NSFC-61261130588

ANR-12-IS02-002-01

Lindicle

Linked data interlinking in a cross-lingual environment
跨语言环境中语义链接关键技术研究
Liage des données dans un environnement interlingue

D1.1 Boosting Cross-lingual Knowledge Linking via Concept Annotation

---

Coordinator: Zhichun Wang
With contributions from: Zhichun Wang, Juanzi Li, Jie Tang

Quality reviewer:	Specify quality controller
Reference:	Lindicle/D1.1/v2
Project:	Lindicle ANR-NSFC Joint project
Date:	March 25, 2015
Version:	2
State:	draft
Destination:	public

Executive Summary

Automatically discovering cross-lingual links (CLs) between wikis can largely enrich the cross-lingual knowledge and facilitate knowledge sharing across different languages. In most existing approaches for cross-lingual knowledge linking, the seed CLs and the inner link structures are two important factors for finding new CLs. When there are insufficient seed CLs and inner links, discovering new CLs becomes a challenging problem. In this work, we propose an approach that boosts cross-lingual knowledge linking by concept annotation. Given a small number of seed CLs and inner links, our approach first enriches the inner links in wikis by using concept annotation method, and then predicts new CLs with a regression-based learning model. These two steps mutually reinforce each other, and are executed iteratively to find as many CLs as possible. Experimental results on the English and Chinese Wikipedia data show that the concept annotation can effectively improve the quantity and quality of predicted CLs. With 50,000 seed CLs and 30% of the original inner links in Wikipedia, our approach discovered 171,393 more CLs in four runs when using concept annotation.

Document Information

Project number	ANR-NSFC Joint project	**Acronym**	Lindicle
Full Title	跨语言环境中语义链接关键技术研究 Linked data interlinking in a cross-lingual environment Liage des données dans un environnement interlingue		
Project URL	`http://lindicle.inrialpes.fr/`		
Document URL			

Deliverable	**Number**	1.1	**Title**	Boosting Cross-lingual Knowledge Linking via Concept Annotation
Work Package	**Number**	1	**Title**	Dummy Lindicle workpackage

Date of Delivery	**Contractual**	M12	**Actual**	20-04-2014
Status	draft		final ☐	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Zhichun Wang, Juanzi Li, Jie Tang			
Resp. Author	**Name**	Zhichun Wang	**E-mail**	zcwang@bnu.edu.cn
	Partner	Beijing Normal University		

Abstract (for dissemination)	This report is a small abstract of the deliverable purposes.
Keywords	data interlinking, linked data, instance matching

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
20/09/2011	1	J. Euzenat	A simple entry example
21/09/2011	2	J. David	Added info

Table of Contents

1. Introduction

With the rapid evolving of the Web to be a world-wide global information space, sharing knowledge across different languages becomes an important and challenging task. Wikipedia is a pioneer project that aims to provide knowledge encoded in various different languages. According to the statistics in Jun 2012, there are more than 20 million articles written in 285 languages in Wikipedia. Articles describing the same subjects in different languages are connected by cross-lingual links (CLs) in Wikipedia. These CLs serve as a valuable resource for sharing knowledge across languages, which have been widely used in many applications, including machine translation [16], cross-lingual information retrieval [10], and multilingual semantic data extraction [1, 6], etc.

However, a large portion of articles are still lack of the CLs to certain languages in Wikipedia. For example, the largest version of Wikipedia, English Wikipedia, has over 4 million English articles in Wikipedia, but only 6% of them are linked to their corresponding Chinese articles, and only 6% and 17% of them have CLs to Japanese articles and German articles, respectively. Recognizing this problem, several cross-lingual knowledge linking approaches have been proposed to find new CLs between wikis. Sorg et al. [12] proposed a classification-based approach to infer new CLs between German Wikipedia and English Wikipedia. Both link-based features and text-based features are defined for predicting new CLs. 5,000 new English-German CLs were reported in their work. Oh et al. [9] defined new link-based features and text-based features for discovering CLs between Japanese articles and English articles, 100,000 new CLs were established by their approach. Wang et al. [15] employed a factor graph model which leverages link-based features to find CLs between English Wikipedia and a large scale Chinese wiki, Baidu Baike[1]. Their approach was able to find about 200,000 CLs.

In most aforementioned approaches, language-independent features are mainly defined based on the link structures in wikis and the already known CLs. Two most important features are the number of seed CLs between wikis and the number of inner links in each wiki. Large number of known CLs and inner links are required for accurately finding sufficient number of new CLs. However, the number of seed CLs and inner links in wikis varies across different knowledge linking tasks. Figure 1.1 shows the information of wikis that are involved in previous knowledge linking approaches, including the average numbers of outlinks and the percentages of articles having CLs to English Wikipedia. It shows that English Wikipedia has the best connectivity. On average each English article has more than 18 outlinks on average. The average numbers of outlinks in German Wikipedia, Japanese Wikipedia and Chinese Wikipedia range from 8 to 14. 35% to 50% of these articles have CLs to English Wikipedia in these three wikis. While we also note that in Baidu Baike, both the average number of outlinks and the percentage of articles linked to English Wikipedia are much smaller than that in Wikipedia. Therefore, discovering all the possible missing CLs between wikis (such as Baidu Baike) that have sparse connectivity and limited seed CLs is a challenging problem. As mentioned above, more than 200,000 Chinese articles in Baidu Baike are linked to English Wikipedia, but they only account for a small portion of more than 4,000,000 Chinese articles in Baidu Baike. How to find large scale CLs between wikis based on small number of CLs and inner links? To the best of our knowledge, this problem has not been studied yet.

In order to solve the above problem, we propose an approach to boost cross-lingual knowledge linking by concept annotation. Before finding new CLs, our approach first identifies important concepts in each wiki article, and links these concepts to corresponding articles

[1] http://baike.baidu.com/

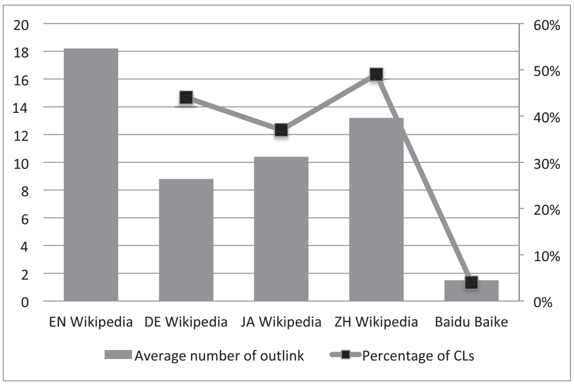


Figure 1.1: Information of the inner links and CLs in different wikis.

in the same wiki. This concept annotation process can effectively enrich the inner links in a wiki. Then, several link-based similarity features are computed for each candidate CL, a supervised learning method is used to predict new CLs. Specifically, our contributions include:

- A concept annotation method is proposed to enrich the inner links in a wiki. The annotation method first extracts key concepts in each article, and then uses a greedy disambiguation algorithm to match the corresponding articles of the key concepts. In the concept disambiguation process, the available CLs are used to merge the inner link structures of two wikis so as to build an integrated concept network, then the semantic relatedness between articles can be more accurately approximated based on the integrated concept network, which improves the performance of concept annotation.
- Several link-based features are designed for predicting new CLs. The inner links appearing in the abstracts and infoboxes of articles are assumed to be more important than other inner links, therefore they are treated differently in similarity features. A regression-based learning model is proposed to learn the weights of different similarity features, which are used to rank the candidate CLs to predict new CLs.
- A framework is proposed to allow the concept annotation and CL prediction to perform in a mutually reinforcement manner. The concept annotation method consumes seed CLs and inner links to produce new inner links. Then the seed CLs and the enriched inner links are used in the CL prediction method to find new CLs. The new CLs are again used to help concept annotation.

We evaluate our approach on the data of Wikipedia. For concept annotation, our approach obtains a precision of 82.1% with a recall of 70.5% without using CLs, and when using 0.2 million CLs, the precision and recall are increased by 3.5% and 3.2%. For the CL prediction, both the precision and recall increase when more seed CLs are used or the concept annotation is performed. Our approach results in a high accuracy (95.9% by precision and 67.2% by recall) with the help of concept annotation with 0.2 million seed CLs. In the four runs of our approach, we are able to incrementally find more CLs in each run. Finally, 171,393 more CLs are discovered when the concept annotation is used, which demonstrates the effectiveness of the proposed cross-lingual knowledge linking approach via concept annotation.

The rest of this report is organized as follows, Chapter 2 describes the proposed approach in detail; Chapter 3 presents the evaluation results; Chapter 4 discusses some related work and finally Chapter 5 concludes this work.

2. The Proposed Approachr

(1) Overview

Figure 2.1 shows the framework of our proposed approach. There are two major components: *Concept Annotator* and *CL Predictor*. *Concept Annotator* identifies extra key concepts in each wiki article and links them to the corresponding article in the same wiki. The goal of the *Concept Annotator* is to enrich the inner link structures within a wiki. *CL Predictor* computes several link-based features for each candidate article pair, and trains a prediction model based on these features to find new CLs. In order to find as many CLs as possible based on a small set of seed CLs, the above two components can be executed iteratively. New discovered CLs are used as seed CLs in the next running of our approach, therefore the number of CLs can be gradually increased. In the following subsections, the *Concept Annotator* and *CL Predictor* are described in detail.

(2) Concept Annotation

Important concepts in a wiki article are usually annotated by editors with hyperlinks to their corresponding articles in the same wiki database. These inner links guide readers to articles that provide related information about the subject of current article. The linkage structure was often used to estimate similarity between articles in traditional knowledge linking approaches, e.g., [15]. However, there might be missing inner links in the wiki articles, especially when the articles are newly created or have fewer editors. In this circumstance, link-based features cannot accurately assess the structure-based similarity between articles. Therefore, our basic idea is to first uses a concept annotator to enrich the inner links in wikis before finding new CLs.

Recently, several approaches have been proposed to annotating documents with concepts in Wikipedia [7, 8, 4, 11]. These approaches aim to identify important concepts in the given documents and link them to the corresponding articles in Wikipedia. The concept annotator in our approach provides the similar function as existing approaches. However, instead of discovering links from an external document to a wiki, concept annotator focuses on enriching the inner links in a wiki. Therefore, there will be some existing links in the articles before annotating. In our approach, the inputs of concept annotator are two wikis and a set of CLs between them. The concept annotation process consists of two basic steps: (1) concept extraction, and (2) annotation.

Concept Extraction. In this step, possible concepts that might refer to other articles in the same wiki are identified in an input article. The concepts are extracted by matching all the n-grams in the input article with the elements in a controlled vocabulary. The vocabulary

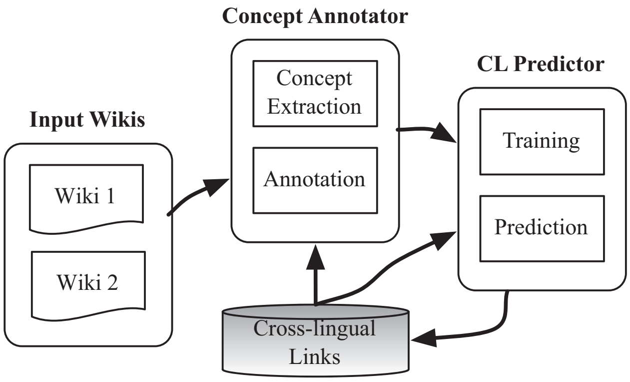


Figure 2.1: Framework of the proposed approach.

contains the titles and the anchor texts of all the articles in the input wiki. For an input article a, the results of concept extraction contain a set of concepts $C = \{c_i\}_{i=1}^{k}$ and the sets of candidate articles $\{T_{c_1}, T_{c_2}, ..., T_{c_k}\}$ that each concept might link to.

Annotation. In the annotation step, links from the concepts to the destination articles are identified. Here we use two metrics, the *Link Probability* and the *Semantic Relatedness*, to decide the correct links for each concept. The *Link Probability* measures the possibilities that an article is the destination of a given concept, which is approximated based on the already known annotations in a wiki. The *Semantic Relatedness* measures relatedness between the candidate destination article and the surrounding context of the concept. Formally, these two metrics are defined as follows.

Definition 1. ***Link Probability.*** *Given a concept c identified in an article, the Link Probability from this concept c to another article a is defined by the approximated conditional probability of a given c:*

$$LP(a|c) = \frac{count(a, c)}{count(c)} \tag{2.1}$$

where $count(a, c)$ denotes the number of times that c links to a and $count(c)$ denotes the number of times that c appears in the whole wiki.

Definition 2. ***Semantic Relatedness.*** *Given a concept c, let N_c be the existing annotations in the same article with c, the Semantic Relatedness between a candidate destination article a and the context annotations of c is computed as:*

$$SR(a, c) = \frac{1}{|N_c|} \sum_{b \in N_c} r(a, b) \tag{2.2}$$

$$r(a, b) = 1 - \frac{log(max(|I_a|, |I_b|)) - log(|I_a \cap I_b|)}{log(|W|) - log(min(|I_a|, |I_b|))} \tag{2.3}$$

where I_a and I_b are the sets of inlinks of article a and article b, respectively; and W is the set of all articles in the input wiki.

The *Semantic Relatedness* metric has been used in several other approaches for linking documents to Wikipedia. The inner link structures in the wiki are used to compute the *Semantic Relatedness*. One thing worth mentioning is that when there are insufficient inner links, this metric might be not accurate enough to reflect the relatedness between articles. To this end, our approach first utilizes the known CLs to merge the link structures of two wikis, which results in a *Integrated Concept Network*. The *Semantic Relatedness* is computed based on the links in the *Integrated Concept Network*, which is formally defined as follows.

Definition 3. ***Integrated Concept Network.*** *Given two wikis $W_1 = (A_1, E_1)$, $W_2 = (A_2, E_2)$, where A_1 and A_2 are sets of articles, E_1 and E_2 are sets of links in W_1 and W_2, respectively. Let $CL = \{(a_i, b_i)|a_i \in B_1, b_i \in B_2\}_{i=1}^{k}$ be the set of cross-lingual links between W_1 and W_2, where $B_1 \subseteq A_1$ and $B_2 \subseteq A_2$. The Integrated Concept Network of W_1 and W_2 is $ICN(W_1, W_2) = (V, L)$, where $V = A_1 \cup A_2$; L is a set of links which are established as follows:*
$(v, v^{'}) \in E_1 \vee (v, v^{'}) \in E_2 \rightarrow (v, v^{'}) \in L$
$(v, v^{'}) \in E_1 \wedge (v, b) \in CL \wedge (v^{'}, b^{'}) \in CL \rightarrow (b, b^{'}) \in L$
$(v, v^{'}) \in E_2 \wedge (v, a) \in CL \wedge (a^{'}, v^{'}) \in CL \rightarrow (a, a^{'}) \in L$

In the *Integrated Concept Network*, the link structures of two wikis are merged by using CLs. The *Semantic Relatedness* computed based on *Integrated Concept Network* is more reliable than on a single wiki. In order to balance the *Link Probability* and the *Semantic Relatedness*, our approach uses an algorithm that greedily selects the articles that maximize the product of the two metrics. Algorithm 1 outlines the algorithm in the concept annotator.

Algorithm 1: Concept annotation algorithm.

```
Input: A wiki W = (A, E)
Output: Annotated wiki W = (A, E')

for each article a ∈ A do
    Extract a set of concepts C_a in a;
    Get the existing annotations N_a in a;
    for each c_i ∈ C_a do
        Find the set of candidate articles T_{c_i} of c_i;
        Get b* = arg max_{b∈T_{c_i}} LP(b|c_i) × SR(b, N_a);
        N_a = N_a ∪ {b*};
    end
    for each b_j ∈ N_a do
        if < a, b_j > ∉ E then
            E = E ∪ {< a, b_j >}
        end
    end
end
return N_d
```

(3) Cross-lingual Link Prediction

As shown in Figure 2.1, the CL predictor takes two wikis and a set of seed CLs between them as inputs, then it predicts a set of new CLs. This subsection first introduces the definitions of different similarity features, and then describes the learning method for predicting new CLs.

BIBLIOGRAPHY

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. "DBpedia - A crystallization point for the Web of Data". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3 (2009), pp. 154–165 (cit. on p. 5).

[2] Bo Fu, Rob Brennan, and Declan O'Sullivan. "Using pseudo feedback to improve cross-lingual ontology mapping". In: *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I*. ESWC'11. Heraklion, Crete, Greece: Springer-Verlag, 2011, pp. 336–351. ISBN: 978-3-642-21033-4 (cit. on p. 18).

[3] Xianpei Han, Le Sun, and Jun Zhao. "Collective entity linking in web text: a graph-based method". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. SIGIR '11. Beijing, China: ACM, 2011, pp. 765–774. ISBN: 978-1-4503-0757-4 (cit. on p. 18).

[4] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. "Collective annotation of Wikipedia entities in web text". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '09. Paris, France: ACM, 2009, pp. 457–466. ISBN: 978-1-60558-495-9 (cit. on pp. 8, 18).

[5] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. "RiMOM: A Dynamic Multistrategy Ontology Alignment Framework". In: *Knowledge and Data Engineering, IEEE Transactions on* 21.8 (2009), pp. 1218–1232. ISSN: 1041-4347 (cit. on p. 18).

[6] 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. Toronto, ON, Canada: ACM, 2010, pp. 1099–1108. ISBN: 978-1-4503-0099-5 (cit. on p. 5).

[7] Rada Mihalcea and Andras Csomai. "Wikify!: linking documents to encyclopedic knowledge". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. CIKM '07. Lisbon, Portugal: ACM, 2007, pp. 233–242. ISBN: 978-1-59593-803-9 (cit. on pp. 8, 18).

[8] David Milne and Ian H. Witten. "Learning to link with wikipedia". In: *Proceedings of the 17th ACM conference on Information and knowledge management*. CIKM '08. Napa Valley, California, USA: ACM, 2008, pp. 509–518. ISBN: 978-1-59593-991-3 (cit. on pp. 8, 18).

[9] Jong-Hoon Oh, Daisuke Kawahara, Kiyotaka Uchimoto, Jun'ichi Kazama, and Kentaro Torisawa. "Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia". In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 322–328. ISBN: 978-0-7695-3496-1. DOI: 10.1109/WIIAT.2008.317 (cit. on pp. 5, 18).

[10] Martin Potthast, Benno Stein, and Maik Anderka. "A Wikipedia-based multilingual retrieval model". In: *Proceedings of the IR research, 30th European conference on Advances in information retrieval*. ECIR'08. Glasgow, UK: Springer-Verlag, 2008, pp. 522–530. ISBN: 3-540-78645-7, 978-3-540-78645-0 (cit. on p. 5).

[11] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. "LINDEN: linking named entities with knowledge base via semantic knowledge". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. Lyon, France: ACM, 2012, pp. 449–458. ISBN: 978-1-4503-1229-5 (cit. on pp. 8, 18).

[12] Lipp Sorg and Philipp Cimiano. "Enriching the crosslingual link structure of Wikipedia - A classification-based approach". In: *AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*. 2008 (cit. on pp. 5, 18).

[13] Dennis Spohr, Laura Hollink, and Philipp Cimiano. "A machine learning approach to multilingual and cross-lingual ontology matching". In: *Proceedings of the 10th international conference on The semantic web - Volume Part I*. ISWC'11. Bonn, Germany: Springer-Verlag, 2011, pp. 665–680. ISBN: 978-3-642-25072-9 (cit. on p. 18).

[14] Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. "Using Bayesian decision for ontology mapping". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 4.4 (2006), pp. 243–262 (cit. on p. 18).

[15] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. "Cross-lingual knowledge linking across wiki knowledge bases". In: *Proceedings of the 21st international conference on World Wide Web*. Lyon, France, 2012, pp. 459–468. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187899 (cit. on pp. 5, 8, 18).

[16] Wolodja Wentland, Johannes Knopp, Carina Silberer Johannes Knopp, Carina Silberer, and Matthias Hartung. "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration". In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco, May 2008 (cit. on p. 5).

Feature Definition

Six features are defined to assess the similarities between articles by using different kinds of information, all of these features are based on the link structures and therefore are language-independent.

Feature 1: *Outlink similarity*

Outlinks of an article correspond to a set of other articles that it links to. The outlink similarity computes the similarities between articles by comparing elements in their outlinks. Given two articles $a \in W_1$ and $b \in W_2$, let $O(a)$ and $O(b)$ be the outlinks of a and b respectively. The outlink similarity is computed as

$$f_1(a,b) = \frac{2 \cdot |\phi_{1\to 2}(O(a)) \cap O(b)|}{|\phi_{1\to 2}(O(a))| + |O(b)|} \tag{2.4}$$

where $\phi_{1\to 2}(\cdot)$ is a function to maps articles in W_1 to their corresponding articles in W_2 if there are CLs between them.

Feature 2: *Outlink$^+$ similarity*

Given two articles $a \in W_1$ and $b \in W_2$, let $O^+(a)$ and $O^+(b)$ be the outlinks in the abstracts and infoboxes of a and b respectively. The outlink$^+$ similarity is computed as

$$f_2(a,b) = \frac{2 \cdot |\phi_{1\to 2}(O^+(a)) \cap O^+(b)|}{|\phi_{1\to 2}(O^+(a))| + |O^+(b)|} \tag{2.5}$$

Feature 3: *Inlink similarity*

Inlinks of an article correspond to a set of other articles linking to it, let $I(a)$ and $I(b)$ denote inlinks of two articles, the inlink similarity is computed as

$$f_3(a,b) = \frac{2 \cdot |\phi_{1\to 2}(I(a)) \cap I(b)|}{|\phi_{1\to 2}(I(a))| + |I(b)|} \tag{2.6}$$

Feature 4: *Inlink$^+$ similarity*

Let $I^+(a)$ and $I^+(b)$ denote two articles' inlinks that are from other articles' abstracts and infoboxes, the inlink$^+$ similarity is computed as

$$f_4(a,b) = \frac{2 \cdot |\phi_{1\to 2}(I^+(a)) \cap I^+(b)|}{|\phi_{1\to 2}(I^+(a))| + |I^+(b)|} \tag{2.7}$$

Feature 5 *Category similarity*

Categories are tags attached to articles, which represent the topics of the articles' subjects. Let $C(a)$ and $C(b)$ denote categories of two articles, the category similarity is computed as

$$f_5(a,b)(= \frac{2 \cdot |\phi_{1\to 2}(C(a)) \cap C(b)|}{|\phi_{1\to 2}(C(a))| + |C(b)|} \tag{2.8}$$

Here $\phi_{1\to 2}(\cdot)$ maps categories from one wiki to another wiki by using the CLs between categories, which can be obtained from Wikipedia.

Feature 6 *Category$^+$ similarity*

Given two articles a and b, let $C(a)$ and $C(b)$ be the categories of a and b respectively. Category$^+$ similarity computes similarities between categories by using CLs between articles. Let $E(c_1)$ and $E(c_2)$ be the set of articles belonging to category c_1 and c_2 respectively, the similarity between two categories is computed as

$$f_6(a,b) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \lambda(c_i^a, c_j^b) \tag{2.9}$$

$$\lambda(c_1, c_2) = \frac{2|\phi_{1\to 2}(E(c_1)) \cap E(c_2)|}{|\phi_{1\to 2}(E(c_1))| + |E(c_2)|} \tag{2.10}$$

where $c_i^a \in C(a)$, $c_j^b \in C(b)$, $n = |C(a)|$, and $m = |C(b)|$.

Regression-based Model for Predicting new CLs

The CL predictor in our approach computes the weighted sum of different similarities between articles, and applies a threshold ω_0 to decide whether an article pair should have a CL. For this purpose, a scoring function is defined:

$$\begin{aligned} s(a,b) &= \omega_0 + \vec{\omega} \cdot \vec{f}_{a,b} \\ &= \omega_0 + \omega_1 \times f_1(a,b) + ... + \omega_6 \times f_6(a,b) \end{aligned} \tag{2.11}$$

For each article a in W_1, the article b^* in W_2 that maximizes the score function $s(a,b^*)$ and satisfies $s(a,b^*) > 0$ is predicted as the corresponding article of a. The idea of CL prediction is simple and straightforward, but how to appropriately set the weights of different similarity features is a challenging problem.

Here, we propose a regression-based model to learn the weights of features based on a set of known CLs. Given a set of CLs $\{(a_i, b_i)\}_{i=1}^n$ as training data, let $A = \{(a_i)\}_{i=1}^n$ and $B = \{(b_i)\}_{i=1}^n$, our model tries to find the optimal weights to ensure:

$$\forall a_i \in A, \forall b^{'} \in (B - \{b_i\}), s(a_i, b_i) - s(a_i, b^{'}) > 0$$

which also means

$$\vec{\omega} \cdot (\vec{f}_{a_i,b_i} - \vec{f}_{a_i,b^{\cdot}}) > 0$$

Therefore, we generate a new dataset $D = \{(x_i, y_i)\}_{i=1}^n$, where the input vector $x_i = (\vec{f}_{a_i,b_i} - \vec{f}_{a_i,b_{j\neq i}})$, the target output y_i is always set to 1. Then we train a linear regression model on the dataset D to get the weights of different features. The threshold ω_0 is set to the value that maximizes the F1-measure on the training CLs.

3. Experiments

Table 3.1: Results of cross-lingual link prediction (%).

		Before Annotation			After Annotation		
#Seed CLs	Model	Precision	Reccall	F1-measure	Precision	Recall	F1-measure
0.05 Mil. CLs	SVM	92.1	35.0	50.7	78.5	37.2	50.5
	RM	93.3	36.0	**52.0**	92.4	38.6	**54.5**
0.10 Mil. CLs	SVM	79.7	35.0	48.6	86.9	50.4	63.8
	RM	84.6	37.4	**51.9**	96.6	49.3	**65.3**
0.15 Mil. CLs	SVM	80.9	35.9	49.7	88.1	57.3	69.5
	RM	93.5	38.2	**54.2**	93.7	56.2	**70.2**
0.20 Mil. CLs	SVM	84.7	37.3	51.8	88.8	68.1	77.1
	RM	94.5	37.9	**54.1**	95.9	67.2	**79.0**

The data of English Wikipedia and Chinese Wikipedia (archived in August 2012) has been used to evaluate our proposed approach. The English Wikipedia contains 4 million articles, and the Chinese Wikipedia contains 499 thousand articles, and there are already 239,309 cross-lingual links between English Wikipedia and Chinese Wikipedia. Using these data, we first evaluate the effectiveness of concept annotation and CL prediction methods separately, and then evaluate the final result obtained by the proposed approach.

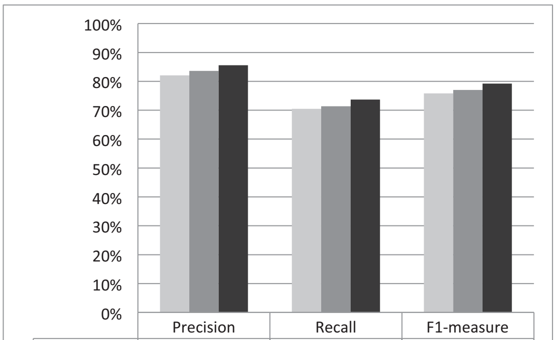
(1) Concept Annotation

For the evaluation of the Concept Annotation method, we randomly selected 1000 articles from Chinese Wikipedia. There are 12,723 manually created inner links in these selected articles. 70% of the existing inner links in the selected articles were randomly chosen as the ground truth for the evaluation, which were removed from the articles before the articles are fed to the annotation algorithm. After the annotation, we collected the new discovered inner links, and computed the Precision, Recall and F1-measure of the new discovered inner links based on the ground truth links.

One advantage of our concept annotation method is to leverage the cross-lingual information to enhance the concept annotation. Therefore, we did experiments with different number of CLs used in the concept annotation. Figure 3.1 shows the experimental results of concept annotation. In the experiments, we respectively used 0, 0.1million and 0.2 million CLs, respectively. As shown in Figure 3.1, our approach can obtain a precision of 82.1% and a recall of 70.5% without using any CLs. The performance can be significantly boosted when using some seed CLs in our approach. For example, when 0.2 million CLs are used, the obtained precision is increased by 3.5% and the recall is increased by 3.2%. Moreover, a large part of missing inner links can be discovered by our concept annotation method. The newly discovered CLs can be again used to improve the quality of the discovered inner links.

(2) Cross-lingual Links Prediction

To evaluate the CL prediction method, we randomly selected 3,000 articles having CLs to English articles from Chinese Wikipedia. The existing CLs between these articles were used as the ground truth for the evaluation. The seed CLs were extracted from the rest of 236,309 existing CLs between English Wikipedia and Chinese Wikipedia. In the experiments, we aimed to investigate how the CL prediction method performed before and after the concept annotation, and how the CL prediction method performed with different numbers of seeding CLs. Therefore, we did three groups of experiments, with each group using different number of seeding cross-lingual links. In each group of experiments, we also compared the



	Precision	Recall	F1-measure
No CLs	82.1%	70.5%	75.9%
0.1 Mil. CLs	83.6%	71.4%	77.0%
0.2 Mil. CLs	85.6%	73.7%	79.2%

Figure 3.1: Results of concept annotation.

performance of our method with the SVM classification model, which was used in previous knowledge linking approaches. Table 3.1 shows the experimental results.

In each group of the experiment, our regression-based learning model (RM) always performed better than the SVM model in terms of F1-measure. It shows that our RM model is more suitable for the knowledge linking problems than the traditional classification model SVM.

Before the concept annotation being performed, the performance (e.g., by F1-measure) of our approach does not always increase with more seed CLs being used. On the other hand, after the concept annotation process, the F1-score increases clearly as the number of seed CLs increases. For example, there is a 24.5% increment of the F1-measure when the number of seed CLs increases from 0.05 million to 0.20 million. This might be because the equivalent relations of seed CLs cannot be propagated to other article pairs when there are few inner links in wikis. Therefore, enriching the inner links is important for knowledge linking and can effectively improve the performance of CL prediction.

(3) Incremental Knowledge Linking

Finally, we evaluated our approach as a whole. Since the proposed approach aims to discover CLs with a small set of seeding links, we did not use all the available CLs as seed links. Instead, we randomly selected 50 thousand of them as the seed links. We also randomly removed 70% inner links in the Chinese Wikipedia (for both articles having cross-lingual links and having no cross-lingual links). We fed the seed links and two wikis to our approach and run the proposed approach iteratively to incrementally find new CLs. Figure 3.2 shows the number of newly discovered CLs in each run. The gray bars plot the results without concept annotation and the black bars indicate the results when the concept annotation was performed.

According to the results, new CLs can be discovered in each run no matter whether the concept annotation was performed or not. However, there would be more CLs discovered when the concept annotation was first performed. In the fourth run, more CLs were discovered than in the third run when using concept annotation, but less CLs were found in the fourth run than in the third run without concept annotation. In summary, 171,393 more CLs were discovered when the concept annotation were performed.

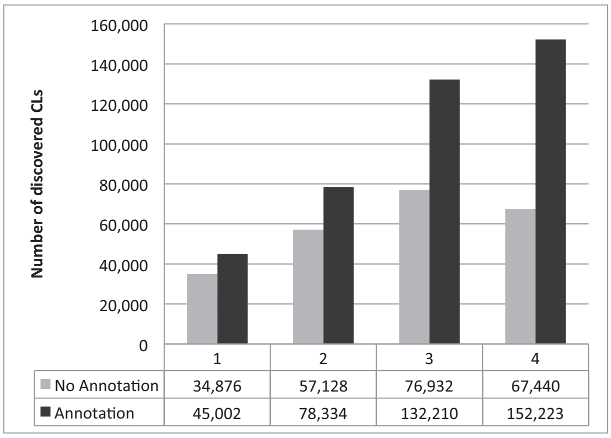


Figure 3.2: Results of incremental knowledge linking.

4. Related Work

Concept Annotation.

Recently, several approaches have been proposed to link documents to Wikipedia. Wikify! [7] is a system which automatically performs the annotation task following the Wikipedia guidelines. Wikify! first extracts keywords in documents, then find the links from these keywords to their corresponding articles by combining both knowledge-based approach and data-driven method. Milne et al. [8] proposed a learning based approach for linking entities in text to Wikipedia. A classification algorithm is also used in the candidate link detection. Shen et al. [11] proposed a system LINDEN, which explores semantic information in YAGO ontology to predict correct links from documents to Wikipedia. Kaulkarni et al. [4] and Han et al. [3] proposed collective approaches that aim to take the relations between annotation results into account. Different from these approaches, our concept annotation method uses the cross-lingual links to improve the annotation results. Only two measures, the link probability and the semantic relatedness, are used to decide the links from concepts to their corresponding wiki articles, which guarantees the efficiency of annotation process.

Cross-lingual Knowledge Linking.

There have been several approaches for cross-lingual knowledge linking. Sorg and Cimiano [12] proposed a method to find missing CLs between English and German. Their method uses the link structure of articles to find candidates of missing links. A classifier is trained based on several graph-based features and text-based features to predict CLs. Oh et al. [9] proposed a method for discovering missing cross-lingual links between English and Japanese. Wang et al. [15] employed a factor graph model which only used link-based features to find CLs between a Chinese wiki and English Wikipedia. In our approach, inner links appearing in different parts of articles are used separately in different similarity features, which is different from the existing approaches. And our approach uses a regression-based learning model to learn weights to aggregate similarities. From a broad viewpoint, ontology matching is also relevant. Approaches such as RiMOM [5, 14] have been proposed. Cross-lingual ontology matching is a more relevant problem to cross-lingual knowledge linking. However, existing approaches, e.g., [13, 2], mainly use the machine translation tools to bridge the gap between languages, which makes the approaches heavily dependent on the language itself. In our approach, all the defined features are language-independent.

5. Conclusions and Future Work

In this work, we propose an approach that boosts cross-lingual knowledge linking by concept annotation. New inner links are first found in wikis by a concept annotation method, which are then used for predicting new CLs. The concept annotation and the CL prediction are designed to mutually reinforce each other, which allows new CLs to be incrementally discovered. Experiments show that the concept annotation can effectively improve the results of CL prediction. The proposed approach can also effectively find new CLs based on a small set of seed CLs and inner links.

The knowledge linking is a core problem in knowledge base and represents a new research direction. There are a number of potential future directions of this work, for example, to design an online algorithm for knowledge linking, or to design a scalable algorithm to handle large knowledge base with billions of entities.

BIBLIOGRAPHY

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. "DBpedia - A crystallization point for the Web of Data". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3 (2009), pp. 154–165 (cit. on p. 5).

[2] Bo Fu, Rob Brennan, and Declan O'Sullivan. "Using pseudo feedback to improve cross-lingual ontology mapping". In: *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I*. ESWC'11. Heraklion, Crete, Greece: Springer-Verlag, 2011, pp. 336–351. ISBN: 978-3-642-21033-4 (cit. on p. 18).

[3] Xianpei Han, Le Sun, and Jun Zhao. "Collective entity linking in web text: a graph-based method". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. SIGIR '11. Beijing, China: ACM, 2011, pp. 765–774. ISBN: 978-1-4503-0757-4 (cit. on p. 18).

[4] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. "Collective annotation of Wikipedia entities in web text". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '09. Paris, France: ACM, 2009, pp. 457–466. ISBN: 978-1-60558-495-9 (cit. on pp. 8, 18).

[5] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. "RiMOM: A Dynamic Multistrategy Ontology Alignment Framework". In: *Knowledge and Data Engineering, IEEE Transactions on* 21.8 (2009), pp. 1218–1232. ISSN: 1041-4347 (cit. on p. 18).

[6] 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. Toronto, ON, Canada: ACM, 2010, pp. 1099–1108. ISBN: 978-1-4503-0099-5 (cit. on p. 5).

[7] Rada Mihalcea and Andras Csomai. "Wikify!: linking documents to encyclopedic knowledge". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. CIKM '07. Lisbon, Portugal: ACM, 2007, pp. 233–242. ISBN: 978-1-59593-803-9 (cit. on pp. 8, 18).

[8] David Milne and Ian H. Witten. "Learning to link with wikipedia". In: *Proceedings of the 17th ACM conference on Information and knowledge management*. CIKM '08. Napa Valley, California, USA: ACM, 2008, pp. 509–518. ISBN: 978-1-59593-991-3 (cit. on pp. 8, 18).

[9] Jong-Hoon Oh, Daisuke Kawahara, Kiyotaka Uchimoto, Jun'ichi Kazama, and Kentaro Torisawa. "Enriching Multilingual Language Resources by Discovering Missing Cross-Language Links in Wikipedia". In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 322–328. ISBN: 978-0-7695-3496-1. DOI: 10.1109/WIIAT.2008.317 (cit. on pp. 5, 18).

[10] Martin Potthast, Benno Stein, and Maik Anderka. "A Wikipedia-based multilingual retrieval model". In: *Proceedings of the IR research, 30th European conference on Advances in information retrieval*. ECIR'08. Glasgow, UK: Springer-Verlag, 2008, pp. 522–530. ISBN: 3-540-78645-7, 978-3-540-78645-0 (cit. on p. 5).

[11] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. "LINDEN: linking named entities with knowledge base via semantic knowledge". In: *Proceedings of the 21st international conference on World Wide Web*. WWW '12. Lyon, France: ACM, 2012, pp. 449–458. ISBN: 978-1-4503-1229-5 (cit. on pp. 8, 18).

[12] Lipp Sorg and Philipp Cimiano. "Enriching the crosslingual link structure of Wikipedia - A classification-based approach". In: *AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*. 2008 (cit. on pp. 5, 18).

[13] Dennis Spohr, Laura Hollink, and Philipp Cimiano. "A machine learning approach to multilingual and cross-lingual ontology matching". In: *Proceedings of the 10th international conference on The semantic web - Volume Part I*. ISWC'11. Bonn, Germany: Springer-Verlag, 2011, pp. 665–680. ISBN: 978-3-642-25072-9 (cit. on p. 18).

[14] Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. "Using Bayesian decision for ontology mapping". In: *Web Semantics: Science, Services and Agents on the World Wide Web* 4.4 (2006), pp. 243–262 (cit. on p. 18).

[15] Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang. "Cross-lingual knowledge linking across wiki knowledge bases". In: *Proceedings of the 21st international conference on World Wide Web*. Lyon, France, 2012, pp. 459–468. ISBN: 978-1-4503-1229-5. DOI: `10.1145/2187836.2187899` (cit. on pp. 5, 8, 18).

[16] Wolodja Wentland, Johannes Knopp, Carina Silberer Johannes Knopp, Carina Silberer, and Matthias Hartung. "Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration". In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco, May 2008 (cit. on p. 5).