ is selected. Within this set, we check which categories have CLLs such that $CAT_a \xrightarrow{CLL} CAT'_a$, forming the set $CAT'(a)$. Then, the set of candidates for a, denoted by $C'(a)$ will consist of all articles $c|c \xrightarrow{is\ in\ category} CAT'(a)$ and $c \in WP_\beta$.

To the set of candidates described above, we add the set of candidates produced by the *Chain Link Hypothesis* [Sorg and Cimiano 2008b]. It is important to notice that the same article c can be repeated many times within $C'(a)$ as it can be part of more than one category in the set $CAT'(a)$. Similarly, the set of candidates generated by the chain link hypothesis could also have repetitions. In fact, the more an article is repeated within $C'(a)$, the more likely it is that it matches a. The number of occurrences of an article can be used to sort the candidates in $C'(a)$. Table III shows the set of candidates $C'(a)$ of articles in English generated for the article *Aves* (birds) in Portuguese.

Table III. Candidates for the article *Aves* sorted by their number of occurrences

Article	Candidate Article	No. of Occurrences	Rank
Aves	Bird	288	1
Aves	List_of_birds	184	2
Aves	Archaeopteryx	172	3
Aves	Palaeognathae	168	4
Aves	...	...	...
Aves	Odonata	10	1913
Aves	...	...	...
Aves	Pesticide	6	7619

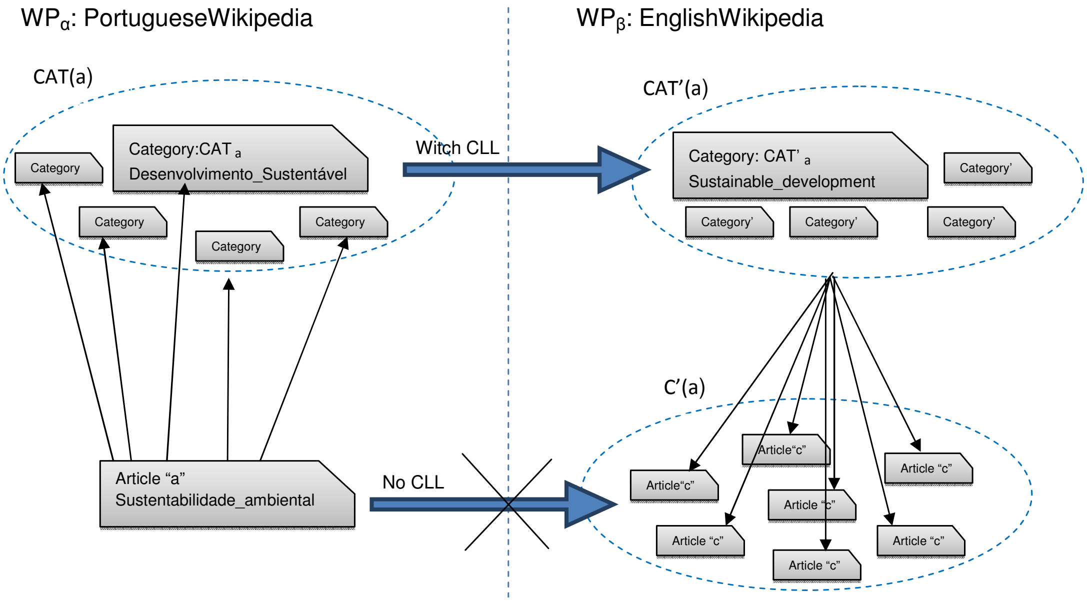


Fig. 3. CategoryLink Method

Preliminary experiments have shown that the set of candidates for a given article could contain over 150K articles. This number is still too high to compute similarity scores and perform comparisons. Thus, it is still necessary to further reduce the number of candidates by taking the N top-ranked candidates. In Section 4, we show the recall for several values of N.

3.2 Identifying Equivalent Articles

The module which identifies equivalent articles has two inputs: (i) the article a in the source language α for which a corresponding article b in the target language β is sought and (ii) the set of candidates generated by the previous module. Thus, for each pair $\langle a, c\rangle \mid c \in C'(a)$, where $C'(a)$ is the of candidate matches for a, four similarity features are computed and used by the next module in order to train a classifier. These features are described in the next subsections.

3.2.1 *Cross-language Link Transitivity.* CLLTransitivity is a feature developed in this work which explores the transitivity in CLLs from other Wikipedia languages. The rationale is that a missing CLL can be the result of an author of an article in language α, who did not add a link to the corresponding version of the article in language β, but added a link to the version of the article in language γ. In turn, there is a chance that there exists a CLL between the article in language γ and the article in language α. This enables leveraging CLL transitivity as an evidence of similarity.

Let α, β, and γ be three distinct languages. For each candidate c, in the pair $\langle a, c\rangle$, where $a \in WP_\alpha$, $c \in C'(a)$ and $C'(a) \subset WP_\beta$, the algorithm seeks for the existence of two CLL $a \xrightarrow{CLL} g \mid g \in WP_\gamma$ and $g \xrightarrow{CLL} j \mid j \in WP_\beta$ and $j = c$. If these two conditions are satisfied, it means we can navigate from the source article a to the corresponding article c through another language denoted by γ.

The type of CLL in both relationships ($a \xrightarrow{CLL} g$ and $g \xrightarrow{CLL} j$) is also taken into consideration. According to Rinser et al. [2013], there are three possibilities for CLLs: bidirectional, unidirectional, and conflicting. Bidirectional links indicate consistency, while conflicting links happen when one of the links points to another article. In this article, we only work with the first two types, assigning them weights that reflect how reliable an evidence they are. These weights are given in Table IV and were manually defined. No tuning was performed to find the values that yield the best results.

Algorithm 1 Cross-language Link Transitivity

```
function CLLTransitivityFunction($a$, $C'(a)$, $\beta$, $L$,$P$)
    $CLLTransitivity$ = null
    for $C_n \in C'(a)$ do
        for $L_n \in L$ do
            $transitivityVector$ = null
            if $existsCLL(a, L_n)$ then
                $intermediateArticle = returnArticleCLL(a, L_n)$
                if $existsCLL$(intermediateArticle,$\beta$) then
                    $targetArticle = returnArticleCLL$(intermediateArticle,$\beta$)
                    if ($targetArticle == C_n$) then
                        $transitivityVector[0] = 1$
                        if ($biDirection$(a,intermediateArticle) then
                            $transitivityVector[1] = 1$
                        end if
                        if ($biDirection$(intermediateArticle,targetArticle) then
                            $transitivityVector[2] = 1$
                        end if
                    end if
                end if
            end if
            $CLLTransitivity$+=($transitivityVector[0]*P[0] + transitivityVector[1]*P[1] + transitivityVector[2]*P[2])*L/$sizeOf($L$);
        end for
    end for
    return $CLLTransitivity$
end function
```

Table IV. Weights for Cross-language Links Connectivity

Relation	Value
$\alpha \xrightarrow{CLL} \gamma$ and $\gamma \xrightarrow{CLL} \beta$ (there is a path)	0.8
$\alpha \xleftrightarrow{CLL} \gamma$ (first bidirectional link)	0.1
$\gamma \xleftrightarrow{CLL} \beta$ (second bidirectional link)	0.1
Both bidirectional link	1.0

To explain the method, the previous example used a single intermediate language denoted by γ. However, more intermediate languages can be used (three intermediate languages were used in our experiments in Section 4). The idea is that the higher is the number of languages, the more reliable is the similarity score generated. Thus, CLLTransitivity weights the score for each language depending on the number of languages used. Algorithm 1 is described below. Its inputs are the source article a, the set of candidates $C'(a)$, the language of the candidates β, the set L of languages used, and the set of weights P given in Table IV.

3.2.2 *Other similarity features.* Besides CLLTransitivity, other features were implemented to quantify the similarity between articles. These are described next.

Title Similarity

String similarity between the titles of the articles has been used in related work [Oh et al. 2008; Sorg and Cimiano 2008b; Adar et al. 2009]. The idea is that, after translated, the titles of corresponding articles should be similar. Considering an article a in the source language α for which a candidate article $c \in C'(a)$ in the target language β, this measure was implemented as follows.

(1) the titles in languages α and β are tokenized
(2) the tokenized title in language α is translated into the target language β using the dictionary from Microsoft Developer Network[2].
(3) with both titles in the same language (β), stopwords are removed stemming is performed.
(4) similarity is then computed as the Dice Coefficient (Eq. 1)

$$DiceCoefficient(DC) = \frac{T_a \cap T_c}{max(T_a, T_c)} \tag{1}$$

where T_a and T_c are the number of tokens in articles a and c, respectively.

Notice that only the titles of the articles are translated. Since they are very short (typically 1 to 3 words), this does not impose a high cost on our method. Furthermore, this is the only feature that relies on external tools (translator and stemmer) which are language dependent. As the experiments show, CLLFinder still performs well even without this feature.

Levenshtein's Edit Distance
Levenshtein's edit distance was applied directly to the titles of the articles without translation. This feature is useful in cases where the article describes a person, a location, or a company since, in many cases, their names are not translated.

Text Overlap
For article a and his candidate $c \in C'(a)$, this feature counts the number of words in common between a and c divided by the number of words in the longest article. The text is tokenized and stemmed, however, no translation is performed. TextOverlap is calculated in the same fashion as TitleSimilarity, using (Eq. 1). For morphologically similar languages, this feature will have a higher score than for languages that are very different. However, there is a good chance it will be greater than zero since even morphologically diverse languages may share proper nouns, numbers and dates, which are not translatable. This feature has also been explored by Sorg and Cimiano [2008b].

3.3 Classifier

The output of the second module is a set of pairs of articles $\langle a, c\rangle$ followed by four similarity features which try to capture different sources of similarity between the source and the candidate articles. In this third module, these features are submitted to a classifier which will predict whether the articles match.

In order to train the classifier, articles in language α for which the corresponding article in language β is known, will be used as positive examples. Articles which knowingly do not match are used as negative examples. Notice that there will be far fewer positive examples than negative examples. So, in order to avoid the class imbalance problem (which would cause the classifier to be biased by the negative class) we performed random undersampling by choosing k negative instances, where k is the number of positive instances. Once the classification model is obtained, the classifier is ready to analyze different instances for which the class is not known.

Our method was not designed to work with a specific classification technique. In our experiments, a decision tree classifier was employed. However, other techniques could have been used with similar results.

[2] http://msdn.microsoft.com

4. EXPERIMENTS

This section presents the experiments we performed to evaluate CLLFinder, proposed in this work. We start by detailing the languages used, the environment and the way in which the articles were selected. Then we present the results of CategoryLink for candidate selection and the results of CLLFinder for identifying missing CLLs. We also provide a comparison to the method proposed by Sorg and Cimiano [2008b] and an evaluation on the contribution of each feature.

4.1 Experimental Setup

Environment: Our experiments aim at finding missing CLLs between articles in Portuguese and English. In order to apply CLLTransitivity, French, Italian, and Spanish were chosen as intermediate languages. Note that the languages do not have to be morphologically similar, but it helps if they have a similar cultural context because then the chance of having more CLLs is higher. In this sense, Basque and Spanish would be considered "similar" even though the languages themselves are very different. The datasets were downloaded from `http://dumps.wikimedia.org/`. For the source and target languages, all tables are needed. However, for the intermediate languages, only tables *page* and *langlinks* are required. The classifier was J48 provided within WEKA [Hall et al. 2009]. The Porter Stemmer [Porter 1980] was used to strip suffixes for the Title Similarity feature.

Article Selection: In order to train the classifier and evaluate our methods, we needed to work with articles for which the CLLs are known. Thus, we selected a dataset, which we will refer to as DS1000, composed of 1000 articles in Portuguese for which their English counterparts are known. DS1000 was collected using a recursive function that, for a given Wikipedia category and a desired number of articles, checks subcategories collecting articles until the desired number of articles has been reached. The following categories were used: *animais* (*animals*), *internet*, *automobilismo* (*auto_racing*), *moda* (*fashion*), *filmes* (*films*), *atores* (*actors*), *biologia* (*biology*), *matemática* (*mathematics*), *física* (*physics*), and *computadores* (*computers*). The idea was to choose diverse categories, avoiding categories which would likely contain country-specific articles. For each of these categories, the first 100 articles were selected.

4.2 Results for Candidate Selection

Recall from Section 3.1 that our method for candidate selection combines the *Chain Link Hypothesis*, proposed by Sorg and Cimiano [2008b] to the CategoryLink algorithm described in Figure 3. For each article a_P in Portuguese from DS1000, the set of candidates $C'(a_P)$ in English was generated. Then, for each article $a_P \in$ DS1000, we check whether the corresponding article $a'_E \in C'(a_P)$. The results are shown in Table V. The second column shows the results for the Chain Link Hypothesis on its own, while the third column shows the results for the combination proposed here. The column entitled *Increase* shows that CategoryLink increases the number of cases in which the corresponding article is present in the candidate set. Ideally, this increase should arise without a significant growth on the number of candidates. The results have shown that CategoryLink increases the number of times the matching article is in the candidate set by 37.3% on average, while the rise on the number of candidates is of 10%. This makes CategoryLink worthwhile as it in helps the corresponding article rank higher in the candidate set and thus increase recall for the fixed cuts (1000, 2000, 5000, and 10000).

When we analyzed the size of the candidate sets, we noticed that there were about 150K candidates for each article. This number is much smaller than the total number of articles in the English Wikipedia, but it is still far too large. Thus, according to the description in Section 3.1, we sort the candidates by their number of occurrences and then take only the top N. Table VI shows the percentage of cases in which the corresponding article is present in the candidate set for different values of N, namely 1000, 2000, 5000, and 10000.

Table V. Cases in which the corresponding article is present in the candidate set.

	Chain Link Hypothesis	**Chain Link Hypothesis + CategoryLink**	**Increase**
Animais	67%	84%	+25%
Internet	52%	64%	+23%
Automobilismo	18%	34%	+88%
Moda	54%	75%	+38%
Filmes	20%	95%	+375%
Atores	82%	98%	+19%
Biologia	73%	83%	+13%
Matemática	52%	55%	+5%
Física	55%	63%	+14%
Computadores	63%	85%	+34%
Average	53.6%	73.6%	+ 37.3%

Table VI. Number of corresponding articles within the N top candidates.

	Chain Link Hypothesis + CategoryLink	N=1000	N=2000	N=5000	N=10000
Animais	84%	67%	73%	81%	82%
Internet	64%	48%	51%	56%	60%
Automobilismo	34%	15%	16%	18%	18%
Moda	75%	45%	50%	63%	66%
Filmes	95%	64%	73%	78%	83%
Atores	98%	68%	74%	86%	93%
Biologia	83%	62%	67%	74%	81%
Matemática	55%	32%	36%	42%	47%
Física	63%	48%	54%	60%	63%
Computadores	85%	66%	75%	84%	88%
Average	73.6%	51.5%	56.9%	64.2%	68.1%

Analyzing the data from Table VI we chose to keep $N = 1000$ as this figure would reduce the number of candidates 150 times, while the presence of the corresponding article in the candidate set is reduced by 30%. This means a significant reduction in computational cost which comes at an acceptable cost in terms of recall.

4.3 Results for CLL Identification

In order to submit the articles to be analyzed by the classifier, we computed the features for each article a in DS1000 paired with all articles from $C'(a)$ (restricted by N=1000). We will refer to this dataset as *WPMAIN*. The number of instances in WPMAIN is 1 million (all articles $a \in$ DS1000 compared against their 1000 candidates $\in C'(a)$). We can see from Table VI that 51.5% of the articles have a counterpart in the candidate set, which means that 515 articles $a \in$ DS1000 have instances formed by $\langle a, c\rangle$ | c $\xrightarrow{CLL}$ a', whereas for the remaining 485 articles $a \in$ DS1000, there is no corresponding article within the candidate set.

Since we know whether each of the instances refers to a positive example (i.e. matching articles) or to a negative example (i.e. unmatching articles), the evaluation measures (Eq. 2), recall (Eq. 3), and F-measure (Eq. 4) can be calculated.

$$Precision(P) = \frac{\#TruePositive}{\#TruePositive + \#FalsePositive} \tag{2}$$

$$Recall(R) = \frac{\#TruePositive}{\#TruePositive + \#FalseNegative} \tag{3}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{4}$$

where True Positive refers to the matching articles which have been identified as such; False Positives are the articles which are not equivalent but which were classified as being equivalent; and False Negatives are the matching articles which have not been identified as such.

From WPMAIN, three datasets have been selected and submitted to the classifier. The first, *TrainingSet*, contains 514 instances (half positive and half negative) and it was used to generate the decision tree model, which will be validated by the next two datasets. The second, *TestSet* has 516 instances (also half positive and half negative). And the last, *LargeTestSet*, has 25,987 instances (258 positive and 25,729 negative). The idea is to assess the quality of the classification on a larger dataset in which more combinations of the values of features can be assessed. The results for the identification of CLLs are given in Figure 4.

The results for precision, recall and F-measure attest that the first dataset was able to generate a classification model which is very accurate in identifying CLLs. The model yielded good results with the other two datasets.

Comparing against the baseline: In order to validate our proposed approach, we compared it against the method developed by Sorg and Cimiano [2008b], hereafter referred to as *S&C Baseline*. Thus, we implemented all of its seven features (*Chain Link Count*, *Normalized Chain Link Count*, *Chain Link Inlink*, *Common Categories*, *CLIA Graph*, *Editing Distance*, and *Text Overlap*) according to their description in the article. We trained their model using the same dataset used for our CLLFinder (*TrainingSet*). However, as described by Sorg and Cimiano [2008b], SVMlight [Joachims 1999] was used instead of J48. After the classification model was generated, the instances in *TestSet* were classified.

The results for this comparison are shown in Figure 5. CLLFinder achieved superior results in terms of precision, recall, and F-measure, despite using fewer features. We attribute this to three reasons: (i) the feature *Common Categories*, used by *S&C Baseline*, does not yield a good precision since a candidate article which is not the correct match (but which is similar) may have a high number of categories in common; (ii) similar to the previous case, the feature *Chain Link Count* used by *S&C Baseline* (which counts the number of occurrences of the candidate article within the candidate set) also does not yield a good precision since an article which is not the correct match may occur more times within the candidate set than the correct match; (iii) in order to compare the similarity between the titles of the articles, *S&C Baseline*, relies on *Levenshtein Distance* without translating one of the titles first. However, we noticed that after translation, the titles of many corresponding articles become identical.

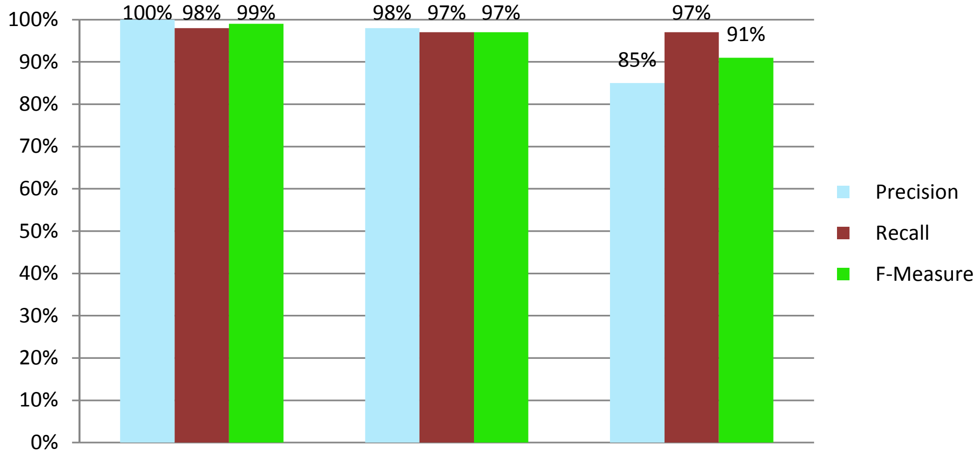


Fig. 4. Precision, Recall, and F-Measure values for CLLFinder

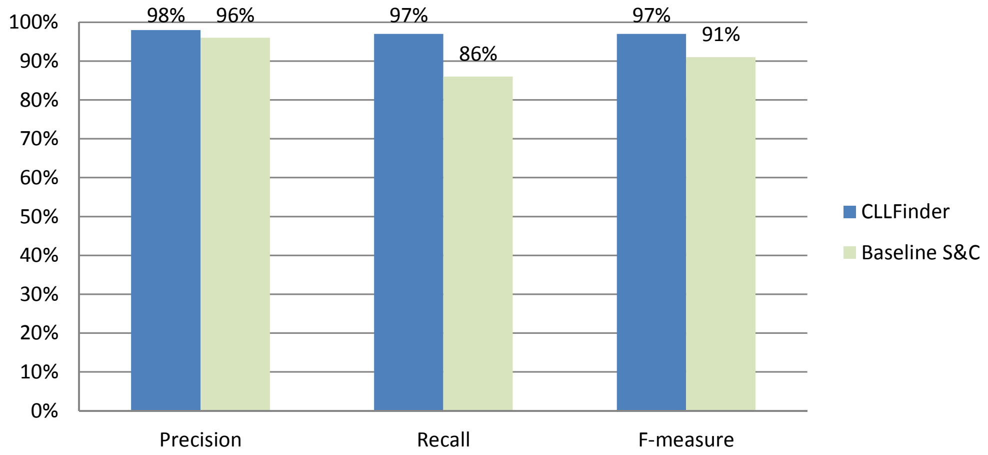


Fig. 5. Comparing CLLFinder against Baseline

We attribute this superior performance to the CLLTransitivity feature. The existence of a path of CLLs from the source article to the target article is a strong evidence which has proven very precise in identifying matching articles. We observed that there a path from the source to the target article could be identified using one intermediate language in about 53% of the cases. Using three intermediate languages increases the number of cases to 76%. The results obtained by *S&C Baseline* were slightly superior to the ones reported by Sorg and Cimiano [2008b]. We believe this happened because here we used a different dataset and a different language pair.

We analyzed the cases in which our classifier predicted the wrong class. Cases of false positives happened for articles which had very similar titles and which shared a good portion of the words in their texts. For example, CLLFinder classified the articles *MacBook* (in Portuguese) and *MacBook Pro* (in English) as being matches, when if fact, they are not. The only low scoring feature for this pair was CLLTransitivity. Cases of false negatives were mostly due to erroneous translations allied to and absence of CLL transitivity. For example, CLLFinder failed to match *Moda Sustentável* and *Sustainable Fashion*. The word *moda* was erroneously translated, which lowered the score for TitleSimilarity, also, there was no CLL transitivity between the articles.

Contribution of each Feature: In order to assess the contribution of each feature, we ran the classifier four times, each time removing one of the features. The result of each of the four runs was compared against the run using all features. Table VII shows the results for 10-fold cross validation on the *TestSet* dataset (so the numbers are slightly different from the ones reported in Figure 4).

CLLTransitivity is the feature that contributes the most in terms of precision, recall, and F-measure. This confirms our hypothesis that the author of the article may not have added the CLL to one language, but may have added it for another language. By leveraging this feature, CLLTransitivity achieves excellent results, contributing more with recall. CLLTransitivity relies on the existence of a path of CLLs between the source and target articles and in some cases, this path does not exist. For

Table VII. Precision, Recall, and F-Measure removing each Feature

	Precision	**Recall**	**F-Measure**
CLLFinder (All Features)	**98%**	**96%**	**97%**
Without CLLFinder-CLLTransitivity	93%	74%	82%
Without CLLFinder-TitleSimilarity	93%	93%	93%
Without CLLFinder-Levenshtein	95%	91%	93%
Without CLLFinder-TextOverlap	95%	93%	94%

Table VIII. Precision, Recall, and F-Measure for each Feature in isolation

	Precision	**Recall**	**F-Measure**
CLLTransitivity	100%	90%	94%
CLLFinder-TitleSimilarity	95%	86%	91%
CLLFinder-Levenshtein	97%	64%	77%
CLLFinder-TextOverlap	81%	79%	80%

these situations, the other three features have shown to be able to identify the matching article. This is the case for articles shown in Figure 1(b). There is no transitivity in the CLLs since no article links to the Portuguese version. However, TitleSimilarity after translation was very high and enabled the discovery of the matching article. TitleSimilarity helps improve precision while Levenshtein increases recall. Overall, all features contribute to CLLFinder, helping achieve high levels of precision, recall, and F-measure. It is worth pointing out that precision is the most important quality measure for the identification of a missing CLL, since creating a link between unmatching articles is worse than having a missing CLL.

Table VIII shows the results for each of the four features used in isolation. Again, CLLTransitivity has shown to be the best feature, achieving perfect precision and a high recall. And, once more we see that the combination of all four features increases F1. Although precision decreases by two points, this is compensated by a gain in recall of 6 points. Also, when an article does not have any CLLs, the other features are fundamental sources of evidence. The combination of features is necessary to guarantee that our approach is robust to deal with cases in which there are no links. This would be the case for newly created articles.

Limitations: Overall we feel that our classification results (Module III) are very good but the selection of candidates (Module I) can still be improved to maximize the presence of the matching article in the candidate set. Because the classifier evaluates each instance independently, in some cases, more than one candidate for the same article is considered as the corresponding article. This could be resolved by adding a post-processing step that chooses from all articles predicted as matches, which one is the most likely.

5. CONCLUSION

This article presents CLLFinder, an approach for discovering missing cross-language links in Wikipedia. CLLFinder is composed of three modules. For the first module, which aims at narrowing down the search space, we devised a method called CategoryLink. It takes the categories of the articles into consideration. By employing CategoryLink, we increased by 37% the presence of the matching article in the candidate set compared to using the Chain Link Hypothesis alone [Sorg and Cimiano 2008b]. The second module computes four features that indicate the similarity between articles in different languages. Amongst the features, we highlight CLLTransitivity which leverages the transitivity in cross-language links present in other languages. This feature has proven very precise, contributing to the quality of our proposed method. The third module employs a decision tree classifier which predicts whether an article and its candidate are matches.

We carried out experiments in which articles from the Portuguese Wikipedia are matched to articles from the English Wikipedia. The Wikipedia versions in Italian, French, and Spanish have been used as intermediate languages by CLLTransitivity. Overall, among training and test instances, one million pairs $\langle article, candidate \rangle$ obtained from a set of 1K articles in Portuguese and its candidates in English have been processed. Our results achieved a recall of 96% and a precision of 98%, outperforming the baseline method.

The classification results are very good, however, we feel that the selection of candidates still has room for improvement. We plan to address that as future work. Also, it would be interesting to try our approach with other language-pairs and intermediate languages. Language pairs could include morphologically similar languages such as Portuguese and Spanish, and also completely different languages, such as Portuguese and Japanese.

REFERENCES

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

Adar, E., Skinner, M., and Weld, D. S. Information Arbitrage across Multi-lingual Wikipedia. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. Barcelona, Spain, pp. 94–103, 2009.

Bouma, G., Duarte, S., and Islam, Z. Cross-lingual Alignment and Completion of Wikipedia Templates. In *Proceedings of the International Workshop on Cross Lingual Information Access: addressing the information need of multilingual societies*. Boulder, Colorado, pp. 21–29, 2009.

de Melo, G. and Weikum, G. Untangling the Cross-lingual Link Structure of Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 844–853, 2010.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA Data Mining Software: an update. *SIGKDD Explorations Newsletter* 11 (1): 10–18, 2009.

Joachims, T. Making Large-scale Support Vector Machine Learning Practical. In *Advances in Kernel Methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola (Eds.). pp. 169–184, 1999.

Nguyen, T., Moreira, V., Nguyen, H., Nguyen, H., and Freire, J. Multilingual Schema Matching for Wikipedia Infoboxes. *Proceedings of the VLDB Endowment* 5 (2): 133–144, 2011.

Oh, J.-H., Kawahara, D., Uchimoto, K., Kazama, J., and Torisawa, K. Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. pp. 322–328, 2008.

Penta, A., Quercini, G., Reynaud, C., and Shadbolt, N. Discovering Cross-language Links in Wikipedia through Semantic Relatedness. In *European Conference on Artificial Intelligence*. Montpellier, France, pp. 642–647, 2012.

Porter, M. F. An Algorithm for Suffix Stripping. *Program* 14 (3): 130–137, 1980.

Potthast, M., Stein, B., and Anderka, M. A Wikipedia-Based Multilingual Retrieval Model. In *Advances in Information Retrieval*. Lecture Notes in Computer Science, vol. 4956. pp. 522–530, 2008.

Rinser, D., Lange, D., and Naumann, F. Cross-lingual Entity Matching and Infobox Alignment in Wikipedia. *Information Systems* 38 (6): 887 – 907, 2013.

Sorg, P. and Cimiano, P. Cross-lingual Information Retrieval with Explicit Semantic Analysis. In *Working Notes for the Cross-Language Evaluation Forum Workshop*. Aarhus, Denmark, 2008a.

Sorg, P. and Cimiano, P. Enriching the Crosslingual Link Structure of Wikipedia - a classification-based approach. In *Proceedings of the Workshop on Wikipedia and Artificial Intelligence*. Chicago, US, 2008b.