Cross-lingual infobox alignment in Wikipedia using Entity-Attribute Factor Graph

Yan Zhang, Thomas Paradis, Lei Hou, Juanzi Li, Jing Zhang and Haitao Zheng

Knowledge Engineering Group, Tsinghua University, Beijing, China
z-y14@mails.tsinghua.edu.cn, thomasparadis@126.com,
greener2009@gmail.com, ljz@keg.cs.tsinghua.edu.cn,
jing-zha15@mails.tsinghua.edu.cn, zheng.haitao@sz.tsinghua.edu.cn

Abstract. Wikipedia infoboxes contain information about article entities in the form of attribute-value pairs, and are thus a very rich source of structured knowledge. However, as the different language versions of Wikipedia evolve independently, it is a promising but challenging problem to find correspondences between infobox attributes in different language editions. In this paper, we propose 8 effective features for cross lingual infobox attribute matching containing categories, templates, attribute labels and values. We propose entity-attribute factor graph to consider not only individual features but also the correlations among attribute pairs. Experiments on the two Wikipedia data sets of English-Chinese and English-French show that proposed approach can achieve high F1-measure: 85.5% and 85.4% respectively on the two data sets. Our proposed approach finds 23,923 new infobox attribute mappings between English and Chinese Wikipedia, and 31,576 between English and French based on no more than six thousand existing matched infobox attributes. We conduct an infobox completion experiment on English-Chinese Wikipedia and complement 76,498 (more than 30% of EN-ZH Wikipedia existing cross-lingual links) pairs of corresponding articles with more than one attribute-value pairs.

1 Introduction

With the rapid evolution of the Internet to be a world-wide global information space, sharing knowledge across different languages becomes an important and challenging task. Cross-lingual knowledge sharing not only benefits knowledge internationalization and globalization, but also has a very wide range of applications such as machine translation [20], information retrieval [19] and multilingual semantic data extraction [9, 7]. Wikipedia is one of the largest multi-lingual encyclopedic knowledge bases on the Web and provides more than 25 million articles in 285 different languages. Therefore, many multilingual knowledge bases (KB) have been constructed based on Wikipedia, such as DBpedia [7], YAGO [9], Bablenet [11] and XLore [18]. Some approaches have been proposed to find cross-lingual links between Wiki articles, e.g., [15], [17] and [16].

There is a large amount of semantic information contained in Wikipedia *infoboxes*, which provide semi-structured, factual information in the form of

attribute-value pairs. Attributes in infoboxes contain valuable semantic information, which play a key role in the construction of a coherent large-scale knowledge bases [9]. However, each language version maintains its own set of infoboxes with their own set of attributes, as well as sometimes providing different values for corresponding attributes. Thus, attributes in different Wikipedia must be matched if we want to get coherent knowledge and develop some applications. For instance, inconsistencies among the data provided by different editions for corresponding attributes could be detected automatically. Furthermore, English Wikipedia is obviously larger and of higher quality than low resource languages, which is why we can use attribute alignments to expand and complete infoboxes in other languages, or at least help Wikipedia communities to do so. In addition, the number of existing attribute mappings is limited, e.g., there are more than 100 thousand attributes in English Wikipedia but only about 5 thousand (less than 5%) existing attribute mappings between English and Chinese.

Being aware of the importance of this problem, several approaches have been proposed to find new cross-lingual attribute mappings between Wikis. Bouma et al. [2] found alignments between English and Dutch infobox attributes based on values. Rinser et al. [13] proposed an instance-based schema matching approach to find corresponding attributes between different language infobox templates. Adar et al. [1] defined 26 features, such as equality, word, translation and n-gram features, then applied logistic regression to train a boolean classifier to detect whether two values are likely to be equivalent. These methods can be split into two categories: similarity-based and learning-based. Both of them mostly use the information of the attributes themselves and ignore the correlations among attributes within one knowledge base.

Based on our observation, there are several challenges involved in finding multilingual correspondences across infobox attributes. Firstly, there are Polysemy-Attributes (a given attribute can have different semantics, e.g., *country* can mean nationality of one person or place of a production) and Synonym-Attributes (different attributes can have the same meaning, e.g., *alias* and *other names*), which leads to worse performance on label similarity or translation based methods. Secondly, there also exist some problems in the values of attributes: 1. different measurement (e.g., *population* of *Beijing* is 21,700,000 in English edition and 2170 *ten thousand* in Chinese). 2. timeliness (e.g., *population* of *Beijing* is 21,150,000 (in 2013) in French edition). In this way, labels and values alone are not credible enough for cross-lingual attribute matching.

In order to solve above problems, we first investigate several effective features considering characteristics of cross-lingual attribute matching problem, and then propose an approach based on factor graph model [6]. The most significant advantage of this model is that it can formalize correlations between attributes explicitly, which is specified in Section 3. Specifically, our contributions include:

- We formulate the problem of attribute matching (attribute alignment) across Wikipedia knowledge bases in different language editions, and analyse several effective features based on categories, templates, labels and values.

– We present a supervised method based on an integrated factor graph model, which leverages information from a variety of sources and utilizes the correlations among attribute pairs.
– We conduct experiments to evaluate our approach on existing attribute mappings in the latest Wikipedia. It achieves a high F1-measure 85.5% between English and Chinese and 85.4% between English and French. Furthermore, we run our model on English, Chinese and French Wikipedia, and successfully identify 23,923 new cross-lingual attribute mappings between English and Chinese, 31,576 between English and French.

The rest of this paper is organized as follows, Section 2 defines the problem of attribute matching and some related concepts; Section 3 describes the proposed approach in detail; Section 4 presents the evaluation results; Section 5 discusses some related work and finally Section 6 concludes this work.

2 Problem Formulation

In this section, we formally define the problem of Wikipedia attribute (property) matching. We define the Wiki knowledge base and elements in it as follows.

Definition 1. ***Wiki Knowledge Base****. We consider each language edition of Wikipedia as a* ***Wiki Knowledge Base*** *, which can be represented as*

$$K = \{A, P\}$$

where $A = \{a_i\}_{i=1}^{n}$ *is the set of articles in* K *and* n *is the size of* A*, i.e., the number of articles.* $P = \{p_i\}_{i=1}^{r}$ *is the set of attributes in* K *and* r *is the size of* P*.*

Definition 2. ***Wiki Article****. A* ***Wiki Article*** *can be formally defined as follows,*

$$a = (Ti(a), Te(a), Ib(a), C(a))$$

where

– $Ti(a)$ *denotes the title of the article* a*.*
– $Te(a)$ *denotes the unstructured text description of the article* a*, in other words, the free-text contents of the article* a*.*
– $Ib(a)$ *is the infobox associated with* a*; specifically,* $Ib(a) = \{p_i, val_i\}_{i=1}^{k}$ *represents the list of attribute-value pairs in this article's infobox,* $P(a) = \{p_i\}_{i=1}^{k}$ *represents the set of attributes which appear in* Ib *of* a*.*
– $C(a)$ *denotes the set of categories of the article* a*.*

Definition 3. ***Attribute****. According to the above definitions, an attribute can be defined as a 5-tuple,*

$$attr = (L(p), SO(p), AU(p), C(p), T(p))$$

where

- $L(p)$ *denotes the label of attribute* p.
- $SO(p) = \{(a, val) \mid \forall a \in A, \exists (p, val) \in Ib(a)\}$ *denotes a set which contains the subject-object pairs of the attribute. For example, in* Fig. 1, *attribute* Alma mater *has a pair* (Mark Zuckerberg, Harvard University) *in* $SO(p_{Alma\ mater})$.
- $AU(p) = \{a \mid \forall a, \exists (a, val) \in SO(p)\}$ *denotes the set of articles which use attribute* p.
- $C(p) = \bigcup_{(p,o) \in Ib(a)} C(a)$ *denotes a set of categories in which the attribute appears. For example,* C *of attribute* Born *contains a category* People.
- $T(p) = \{p_i\}_{i=1}^{m}$ *denotes the Infobox template to which the attribute* p *belongs.*

Definition 4. ***Attribute Matching (Property Matching)****. Given two Wiki Knowledge Bases* $K_1 = \{A_1, P_1\}$ *and* $K_2 = \{A_2, P_2\}$*, attribute matching is a process of finding, for each attribute* $p_i \in P_1$*, one or more equivalent attributes in knowledge base* K_2*. When the two Wiki knowledge bases are in different languages, we call it the cross-lingual attribute matching (infobox alignment) problem. Generally, EL, EC and AL denote the existing cross-lingual links between articles, categories and attributes respectively between different language versions of Wikipedia.*

Here, we say two attributes are equivalent if they *semantically* describe the same type of information about an entity. Fig. 1 shows an example of attribute matching results concerning infoboxes of *Zuckerberg* (CEO of Facebook) in English, Chinese and French Wikipedia.

Fig. 1. An example of attribute matching

As shown in Fig. 1, *Born*, 出生 and *Naissance* are equivalent infobox attributes, which can be easily found according to the values using a translation

tool. However, for attribute *Net worth* and its Chinese corresponding attribute 净资产, they have different values because of timeliness, so we cannot find the alignment using value-based method. Furthermore, English Infobox (the left) has an attribute *Relatives*, which does not exist in other two versions. So we can complete the Chinese infobox of Zuckerberg if we find that 亲人 is the corresponding attribute of *Relatives* in Chinese.

3 The Proposed Approach

In this section, we first describe the motivation and overview of our approach, and then we introduce our proposed model in detail.

3.1 Overview

For the problem of Wikipedia attribute matching, existing works [2, 1, 13] mostly used label- and value-based features. Effectiveness of these direct features has been proved. However, as for cross-lingual attribute matching, text similarity cannot be computed directly and machine translation may induce more errors. In this way, only text feature is not enough. There are some works [15, 17] on a similar problem, Wikipedia cross-lingual entity matching, and in these works some useful language-independent features are proposed, such as text hyperlinks. Furthermore, these works also provide large amounts of cross-lingual article links which are very valuable. Inspired by these works, we try to design a model leveraging text, article, category and template features simultaneously. Thus, there are two questions in front of us.

- How to use existing cross-lingual links as seeds to help us find more attribute mappings?
- How to use other information (e.g. article, category and external text) to deal with the lack of information in attribute itself?

3.2 Entity-Attribute Factor Graph Model

Factor graph model [6] has such an assumption that observation data depends on not only local features, but also on relationships with other instances. The characteristic of this model is fit for our problem intuitively, because:

- A pair of attributes is more likely equivalent if they co-occur with aligned attributes in a pair of equivalent articles.
- Template pairs which contain more equivalent attribute pairs tend to be more semantically similar, and other attribute pairs in such templates are more likely equivalent than the ones in other templates.
- Attribute pairs tend to be equivalent if their synonymous pairs are equivalent.

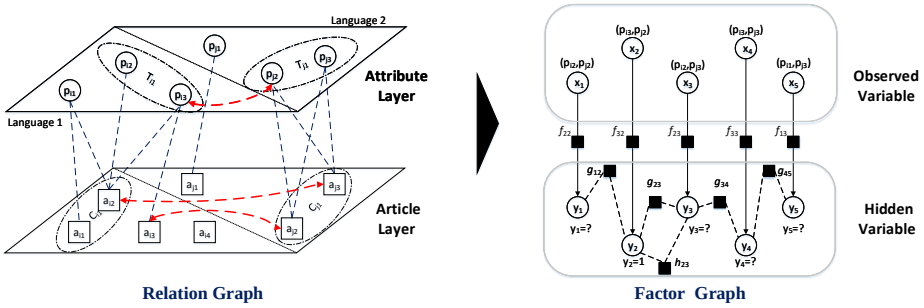

Fig. 2. An illustration of Entity-Attribute Factor Graph (EAFG) model

In this paper, using definitions in Section 2, we formalize the attribute matching problem into a model named Entity-Attribute Factor Graph (EAFG), which is shown in Fig. 2.

Fig. 2 contains two parts, the left one is *relation graph*, which represents several relations in two editions of Wikipedia K_1 and K_2. Different language versions are separated by a diagonal line. The *attribute layer* contains the attributes and template relations among them. Similarly, the *article layer* contains the articles and category relations. The imaginary lines between the two layers denote the relation of usage between articles and attributes, and the red dashed lines denote the existing cross-lingual links. The right one is *factor graph*, the white nodes are *variables*, there are two types of variables, x_i and y_i. Each candidate pair is mapped to an observed variable x_i. The hidden variable y_i represents a Boolean label (equivalent or inequivalent) of the observed variable x_i. For example, x_2 in Fig. 2 corresponds to a candidate attribute pairs (p_{i3}, p_{j2}), and there exists a cross-lingual link between p_{i3} and p_{j2}, so the hidden variable y_2 equals to 1. The black nodes in *factor graph* are *factors*, there are three types, f, g, and h. Each type is associated with a kind of feature function which transforms relations into a computable feature.

Now, we define these feature functions in EAFG model in detail:

- **Local 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 and similarity on observed variables in EAFG;
- **Template feature function**: $g(y_i, CO(y_i))$ denotes the correlation between hidden variables according to template information. $CO(y_i)$ is the set of variables having template co-occurrence relation with y_i.
- **Synonym feature function**:$h(y_i, SY(y_i))$ denotes the correlation between hidden variables according to synonymous information. $SY(y_i)$ is the set of variables being semantically equivalent.

According to these feature functions, we can define joint distribution over the Y on our graph model as

$$p(Y) = \prod_i f(y_i, x_i) g(y_i, CO(y_i)) h(y_i, SY(y_i)) \tag{1}$$

Then we introduce the definition of three feature functions in detail.

1. **Local feature function**

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

where $\mathbf{f}(y_i, x_i) =< f_{label}, f_{we}, f_{so}, f_{au}, f_{cate} >$ is a vector of features; α denotes the corresponding weights of these features; x_i is a variable corresponding to attribute pair (p_a, p_b). Then we describe these five features in detail.

(a) Label similarity feature: it computes the Levenshtein distance [3] after translating non-English attribute labels to English ones, and then get the similarity according to it.

$$f_{label} = 1 - \frac{Leven(L(p_a), L(p_b))}{\max(len(L(p_a)), len(L(p_b)))} \tag{3}$$

where $Leven(L(p_a), L(p_a))$ denotes the Levenshtein distance between two labels, and $len(L(p))$ denotes the length of the label of the attribute p.

Word embedding [10] represents each word as a vector and is able to grasp semantic information. We trained 100-dimension word embeddings on English Wikipedia text and represent each label as a vector (non-English labels are replaced by their translation result). Let $WE(p)$ be the word embedding (a 100-dimension vector) of the label of attribute p, we have

$$f_{we} = 1 - \frac{\arccos(\frac{WE(p_a) \cdot WE(p_b)}{\|WE(p_a)\|_2 \times \|WE(p_b)\|_2})}{\pi} \tag{4}$$

where $\|WE(p_a)\|_2$ denotes the Euclidean norm of the vector $WE(p_a)$, and f_{we} is the cosine similarity between word embeddings of p_a and p_b.

(b) Subject-object similarity feature: according to **Definition 3**, we can get a set SO for each attribute and compute the similarity between the two sets. First, we define an equivalence relation between subject-object pairs as

$$(s_i, o_i) \equiv (s_j, o_j) \iff (s_i, s_j) \in EL \wedge o_i \equiv o_j$$

it denotes pair (s_i, o_i) in SO_i is equivalent with (s_j, o_j) in SO_j if and only if there is a cross-lingual link between subjects, and objects are equivalent. The condition of objects being equivalent depends on the data type. For example, for type *Integer*, the objects should be equal,

and for type *entity*, they should also have a cross-lingual link. f_{so} is defined as

$$f_{so} = \frac{2 \times |\{(s_i, o_i) \equiv (s_j, o_j) \mid (s_i, o_i) \in SO(p_a), (s_j, o_j) \in SO(p_b)\}|}{|SO(p_a)| + |SO(p_b)|} \tag{5}$$

(c) Article-usage feature: according to **Definition 3** and **4**, we can define f_{au} as

$$f_{au} = \frac{2 \times |\{(a, b) \mid (a, b) \in EL, a \in AU(p_a), b \in AU(p_b)\}|}{|AU(p_a)| + |AU(p_b)|} \tag{6}$$

this feature represents the similarity between two article sets which contain the two attributes in their infoboxes respectively.

(d) Category similarity feature: similarly, we can define f_{cate} as

$$f_{cate} = \frac{2 \times |\{(c, c') \mid (c, c') \in EC, c \in C(p_a), c' \in C(p_b)\}|}{|C(p_a)| + |C(p_b)|} \tag{7}$$

where $C(p)$ is defined in **Definition 3** and EC is defined in **Definition 4**. This feature represents the similarity between two category sets related to the two attributes.

2. **Template feature function**

$$g(y_i, CO(y_i)) = \frac{1}{Z_\beta} \exp\{\sum_{y_j \in CO(y_i)} \beta^{\mathrm{T}} \mathbf{g}(y_i, y_j)\} \tag{8}$$

where β denotes the weight remaining to learn, and $\mathbf{g}(y_i, y_j)$ denotes a function to specify whether there exists a template sharing correlation between attribute pairs. Let (p_{a_i}, p_{b_i}) and (p_{a_j}, p_{b_j}) be the attribute pairs related with node y_i and y_j respectively in the factor graph. $\mathbf{g}(y_i, y_j) = 1$ if p_{a_i} and p_{a_j} appear in one common template, and so are p_{b_i} and p_{b_j}, otherwise 0. It should be noticed that this function is used to capture the relations between candidate attribute mappings.

3. **Synonym feature function**

$$h(y_i, SY(y_i)) = \frac{1}{Z_\gamma} \exp\{\sum_{y_j \in S(y_i)} \gamma^{\mathrm{T}} \mathbf{h}(y_i, y_j)\} \tag{9}$$

where γ denotes the weight remaining to learn, and $\mathbf{h}(y_i, y_j)$ denotes the probability of semantically equivalence between y_i and y_j. First we define semantic relatedness between two attributes as,

$$SR(p_i, p_j) = \frac{2 \times |\{(c_i, c_j) \mid c_i \equiv c_j, c_i \in C(p_i), c_j \in C(p_j)\}|}{|C(p_i)| + |C(p_j)|} \tag{10}$$

which is similar with Equation 7, except that p_i and p_j here are from the same language, thus the equivalence between category pairs can be derived directly.

Then let (p_{a_i}, p_{b_i}) and (p_{a_j}, p_{b_j}) be the attribute pairs related with node y_i and y_j respectively, we have

$$\mathbf{h}(y_i, y_j) = SR(p_{a_i}, p_{a_j}) \times SR(p_{b_i}, p_{b_j}) \tag{11}$$

Therefore, the purpose of this feature function is to find more cross-lingual attribute mappings using information of synonym within one language edition of data set.

3.3 Learning and Inference

Given a set of labeled nodes (known attribute mappings) in the EAFG, learning the model is to estimate an optimum parameter configuration $\theta = (\alpha, \beta, \gamma)$ to maxmize the log-likelihood function of $p(Y)$. Based on Equations 1-11, the joint distribution $p(Y)$ can be denoted as

$$p(Y) = \frac{1}{Z} \prod_i \exp\{\theta^{\mathrm{T}}(\mathbf{f}(y_i, y_j), \sum_{y_j} \mathbf{g}(y_i, y_j), \sum_{y_j} \mathbf{h}(y_i, y_j))\} \tag{12}$$

We use log-likelihood function $\log(p(Y^L))$ as the object function, where Y^L denotes the known labels. Then we apply a gradient descent method to estimate the parameter θ. After learning the optimal parameter θ, we can infer the unknown labels by finding a set of labels which maximizes the joint probability $p(Y)$.

4 Experiments

In this paper, the proposed approach is a general model (translation based features are optional), so we use the data from three language editions of Wikipedia (English, Chinese and French) to evaluate our proposed approach. First we evaluate EAFG model on existing cross-lingual attribute mappings, and then we use our approach to find English-Chinese and English-French mappings within Wikipedia.

4.1 Data set

We construct two data sets (English-Chinese and English-French) from existing cross-lingual attribute links in Wikipedia. Table 1 shows the size of the 2 data sets. In each data set, we randomly select 2,000 corresponding attribute pairs which are labeled as positive instances. For each positive instance, we generate 5 negative instances by randomly replacing one of the attribute in the pairs with a wrong one.

Table 1. Size of the 2 data sets

Data set	#Attribute Pairs	#Related Articles	#Related Categories
EN-ZH	2000	EN:96,331 ZH:54,195	EN:13,763 ZH:9,132
EN-FR	2000	EN:103,915 FR:89,012	EN:15,698 FR:12,371

4.2 Comparison Methods

We conduct four existing cross-lingual attribute matching methods. They are translation based method Label Matching (LM), Similarity Aggregation (SA) based method, classification based method Support Vector Matching (SVM) and another logistic regression based method (LR-ADAR) on the work of Adar [1]. As for our proposed approach, in order to evaluate the influence of translation tool, we conduct EAFG-NT (No Translation) which is same as EAFG except that it does not use translation-based features.

- **Label Matching (LM)**. This method first uses Google Translation API to translate the labels of attributes in other languages into English, and then matches them. For each attribute pair, they are considered as equivalent attributes if they have strictly the same English labels.
- **Similarity Aggregation (SA)**. This method aggregates several similarities of each attribute pair into a combined one averagely. Here, we compute 5 similarities same as **local feature function** in Section 3, namely label similarity, subject-object similarity, article-usage similarity, category similarity and word embedding similarity.

$$Sim(p_i, p_j) = \frac{1}{5}(f_{label} + f_{so} + f_{au} + f_{cate} + f_{we})$$

  Then it selects pairs whose similarity is over a threshold ϕ as equivalent pairs. In our experiment, we test the parameter ϕ from 0.05 to 1.00 increasing by 0.05, and this method achieves the best F1-measure when $\phi = 0.75$ on EN-ZH data set, $\phi = 0.80$ on EN-FR data set.
- **Support Vector Machine (SVM)**. This method first computes the five similarities in method SA, and then trains a SVM model [4]. Here, we use Scikit-Learn package [12] in our experiment with a linear kernel and parameter $C = 1.0$. Finally we predict the equivalence of new attribute pairs using this model. Compared with our approach, this method only uses similarities of attributes as features, and it does not take correlations among these instances into consideration.
- **Logistic Regression (LR-ADAR)** In [1], the author defined 26 features and trained a logistic regression model to solve this problem. They obtained good results in their experiments, so we implement this method as a comparison. Here we also use Scikit-Learn package to train a logistic regression model with 17 of their features (removing some language features because they are not suitable for Chinese). In our experiment, it achieves the best result when we use parameter $C = 10$ and L1-regularization.

4.3 Performance Metrics

We use *Precision*, *Recall* and *F1-measure* to evaluate different attribute matching methods. They are defined as usual: *Precision* is the percentage of correct discovered matched in all discovered matches; *Recall* is the percentage of correct discovered matches in all correct matches; *F1-Measure* is the harmonic mean of precision and recall. The data sets we use are described in Section 4.1.

4.4 Settings

For SVM, LR-ADAR and EAFG, we conduct 10-fold cross validation on the evaluation data set. EAFG uses 0.001 learning rate and runs 1000 iterations in all the experiments, and SVM and LR-ADAR runs with settings described in the above. As mentioned before, translation tool is optional in our approach, so we also implement method EAFG-NT for comparison. All experiments are carried out on a Ubuntu 14.04 server with 2.8GHz CPU (8 cores) and 128 GB memory.

Table 2. Perfomance of 5 methods on English-Chinese and English-French data sets.

Method	English-Chinese			English-French		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
LM	**0.973**	0.261	0.412	**0.982**	0.271	0.425
SA	0.749	0.673	0.709	0.764	0.662	0.709
SVM	0.875	0.752	0.809	0.883	0.755	0.814
LR-ADAR	0.907	0.746	0.819	0.917	0.739	0.818
EAFG(NT)	0.863	0.771	0.814	0.877	0.774	0.822
EAFG	0.911	**0.805**	**0.855**	0.913	**0.802**	**0.854**

4.5 Results Analysis

Table 2 shows the performance of these 5 methods on English-Chinese (EN-ZH) and English-French (EN-FR) data sets. For EN-ZH data set, according to the results, the LM method gets a high precision of 97.3%, but its recall is only 26.1%. Apparently, the variety of translation results and too strict matching condition are the main reasons of the result. By using similarities on various information, SA improves recall significantly in comparison to LM, but it does not achieve good precision because averaging strategy is too simple. SVM and LR-ADAR are both learning-based methods. SVM method gets a precision of 87.5% with a recall 75.2%. Compared with SVM, LR-ADAR gets better precision but lower recall, and outperforms SVM by 1.0% in terms of F1-measure. Our method EAFG uses the same training data with SVM, and outperforms SVM by 4.6% in terms of F1-measure. EAFG get similar precision with LR-ADAR, but EAFG is able to discover more attribute mappings by considering the correlation between attribute pairs. EAFG-NT only uses language-independent features, although

it does not work as well as EAFG, it still out performs SVM by 0.5%, which indicates that correlations among attributes are helpful for the problem indeed. As for EN-FR data set, most of these methods get better precision than EN-ZH, and we think it is because English and French are both European languages. Correspondingly, we can get similar conclusions from the experiment on EN-FR data set.

Table 3. Examples of discovered attribute mappings

Classes	English	Chinese	French
Person	Alma mater	母校	Formation
	Spouse(s)	配偶	Conjoint
	Title	头衔	Activité principale
	Nationality	国籍	Nationalité
Location	Party Secretary	书记	Secrétaire du PCC
	Completed	竣工年份	Fin des travaux
	Population	人口	Population
	Mayor	市长	Maire
Film	Directed by	导演	Réalisation
	Screenplay by	编剧	Scénario
	Running time	片长	Durée
	Country	产地	Pays d'origine

4.6 Discovering New Cross-lingual Attribute Mappings in Wikipedia

The motivation of this work is to find more attribute mappings among different language versions of Wikipedia. Therefore, we applied our proposed EAFG to align attributes in English, Chinese and French Wikipedia. First, we extract 107,302, 56,140 and 85,841 attributes from English, Chinese and French Wikipedia respectively. The existing attribute mappings are used for training, and the learned model is employed to predict the correspondence between cross-lingual attribute pairs. Both training and prediction are completed on a server with a 2.8GHz CPU (32 cores) and 384 GB memory, and it costs 13 hours and 21 hours for EN-ZH and EN-FR data set respectively. Finally we get 23,923 new attribute mappings between English and Chinese Wikipedia, and 31,576 mappings between English and French. Table 3 presents a few examples of the discovered mappings.

4.7 Wikipedia Infobox Completion

Apparently, we can transfer infobox information that is missing in one language from other languages in which the information is already present, if we have the alignment of attributes. In this paper, we try to complement Chinese and English

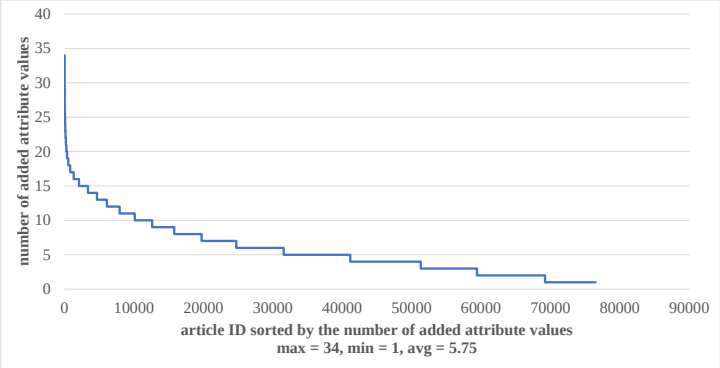

Fig. 3. Statistics of EN-ZH Infobox Complementing

Wikipedia infoboxes from each other using the attribute alignments obtained above EAFG. Firstly, we extract 223,159 existing corresponding English-Chinese article pairs, and finally 76,498 article pairs are replenished by at least 1 attribute value. Fig. 3 shows the number of added attribute values with respect to each article. The maximum number of added attribute values for one article is 34 and the average is 5.75, which indicates that infoboxes in Chinese and English both benefit a lot from the attribute alignments.

We also count the times of each attribute being added into Chinese infoboxes, and list the top 20 attributes in Fig. 4. It should be noticed that most of the attributes are from these categories: *Person* (e.g. Born and Nationality), *Location* (e.g. time-zone and Original language) and *Film* (e.g. Director and Producer). The reason is that entities of these categories tend to have strong local features, and thus lead to imbalance of information among different language versions of Wikipedia. For example, a recent TV play *The Journey of Flower*[1] (花千骨[2] in Chinese) is very popular in China and its Chinese Wikipedia page contains elaborate information. In this experiment, we add 7 attribute values (such as (*editor*, Tianen Su), (*original channel*, Hunan Satellite)) from Chinese to English Wikipedia with respect to this entity (i.e., *The Journey of Flower*).

5 Related Work

In this section, we review some related work.

5.1 Wikipedia Infobox Alignment

Though there have been some works on Wikipedia cross-lingual infobox alignment (attribute matching) and its applications in the real world. Adar [1] used

[1] https://en.wikipedia.org/wiki/The_Journey_of_Flower
[2] https://zh.wikipedia.org/wiki/%E8%8A%B1%E5%8D%83%E9%AA%A8

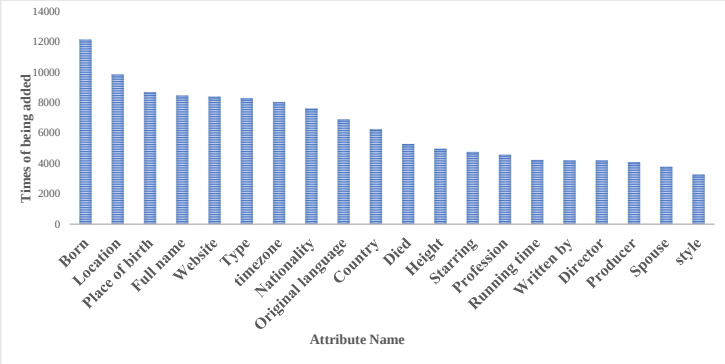

Fig. 4. Statistics of EN-ZH Infobox Complementing

a supervised classifier to identify cross-language infobox alignments. They use 26 features, including equality and n-gram to train the classifier. Through a 10-fold cross-validation experiment on English, German, French and Spanish, they report having achieved 90.7% accuracy. Bouma [2] proposed a value-based method for matching infobox attributes. They first normalized all infobox attribute values, such as numbers, data formats and some units, and then matched the attributes according to the equality between English and Dutch Wikipedia. Rinser [13] proposed an instance-based attribute matching approach. They first matched entities in different language editions of Wikipedia, then they compared the values in attribute pairs and got final results using the entity mappings. However, these works did not consider the correlations among candidate attribute pairs, which is proved to be effective for attribute matching in our work.

5.2 Ontology Schema Matching

Ontology schema matching [14] is another related problem which mainly aims to get alignments of concepts and properties. Currently, some works focus on monolingual matching tasks, such as SOCOM [5] and RiMOM [8, 21]. These Systems deal with the cross-lingual ontology matching problem mainly using machine translation tools to bridge the gap between languages. In our approach, translation-based features are optional.

6 Conclusions and Future Work

In this paper, we propose a cross-lingual attribute matching approach. Our approach integrates several feature functions in a factor graph model (EAFG), including labels, templates, categories and attribute correlations to predict new cross-lingual attribute mappings. Evaluations on existing mappings show that our approach can achieve high F1-measure with 85.5% and 85.4% on English-Chinese and English-French Wikipedia respectively. Using our approach, we have

found 23,923 new attribute mappings between English and Chinese Wikipedia and 31,576 between English and French. It is obvious that article and attribute mappings can benefit each other. Therefore, in the future, we are going to design a framework which can simultaneously and iteratively align all of the elements in Wikipedia.

Acknowledgments

The work is supported by 973 Program (No. 2014CB340504), NSFC key project (No. 61533018, 61661146007), Fund of Online Education Research Center, Ministry of Education (No. 2016ZD102), THUNUS NExT Co-Lab, National Natural Science Foundation of China (Grant No. 61375054), Natural Science Foundation of Guangdong Province (Grant No. 2014A030313745).

References

1. Adar, E., Skinner, M., Weld, D.S.: Information arbitrage across multi-lingual wikipedia. In: International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February. pp. 94–103 (2009)
2. Bouma, G., Duarte, S., Islam, Z.: Cross-lingual alignment and completion of wikipedia templates. In: Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies. pp. 21–29. Association for Computational Linguistics (2009)
3. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks 2003, 73–78 (2003)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
5. Fu, B., Brennan, R., OŚullivan, D.: Cross-lingual ontology mapping–an investigation of the impact of machine translation. In: Asian Semantic Web Conference. pp. 1–15. Springer (2009)
6. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory 47(2), 498–519 (2001)
7. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web 6(2), 167–195 (2015)
8. Li, J., Tang, J., Li, Y., Luo, Q.: Rimom: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering 21(8), 1218–1232 (2009)
9. Mahdisoltani, F., Biega, J., Suchanek, F.: Yago3: A knowledge base from multilingual wikipedias. In: 7th Biennial Conference on Innovative Data Systems Research. CIDR Conference (2014)
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
11. Navigli, R., Ponzetto, S.P.: Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)

12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12(10), 2825–2830 (2013)
13. Rinser, D., Lange, D., Naumann, F.: Cross-lingual entity matching and infobox alignment in wikipedia. Information Systems 38(6), 887–907 (2013)
14. Shvaiko, P., Euzenat, J.: Ontology matching: State of the art and future challenges. IEEE Transactions on Knowledge & Data Engineering 25(1), 158–176 (2013)
15. Sorg, P., Cimiano, P.: Enriching the crosslingual link structure of wikipedia-a classification-based approach. Proceedings of the Aaai Workshop on Wikipedia & Artifical Intelligence (2008)
16. Wang, Z., Li, J., Tang, J.: Boosting cross-lingual knowledge linking via concept annotation. In: International Joint Conference on Artificial Intelligence. pp. 2733–2739 (2013)
17. Wang, Z., Li, J., Wang, Z., Tang, J.: Cross-lingual knowledge linking across wiki knowledge bases. In: International Conference on World Wide Web. pp. 459–468 (2012)
18. Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., Tang, J.: Xlore: A large-scale english-chinese bilingual knowledge graph. In: Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035. pp. 121–124. CEUR-WS. org (2013)
19. Wang, Z., Li, Z., Li, J., Tang, J., Pan, J.Z.: Transfer learning based cross-lingual knowledge extraction for wikipedia. In: ACL (1). pp. 641–650 (2013)
20. Wentland, W., Knopp, J., Silberer, C., Hartung, M.: Building a multilingual lexical resource for named entity disambiguation, translation and transliteration. In: International Conference on Language Resources and Evaluation, Lrec 2008, 26 May - 1 June 2008, Marrakech, Morocco. pp. 3230–3237 (2008)
21. Zhang, Y., Li, J.: Rimom results for oaei 2015. Ontology Matching 185 (2015)