Cross-lingual Knowledge Linking Across Wiki Knowledge Bases

Zhichun Wang, Juanzi Li, Zhigang Wang, and Jie Tang
Department of Computer Science and Technology
Tsinghua University, Beijing, China
{zcwang, ljz, wzhigang, tangjie}@keg.cs.tsinghua.edu.cn

ABSTRACT

Wikipedia becomes one of the largest knowledge bases on the Web. It has attracted 513 million page views per day in January 2012. However, one critical issue for Wikipedia is that articles in different language are very unbalanced. For example, the number of articles on Wikipedia in English has reached 3.8 million, while the number of Chinese articles is still less than half million and there are only 217 thousand cross-lingual links between articles of the two languages. On the other hand, there are more than 3.9 million Chinese Wiki articles on Baidu Baike and Hudong.com, two popular encyclopedias in Chinese. One important question is how to link the knowledge entries distributed in different knowledge bases. This will immensely enrich the information in the online knowledge bases and benefit many applications. In this paper, we study the problem of cross-lingual knowledge linking and present a linkage factor graph model. Features are defined according to some interesting observations. Experiments on the Wikipedia data set show that our approach can achieve a high precision of 85.8% with a recall of 88.1%. The approach found 202,141 new cross-lingual links between English Wikipedia and Baidu Baike.

Categories and Subject Descriptors

E.2 [**Data**]: Data Storage Representations—*Linked representations*; H.3 [**Information Systems**]: Miscellaneous—*Information Storage and Retrieval*

General Terms

Algorithms, Languages

Keywords

Knowledge Linking, Cross-lingual, Wiki knowledge base, Knowledge sharing

1. INTRODUCTION

Cross-lingual knowledge linking is the task of creating links between articles in multiple different languages that reports on the same content. Cross-lingual knowledge linking not only globalizes the knowledge sharing of different languages on the Web, but also benefits many online applications such as information retrieval and machine translation. For example, [20] explores cross-lingual links in machine translation and [30] studies how to improve information retrieval by leveraging cross-lingual knowledge links. The project of DBpedia [6][4] provides a semantic representation of Wikipedia in which multiple language labels are attached to individual concepts. The idea of cross-lingual linking has already become the nucleus for the linked data [6].

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

However, most traditional resources are monolingual, such as Cyc [23] and WordNet [25] in English, HowNet [12] in Chinese. Wikipedia tries to deal with this problem by providing information in different languages. Wikipedia contains 19 million articles in 281 languages. Articles in different languages are interlinked. However, the number the articles in different languages is very unbalanced. Figure 1 shows the number of articles in 12 different languages on Wikipedia. As it can be seen that there are 3.8 million English articles, but only 382,000 Chinese articles on Wikipedia. This makes it infeasible to create cross-lingual links between articles of different languages with a large coverage.

On the other hand, there are several separated Wiki knowledge bases on the Web. For example, Baidu Baike and Hudong.com are two Chinese Wiki knowledge bases containing more than 3.9 million articles. Ideally, automatically creating cross-lingual links between these Chinese Wiki knowledge bases and the English Wikipedia would be very useful. However, at present, the work is mainly taken by manual, which is obviously tedious, time consuming, and error prone. In existing literature, a few approaches have been proposed for finding missing cross-lingual links in Wikipedia [29, 31]. However, as we mentioned before, the number of articles in different languages is very unbalanced. For most English articles, we will be not able to find the corresponding Chinese version. To the best of our knowledge, no previous work has extensively studied the problem of creating cross-lingual links across different knowledge bases (e.g., the English Wikipedia and the Chinese encyclopedia Baidu Baike).

In this paper, we try to systematically study the problem of cross-lingual knowledge linking across multiple Wiki knowledge bases. The problem is non-trivial and poses a set of challenges.

- *Linguistics.* Existing methods for finding cross-lingual links heavily depend on translation tools. Such a method often results in high precisions, but low recalls. Can we find some language-independent features for mining cross-lingual knowledge links?
- *Model.* There are different kinds of information that

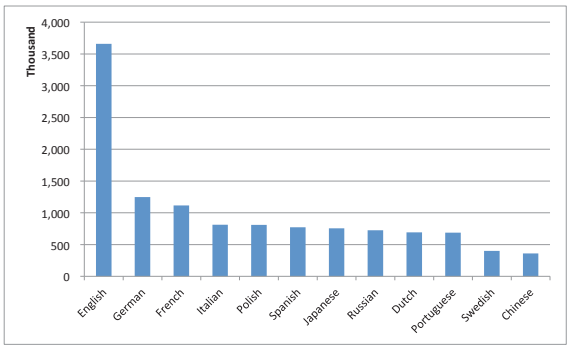


Figure 1: Number of articles in different languages in Wikipedia.

Figure 2: An example of knowledge linking.

could be used in the knowledge linking problem such as article links, categories and authors, and they are correlated with each other. How to define a model to incorporate both the local features of articles and relations of cross-lingual links together?

- *Efficiency.* Wiki knowledge bases contain thousands and millions of articles. How to develop an effective and efficient algorithm that can deal with both complex and large data sets?

In order to solve the above challenges, we first empirically investigate several important factors for cross-lingual knowledge linking and then propose a factor graph model to solve the knowledge-linking problem. Our contributions include:

- We formally formulate the problem of knowledge linking across Wiki knowledge bases in different languages, and analyze important factors for cross-lingual knowledge linking.
- We present a unified model for solving cross lingual knowledge linking problem: Linkage Factor Graph (LFG) model. Effective candidate selection method and distributed learning algorithm enable LFG scale to large data sets.
- We evaluate our proposed approach on existing cross-lingual links in Wikipedia; it achieves high precision of 85.8% with a recall of 88.1%. Using our model, we successfully identify 202,141 new cross-lingual links between English Wikipedia and Baidu Baike, which doubles the number of existing cross-lingual links on Wikipedia.

The rest of this paper is organized as follows, Section 2 formally defines the problem of knowledge linking and some related concepts; Section 3 presents some motivational analysis on collected data sets; Section 4 describes the proposed knowledge linking approach; Section 5 presents the evaluation results; Section 6 outlines some related work and finally Section 7 concludes this work.

2. PROBLEM FORMULATION

In this section, we formally define the knowledge linking problem. Here, we first define the Wiki knowledge base as follows according to mechanism of the Wiki knowledge bases.

DEFINITION 1. *A **Wiki knowledge base** is a collection of collaboratively written articles, each of which defines a specific concept. It can be formally represented as* $K = \{a_i\}_{i=1}^{n}$*, where* a_i *is an article in* K *and* n *is the size of* K*.*

Articles are the key elements in a Wiki knowledge base. Each article describes a specific concept. Articles are connected with categories, authors, and other articles. Thus an article $a_i \in K$ can be represented as a five-tuple $(T(a_i), I(a_i), O(a_i), C(a_i), U(a_i))$, where $T(a_i)$ denotes the title of the article; $O(a_i)$ is the set of outlinks of a_i, which denotes the set of articles that are mentioned in the content of a_i; $I(a_i)$ is the set of inlinks of a_i, which denotes the set of articles that link to a_i; $C(a_i)$ represents category tags of a_i, and $U(a_i)$ represents the article's authors.

DEFINITION 2. ***Knowledge linking.*** *Given two Wiki knowledge bases* K_1 *and* K_2*, knowledge linking is the process of finding, for each article* $a_i \in K_1$ *from knowledge base* K_1*, an equivalent article* $a_j \in K_2$ *in knowledge base* K_2*. When the two Wiki knowledge bases are in different languages, we call it the cross-lingual knowledge linking problem.*

Here, we say two articles are equivalent if they *semantically* describe a same subject or topic. Figure 2 shows an example for a possible knowledge linking result.

As shown in Figure 2, the article "Anaerobic exercise" is from English Wikipedia and the other article "无氧运动" is from Baidu Baike. There is not a cross-lingual link from "Anaerobic exercise" to any article in Chinese on Wikipedia. In the cross-lingual knowledge linking problem, our goal is to find an equivalent article "无氧运动" in Baidu Baike for the English article "Anaerobic exercise" from Wikipedia. In order to find the equivalent relations between articles, different features of articles can be considered. Figure 2 highlights some useful features in the two articles, including title, outlinks, categories and authors. Please note that we are more interested in the link-based features, instead of the linguistic-based features. This is because the former is more general and can be easily adapted to other languages while the latter is heavily dependent on the specific language. Given this, the knowledge base can be represented as a graph of linked articles, which is referred to as *Citation Graph.*

DEFINITION 3. ***Citation Graph.*** *A Wiki knowledge base is represented as a citation graph* $CG(K) = (A, L)$*, where*

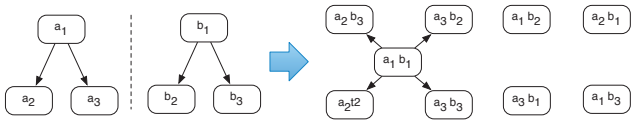


Figure 3: Pair-wise connectivity graph.

the node set A represents all the articles in K, edge $(a_i, a_j) \in L$ denotes a link from a_i to a_j, satisfying $a_j \in O(a_i)$ and $a_i \in I(a_j)$.

In this paper, we try to solve the knowledge linking problem as predicting the label (equivalent or not equivalent) of article pairs between two Wiki knowledge bases. Given two Wiki knowledge bases, we first build their Citation Graphs, and then construct a Pair-wise Connectivity Graph of two Citation Graphs, which is defined as follows.

DEFINITION 4. ***Pair-wise Connectivity Graph.*** *Given two graphs $G_1 = (A_1, L_1)$ and $G_2 = (A_2, L_2)$, the Pair-wise Connectivity Graph (PCG) of them is*

$$PCG(G_1, G_2) = (V, E)$$

where each element in the node set V denotes a node-pair between A_1 and A_2; the set of edges E in $PCG(G_1, G_2)$ is established as follows:

$$(a_1, a_2) \in L_1 \wedge (b_1, b_2) \in L_2 \Longleftrightarrow ((a_1, b_1), (a_2, b_2)) \in E$$

Figure 3 shows an example of the constructed PCG of two graphs. There are two graphs, each of them having 3 nodes. The constructed PCG between them (Cf. the right of Figure 3) contains 9 nodes representing all the possible node-pairs of two graphs. PCG can represent the linking relations of node-pairs between two graphs, we use PCG to capture the interaction of cross-lingual links between two Wiki knowledge bases. Our proposed approach takes PCG of two Wiki knowledge bases as input and solves the knowledge linking problem by predicting the label (equivalent or inequivalent) of nodes in the PCG.

3. DATA OBSERVATIONS

3.1 Data Collection

What are the fundamental factors underlying the formation of cross-lingual knowledge links? We first use existing cross-lingual links in Wikipedia to investigate which factors are of important for knowledge linking. Here, we download English Wikipedia and Chinese Wikipedia dumps from Wikipedia's website and extract cross-lingual links between them. The English Wikipedia dump was archived in April 2011, and the Chinese Wikipedia dump was archived in October 2011. Table 1 shows some statistics of the collected data sets. We have extracted 180,807 cross-lingual links from English Wikipedia to Chinese Wikipedia, and 205,608 cross-lingual links from Chinese to English. Finally, by merging them together, we obtain 217,689 cross-lingual links. We have also extracted 35,294 cross-lingual links between category pages in Chinese and English.

Table 1: Statistics for our data sets

Knowledge base	#Articles	#Categories	#Authors
English Wikipedia	3,786,000	531,771	3,592,495
Chinese Wikipedia	382,000	97,045	91,226

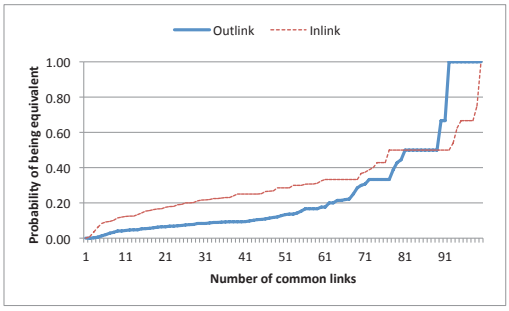


Figure 4: Probability of being equivalent conditioned on the number of common links.

3.2 Observations

Based on the above data sets, we first investigate what factors will be helpful for predicting the cross-lingual links between Wiki articles. In particular, we study the correlation of the following factors with the cross-lingual links: (1) Link homophily: do articles linking to or linked by equivalent articles tend to be equivalent? (2) Category homophily: do articles have semantically equivalent category tags tend to be equivalent? (3) Author interest: are authors' interests useful for finding cross-lingual links? We randomly selected 10,000 English-Chinese article pairs connected by cross-lingual links from Wikipedia. We generate all possible $10,000 \times 10,000$ article pairs from the selected articles; 10,000 of them are equivalent pairs and the others are inequivalent pairs. Considering all these article pairs as a sample set, we make some analyses to figure out whether the above factors are helpful to knowledge linking.

Link homophily. If two articles cite two other articles which have an equivalent relationship, we say the two articles have a common outlink. Similarly, if two articles are cited by two other equivalent articles, we say they have a common inlink. We calculate the probabilities of being equivalent conditioned on the number of common outlinks and inlinks in the data set. As shown in Figure 4, the probabilities of being equivalent grows as the number of common outlinks and inlinks increase. It is obvious that the number of common links is relevant to the equivalent relation of articles, and the inlinks seems more important than outlinks.

Category homophily. If two articles belong to two categories which have an equivalent relationship, we say they have a common category. Figure 5 shows the probability of two articles be equivalent conditioned on the number of common categories between them. It clearly shows a close correlation between common categories and equivalent relationship. When the number of common categories is more than 10, the probability of equivalent relation is close to 1.0, 20 times higher than the probability when they only

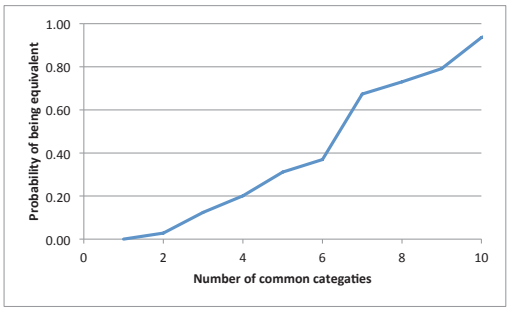


Figure 5: Probability of being equivalent conditioned on the number of common categories.

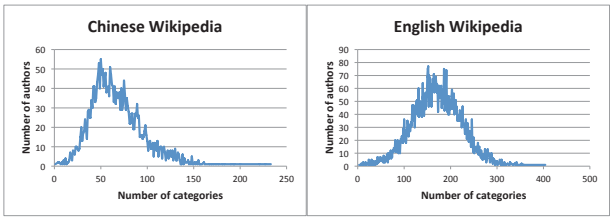


Figure 6: Distribution of author's interests.

have one common category.

Author interest. We presume that if two authors share similar interests, their articles will have a higher probability to be equivalent than that of two articles edited by random authors. Here, we simply define authors' interests by those categories to which the authors' edited articles belong. We choose authors who have edited more than 50 articles in both English and Chinese Wikipedia. Figure 6 shows the distributions of authors over the number of categories in two languages respectively. It appears to be normal distributions that most authors participated in a reasonable number of categories, and only a few have extremely small or large number categories. Therefore, most users concentrate on a fixed number of categories, and we may use author interests in the knowledge linking problem.

We also calculate the percentage of article pairs that have at least one common inlink, outlink, or category respectively. Figure 7 shows that a large portion of equivalent articles have common links and categories, with a probability much higher than that of two random articles. For outlinks, it has a similar pattern: the probability of two equivalent articles sharing a common outlink is 15 times higher than that of two random articles.

According to the above analyses, we have the following summaries:

- Common links and common categories have obvious correlations with cross-lingual links, but it seems that if a factor has higher correlation, it will have low coverage of cross-lingual links.
- Author interest would be an important factor to identify the equivalent relationships.
- Equivalent articles are very likely to share inlinks, outlinks, or categories. The probability is more than 15 times higher than that of two random articles.

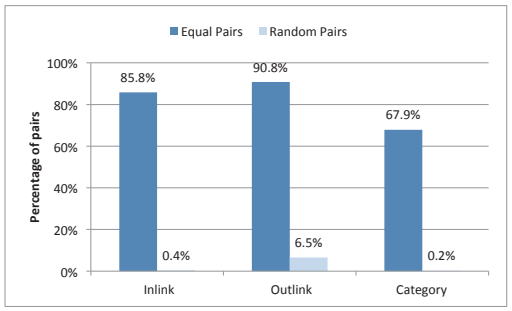


Figure 7: Percentage of article pairs have common inlinks, outlinks and categories.

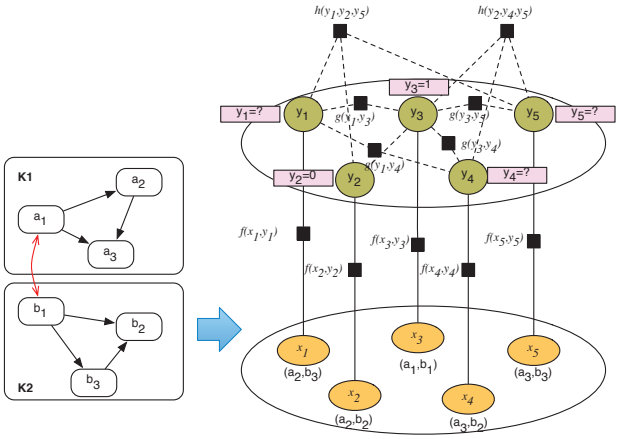


Figure 8: Graphical representation of the linkage factor graph (LFG) model.

4. THE PROPOSED APPROACH

In this section, we describe our proposed model and its learning algorithm in detail.

4.1 Linkage Factor Graph Model

Factor graph [22] assumes observation data are cohesive on both local features and relationships. It has been successfully applied in many applications, such as social influence analysis [34], social relationship mininig[19, 35, 37, 33], and linked data disambiguation [10]. In this work, we formalize the knowledge linking problem into a linkage factor graph model, which is shown in Figure 8. Given two Wiki knowledge bases $K_1 = \{A_1, L_1\}$ and $K_2 = \{A_2, L_2\}$, let $EL = \{e_i = (a_{i_1}, b_{i_2})\}_{i=1}^{p}$ be p existing cross-lingual links between K_1 and K_2, $a_{i_1} \in A_1$, $b_{i_2} \in A_2$, $|A_1| = n$ and $|A_2| = m$. The input of the LFG model is $PCG(K_1, K_2)$. Each node (a_{i_1}, b_{i_2}) in $PCG(K_1, K_2)$ is mapped to a observed variable x_i in LFG. There is also a set of hidden variables $Y = \{y_i\}_{i=1}^{n \cdot m}$, representing the labels (equivalent or inequivalent) of the observed variables.

We define three feature functions in LFG model:

- Node feature function: $f(y_i, x_i)$ is a feature function which represents the posterior probability of label y_i

given x_i; it describes local information on nodes in LFG;

- Edge feature function: $g(y_i, G(y_i))$ denotes the correlation between nodes via the edge on the graph model; $G(y_i)$ is the set of nodes having relations to y_i;
- Constraint feature function: $h(y_i, H(y_i))$ defines constraints on all relationships, where $H(y_i)$ is the set of relationships constrained on y_i.

Based on the LFG model, we can define joint distribution over Y as

$$p(Y) = \prod_i f(y_i, x_i) g(y_i, G(y_i)) h(y_i, H(y_i)) \quad (1)$$

In the following part, we introduce the definition of three feature functions in detail.

(1) Node feature function

$$f(y_i, x_i) = \frac{1}{Z_\alpha} exp\{\alpha^T \mathbf{f}(y_i, x_i)\} \quad (2)$$

where $\mathbf{f} =< f_{out}, f_{in}, f_{cate}, f_{auth} >$ is a vector of feature functions; α defines the corresponding weights; and variable x_i corresponds to article pair (a_{i_1}, b_{i_2}). Functions f_{out}, f_{in}, f_{cate} and f_{auth} are similarity functions based on outlinks, inlinks, categories and authors. The similarity functions are defined as follows:

(a) Outlink similarity function: it computes similarities between articles based on the equivalent articles in their outlinks.

$$f_{out} = \frac{2 \cdot |\{(a^{'}, b^{'}) | (a^{'}, b^{'}) \in EL, a^{'} \in O(a_{i_1}), b^{'} \in O(b_{i_2})\}|}{|O(a_{i_1})| + |O(b_{i_2})|} \quad (3)$$

(b) Inlink similarity function: it computes similarities between articles based on the equivalent articles in their inlinks.

$$f_{in} = \frac{2 \cdot |\{(a^{'}, b^{'}) | (a^{'}, b^{'}) \in EL, a^{'} \in I(a_{i_1}), b^{'} \in I(b_{i_2})\}|}{|I(a_{i_1})| + |I(b_{i_2})|} \quad (4)$$

(c) Category similarity function: it computes similarities between articles based on the equivalent categories between them.

$$f_{cate} = \frac{2 \cdot |\{(c, c^{'}) | (c, c^{'}) \in EC, c \in C(a_{i_1}), c^{'} \in C(b_{i_2})\}|}{|C(a_{i_1})| + |C(b_{i_2})|} \quad (5)$$

Here EC is a set of equivalent categories from two Wiki knowledge bases.

(d) Author interest similarity function: it computes similarities between articles based on their authors' mutual interests. In order to compute interest similarity between two authors, we first represent each author as a vector of categories they have participated, then compute the angle of two authors' feature vectors, as shown in Figure 9. Let $s(u_1, u_2)$ be the interest similarity of two authors, the author interest similarity of two articles is defined as

$$f_{auth} = \frac{1}{|U(a_{i_1})| \cdot |U(b_{i_2})|} \sum_{u_1 \in U(a_{i_1})} \sum_{u_2 \in U(b_{i_2})} s(u_1, u_2) \quad (6)$$

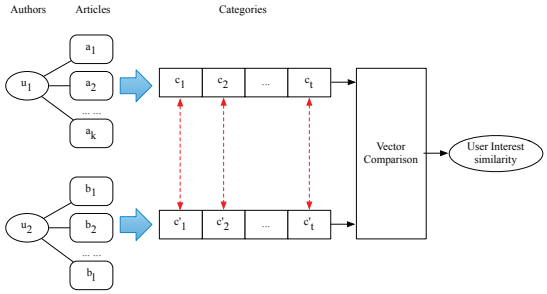


Figure 9: An illustration of computing interest similarity between authors (categories connected by red dash lines are equivalent)

(2) Edge feature function

$$g_(y_i, G(y_i)) = \frac{1}{Z_\beta} exp\{\sum_{y_j \in G(x_i)} \beta^T \mathbf{g}(y_i, y_j)\} \quad (7)$$

where $\mathbf{g}(y_i, y_j)$ is a function to specify whether there is a link from node i to node j in the $PCG(K_1, K_2)$; $\mathbf{g}(y_i, y_j) = 1$ if there is a edge from node i to node j, otherwise 0. Edge feature function is used to consider the relations between nodes in the model, which is based on the assumption that articles links to other two equivalent articles tend to be equivalent, too. We should notice that similarity functions f_{out} and f_{in} capture the relations between candidate cross-lingual links and existing ones, but $g_(y_i, G(y_i))$ is used to model the relations within candidate cross-lingual links.

(3) Constraint feature function

Here, we set a rule that one article from K_1 can only have cross-lingual link with one article from K_2, which is consistent with real circumstances. Therefore, we define the constrain feature function as

$$h(y_i, H(y_i)) = \frac{1}{Z_\gamma} exp\{\sum_{y_j \in H(y_i)} \gamma^T \mathbf{h}(y_i, y_j)\} \quad (8)$$

where $H(y_i)$ denotes the set of labels conflicting with y_i according to the 1-to-1 linking constraint. h is the contraint function, $h(y_i, y_j) = 0$ if $y_i = 1$ and $y_j = 1$, otherwise 1.

4.2 Model Learning and Inference

Given a set of labeled nodes in the LFG, learning the model is to estimate a optimum parameter configuration $\theta = (\alpha, \beta, \gamma)$ to maximize the log-likelihood function of $p(Y)$. Based on Equtions 1-8, the joint distribution $p(Y)$ can be written as

$$\begin{aligned} p(Y) &= \frac{1}{Z} \prod_i \exp\{\theta^T (\mathbf{f}(y_i, y_j), \sum_{y_j} \mathbf{g}(y_i, y_j), \sum_{y_j} \mathbf{h}(y_i, y_j))\} \\ &= \frac{1}{Z} \exp\{\theta^T \sum_i \mathbf{s}(y_i)\} = \frac{1}{Z} \exp\{\theta^T \mathbf{S}\} \end{aligned} \quad (9)$$

where all feature functions for a node y_i is briefly written as $\mathbf{s}(y_i) = (\mathbf{f}(y_i, y_j)^T, \sum_{y_j} \mathbf{g}(y_i, y_j)^T, \sum_{y_j} \mathbf{h}(y_i, y_j)^T)^T$; $Z = Z_\alpha Z_\beta Z_\gamma$, and $\mathbf{S} = \sum_i \mathbf{s}(y_i)$. Thus, the log-likelihood objective function is defined as

$$O(\theta) = \log p(Y^L) = \log \sum_{Y|Y^L} \frac{1}{Z} \exp\{\theta^T \mathbf{S}\} \\ = \log \sum_{Y|Y^L} \exp\{\theta^T \mathbf{S}\} - \log \sum_{Y} \exp\{\theta^T \mathbf{S}\} \tag{10}$$

where Y^L denotes the known labels and $Y|Y^L$ is a labeling configuration of Y inferred from Y^L. In order to maximize the object function, we adopt a gradient decent method. We calculate the gradient for each parameter θ

$$\frac{\partial O(\theta)}{\partial \theta} = \frac{\partial(\log \sum_{Y|Y^L} \exp\{\theta^T \mathbf{S}\} - \log \sum_Y \exp\{\theta^T \mathbf{S}\})}{\partial \theta} \\ = \frac{\sum_{Y|Y^L} \exp \theta^T \mathbf{S} \cdot \mathbf{S}}{\sum_{Y|Y^L} \exp \theta^T \mathbf{S}} - \frac{\sum_Y \exp \theta^T \mathbf{S} \cdot \mathbf{S}}{\sum_Y \exp \theta^T \mathbf{S}} \\ = \mathbb{E}_{p\theta(Y|Y^L)} \mathbf{S} - \mathbb{E}_{p\theta(Y)} \mathbf{S} \tag{11}$$

where $\mathbb{E}_{p\theta(Y|Y^L)}\mathbf{S}$ and $\mathbb{E}_{p\theta(Y)}\mathbf{S}$ are two expectations of $\mathbf{S}$, which cannot be directly calculated. Here, we use an extended version of the Loopy Belief Propagation algorithm [35] to approximate marginal probabilities $p(y_i|\theta)$ and $p(y_i, y_j|\theta)$. The general idea is to use two steps, one step for calculating $\mathbb{E}_{p\theta(Y|Y^L)}\mathbf{S}$ and the other step for calculating $\mathbb{E}_{p\theta(Y)}\mathbf{S}$, to estimate the gradient of a parameter θ wrt the objective function (Eq. 10). Interested readers please refer to [35] for details of the algorithm.

After learning the optimal parameters $\{\theta\}$, we can infer the unknown labels by finding a label configuration which maximizes the joint probability $p(Y)$

$$Y^* = \text{argmax}_{Y|Y^L} p(Y) \tag{12}$$

To do the inference in the above equation, we again perform the two-step LBP to compute marginal probabilities. Finally, each node in LFG is assigned with label that maximizes the marginal probability.

4.3 Candidate Selection and Distributed Learning

Finding cross-lingual links between two large Wiki knowledge bases is a challenging problem, because the number of nodes in LFG model will increase sharply. In order to handle large scale knowledge linking problems, we first use a candidate selection strategy to reduce the number of nodes in LFG, that is only article pairs that have at least one common outlink are mapped to nodes in the LFG model. According to the observation on existing cross-lingual links in Wikipedia, this candidate selection criterion can eliminate a large number of unnecessary nodes. Thus complexity of the resultant LFG can be effectively reduced within a small loss of recall.

We also implement the learning algorithm of LFG based on MPI to enable distributed learning. In the process of distributed learning, the LFG is first divided into several subgraphs that are assigned to slave computing nodes. Then LBP is performed on each slave nodes to compute the marginal probabilities and the parameter gradient. There is a master node collects and sums up all gradients from subgraphs, and updates parameters by gradient descent method. For details, please refer to [35].

5. EXPERIMENTS

In this paper, the proposed approach for cross-lingual knowledge linking is a general model. It can be used to find cross-lingual links between any Wiki knowledge bases in different languages. In this section, we first evaluate our approach on existing Chinese-English cross-lingual links within Wikipedia. And then we use our approach to find English-Chinese links between Wikipedia and Baidu Baike.

5.1 Experiment Settings

5.1.1 Dataset

In order to evaluate our approach, we construct a dataset that contains article pairs from English Wikipedia and Chinese Wikipedia. We randomly select 2,000 English articles with cross-lingual links to Chinese articles from Wikipedia, and then pick out the corresponding 2,000 Chinese articles. $2,000 \times 2,000$ article pairs are generated from the selected Chinese and English articles. Among all these article pairs, those 2,000 pairs linked by cross-lingual links are labeled as positive examples, and the rest of article pairs are labeled as negative examples.

5.1.2 Comparison Methods

We define four state-of-the-art cross-lingual linking methods as the comparison methods. They are translation based method Title Matching (TM), Similarity Aggregation (SA) based method, classification based method Support Vector Machine (SVM) and another classification based method (SVM-SC) based on the work of Sorg and Cimiano [31].

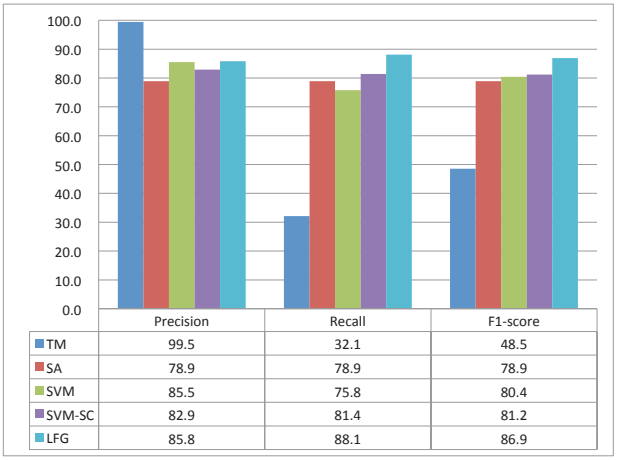
- **Title Matching (TM)**. This method first translates the titles of Chinese articles into English by Google Translation API [1], then matches the translated titles with English articles. For each article pair, if two articles have strictly the same English titles, they are considered as equivalent articles.

- **Similarity Aggregation (SA)**. This method aggregates different similarities of each article pair into a combined one. Then for each Chinese article, select the English article having the largest similarity with it to establish its cross-lingual link. Here, we compute outlink similarity, inlink similarity, category similarity and author interest similarity for each article pair, which are the same as defined in Section 4. For each article pair (a_{i_1}, b_{i_2}), its different similarities are aggregated by computing their average value:

$$Sim(a_{i_1}, b_{i_2}) = \frac{1}{4}(f_{out} + f_{in} + f_{cate} + f_{auth}) \tag{13}$$

- **Support Vector Machine (SVM)**. This method first computes the four similarities defined in Section 4 for each article pair, and then train a SVM [9] classification model on the known cross-lingual links, and predict the relationships of new article pairs. Compared to our approach, SVM only consider the similarity of articles' local features, it does not take the relations of predictions and any constraints into account. Here, we use SVM-Light package [3] in our experiment.

- **SVM-SC**. Sorg and Cimiano [31] defined several graph-based and text-based features between Wiki ar-



	Precision	Recall	F1-score
TM	99.5	32.1	48.5
SA	78.9	78.9	78.9
SVM	85.5	75.8	80.4
SVM-SC	82.9	81.4	81.2
LFG	85.8	88.1	86.9

Figure 10: Performance of knowledge linking with different methods (%).

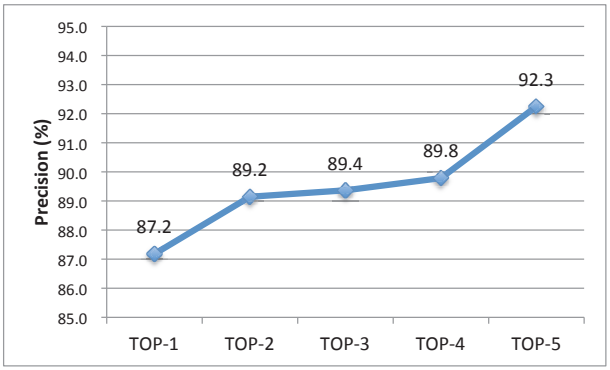


Figure 11: TOP-k precision of LFG.

ticles, and also trained a classifier to find missing cross-lingual links between German Wikipedia and English Wikipedia. Here we train a SVM with their features on evaluation dataset, and compare the results with our approach.

We use precision, recall and F1-score to evaluate different knowledge linking methods. For SVM, SVM-SC and LFG, we conduct 3-fold cross validation on the evaluation dataset. LFG uses 0.001 learning rate and runs 3000 iterations in all the experiments, SVM runs with the default settings in SVM-Light package. All experiments are carried out on a Windows 2008 server with 1.87GHz CPU (4 cores) and 6 GB memory.

5.2 Results Analysis

5.2.1 *Performance Comparison*

Figure 10 shows the performance of 5 different methods. According to the result, the TM method gets a really high precision of 99.5%, but its recall is only 32.1%. SA does not achieve good results by using simple averaging strategy. SVM and SVM-SC both use the same classification model but with different features. SVM gets better precision while SVM-SC gets better recall, SVM-SC outperforms SVM by 0.8% in terms of F1-score. LFG achieves the best recall and F1-score among all these methods. Compared to the SA method, LFG outperforms it by 8.0% in terms of F1-score. LFG and SVM both using the same training data, they have similar precisions, but LFG outperforms SVM by 12.3% in terms of recall, and has a 6.5% increase of F1-score. Therefore, our LFG model can discover more cross-lingual links by considering the relations between article pairs. LFG also performs better than SVM-SC in terms of both precision and recall.

We also evaluate LFG by *TOP-k* evaluation. For each Chinese article, we find all the nodes in LFG model related to it. Then *TOP-k* nodes are selected according to a ranking determined by the marginal probabilities $p(y_i = 1)$. We define *TOP-k-Precision* as the percentage of Chinese articles that have correct equivalent articles in its Top-k candidate set. We set $k = 1, 2, ..., 5$ respectively and calculate the *TOP-k* precisions of LFG.

Figure 11 shows the result of *TOP-k* evaluation. The precision grows as k increases, LFG achieves 92.3% precision when $k = 5$. Therefore, if we do not want to find the exact cross-lingual links, LFG can also provide candidates of cross-lingual links of high precision.

Table 2: Contribution analysis of different factors (%).

Ignored Factor	Pre.	Rec.	F1-score
Outlinks	82.2	83.8	83.0 (-3.9)
Inlinks	82.6	82.0	82.3 (-4.6)
Categories	84.2	84.9	84.6 (-2.3)
Authors' Interests	82.0	88.5	85.1 (-1.8)
Relations	83.3	84.2	83.8 (-3.1)
LFG	85.8	88.1	86.9

5.2.2 *Factor contribution analysis*

How much does each factor contribute to the LFG model? In order to get some insight to this question, we perform an analysis to evaluate the contribution of different factors. Here, we run LFG 5 times on the evaluation data, and each time remove one factor from LFG. Table 2 lists the results of ignoring different factors.

According to the decrement of F1-scores, all these factors are useful in predicting new cross-lingual links. It is reasonable to evaluate the importance of each factor by the decrease of F1-score without that factor. So we can rank these factors in a descending order of importance as inlinks, outlinks, relations, categories and authors' interests. Although the factor of authors' interests is less important than other factors, we find it is indeed helpful to improve the performance of LFG. Also, LFG achieves a 3.1% increase of F1-score by considering the relations among article pairs.

5.3 Discover new links between Wikipedia and Baidu

The motivation of our work is to find cross-lingual links across Wiki knowledge bases to get more equivalent article pairs in different languages. Therefore, we use LFG to discover cross-lingual links between English Wikipedia and Baidu Baike (a large scale Chinese Wiki knowledge base).

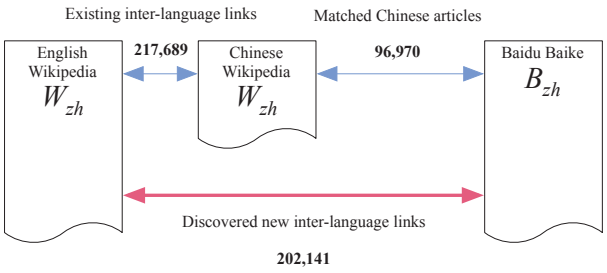


Figure 12: Illustration of existing and discovered inter-language links between articles.

We crawled 3,941,659 articles from Baidu Baike, which are edited by 1,454,204 authors and organized in 599,463 categories.

There have already been 217,689 cross-lingual links of articles and 35,294 cross-lingual links of categories between English Wikipedia and Chinese Wikipedia. Among those linked Chinese articles and categories, we have found 96,970 equivalent articles and 10,350 equivalent categories in Baidu Baike. Therefore, we establish 96,970 initial cross-lingual links of articles between English Wikipedia and Baidu Baike, and 10,350 initial cross-lingual links of categories between English Wikipedia and Baidu Baike. Based on our candidate selection method, we choose 3,082,751 English articles and 963,788 Chinese articles for LFG to discover cross-lingual links between them. We run LFG on a server with 1.87GHz CPU (4 cores) and 6 GB memory. It costs 17 hour 32 minutes to finally get 202,141 cross-lingual links between English Wikipedia and Baidu Baike. Figure 12 shows the relation of existing and discovered inter-language links between articles. Table 3 lists some of these cross-lingual links. There are generally four types these links, including persons (Per.), locations (Loc.), organizations (Org.) and scientific terms (Ter.). The discovered links are not available in Wikipedia. By using our approach, the number of cross-lingual links has been doubled.

Table 3: Examples of discovered cross-lingual links.

Types	English articles	Chinese articles
Per.	Kenneth Clark	肯尼斯·克拉克
	Heather Graham	海瑟·格拉汉姆
	Jeff Daniels	杰夫丹·尼尔斯
	Daniel Craig	丹尼尔·克雷格
	Mick McCarthy	杰克·麦卡锡
	Ralph Bellamy	拉尔·夫贝拉米
Loc.	Anticosti Island	安蒂科斯蒂岛
	Huehuetenango	韦韦特南戈
	San Juan Islands	圣胡安群岛
	Alabama River	亚拉巴马河
	Mandalay Hill	曼德勒山
Org.	Oslo City Hall	奥斯陆市政厅
	Yale University Library	耶鲁大学图书馆
	University of Troms	特罗姆瑟大学
	American Mafia	美国黑手党
	America West Airlines	美国西部航空公司
Ter.	Superconductivity	超导电性
	Wave propagation	电波传播
	Basal cell carcinoma	基底细胞癌
	Pleural effusion	胸腔积液
	Mildew	霉菌

6. RELATED WORK

In this section, we review some related work.

6.1 Discovering Missing Cross-lingual Links in Wikipedia

A group of highly related work is to discover missing cross-lingual links within Wikipedia. As more and more researches use the cross-lingual links of Wikipedia to build multilingual lexical resources, the problem of missing cross-lingual links has attracted increasingly attention. The missing of cross-lingual links means that there are corresponding articles within two languages, but there is no direct cross-lingual link between them. In order to solve this problem and enrich the cross-lingual links in Wikipedia, several approaches have been proposed. Sorg and Cimiano [31] proposed a method to find missing cross-lingual links between English and German. Their method makes use of the link structure of articles to find candidates of missing links. And then a classifier is trained based on several graph-based features and text-based features to predict the missing links. Oh et al. [29] proposed a method for discovering missing cross-lingual links between English and Japanese. Their method works in two steps, it first selects candidates of missing links based on cross-lingual similarities between English and Japanese Wikipedia articles, and then trains a classifier to predict whether a given candidate of missing links is correct or not. These two methods try to find the missing links within Wikipedia, while our proposed approach aims to find cross-lingual links across different Wiki knowledge bases. Furthermore, some features used by these two methods are not available in the task of finding cross-lingual links between Wikipedia and other Wiki knowledge bases. For example, Sorg and Cimiano used the orthographical similarity between English and German, which cannot be calculated between other language pairs, such as English and Chinese. The method proposed by Oh et al. used common images as a feature, which cannot be used across Wiki knowledge bases. Although both of the two methods train SVM to predict new cross-lingual links, they did not consider the relations between predictions.

Recently, many projects based on the cross-lingual links of Wikipeida have been proposed. DBpedia [6][4] is a knowledge based built by extracting structured information from Wikipedia. Currently DBpedia has described more than 3.5 million things, and 1.67 million of these things are classified in a consistent ontology. The DBpedia ontology organized things into persons, places, music albums, films, video games, organizations, species and diseases. Erdmann et al. [15] extracted a dictionary from Wikipedia by analyzing the link structure of Wikipdia. In addition to the cross-lingual links, they also explore the redirect page, link text to extend the coverage of the built dictionary. They first constructed a baseline dictionary by exploring the cross-lingual links in Wikipedia; and then extracted more translation candidates from redirect page and link text information. MENTA [11] is a multilingual entity taxonomy built from Wikipedia and WordNet. By aggregating unreliable taxonomic links between entities from different language ver-

sions of Wikipedia, a single more reliable and coherent taxonomy is build. HeiNER [39] is a multilingual Heidelberg Named Entity Resource, which translates Named Entities into the various target languages by exploiting cross-lingual information contained in the online encyclopedia Wikipedia. BabelNet [26] is a large multilingual semantic network built from Wikipedia and WordNet, which provides concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. Hassan et al. [18] address the task of cross-lingual semantic relatedness by exploiting the cross-lingual links available between Wikipedia versions in multiple languages. Ye et al. [40] proposed a graph-based approach to constructing a multilingual association dictionary from Wikipedia. The extracted association dictionary is applied in cross language information retrieval.

6.2 Ontology and Instance Matching

Ontology and instance matching is another related problem. As the development of the Linked Data project [2], ontology and instance matching is attracting more and more interests. The goal of otology and instance matching is to find equivalent elements between two heterogeneous semantic data sources. Currently, most work focus on monolingual matching tasks, such as Silk [36],idMesh [10],KnoFuss [28]. Some approaches such as SOCOM [17], RiMOM [24][32][41] and [16] deal with the cross-lingual ontology matching, they mainly use the machine translation tools to bridge the gap between languages. Our approach uses only language independent features of Wiki articles, which does not need any translation tools.

6.3 Record linkage

Record linkage is to identify records in the same or different databases that refer to the same real-world entity [14]. An record linkage approach typically compares various fields of database records, and either matches records based on domain knowledge and generic distance metrics, or applies supervised machine learning techniques to learn how to match the records [21]. Approaches based on supervised learning techniques include [7][5][8] . Unsupervised record linkage approaches include [27][38]. Some tools such as TAILOR [13] have been proposed for record linkage applications. Researches on record linkage try to find equivalent objects in databases, while our approach aims to find equivalent articles across Wiki knowledge bases in two different languages. Although the problem of record linkage has been studied for decades, few works on cross-lingual record matching have been proposed.

7. CONCLUSION AND FUTURE WORK

In this paper, we propose a cross-lingual knowledge linking approach for building cross-lingual links across Wiki knowledge bases. Our approach uses only language-independent features of article, and employs a graph model to predict new cross-lingual links. Evaluations on existing cross-lingual links in Wikipedia shows that our approach can achieve high precision 85.8% with a recall of 88.1%. Using our approach, we are able to find 202,141 new cross-lingual links between English Wikipedia and Baidu Baike.

Because the number of initial links is relative small based on the existing cross-lingual links in Wikipedia, we can find only part of new links between Wikipedia and Baidu Baike. If we add discovered new links to existing links, and use the merged links as seed, our approach can iteratively discover new linked articles. Therefore, our future work is to extending our approach to an iterative one, to find more cross-lingual links.

8. ACKNOWLEDGEMENTS

We thank Honglei Zhuang and Wenbin Tang for providing their software and useful suggestions. We also thank anonymous reviewers for valuable suggestions and comments. This work is supported by the National Natural Science Foundation of China (No. 61073073, 61035004, 661035004, 60973102), 863 high technology program (2011AA01A207) and European Union 7th framework project FP7-288342. It is also partially supported by THU-NUS Next research center.

9. REFERENCES

[1] http://code.google.com/intl/zh-cn/apis/language/translate/overview.html.

[2] http://linkeddata.org/.

[3] http://svmlight.joachims.org/.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC'07*, pages 722–735, 2007.

[5] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg. Adaptive name matching in information integration. *Intelligent Systems, IEEE*, 18(5):16 – 23, 2003.

[6] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154 – 165, 2009.

[7] M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha. Efficient data reconciliation. *Information Sciences*, 137(1-4):1 – 15, 2001.

[8] W. W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of SIGKDD'02*, pages 475–480, 2002.

[9] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, Sept. 1995.

[10] P. Cudre-Mauroux, P. Haghani, M. Jost, K. Aberer, and H. De Meer. idmesh: graph-based disambiguation of linked data. In *Proceedings of WWW '09*, pages 591–600, 2009.

[11] G. de Melo and G. Weikum. Menta: inducing multilingual taxonomies from wikipedia. In *Proceedings of CIKM'10*, pages 1099–1108, 2010.

[12] Z. Dong and Q. Dong. *Hownet And the Computation of Meaning*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2006.

[13] M. Elfeky, V. Verykios, and A. Elmagarmid. Tailor: a record linkage toolbox. In *Proceedings of ICDE'02*, pages 17 –28, 2002.

[14] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19:1–16, 2007.

[15] M. Erdmann, K. Nakayama, T. Hara, and S. Nishio. Improving the extraction of bilingual terminology from wikipedia. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5:31:1–31:17, November 2009.

[16] B. Fu and R. Brennan. Cross-lingual ontology mapping and its use on the multilingual semantic web. In *Proceedings of WWW Workshop on Multilingual Semantic Web*, 2010.

[17] B. Fu, R. Brennan, and D. O'Sullivan. Cross-lingual ontology mapping – an investigation of the impact of machine translation. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *Proceedings of ASWC '09*, volume 5926, pages 1–15, 2009.

[18] S. Hassan and R. Mihalcea. Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of EMNLP '09*, volume 3, pages 1192–1201, 2009.

[19] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *Proceedings of CIKM'11*, 2011.

[20] G. J. Jones, F. Fantino, E. Newman, and Y. Zhang. Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from wikipedia. In *Proceedings of CLIA '08*, 2008.

[21] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *Proceedings of SIGMOD '06*, pages 802–803, 2006.

[22] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[23] D. B. Lenat. Cyc: a large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38:33–38, November 1995.

[24] J. Li, J. Tang, Y. Li, and Q. Luo. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1218–1232, 2009.

[25] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38:39–41, November 1995.

[26] R. Navigli and S. P. Ponzetto. Babelnet: building a very large multilingual semantic network. In *Proceedings of ACL '10*, pages 216–225, 2010.

[27] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.

[28] A. Nikolov, V. S. Uren, E. Motta, and A. N. D. Roeck. Handling instance coreferencing in the knofuss architecture. In *In Proceedings of IRSW'08*, volume 422, 2008.

[29] J.-H. Oh, D. Kawahara, K. Uchimoto, J. Kazama, and K. Torisawa. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *Proceedings of WI-IAT '08*, volume 1, pages 322 –328, 2008.

[30] M. Potthast, B. Stein, and M. Anderka. A wikipedia-based multilingual retrieval model. In *Proceedings of ECIR'08*, pages 522–530, 2008.

[31] L. Sorg and P. Cimiano. Enriching the crosslingual link structure of Wikipedia - A classification-based approach. In *AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[32] J. Tang, J. Li, B. Liang, X. Huang, Y. Li, and K. Wang. Using bayesian decision for ontology mapping. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(4):243–262, 2006.

[33] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of WSDM'12*, pages 743–752, 2012.

[34] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *Proceedings of SIGKDD'09*, pages 807–816, 2009.

[35] W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *Proceedings of ECML/PKDD'11*, pages 381–397, 2011.

[36] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of ISWC '09*, pages 650–665, 2009.

[37] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo. Mining advisor-advisee relationships from research publication networks. In *Proceedings of KDD'10*, pages 203–212, 2010.

[38] W. E. Winkler. Methods for record linkage and bayesian networks. Technical report, Series RRS2002/05, U.S. Bureau of the Census, 2002.

[39] C. S. Wolodja Wentland, Johannes Knopp and M. Hartung. Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In *Proceedings of LREC'08*, 2008.

[40] Z. Ye, X. Huang, and H. Lin. A graph-based approach to mining multilingual word associations from wikipedia. In *Proceedings of SIGIR'09*, pages 690–691, 2009.

[41] X. Zhang, Q. Zhong, F. Shi, J. Li, and J. Tang. Rimom results for oaei 2009. In *Proceedings of ISWC Workshop on Ontology Matching*, 2009.