Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes

Wen-Pin Lin, Matthew Snover, Heng Ji
Computer Science Department
Queens College and Graduate Center
City University of New York
New York, NY 11367, USA
danniellin@gmail.com, msnover@qc.cuny.edu, hengji@cs.qc.cuny.edu

Abstract

The automatic generation of entity profiles from unstructured text, such as Knowledge Base Population, if applied in a multi-lingual setting, generates the need to align such profiles from multiple languages in an unsupervised manner. This paper describes an unsupervised and language-independent approach to mine name translation pairs from entity profiles, using Wikipedia Infoboxes as a stand-in for high quality entity profile extraction. Pairs are initially found using expressions that are written in language-independent forms (such as dates and numbers), and new translations are then mined from these pairs. The algorithm then iteratively bootstraps from these translations to learn more pairs and more translations. The algorithm maintains a high precision, over 95%, for the majority of its iterations, with a slightly lower precision of 85.9% and an f-score of 76%. A side effect of the name mining algorithm is the unsupervised creation of a translation lexicon between the two languages, with an accuracy of 64%. We also duplicate three state-of-the-art name translation mining methods and use two existing name translation gazetteers to compare with our approach. Comparisons show our approach can effectively augment the results from each of these alternative methods and resources.

1 Introduction

A shrinking fraction of the world's web pages are written in English, while about 3,000 languages are endangered (Krauss, 2007). Therefore the ability to access information across a range of languages, especially low-density languages, is becoming increasingly important for many applications. In this paper we hypothesize that in order to extend cross-lingual information access to all the language pairs on the earth, or at least to some low-density languages which are lacking fundamental linguistic resources, we can start from the much more scalable task of "information" translation, or more specifically, new name translation.

Wikipedia, as a remarkable and rich online encyclopedia with a wealth of general knowledge about varied concepts, entities, events and facts in the world, may be utilized to address this need. As of March 2011 Wikipedia contains pages from 275 languages[1], but statistical machine translation (MT) techniques can only process a small portion of them (e.g. Google translate can only translate between 59 languages). Wikipedia infoboxes are a highly structured form of data and are composed of a set of subject-attribute-value triples that summarize or highlight the key features of the concept or subject of each article. A large number of instance-centered knowledge-bases that have harvested this structured data are available. The most well-known are probably DBpedia (Auer et al., 2007), Freebase (Bollacker et al., 2007) and YAGO (Suchanek et al., 2007). However, almost all of these existing knowledge bases contain only one language. Even for high-density languages, more than 70% of Wikipedia pages and their infobox entries do not contain cross-lingual links.

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 43–52,
Edinburgh, Scotland, UK, July 27–31, 2011.

Recent research into Knowledge Base Population, the automatic generation of profiles for named entities from unstructured text has raised the possibility of automatic infobox generation in many languages. Cross-lingual links between entities in this setting would require either expensive multilingual human annotation or automatic name pairing. We hypothesize that overlaps in information across languages might allow automatic pairing of profiles, without any preexisting translational capabilities. Wikipedia infoboxes provide a proxy for these high quality cross lingual automatically generated profiles upon which we can explore this hypothesis.

In this paper we propose a simple and general unsupervised approach to discover name translations from knowledge bases in any language pair, using Wikipedia infoboxes as a case study. Although different languages have different writing systems, a vast majority of the world's countries and languages use similar forms for representing information such as time/calendar date, number, website URL and currency (IBM, 2010). In fact most languages commonly follow the ISO 8601 standard[2] so the formats of time/date are the same or very similar. Therefore, we take advantage of this language-independent formatting to design a new and simple bootstrapping based name pair mining approach. We start from language-independent expressions in any two languages, and then extract those infobox entries which share the same slot values. The algorithm iteratively mines more name pairs by utilizing these pairs and comparing other slot values. In this unsupervised manner we don't need to start from any name transliteration module or document-wise temporal distributions as in previous work.

We conduct experiments on English and Chinese as we have bi-lingual annotators available for evaluating results. However, our approach does not require any language-specific knowledge so it's generally applicable to any other language pairs. We also compare our approach to state-of-the-art name translation mining approaches.

1.1 Wikipedia Statistics

A standard Wikipedia entry includes a title, a document describing the entry, and an "infobox" which is a fixed-format table designed to be added to the top right-hand corner of the article to consistently present a summary of some unifying attributes (or "slots") about the entry. For example, in the Wikipedia entry about the singer *"Beyonce Knowles"*, the infobox includes information about her birth date, origin, song genres, occupation, etc. As of November 2010, there were 10,355,225 English Wikipedia entries, and 772,826 entries. Only 27.2% of English Wikipedia entries have cross-lingual hyperlinks referring to their corresponding Chinese entries.

Wikipedia entries are created and updated exponentially (Almeida et al., 2007) because of the increasing number of contributors, many of whom are not multi-lingual speakers. Therefore it is valuable to align the cross-lingual entries by effective name mining.

1.2 Motivating Example

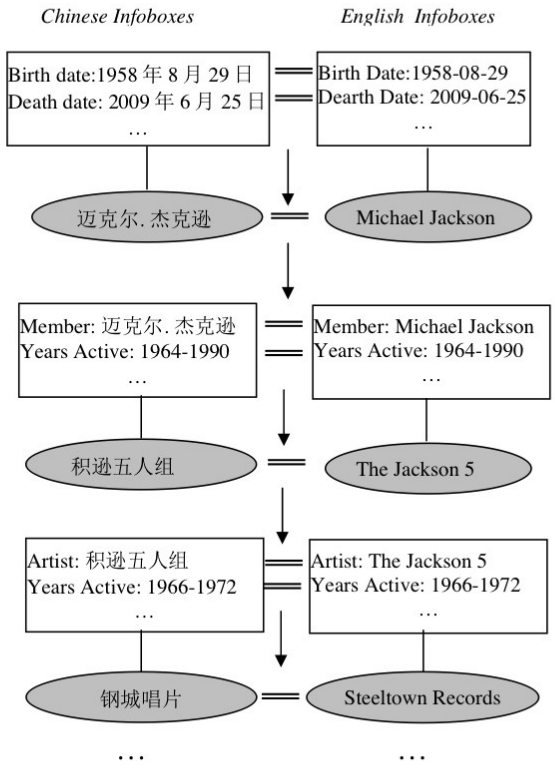

Figure 1: A Motivating Example

Figure 1 depicts a motivating example for our approach. Based on the assumption that if two person entries had the same birth date and death date,

[2] http://en.wikipedia.org/wiki/ISO_8601

they are likely to be the same person, we can find the entity pair of (*Michael Jackson* / 迈克尔.杰克逊). We can get many name pairs using similar language-independent clues. Then starting from these name pairs, we can iteratively get new pairs with a large portion of overlapped slots. For example, since "积逊五人组" and "*The Jackson 5*" share many slot values such as '*member*' and '*years active*', they are likely to be a translation pair. Next we can use the new pair of (*The Jackson 5* / 积逊五人组) to mine more pairs such as "钢城唱片" and *"Steeltown Records."*

2 Data and Pre-Processing

Because not all Wikipedia contributors follow the standard naming conventions and date/number formats for all languages, infoboxes include some noisy instances. Fortunately the NIST TAC Knowledge Base Population (KBP) task (Ji et al., 2010) defined mapping tables which can be directly used to normalize different forms of slot types[3]. For example, we can group '*birthdate*', '*date of birth*', '*datebirth*' and '*born*' to '*birth_date.*' In addition, we also normalized all date slot values into one standard format as "YYYY MM DD." For example, both "1461-8-5" and "5 August, 1461" are normalized as "1461 08 05." Only those Wikipedia entries that have at least one slot corresponding to the Knowledge Base Population task are used for name mining. Entries with multiple infoboxes are also discarded as these are typically "List of ___" entries and do not correspond to a particular named entity. The number of entries in the resulting data set are shown in Table 1. The set of slots were finally augmented to include the entry's name as a new slot. The cross-lingual links between Chinese and English Wikipedia pages were used as the gold standard that the unsupervised algorithm attempted to learn.

Language	Entries	Slot Values	E-Z Pairs
English (E)	634,340	2,783,882	11,109
Chinese (Z)	21,152	110,466	

Table 1: Processed Data Statistics

3 Unsupervised Name Pair Mining

The name pair mining algorithm takes as input a set of English infoboxes E and Chinese infoboxes Z. Each infobox consists of a set of slot-value pairs, where each slot or value may occur multiple times in a single infobox. The output of the algorithm is a set of pairs of English and Chinese infoboxes, matching an infobox in one language to the corresponding infobox in the other language. There is nothing inherently designed in the algorithm for English and Chinese, and this method could be applied to any language pair.

Because the algorithm is unsupervised, it begins with no initial pairs, nor is there any initial translation lexicon between the two languages. As the new pairs are learned, both the entries titles and the values of their infoboxes are used to generate new translations which can be used to learn more cross-lingual name pairs.

3.1 Search Algorithm

The name pair mining algorithm considers all pairs of English and Chinese infoboxes[4], assigns a score, described in Section 3.2, to each pair and then greedily selects the highest scoring pairs, with the following constraints:

1. Each infobox can only be paired to a single infobox in the other language, with the highest scoring infobox being selected. While there are some instances of two entries in one language for one entity which both have translation links to the same page in another language, these are rare occurrences and did not occur for the KBP mapped data used in these experiments.

2. An pair (e, z) can only be added if the score for the pair is at least 95%[5] percent higher than the score for the second best pair for both e and z. This eliminates the problem of ties in the data, and follows the intuition that if there are

[3] It is important to note that the vast majority of Chinese Wikipedia pages store slot types in English in the underlying wiki source, removing the problem of aligning slot types between languages.

[4] The algorithm does not need to compare all pairs of infoboxes as the vast majority will have a score of 0. Only those pairs with some equivalent slot-value pairs need to be scored. The set of non-zero scoring pairs can thus be quickly found by indexing the slot-value pairs.

[5] The value of 95% was arbitrarily chosen; variations in this threshold produce only small changes in performance.

multiple pairs with very similar scores it is beneficial to postpone the decision until more evidence becomes available.

To improve the speed of the algorithm, the top 500 scoring pairs, that do not violate these constraints, are added at each iteration. The translation lexicon is then updated. The translation lexicon is updated each iteration from the total set of pairs learned using the following procedure. For each pair (e, z) in the learned pairs, new translations are added for each of the following conditions:

1. A translation of the name of e to the name z is added.

2. If a slot s in e has one value, v_e, and that slot in z has one value, v_z, a translation $v_e \rightarrow v_z$ is added.

3. If a slot s has multiple values in e and z, but all but one of these values, for both e and z, have translations to values in the other entry, then a translation is learned for the resulting untranslated value.

These new translations are all given equal weight and are added to the translation lexicon even if the evidence for this translation occurs in only a single name pair[6]. These translations can be used to align more name pairs in subsequent iterations by providing more evidence that a given pair should be aligned. After a translation is learned, we consider the English side to be equivalent to the Chinese side when scoring future infobox pairs.

The algorithm halts when there are no longer any new name pairs with non-zero score which also satisfy the search constraints described above.

3.2 Scoring Function

A score can be calculated for the pairing of an English infobox, e and a Chinese infobox, z according to the following formula:

$$\sum_{s \in \text{slots}} \begin{cases} I_Z(s) + I_E(s) & \exists v_1, v_2 : z.s.v_1 \approx e.s.v_2 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A slot-value pair in Chinese, $z.s.v_1$, is considered equivalent to a slot-value pair in English, $e.s.v_2$, if the values are the same (typically only the case with numerical values) or if there is a known translation from v_1 to v_2. These translations are automatically learned during the name-mining process. Initially there are no known translations between the two languages.

The term $I_L(s)$ in equation 1 reflects how informative the slot s is in either English (E) or Chinese (Z), and is calculated as the number of unique values for that slot for that language divided by the total number of slot-value pairs for that language, as shown in equation 2.

$$I_L(\text{slot } s) = \frac{|\{v | i \in L \wedge \exists i.s.v\}|}{|\{i.s.v | i \in L\}|} \tag{2}$$

If a slot s contains unique values such that a slot and value pair is never repeated then $I_L(s)$ is 1.0 and indicates that the slot distinguishes entities very well. Slots such as '*date_of_birth*' are less informative since many individuals share the same birthdate, and slots such as '*origin*' are the least informative since so many people are from the same countries. A sampling of the $I_L(s)$ scores is shown in Table 2. The slots '*origin*' and '*religion*' are the two lowest scoring slots in both languages, while '*infobox_name*' (the name of wikipedia page in question), '*website*', '*founded*' are the highest scoring slot types.

Slot	I_Z	I_E
origin	0.21	0.03
religion	0.24	0.08
parents	0.57	0.60
date_of_birth	0.84	0.33
spouse	0.97	0.86
founded_by	0.97	0.94
website	0.99	0.96
infobox_name	1.00	1.00

Table 2: Sample $I(s)$ Values

4 Evaluation

In this section we present the evaluation results of our approach.

[6] Assigning a probability to each translation learned based upon the number of entries providing evidence for the translation could be used to further refine the predictions of the model, but was not explored in this work.

4.1 Evaluation Method

Human evaluation of mined name pairs can be difficult as a human assessor may frequently need to consult the infoboxes of the entries along with contextual documents to determine if a Chinese entry and an English entry correspond to the same entity. This is especially true when the translations are based on meanings instead of pronunciations. An alternative way of mining name pairs from Wikipedia is to extract titles from a Chinese Wikipedia page and its corresponding linked English page if the link exists (Ji et al., 2009). This method results in a very high precision but can miss pairs if no such link between the pages exists. We utilized these cross-lingual page links as an answer key and then only performed manual evaluation, using a bilingual speaker, on those pairs generated by our algorithm that were not in the answer key.

4.2 Results

Figure 2 shows the precision, recall and f-score of the algorithm as it learns more pairs. The final output of the mining learned 8799 name pairs, of which 7562 were correct according to the cross-lingual Wikipedia links. This results in a precision of 85.94%, a recall of 68.07% and a F1 score of 75.9%. The precision remains above 95% for the first 7,000 name pairs learned. If highly precise answers are desired, at the expense of recall, the algorithm could be halted earlier. The translation lexicon contained 18,941 entries, not including translations learned from the entry names themselves.

Assessment	Number	
Link Missing From Wikipedia	35	2.8%
Same Name, Different Entity	17	1.4%
Partially Correct	98	7.9%
Incorrect	1,087	87.9%

Table 3: Human Assessment of Errors

Because the answer key for name mining is automatically extracted from the cross-lingual links in Wikipedia, it is possible that correct name pairs could be missing from the answer key if no cross-lingual link exists. To examine if any such pairs were learned, a manual assessment of the name pairs that were not in the answer key was performed, as shown in Table 4.2. This assessment was performed by bilingual speakers with an inter-annotator agreement rate of 93.75%.

The vast majority, 87.9%, of the presumably erroneous name pairs assessed that were missing from the answer-key were actually incorrect pairs. However, 35, or 2.8%, of the name pairs were actually correct with their corresponding Wikipedia pages lacking cross-lingual links (these corrections are not reflected in the previous results reported above, which were based solely on the pairs in the answer key). For a small portion, 1.4%, of the errors, the name translation is correct but the entries actually refer to different entities with the same name. One such example is (*Martin Rowlands* / 羅能士). The English entity, *"Martin Rowlands"* is an athlete (an English football player), while the Chinese entity is a former Hong Kong government official, whose name translates to English as *"Martin Rowlands"*, as revealed on his Wikipedia page. Neither entity has an entry in the other language. The final category are partially correct answers, such as the pair (*Harrow, London* / 哈羅區), where the English entry refers to an area within the London Borough of Harrow, while the Chinese entry refers to the London Borough of Harrow as a whole. The English entry *"Harrow, London"* does not have a corresponding entry in Chinese, although there is an entry in both language for the larger Borough itself. All of these cases represent less 15% of the learned name pairs though as 85.94% of the name pairs were already determined to be correct based on cross-lingual Wikipedia links.

Judgement	Percent
Correct	64.4%
Partial	18.4%
Incorrect	15.1%
Not Translations	2.1%

Table 4: Slot Value Translation Assessment from Random Sample of 1000

The name mining algorithm bootstraps many name pairs by using possible translations between the slot values in previously learned pairs. The final translation lexicon learned had 18,941 entries. A random sample of 1,000 entries from the trans-

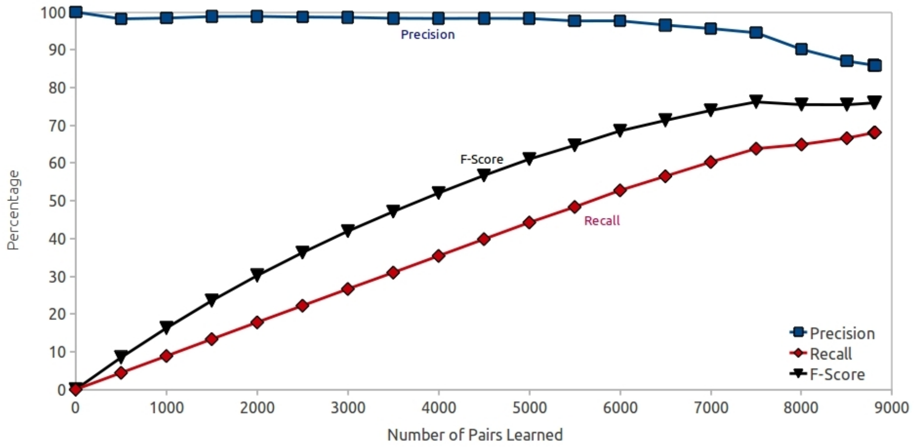

Figure 2: Performance of Unsupervised Name Mining

lation lexicon was assessed by a human annotator, and judged as correct, partial, incorrect or not translations, as shown in Table 4.2. Partial translations were usually cases where a city was written with its country name in language and as just the city name in the other languages, such as *"Taipei Taiwan Republic of China"* and "臺北市" (*Taipei*). Cases are marked as "not translations" if both sides are in the same language, typically English, such as *"Eric Heiden"* in English being considered a translation of *"Eric Arthur Heiden"* from a Chinese entry (not in Chinese characters though). This normally occurs if the Chinese page contained English words that were not translated or transliterated.

An example[7] of the name mining is shown in Figure 3, where the correct name pair for (*George W. Bush* / 乔治· 沃克· 布什) is learned in iteration i, is mined for additional translations and then provides evidence in iteration $i+1$ for the correct name pair (*Laura Bush* / 劳拉· 威尔士· 布什). When learning the name pair for *"George W. Bush"*, evidence is first found from the slots marked as equivalent ($approx$). Translations for *"Harvard Business School"* and *" Republican Party"* were learned in previous iterations from other name pairs and now provide evidence, along with the identical values in the '*date_of_birth*' slot for the pair (*George W. Bush* / 乔治· 沃克· 布什). After learning this pair, new translations are extracted from the pair for *"George W. Bush"*, *"George Walker Bush"*, *"President of the United States"*, *"Laura Bush"*, and *"Yale University"*. The translations for *"Laura Bush"* and *"George W. Bush"* provide crucial information in the next iteration that the pair (*Laura Bush* / 劳拉· 威尔士· 布什) is correct. From this, more translations are learned, although not all of these translations are fully correct, such as *"Author Teacher Librarian First Lady"* which is now postulated to be a translation of 图书管理员 *(Librarian)*, which is only partially true, as the other professions are not represented in the translation. While such translations may not be fully correct, they still could prove useful for learning future name pairs (although this is unlikely in this case since there are very few entries with *"first lady"* as part of their title.

5 Discussion

Besides retaining high accuracy, the final list of name pairs revealed several advantages of our approach.

Most previous name translation methods are limited to names which are phonetically transliterated (e.g. translate Chinese name "尤申科 (*You shen ke*)" to "*Yushchenko*" in English). But many other types of names such as organizations are often rendered semantically, for example, the Chinese name "解放之虎 (*jie fang zhi hu*)" is translated into *"Liberation Tiger"* in English. Some other names in-

[7] Many slot value pairs that were not relevant for the calculation are not shown to save space. Otherwise, this example is as learned in the unsupervised name mining.

Iteration i

George W. Bush			乔治· 沃克· 布什 *(George Walker Bush)*	
alt_names	George Walker Bush		alt_names	乔治· · 布什 *(George Bush)*
title	President of the United States		title	美國總統 *(President of the USA)*
date_of_birth	1946-7-6	≈	date_of_birth	1946-7-6
member_of	Republican Party	≈	member_of	共和黨 *(Republican Party)*
spouse	Laura Bush		spouse	劳拉· 威尔士· 布什 *(Laura Welch Bush)*
schools_attended	Yale University		schools_attended	耶魯大學 *(Yale University)*
schools_attended	Harvard Business School	≈	schools_attended	哈佛商学院 *(Harvard Business School)*

Iteration $i + 1$

Laura Bush			劳拉· 威尔士· 布什 *(Laura Welch Bush)*	
alt_names	Laura Bush	≈	alt_names	劳拉· 威尔士· 布什 *(Laura Welch Bush)*
			alt_names	劳拉· 莲恩· 威尔士 *(Laura Lane Welch)*
date_of_birth	1946-11-4	≈	date_of_birth	1946-11-4
place_of_birth	Midland Texas		place_of_birth	得克萨斯州米德兰 *(Texas Midland)*
title	Author Teacher Librarian First Lady		title	图书管理员 *(Librarian)*
title	First Lady of the United States	≈	title	美國第一夫人*(First Lady of USA)*
spouse	George W. Bush	≈	spouse	乔治· 沃克· 布什 *(George Walker Bush)*

Figure 3: Example of Learned Name Pairs with Gloss Translations in Parentheses

volve both semantic and phonetic translations, or none of them. Our approach is able to discover all these different types, regardless of their translation sources. For example, our approach successfully mined a pair (*Tarrytown* / 柏油村) where *"Tarrytown"* is translated into "柏油村" neither by its pronunciation *"bai you cun"* nor its meaning *"tar village."*

Name abbreviations are very challenging to translate because they need expansions based on contexts. However our approach mined many abbreviations using slot value comparison. For example, the pair of (*Yctc* / 业强科技) was successfully mined although its English full name *"Yeh-Chiang Technology Corp."* did not appear in the infoboxes.

Huang (2005) also pointed out that name translation benefited from origin-specific features. In contrast, our approach is able to discover name pairs from any origins. For example, we discovered the person name pair (*Seishi Yokomizo* / 横溝正史) in which *"Seishi Yokomizo"* was transliterated based on Japanese pronunciation.

Furthermore, many name translations are context dependent. For example, a person name in Chinese "亚西尔•阿拉法特" could be translated into *"Yasser Arafat" (PLO Chairman)* or *"Yasir Arafat" (Cricketer)* based on different contexts. Our method can naturally disambiguate such entities based on slot comparison at the same time as translation mining.

More importantly, our final list includes a large portion of uncommon names, which can be valuable to address the out-of-vocabulary problem in both MT and cross-lingual information processing. Especially we found many of them are not in the name pairs mined from the cross-lingual Wikipedia title links, such as (*Axis Communications* / 安讯士), (*Rowan Atkinson* / 路雲· 雅堅遜), (*ELSA Technology* / 艾爾莎科技) and (*Nelson Ikon Wu* / 吳訥孫).

6 Comparison with Previous Methods and Resources

There have been some previous methods focusing on mining name translations using weakly-supervised learning. In addition there are some existing name translation gazetteers which were manually constructed. We duplicated a variety of alternative state-of-the-art name translation mining methods and mined some corresponding name pair sets for comparison. In fact we were able to implement the techniques in previous approaches but could not duplicate the same number of results because we could not access the same data sets. Therefore the main purpose of this experiment is not to claim our approach outperforms these existing methods, rather to investigate whether we can mine any new information on top of these methods from reasonable amounts of data.

1. **Name Pair Mining from Bitexts**
   Within each sentence pair in a parallel corpus, we ran an HMM based bilingual name tagger (references omitted for anonymous review). If the types of the name tags on both sides are identical, we extract the name pairs from this sentence. Then at the corpus-wide level, we count the frequency for each name pair, and only keep the name pairs that are frequent enough. The corpora used for this approach were all DARPA GALE MT training corpora.

2. **Comparable Corpora**
   We implemented an information extraction driven approach as described in Ji (2009) to extract name pairs from comparable corpora. This approach is based on extracting information graphs from each language and align names by a graph traverse algorithm. The corpora used for this approach were 2000 English documents and 2000 Chinese documents from the Gigaword corpora.

3. **Using patterns for Web mining**
   We constructed heuristic patterns such as parenthetical structure "Chinese name (English name)" (Lin et al., 2008) to extract name pairs from web data with mixed Chinese and English. We used about 1,000 web pages for this experiment.

4. **Bilingual Gazetteer**
   We exploited an LDC bilingual name dictionary (LDC2005T34) and a Japanese-English person name dictionary including 20126 Japanese names written in Chinese characters (Kurohashi et al., 1994).

5. **ACE2007 Entity Translation Training Data**
   We also used ACE 2007 entity translation training corpus which includes 119 Chinese-English document pairs.

Table 5 shows the number of correct and unique pairs mined pairs from each of the above approaches, as well as how these name mining methods can be augmented using the infobox name mining described in this paper. The names mined from our approach greatly extend the total number of correct translations with only a small number of conflicting name translations.

7 Related Work

Most of the previous name translation work combined supervised transliteration approaches with Language Model based re-scoring (Al-Onaizan and Knight, 2002; Huang et al., 2004; Huang, 2005). Our goal of addressing name translation for a large number of languages is similar to the panlingual lexical translation project (Etzioni et al., 2007). Some recent research used comparable corpora to re-score name transliterations (Sproat et al., 2006; Klementiev and Roth, 2006) or mine new word translations (Udupa et al., 2009; Ji, 2009; Fung and Yee, 1998; Rapp, 1999; Shao and Ng, 2004; Hassan et al., 2007). However, most of these approaches needed large amount of seeds and suffered from information extraction errors, and thus relied on phonetic similarity or document similarity to re-score candidate name translation pairs.

Some recent cross-lingual information access work explored attribute mining from Wikipedia pages. For example, Bouma et al. (2009) aligned attributes in Wikipedia infoboxes based on cross-page links. Navigli and Ponzetto (2010) built a multilingual semantic network by integrating the cross-lingual Wikipedia page links and WordNet. Ji et

		# Name	Infobox Mining	
Method		Pairs	# New	# Conflicting
Automatic	(1) Bitexts	2,451	8,673	78
	(2) Comparable Corpora	288	8,780	13
	(3) Patterns for Web Mining	194	8799	0
Manual	(4) Bilingual Gazetteer	59,886	8,689	74
	(5) ACE2007 Training Data	1,541	8,718	52

Table 5: Name Pairs Mined Using Previous Methods

al. (2009) described various approaches to automatically mine name translation pairs from aligned phrases (e.g. cross-lingual Wikipedia title links) or aligned sentences (bi-texts). G et al. (2009) mined candidate words from Wikipedia and validated translations based on parallecl corpora. Some other work mined name translations from monolingual documents that include foreign language texts. For example, Lin et al. (2008) described a parenthesis translation mining method; You et al. (2010) applied graph alignment algorithm to obtain name translation pairs based on co-occurrence statistics. This kind of data does not commonly exist for low-density languages. Sorg and Cimiano (2008) discovered cross-lingual links between English and German using supervised classification based on support vector machines. Adar et al. (2009) aligned cross-lingual infoboxes using a boolean classifier based on self-supervised training with various linguistic features. In contrast, our approach described in this paper is entirely based on unsupervised learning without using any linguistic features. de Melo and Weikum (2010) described an approach to detect imprecise or wrong cross-lingual Wikipedia links based on graph repair operations. Our algorithm can help recover those missing cross-lingual links.

8 Conclusion and Future Work

In this paper we described a simple, cheap and effective self-boosting approach to mine name translation pairs from Wikipedia infoboxes. This method is implemented in a completely unsupervised fashion, without using any manually created seed set, training data, transliteration or pre-knowledge about the language pair. The underlying motivation is that some certain expressions, such as numbers and dates, are written in language-independent forms among a large majority of languages. Therefore our approach can be applied to any language pairs including low-density languages as long as they share a small set of such expressions. Experiments on English-Chinese pair showed that this approach is able to mine thousands of name pairs with more than 85% accuracy. In addition the resulting name pairs can be used to significantly augment the results from existing approaches. The mined name pairs are made publicly available.

In the future we will apply our method to mine other entity types from more language pairs. We will also extend our name discovery method to all infobox pairs, not just those that can be mapped into KBP-like slots. As a bi-product, our method can be used for automatic cross-lingual Wikipedia page linking, as well as unsupervised translation lexicon extraction, although this might require confidence estimates on the translations learned. Once our approach is applied to a panlingual setting (most languages on the Wikipedia), we can also utilize the voting results across multiple languages to automatically validate information or correct potential errors in Wikipedia infoboxes. Finally, as automatic name profile generation systems are generated cross-lingually, our method could be attempted to automatic cross-lingual mappings between entities.

Acknowledgement

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149 and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Eytan Adar, Michael Skinner, and Daniel S. Weld. 2009. Information arbitrage across multi-lingual wikipedia. In *Second ACM International Conference on Web Search and Data Mining (WSDM'09), Barcelona, Spain, February 2009*, February.

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *ACL 2002*.

Rodrigo B. Almeida, Barzan Mosafari, and Junghoo Cho. 2007. On the evolution of wikipedia. In *Int. Conf. on Weblogs and Social Media*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The 6th International Semantic Web Conference*.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *The National Conference on Artificial Intelligence (Volume 2)*.

Gosse Bouma, Sergio Duarte, and Zahurul Islam. 2009. Cross-lingual alignment and complettion of wikipedia templates. In *The Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*.

Gerard de Melo and Gerhard Weikum. 2010. Untangling the cross-lingual link structure of wikipedia. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden*.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel and comparable texts. In *COLING-ACL*.

Rohit Bharadwaj G, Niket Tandon, and Vasudeva Varma. 2009. An iterative approach to extract dictionaries from wikipedia for under-resourced languages. In *Proc. ICON2010*, February.

Ahmed Hassan, Haytham Fahmy, and Hany Hassan. 2007. Improving named entity translation by exploiting comparable and parallel corpora. In *RANLP*.

Fei Huang, Stephan Vogel, and Alex Waibel. 2004. Improving named entity translation combining phonetic and semantic similarities. In *HLT/NAACL2004*.

Fei Huang. 2005. Cluster-specific name transliteration. In *HLT-EMNLP 2005*.

IBM. 2010. Ibm globalization library.

Heng Ji, Ralph Grishman, Dayne Freitag, Matthias Blume, John Wang, Shahram Khadivi, Richard Zens, and Hermann Ney. 2009. Name translation for distillation. Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation.

Heng Ji, Ralph Grishman, Hoa Trang Dang, and Kira Griffitt. 2010. An overview of the tac2010 knowledge base population track. In *Text Analytics Conference (TAC2010)*.

Heng Ji. 2009. Mining name translations from comparable corpora by creating bilingual information networks. In *ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora (BUCC 2009): from Parallel to Non-parallel Corpora*.

Michael E. Krauss. 2007. *Keynote-mass Language Extinction and Documentation: The Race Over Time. The Vanishing Languages of the Pacific Rim.* Oxford University Press.

Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. 1994. Improvements of japanese morphological analyzer juman. In *The International Workshop on Sharable Natural Language Resources and pp.22-28*.

Dekang Lin, Shaojun Zhao, Benjamin Van Durme, and Marius Pasca. 2008. Mining parenthetical translations from the web by word alignment. In *ACL2008*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *ACL 1999*.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *COLING2004*.

Philipp Sorg and Philipp Cimiano. 2008. Enriching the crosslingual link structure of wikipedia - a classification-based approach. In *AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, June.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *The 16th International World Wide Web conference*.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2009. Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora. In *EACL2009*.

Gae-won You, Seung won Hwang, Young-In Song, Long Jiang, and Zaiqing Nie. 2010. Mining name translations from entity graph mapping. In *EMNLP2010*.