ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-642

Discovering Cross-language Links in Wikipedia through Semantic Relatedness

Antonio Penta[1] and **Gianluca Quercini**[2] and **Chantal Reynaud**[3] and **Nigel Shadbolt**[4]

Abstract. *Wikipedia* is a large multilingual collection of interlinked articles, used and contributed by millions of users over the Internet, that provides editions in up to 283 languages. Two articles in different language versions of *Wikipedia* may have information on the exactly the same concept, in which case they are often connected through a cross-language link. However, many cross-language links are either missing or incorrect and this negatively affects both the readers of *Wikipedia* and multilingual information retrieval applications. In this paper, we propose WIKICL, an algorithm for discoverinrg cross-language links using the semantic relatedness of two articles derived from the *Wikipedia* graph structure. Our evaluation shows that we achieve comparable, and in some cases, better results than previous methods with much less computational time .

1 Introduction

Wikipedia is the largest online encyclopedia to date; its articles are accessed and contributed by millions of users over the Internet and cover diverse topics, such as, but not limited to, arts, history, events, geography, mathematics and technology. Typically, a *Wikipedia* article describes a specific concept and has always a *title* and an *introduction*, which sums up its content in few words. Throughout this paper the term "article" does not refer to *disambiguation* nor *redirect* pages, which *Wikipedia* includes to respectively resolve the ambiguity between articles sharing similar titles (e.g. "Mercury (planet)" , "Mercury (mythology)") and identify an article (e.g. "Paris") through synonymous titles (e.g. "Ville Lumière", "City of Paris").

Due to the ever growing richness of its content, *Wikipedia* has been increasingly gaining attention as a precious knowledge base for a wide range of applications, including, among others, document classification [18], ontology creation and enrichment [2, 16], entity ranking [9] and information extraction [19, 14]. One important aspect of *Wikipedia* is its multilingualism; as of January 2012, there are 283 language versions, of which the English edition is the biggest with up to 4,000,000 articles. Articles that provide information on the same concept in different languages may be connected through *cross-language links*. For instance, the article titled "Computer Science" in the English *Wikipedia* has a cross-language link to the corresponding article in the Italian *Wikipedia* titled "Informatica". The interconnection of articles in different languages is used by many applications to evaluate the similarity between sentences in different languages [1], mine multilingual topics [13] and create bilingual dictionaries [6]. All these applications would benefit from having all articles connected across all language versions of *Wikipedia*, which unfortunately is not the case, due to the fact that cross-language links are manually created by the *Wikipedia* contributors. This is exemplified in Table 1, where we show the percentage of cross-language links over the total number of articles for four language versions of *Wikipedia* as of October 2010. Only 15% of the articles in the English *Wikipedia* are connected to the corresponding articles in the Italian version and similar statistics, with few exceptions, are observed in the other pairs of language versions. Note that in the Italian version up to 71.5% of articles have a cross-language link to the English version; although this number is high, there is still 30% of articles (around 200,000) that need to be manually inspected to find their corresponding article in the English version, if any. Hence, the need of an automatic approach.

In this paper we propose WIKICL, an algorithm for discovering cross-language links in *Wikipedia*. More specifically, given an article in one language version, we want to determine the corresponding articles, if any, in other language versions. The main contribution of WIKICL over a previous approach to this problem [15] is that it is unsupervised and language-independent, as we explain in greater detail in Section 2. We note that WIKICL can also be used to find erroneous cross-language links in *Wikipedia*, as it is the same as predicting new ones. However, we do not expand on this point, as our evaluation only concerns the discovery of missing links.

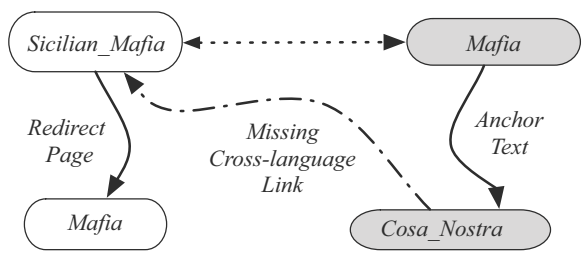

Figure 1. Example of missing cross-language link.

Figure 1 shows an example of a missing cross-language link between an English and an Italian article of *Wikipedia* as of February 2012. The blank and shadowed rectangles represent English and Italian articles respectively; the dashed and the full lines denote respectively cross-language links and page links between articles in the same language version. Although the English article *Sicilian_Mafia* and the Italian article *Cosa_Nostra* denote the same concept, they are not cross-language linked; there is just a cross-language link between the English article *Sicilian_Mafia* and the Italian article *Mafia*, which in turn has a page link to the Italian article titled *Cosa_Nostra*, which

[1] University of Southampton, UK, email: ap7@ecs.soton.ac.uk
[2] Université Paris-Sud XI, France, email:gianluca.quercini@lri.fr
[3] Université Paris-Sud XI, France, email:chantal.reynaud@lri.fr
[4] University of Southampton, UK, email: nrs@ecs.soton.ac.uk

	Cross-language links to English (%)	Cross-language links to Italian (%)	Cross-language links to France (%)	Cross-language links to German (%)
English Version	–	15.0	19.3	18.2
Italian Version	71.5	–	48.3	41.2
French Version	66.9	35.2	–	38.2
German Version	58.2	27.8	36.8	–

Table 1. Percentage of cross-language links in four *Wikipedia* language versions as of October 2010.

represents the Mafia in Sicily.

The remainder of this paper is organized as follows. Section 2 presents an overview of related work. In Section 3 we introduce basic notation and definitions to describe WIKICL, which is presented in Section 4. Finally, we present the evaluation of the methodology in Section 5 and conclude the presentation in Section 6.

2 Related Work

Wikipedia cross-language links have received a lot of attention in the last years from different communities such as information retrieval and natural language processing. In particular, several researchers exploited the cross-language structure of the *Wikipedia* to support different cross-language applications.

Adafre and de Rijke [1] use the cross-language links to find similarities between sentences in different languages, while Ni et al. [13] use them to mine multi-language topics. Erdmann et al. [6] leverage the multilingual structure of *Wikipedia* to create bilingual dictionaries that are very useful to emerging research area such as machine translation, human-aided translation and cross-language information retrieval. They define a set of features to train a SVM classifier using the *Wikipedia* structure such as anchor text, redirect pages, forward and backwards links.

Nguyen et al. [12] propose a method for identifying mappings between attributes from *Wikipedia* Infoboxes of pages in different languages that can be useful to answer multilingual structured queries. Their approach does not require any training data and could be effective for languages that are under-represented.

Navigli et al. [11] describe a system that aims at producing a large multilingual semantic network by linking *Wikipedia* to *WordNet* via an automatic mapping strategy. *Menta* (Multilingual Entity Taxonomy) [4] is another multilingual knowledge base that also integrates *Wikipedia* and *Wordnet*, based on linking heuristics to discover potential taxonomic relationships and graph and statistical techniques to discover the final multilingual taxonomy.

While these researches show what can be done by using the multilingual structure of *Wikipedia*, our paper focuses on improving that structure to support applications such as the ones described above. In literature there some papers that deal with the same problem. Hassan and Mihalcea [8] describe a simple heuristics that discovers new cross-language links through reflexivity and transitivity. De Melo and Weikum [5] tackle the problem of detecting incorrect cross-language links in *Wikipedia*. They model the removal of incorrect links as an optimization problem and describe an algorithm that computes a suboptimal solution. The algorithm is based on enforcing a consistency by applying repair operation in order to reconcile the graph link structure. Our approach differs from this because it can discover new cross-language links rather than just correcting the existing ones. Finally, in [17] a comparative study is presented on the problem of discovering missing links between pages of the same language version of *Wikipedia*.

Finally, the approach described by Sorg and Cimiano [15] is the most related to ours as it is intended to find missing cross-language links in *Wikipedia*. To this extent, given an article a in a source *Wikipedia* language version, it first selects a set of candidate articles in a target *Wikipedia* language version, and then predicts, by using a SVM-based classifier, the one that is to be connected to a through a cross-language link. Although WIKICL consists of the same two-steps structure, it introduces significant differences. As for the first step, it uses information provided by *Wikipedia* to narrow down the candidate search space (cf. Section 4.1). As for the second, WIKICL has two main novelties. First, it is unsupervised, which means that it does not require any training for learning how to connect two articles with a cross-language link. This is particularly important because it does not need to be trained for each pair of *Wikipedia* language versions under consideration and because it does not need to go through an expensive feature extraction step, as the approach proposed by Sorg and Cimiano. Second, WIKICL does not use any text-based features, as done by Sorg and Cimiano, meaning that WIKICL can be really claimed to be language independent; note that it is the use of text-based features that mostly contributes to the highest values of precision and recall of the Sorg-Cimiano approach.

3 Preliminaries

Any version of *Wikipedia* in language l can be modelled as a directed graph W_l, with a *node* v_α for each article α and an *edge* (v_α, v_β) if and only if α has a link to β. To ease the notation, greek letters always refer to articles while small latin letters refer to nodes in the *Wikipedia* graph; moreover, we omit the subscript when we refer to a node in W_l, unless we need to explicitly mention its corresponding article. We denote by $Label(v_\alpha)$ the *label* of node v_α, which is the title of its corresponding article; moreover, $CL_t(v_\alpha)$ refers to a node w_β in the *Wikipedia* version W_t in a target language t, such that β and α are connected through a cross-language link. If α is not connected to any node in W_t through a cross-language link, $CL_t(v_\alpha) = nil$. Finally, we define as $N_{out}(v)$ (respectively, $N_{in}(v)$) the set of the nodes in W_l to which v links (which link to v); $N(v) = N_{out}(v) \cup N_{in}(v)$ is the set of the *neighbours* of v.

For this paper we downloaded from the Wikimedia website (`download.wikimedia.org`) four *Wikipedia* language versions, as of October 2010: English (October, 11th), which contains 3,265,081 articles, German (October, 13th), with 1,022,944 articles, French (October, 17th), with 955,010 articles, and Italian (October, 20th), including 700,884 articles. We imported their content to a single local PostgreSQL database and we pre-processed it in order to obtain the following information useful for WIKICL:

- lat/long values associated to any article that refers to a geographic entity;
- A flag to indicate whether an article refers to a named entity;

Although lat/long values if any, are usually displayed on the top-right corner of an article, their extraction is difficult, as *Wikipedia* allows multiple ways to specify them through the inclusion of predefined templates. For this reason, we obtained them from *Yago* [16]

and *DBpedia* [2], which both provide structured information on all *Wikipedia* articles. As a result, we associated lat/long values to 568,005, 124,034, 127,569 and 98,143 articles in the English, German, French and Italian *Wikipedia* respectively.

In order to determine whether the title t of an article α refers to a named entity, we used the heuristics presented by [3], which are quite accurate. Based on the observation that the title of a *Wikipedia* article always begins with a capital letter, the heuristics considers t as a reference to a named entity if and only if:

- t is composed of multiple words, each one beginning with a capital letter, with the exception of prepositions, determiners, conjunctions, relative pronouns and negations;
- t is composed of one word with more than one capital letter (t is an acronym);
- t is composed of one word and 75% of its occurrences in the text of α begin with a capital letter.

As in German all nouns begin with a capital letter, and thus this heuristics cannot be used, we resorted to the Stanford NLP named entity recognition tool [7].

4 WIKICL

Let W_s and W_t be the graphs corresponding respectively to the *Wikipedia* version in language s, called the *source* Wikipedia, and the *Wikipedia* version in language t, called the *target* Wikipedia. The problem of detecting a cross-language link between a node v_α in the source *Wikipedia* and a node v_β in the target *Wikipedia*, if any, can be formalized as follows:

Input: W_s, W_t, $v_\alpha \in W_s$.
Output: $v_\beta \in W_t$ such that α and β provide information on exactly the same topic, or nil if β does not exist.

Henceforth, node v_β is referred to as the *translation* of v_α. The rationale of our algorithm WIKICL that we propose to solve this problem is as follows. Since β, if it exists, focuses on exactly the same topic as α, β must have the highest semantic correlation to α across all the articles in the target *Wikipedia*. WIKICL uses a *graph-based semantic relatedness measure* to assign a *relatedness score* to the nodes in the target *Wikipedia*, reflecting their relatedness to v_α, and selects the one with the highest score as the translation of v_α.

Since it would be computationally expensive to assign a relatedness score to each node of the target *Wikipedia*, WIKICL proceeds in two steps:

- Determines a subset C of nodes of W_t that are likely to be the translation of v_α. C is termed the *candidate set*.
- Assigns a relatedness score to each node in C and selects the one with the highest score as the translation of v_α.

In the remainder of this section we explain these two steps in greater detail.

4.1 Candidate Set Determination

The first step of WIKICL is the determination of a set C, which contains the translation v_β of v_α, if any, while including only a small subset of nodes of the target *Wikipedia*. A key observation here is that we can partition the nodes of a *Wikipedia* graph into three classes, based on their labels:

1. Nodes that refer to *non-geographic named entities*.
2. Nodes that refer to *geographic named entities*.
3. Nodes that do not denote a named entity.

Based on the class of the input node v_α, we adopt a different strategy to identify the candidate set C.

Case 1: v_α is a non-geographic named entity. Since v_α refers to a non-geographic named entity, the translation of v_α must be a node which also refers to a non-geographic named entity. This restricts the search space only sensibly, as most of the articles in all *Wikipedia* language versions are about named entities.

As (non-geographic) named entities denote, among others, names of people, companies and events, chances are that most of them (e.g. "the president of the United States") are referred to by using the same phrase (e.g. "Barack Obama") across different languages. Therefore, if the target *Wikipedia* has a node v_β that denotes a non-geographic named entity and $Label(v_\beta) = Label(v_\alpha)$ then v_β can be directly selected as the translation of v_α. We note that we compare the labels of the two nodes only after checking that both nodes refer to non-geographic named entities. In fact, blindly selecting the translation of v_α as the node v_β having the same label without first checking that v_α and v_β belong to the same class can lead to errors. For example, the English *Wikipedia* contains a node with label "Emperor", which by no means corresponds to the node in the Italian *Wikipedia* with the same label. In fact, the first (the ruler of an empire) does not refer to a named entity, while the second (the name of a Norwegian black metal band) does.

Obviously, we cannot always expect that the label of a named entity is independent of the language; for example, "Francis Bacon", an English philosopher, is often referred to as "Francesco Bacone" in Italian. Moreover, two languages may have different alphabets, in which case any label comparison is completely pointless. Therefore, if we cannot find any node in the target *Wikipedia* with the same label as v_α, we select the candidates in C by using the procedure described below for case 3, while filtering out all the nodes that are not non-geographic named entities. We note that in the 3 datasets ITA, GER and FRA that we use in our evaluation, 40% of articles on average fall in this case.

Case 2: v_α is a geographic named entity. Similarly to case 1, we first check whether the target *Wikipedia* has a node v_β that refers to a geographic named entity with the same label as v_α, in which case v_β is selected as the translation of v_α.

If such a node cannot be found, we can identify a set of candidates based on the geographic coordinates of v_α, with some caveats. First of all, the coordinates of two corresponding articles in two different languages might not be exactly the same, as they are usually specified independently by two different *Wikipedia* users. Second, v_β may not have any geographic coordinates at all, because no *Wikipedia* contributors set them, either by mistake, carelessness or forgetfulness; after all, *Wikipedia*, being a human artifact, cannot be expected to be perfect. In order to address these two problems, we identify a set of candidates to be the translation node of v_α as follows:

- We add to C all nodes in the target *Wikipedia* that have lat/long values within a $\delta = 3$ miles radius of the coordinates of v_α. The value of δ is manually tuned. This addresses the first problem.
- For each node w selected on first step, we add to C all nodes in $N(w)$, the set of the neighbours of w. This addresses the second problem. In fact, if v_β has no coordinates, it is anyway likely to

link to, or be linked by, nodes that refer to nearby geographic entities.

Case 3: v_α is not a named-entity. In order to select a candidate set when v_α does not refer to a named entity, we use an approach similar to the one proposed by Sorg and Cimiano [15]. They observe that if v_β is the translation of v_α, the neighbour set $N(v_\alpha)$ is likely to contain many nodes that are the translation of nodes in $N(v_\beta)$. For instance, the node "Chocolat" in the French *Wikipedia* links to nodes "Chocolat chaud" and "Chocolat au lait"; both are the translations of nodes "Hot chocolate" and "Milk chocolate" respectively, which belong to the neighbour set of the node "Chocolate", which is the translation of "Chocolat" in the English *Wikipedia*. In other words, v_β is likely to be connected to v_α through at least one *chain link*, defined as:

$$v_\alpha \xrightarrow{pl} w \xrightarrow{cl} z \xleftarrow{pl} v_\beta$$

where $\xrightarrow{pl}$ denotes a link between two nodes within the same language version, while $\xrightarrow{cl}$ denotes a cross-language link. Since there are potentially many chain links between v_α and v_β, we need to consider only those that are more likely than others to lead to the inclusion of node v_β in the candidate set C. Sorg and Cimiano propose to rank the nodes of the target *Wikipedia* based on the number of chain links that connect them to v_α and include in C only the top-1000.

Our approach also limits the number of candidates to 1000, but ranks them based on their relatedness to v_α. We observe that some nodes in the neighbour set $N(v_\alpha)$ are more related to v_α than the others and we can use simple graph properties to identify them. For example, any node $w_\gamma \in N(v_\alpha)$ that mutually link to v_α (w_γ links to v_α and the other way round) has a high probability of being important to v_α. Consequently, also the node $CL_t(w_\gamma)$ in the target *Wikipedia* is likely to be important to v_β and thus to link to v_β. Therefore, the nodes in the neighbour set of $CL_t(w_\gamma)$ are good candidates to be the translation of v_α. Thus, by denoting by $M(v_\alpha)$ the set $N_{out}(v_\alpha) \cap N_{in}(v_\alpha)$ of nodes that mutually link to v_α, we add to C up to 1000 nodes of the target *Wikipedia* in the following order, which reflects their importance to the target translation v_β:

1. All nodes in $N(CL_t(w_\gamma))$, for each $w_\gamma \in M(v_\alpha)$;
2. All nodes in $N(CL_t(w_\gamma))$, for each $w_\gamma \in N_{out}(v_\alpha)$.

4.2 Translation Selection

The second step of WIKICL consists of selecting one of the nodes in C as the translation of v_α. To this extent, WIKICL assigns a relatedness score $S(v_\gamma)$ to each node $v_\gamma \in C$ and selects the one with the highest score as the translation of v_α, provided that its score is above a predetermined threshold τ, which we experimentally set to 0.15. Here the problem is to assign a score to node v_γ in the target *Wikipedia* that reflects its relatedness to node v_α in the source *Wikipedia*. Since α and γ are written in two different languages, the evaluation of the semantic relatedness between them requires the translation of the content of one of them. Since automatic translation is far from perfect, we do not investigate this option. Instead, we exploit the structure of the *Wikipedia* graph to describe v_α and any concept $v_\gamma \in C$ in terms of the nodes of the source (respectively, target) *Wikipedia* that are related to v_α (respectively, v_γ).

Given a *Wikipedia* language version W in language l, we represent any node v of W as a *similarity vector* $Sim^l(v)$, which is created as follows. For each node $i \in N(v)$, the value of $Sim^l(v)[i]$ is a measure of the semantic relatedness of i to v. The relatedness between v and i is computed by using the measure WLM (*Wikipedia Link-based Measure*), presented by Milne and Witten [10]. WLM is based on the *Normalized Google Distance*:

$$WLM(v,i) = \frac{max\{\log f(v), log f(i)\} - \log f(v,i)}{\log |W| - min\{\log f(v), \log f(i)\}}$$

where v and i denote two *Wikipedia* articles, $f(v)$ ($f(i)$) is the number of articles that link to v (respectively, i), $f(v,i)$ is the number of articles that link to both v and i and $|W|$ is the total number of articles in *Wikipedia*. Note that the name *Normalized Google Distance* may mislead the reader to think that the measure needs to access some search engine in order to determine a relatedness score of two articles. However, this is not the case. In fact, in order to compute the previous formula, we easily retrieve $f(v)$ and $f(i)$ in constant time from our graph structure while the computation of $f(v,i)$ is linear to the number of inlinks of v and i, which is usually small (around 20 on average across the four *Wikipedia* language editions that we consider in this paper).

We note that the computation of the similarity vector of a node, which is done on-the-fly when needed, dominates the running time of WIKICL, which amounts to 2 minutes on average for every node to translate, as best described in Section 5. However, the similarity vectors can be easily precomputed and stored so as to considerably drive down the running time. We did not investigate further this aspect, although we reserve it for immediate future work.

Since node v_α and all nodes $v_\gamma \in C$ can be described through their similarity vectors $Sim^s(v_\alpha)$ and $Sim^t(v_\gamma)$, we compute the relatedness score $S(v_\gamma)$ as the cosine similarity between $Sim^s(v_\alpha)$ and $Sim^t(v_\gamma)$. In order to do that, we need to recall that $Sim^t(v_\gamma)$ is created on the nodes of the target *Wikipedia*, while $Sim^s(v_\alpha)$ is created on the nodes of the source *Wikipedia*. Therefore, we need to map $Sim^t(v_\gamma)$ to a new vector $Sim^s(v_\gamma)$ in the source language s as follows. For each node $i \in N(v_\gamma)$, we obtain its corresponding node $CL_s(i)$ in W_s; if $CL_s(i) \neq nil$, then $Sim^s(v_\gamma)[CL_s(i)] = Sim^t(v_\gamma)[i]$, otherwise $Sim^s(v_\gamma)[CL_s(i)] = 0$.

5 Evaluation

We evaluated WIKICL in three steps. First, we assessed its ability to determine the correct cross-language links between articles in two different language versions of *Wikipedia*. To this extent, we randomly selected three sets of nodes in the English *Wikipedia* (the source *Wikipedia*) and used WIKICL to find their corresponding nodes in the Italian, French and German *Wikipedia* (the target *Wikipedia*) respectively. We opted for these language versions of *Wikipedia* because they include more articles than any other version, which allows for better evaluation. Moreover, we focus on the translation from English to the other three languages (and not the other way round) because, as shown in Table 1, the cross-language links from the English version are sparser than from the other versions, which may negatively affect the candidate selection step. In other words, we consider the worst case scenario. Second, we compare WIKICL against the

Dataset	# NonGeo NE	# Geo NE	# Others
ITA	276	128	96
FRA	301	106	93
GER	279	132	89

Table 2. Properties of the three datasets under evaluation.

one described by Sorg and Cimiano (S&C) [15]. Note that we do

Evaluation	Dataset	Precision	Recall	F-measure	Top-1	Top-2	Top-3	Top-4	Top-5
Whole dataset	ITA	0.92	0.87	0.89	0.88	0.89	0.90	0.91	0.91
	FRA	0.86	0.84	0.85	0.85	0.89	0.90	0.92	0.92
	GER	0.87	0.84	0.86	0.85	0.87	0.88	0.89	0.89
Restricted dataset	ITA	0.94	0.92	0.93	0.93	0.94	0.95	0.97	0.97
	FRA	0.89	0.88	0.88	0.88	0.93	0.94	0.95	0.96
	GER	0.89	0.88	0.89	0.89	.0.91	0.92	0.93	0.94

Table 3. Results for the cross-language link prediction test.

not compare WIKICL against baseline methods as they are already discussed by Sorg and Cimiano.

Finally, we discuss the computational cost of WIKICL. All experiments are conducted on a desktop computer running Ubuntu 11.04 with a Intel Core i7-2600 3.40 Ghz and 8 GB of RAM. We note that the source code of WIKICL as well as the datasets used in our evaluation are available at the following URL: http://sites.google.com/site/gquercini/software/wikicl.

5.1 Cross-language Links Determination

From the English *Wikipedia*, we randomly selected 3 (not-necessarily disjoint) test sets GER, FRA and ITA, each containing 500 nodes for which we already know their actual translations in the German, French and Italian *Wikipedia* respectively.

We used WIKICL to find the translation of each node v in GER (respectively, FRA and ITA) in the German (respectively, French and Italian) *Wikipedia*. The translation is correct if it coincides with the node that is the actual translation of v, which we already know. Observe that WIKICL may not be able to determine a translation, when the score of the highest-ranked node in the candidate set is below the threshold τ, as discussed in Section 4.2. Let X denote any of GER, FRA and ITA, $D_X \subseteq X$ the nodes for which WIKICL determined a translation and $C_X \subseteq X$ the nodes for which WIKICL determined the correct translation. WIKICL is evaluated based on *precision* P, *recall* R and *F-measure* F, which are defined as follows:

$$P = \frac{C_X}{D_X} \quad R = \frac{C_X}{X} \quad F = \frac{2 * P * R}{P + R}$$

The precision measures how many of the translations that WIKICL reported are correct; the recall measures how many translations WIKICL can correctly determine over the whole dataset; finally, the F-measure is an harmonic mean between precision and recall, and is a measure of the overall *accuracy* of WIKICL. We also report the *top-k accuracy*, $k \leq 5$, as defined by Sorg and Cimiano [15].

In Table 2 we show the properties of the three datasets: the number of non-geographic named entitites (# NonGeo NE), geographic named entities (# Geo NE) and the remaining concepts (# Others). Not surprisingly, the named entities dominate the three datasets, as *Wikipedia* has mostly articles on named entities.

Precision, recall, f-measure and top-k accuracy are presented in Table 3. The table reports two evaluations for each dataset. The first (*Whole dataset*) takes into account all the concepts in each dataset, while the second (*Restricted dataset*) only considers those concepts for which WIKICL manages to include the correct translation in the candidate set.

We observe that the overall accuracy (F-measure) is stable across the three data sets, which suggests that WIKICL is robust enough to handle different *Wikipedia* language versions. We also note a good balance between precision and recall, which indicates that WIKICL effectively determines many cross-language links (high recall) that are also correct (high precision). Not surprisingly, all metrics have larger values when considering the restricted datasets, as they do not include the concepts for which the candidate selection fails. This is an indication of what the performances of WIKICL would be if the correct translation was always included in the candidate set.

Moreover, we point out that the restricted datasets of ITA FRA and GER have 473, 480, 477 concepts respectively, that amount to more than 90% of the whole dataset, which witnesses the ability of the candidate selection step of WIKICL to include the correct translation in the candidate set.

Finally, the values of the top-k accuracy show that the correct translation is often found within the top-5 translations reported by WIKICL. More specifically, the top-5 accuracy has always values greater then 0.9.

5.2 Comparison

Sorg and Cimiano [15] evaluated their methodology on a dataset called *RAND1000*, which includes 1000 nodes randomly selected from the German *Wikipedia*. For each node in *RAND1000*, we know the corresponding node in the English *Wikipedia*, which is used as the ground truth. Of the 1000 nodes in *RAND1000*, Sorg and Cimiano use 750 for training their SVM and 250 for testing it. Since we do not know which subset of *RAND1000* was used for testing, we decided to evaluate WIKICL on four randomly selected subsets, each containing 250 nodes. The results are shown in Table 4. We observe that across the four subsets of *RAND1000*, WIKICL suffers from a slight drop in precision with respect to S&C; however, WIKICL consistently achieves a much higher recall than S&C, which results in an clear improvement of its accuracy (the highest value of the f-measure is 0.89). As for the top-k accuracy, WIKICL clearly outperforms S&C. We point out that the improvements of WIKICL over S&C are consistent across the evaluations on the four randomly selected subsets of *RAND1000*. This suggests that the improvements are not due to chance and therefore are statistical significant.

5.3 Performance

The time required by WIKICL to determine the translation of a node largely depends on its degree; not surprisingly, nodes with high degree are likely to have a large candidate set. Therefore, the computation for these nodes can take up to 4-5 minutes on the computer we used for our evaluation. On average, the time required to compute the translations of 500 nodes is 16 hours, which corresponds to an average of 2 minutes for each node. The running time is largely dominated by the creation of the similarity vectors; in other words, by pre-computing the similarity vectors of the nodes we expect to dramatically decrease the average running time and also the maximum running time. We did not try this option yet, as the pre-computation of the similarity vectors is likely to take some days, but we will consider this option in the future. As a side note, we point out that the

Algorithm	Precision	Recall	F-measure	Top-1	Top-2	Top-3	Top-4	Top-5
S&C	**0.94**	0.70	0.80	0.76	0.78	0.79	0.80	0.81
WIKICL	0.83	0.80	0.81	0.80	0.83	0.85	0.85	0.86
	0.87	0.86	0.87	0.86	0.88	**0.91**	0.91	0.91
	0.89	**0.88**	**0.89**	**0.88**	**0.90**	**0.91**	**0.92**	**0.92**
	0.86	0.84	0.85	0.84	0.87	0.88	0.88	0.89

Table 4. Comparison between WIKICL and S&C

approach described by Sorg and Cimiano requires up to 26 hours to extract the features.

6 Concluding Remarks

In this paper we described WIKICL, an algorithm for discovering cross-language links in *Wikipedia*. We explained how this is an important issue for different applications in the area of cross-language information retrieval. WIKICL has many advantages with respect to previous approaches that addressed the similar problem. Among others, WIKICL is an unsupervised method, which requires no training at all, and is language independent, as it only uses the graph structure of *Wikipedia* and no text-based features. Our evaluation shows that WIKICL consistently achieves high accuracy across all datasets that we selected. However, further research is still needed to assess (a) the accuracy of WIKICL while determining missing cross-language links of thousand articles in *Wikipedia* and (b) the robustness of WIKICL in *Wikipedia* language versions that have few articles and existing cross-language links. This will be part of our immediate future work.

REFERENCES

[1] S. F. Adafre and M. de Rijke, '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, (2006).

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, 'Dbpedia: A nucleus for a web of open data', *The Semantic Web*, 722–735, (2007).

[3] Razvan C. Bunescu and Marius Pasca, 'Using Encyclopedic Knowledge for Named entity Disambiguation', in *EACL*. The Association for Computer Linguistics, (2006).

[4] Gerard de Melo and Gerhard Weikum, 'Menta: inducing multilingual taxonomies from wikipedia', in *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 1099–1108. ACM, (2010).

[5] Gerard de Melo and Gerhard Weikum, 'Untangling the cross-lingual link structure of wikipedia', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pp. 844–853. Association for Computational Linguistics, (2010).

[6] Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio, 'Improving the extraction of bilingual terminology from wikipedia', *ACM Trans. Multimedia Comput. Commun. Appl.*, **5**, 31:1–31:17, (November 2009).

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning, 'Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling', in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363–370, Stroudsburg, PA, USA, (2005). Association for Computational Linguistics.

[8] Samer Hassan and Rada Mihalcea, 'Cross-lingual semantic relatedness using encyclopedic knowledge', in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP '09, pp. 1192–1201. Association for Computational Linguistics, (2009).

[9] Rianne Kaptein, Pavel Serdyukov, Arjen De Vries, and Jaap Kamps, 'Entity ranking using wikipedia as a pivot', in *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pp. 69–78. ACM, (2010).

[10] D. Milne and I. H. Witten, 'An effective, low-cost measure of semantic relatedness obtained from Wikipedia links', in *WikiAI'08: Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pp. 25–30, Chicago, (July 2008).

[11] Roberto Navigli and Simone Paolo Ponzetto, 'BabelNet: Building a very large multilingual semantic network', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* Uppsala, Sweden, 11–16 July 2010, pp. 216–225, (2010).

[12] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and Juliana Freire, 'Multilingual schema matching for wikipedia infoboxes', *Proc. VLDB Endow.*, **5**(2), 133–144, (2011).

[13] Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen, 'Cross lingual text classification by mining multilingual topics from wikipedia', in *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pp. 375–384, New York, NY, USA, (2011). ACM.

[14] Eyal Shnarch, Libby Barak, and Ido Dagan, 'Extracting lexical reference rules from wikipedia', in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pp. 450–458. Association for Computational Linguistics, (2009).

[15] P. Sorg and P. Cimiano, 'Enriching the crosslingual link structure of wikipedia-a classification-based approach', *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, (2008).

[16] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum, 'Yago: a core of semantic knowledge', in *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pp. 697–706. ACM, (2007).

[17] Omer Sunercan and Aysenur Birturk, 'Wikipedia missing link discovery: A comparative study', in *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, (2010).

[18] Pu Wang and Carlotta Domeniconi, 'Building semantic kernels for text classification using wikipedia', in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 713–721. ACM, (2008).

[19] Fei Wu and Daniel S. Weld, 'Open information extraction using wikipedia', in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pp. 118–127. Association for Computational Linguistics, (2010).