# Utilization of DBpedia Mapping in Cross Lingual Wikipedia Infobox Completion

Megawati, Saemi Jang, and Mun Yong Yi(✉)

Department of Industrial and Systems Engineering,
Graduate School of Knowledge Service Engineering,
KAIST, Daejeon, South Korea
{megawati, sammyl22l, munyi}@kaist.ac.kr

**Abstract.** Wikipedia plays a central role in the web as one of the biggest knowledge source due to its large coverage of information that comes from various domains. However, due to the enormous number of pages and limited number of contributors to maintain all of the pages, the problem of missing information among Wikipedia articles has emerged, especially articles in multiple language versions. Several approaches have been studied to fix information gap in between cross- language Wikipedia articles. However, they can only be applied for languages that came from the same root. In this paper, we propose an approach to generate new information for Wikipedia infoboxes written in different languages with different roots by utilizing the existing DBpedia mappings. We combined mapping information from DBpedia with an instance-based method to align the existing Korean-English infobox attribute-value pairs as well as to generate new pairs from the Korean version to fill missing information in the English version. The results showed that we could expand up to 38% of the existing English Wikipedia attribute-value pairs from our datasets with 61% of accuracy.

**Keywords:** Infobox alignment · Infobox completion · DBpedia · Cross language Wikipedia

## 1 Introduction

Wikipedia is an online encyclopedia, which contains a massive amount of information from various domains provided collaboratively by its contributors. The information is easily accessible through the website[1] and is being continually updated by the contributors. Moreover, Wikipedia pages are also available in several languages (in May 2016, there are 282 different active Wikipedia language editions), creating an opportunity for people around the world to make contributions despite the language barrier. Thus, many practices have relied on Wikipedia as a knowledge source, such as Q&A systems, Linked Open Data (LOD), and intelligent agents.

Many Wikipedia pages usually contain an infobox, a small box located at the right side of the page, providing a summary of the page content in a structured manner. Due to its structure, the infoboxes are useful if we want to mine key information from a

[1] https://www.wikipedia.org/.


B.H. Kang and Q. Bai (Eds.): AI 2016, LNAI 9992, pp. 303–316, 2016.
DOI: 10.1007/978-3-319-50127-7_25

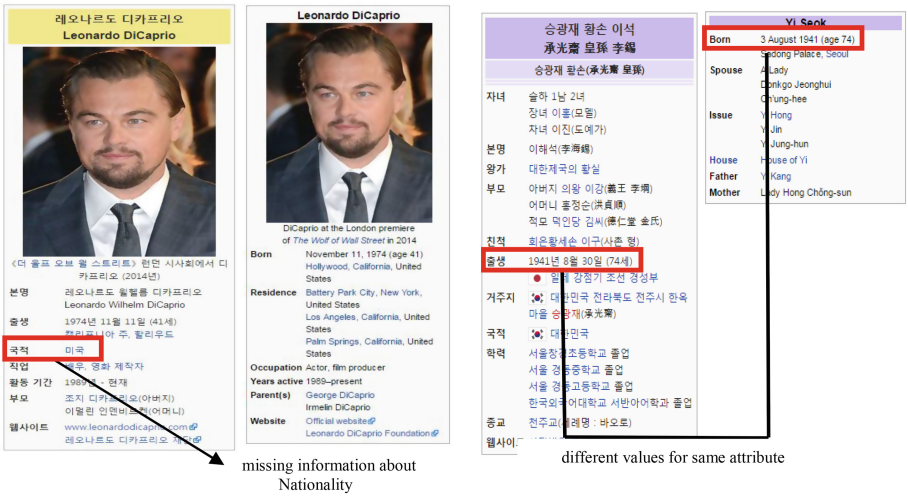


**Fig. 1.** Example of error Type II and Type III

particular page, which can take a lot more effort if we mine from free texts (the article itself). Generally, the infobox consists of three parts: (1) **template** represents type/category of the entity that is being discussed in the page (e.g. template Infobox Person is used in those pages related to a person, such as presidents, actors, soldiers), (2) **attributes** represents characteristics of the template (e.g. a person has a name, birth date, birth place, and parents), and (3) **values** are the instances of the attribute.

Although Wikipedia is considered as a reliable knowledge source, its information is not flawless, given the fact that it is entered by people. Problems such as the use of different names for an entity are common to be found across the pages. Moreover, the number of Wikipedia pages is growing rapidly, making it hard to maintain all the existing pages, including their corresponding pages in different languages. Consequently, information incompleteness and inconsistencies have emerged as serious issues and must be tackled to maintain information quality in Wikipedia. This paper focused on maintaining completeness and consistencies between infoboxes in multilingual Wikipedia pages. In terms of infoboxes, we did an observation with some pairs of random Wikipedia pages in different language and found three types of errors regarding the information in the infoboxes.

1. Type I
   This error is related to the missing infobox in one of the pages. For example, articles about Takeo Takagi has an infobox in English version of Wikipedia while the infobox does not exist in the Korean version.
2. Type II
   This error happens when both pages have infoboxes but one of them have missing attributes that exist in the other version. Figure 1 (left) shows comparison between infoboxes Leonardo DiCaprio from Korean and English version. In the Korean

version there is an attribute that describes his Nationality which should also be presents in the English version.

3. Type II
   Similar with error Type II, error Type III can be found when both pages have infoboxes. This error happens when the same attribute have different values, such as values for attribute 출생 and Born that are being shown in Fig. 1 (right).

As a knowledge source, Wikipedia should maintain the quality of the available information at a high level to ensure its reliability. However, huge efforts will be needed to go through all the existing pages and check for error one by one. Therefore, we conducted a study whose aim is to enhance information quality of Wikipedia infoboxes by correcting the Type I and II errors. We developed a model that is able to automatically generate new infoboxes for pages that do not have any infobox or adding more information to the existing infoboxes with the help from DBpedia mappings. We found the possibility that the DBpedia mappings might be useful as a translation tool, eliminating the need to involve any bilingual dictionary. Instead, the mappings can map an attribute from one language to another since DBpedia facts are also available in multilingual environment. If the two attributes are mapped to the same property of the same entity, then they are very likely matching words.

The purpose of this study is to evaluate the use of DBpedia mappings to align infobox attributes and templates in multilingual environments. First, we evaluated the capability of the existing mappings to translate infobox templates and attributes from Korean to English and we measured how many new pairs could be generated. Then, we tried to search other possible translations that are not covered in the mapping by using instance-based method that were used in [2, 3] to expand the number of generated pairs.

The organization of this paper is as follows. Section 2 will explain about the related studies about various approaches for cross-lingual schema matching and DBpedia enrichment that have already been done by other researchers. In Sect. 3, we will explain in detail about our proposed model and the techniques that we used in aligning infoboxes. Section 4 will describe the detail of our experiment and the results. Section 5 presents the conclusion that we could draw from the experiment and discusses some possibilities for future research.

## 2 Related Work

*Schema Matching.* The infobox alignment problem can be considered as a schema matching problem. Studies related to this area have already been done by many researchers [6]. Several studies have been conducted to develop approaches for aligning multilingual schemas as well as ontologies. Wang et al. [12] tried to identify correspondences between Chinese and English attributes from multilingual schemas by transliterating Chinese characters into alphabets and took the first letter of each syllable to replace the original attribute name. A domain ontology built by human was used to determine the mapping between translated attribute and English attribute. [13, 14] proposed models that can align multilingual ontologies by translating the source

ontology into the target language and match them by using the existing monolingual ontology matching approaches. Unfortunately, the proposed approaches mentioned above are hardly applicable to matching infobox data. Schemas and ontologies have a well-defined structure and metadata while infoboxes are much looser on their data type constraints. Thus, for infobox matching, comparing only metadata and structures might be insufficient to solve the problem. Moreover, the approaches to align multilingual schemas were fully based on translation tools, which have clear limitations when applied to aligning infoboxes, as we discussed in the previous chapter.

*Cross-Language Infobox Alignment.* Studies about aligning Wikipedia infoboxes in different language have been done by several researches. [15] proposed an approach to align Dutch and English Wikipedia templates and attributes by utilizing multilingual nature of Wikipedia as well as cross-language links between pages with precision 65%. Moreover, the approach can be used to generate new attribute-value pairs in Dutch Wikipedia by 50%. [16] developed WikiMatch, a tool that can align two infoboxes in different language without using dictionary or translator. They combined three similarity measures to determine the similarity between two attributes; value similarity, link similarity, and cosine similarity from attribute co-occurrence vectors, which were decomposed by using Latent Semantic Indexing (LSI). [2] used a three stages approach to align multilingual infoboxes from six languages; entity matching, template matching, and attribute matching. In our work, we adapted a similar template matching from the paper to determine the template for the generated infoboxes. We also adapted attribute matching techniques that had been used in [2] to determine potential new attribute mappings. Another approach was used by [4] who exploited machine learning approach to align Wikipedia infoboxes while [5] has developed an information extraction tool, Kylin, to extract information from Wikipedia text and predict possible attribute-value pairs from the sentences by using CRF classifier. However, Kylin was tested by using only English Wikipedia articles and, as for our knowledge, there has not been a study yet that try to measure Kylin's capability of generating new infoboxes from different language.

*Cross-Language DBpedia Enrichment.* DBpedia, which is essentially structured information extracted from Wikipedia, is also dealing with inconsistencies and incompleteness issues due its multilingual nature. These problems are being solved by crowd sourcing effort from the community members. Various approaches have been studied to develop an automatic system that can better align multilingual DBpedia. [17] proposed an approach that can automatically extend the existing alignment of multilingual DBpedia chapters by using mapping frequencies of two properties and then integrate the results to a question answering system over linked data. A study in [18] has been conducted to find semantically corresponding properties from Korean and English DBpedia datasets by using the triple-conceptualization technique. The enrichment can also be done by using the existing Wikipedia data and map them to DBpedia ontology, like what have been done by [3, 20].

## 3 Cross Language Infobox Completion

We developed a model that compares two infoboxes from the Korean Wikipedia and English Wikipedia to find which information should be added from the Source to the Target infobox. Later in this paper, we refer the Korean Wikipedia infoboxes as the Source infoboxes and English Wikipedia infoboxes as the Target infoboxes. We used Korean infoboxes as source because the localized version might still cover more information about the topics related to the local culture though English version has the largest information coverage. Therefore, we could introduced such information to the people outside the culture as well as contribute to expanding the information coverage in the English Wikipedia. The overview of the model is shown in Fig. 2. It basically consists of 4 main parts: Mapping Table, Infobox Alignment, Infobox Generator, and Infobox Populator. Details of each part will be elaborated in the following subsections.

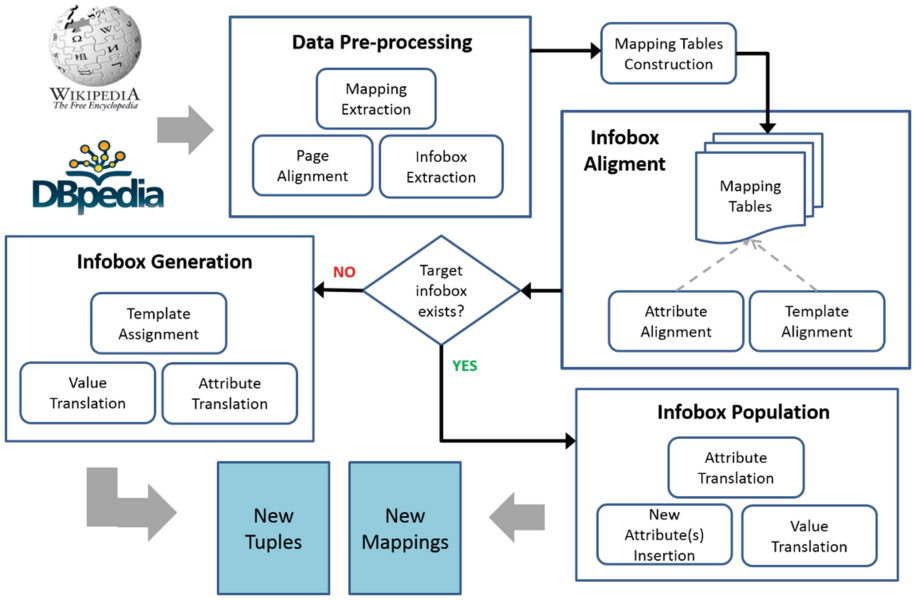


**Fig. 2.** Overview of cross language infobox completion model

### 3.1 Mapping Tables

The mapping tables contain mapping information extracted from DBpedia. DBpedia is a knowledge base built based on structured information from Wikipedia, i.e. infoboxes. Up to this day, DBpedia community members manually map Wikipedia infobox templates to DBpedia ontology classes as well as Wikipedia infobox attributes to DBpedia ontology properties. The results are available on the Web[2] and can also be downloaded as xml files.

[2] http://wiki.dbpedia.org/Downloads2015-10.

**Table 1.** Examples of attribute mapping tables

| Attribute_ko | Attribute_en | DBpedia_property |
|---|---|---|
| 이름 | name | foaf:name |
| 출생지 | birth_place | dbo:birthPlace |
| 사망지 | death_place | dbo:deathPlace |
| 개교 | established | dbo:established |
| 학생수 | students | dbo:numberOfStudents |

For each language, we built two kinds of mapping tables; the template mapping table and the attribute mapping table. From these tables, we could find pairs of attributes/templates that are semantically similar. We tackled one-to-many mappings by only picking one common attribute to be included in the mapping table. However, we kept all attributes that appear in the infobox in another table along with their corresponding common attribute and took them into account in the matching process. Table 1 shows the examples of the mapping tables.

### 3.2 Template Alignment

To generate a new infobox for fixing the error Type I, we need to define the three components of an infobox: template, attributes, and values. The template alignment process defines which template will be used in the new infobox by aligning them with the template used in the existing infobox. There are two cases that might happen while mapping the Source template T to the Target template T′. First, when T is already mapped to a DBpedia ontology class. Second, T does not have mapping to any DBpedia ontology class. For the first case, we can find the template(s) from the English infobox(es) that was (were) also mapped to the same class by looking at the template mapping table. For example, both template 군인 from the Korean Wikipedia and template military person from the English Wikipedia are mapped to class MilitaryPerson. Therefore, we can use the template military person in the creation of new infoboxes. However, if the second case happened, we have to pick a template that is the most suitable for the new infobox. To solve the problem, we looked at the number of template co-occurrence in both infoboxes [2]. The steps are as follows.

1. Let $P_S$ be a set of articles in Source languages and $P_{S'}$ be a set of articles in Target language that are connected to element in $P_S$ through interlanguage links. Let $T_S$ be the Source template and $T_{S'}$ be the Target template that we are trying to define
2. Calculate the total occurrence number of each template that is being used by the members in $P_{S'}$
3. Template with the highest occurrence number will become $T_{S'}$

### 3.3 Attribute Alignment

The purpose of the attribute alignment process was to find pairs of cross language attributes that are semantically similar. Similar to the template alignment, two cases

might happen while mapping Source attribute $a$ to Target attribute $a'$. The first case is when both $a$ and $a'$ are connected via their mapping to the same DBpedia ontology property. The second case happens when either $a$ or $a'$ does not have mapping information to any property in DBpedia ontology so they are not connected to each other. While in the first case we could easily look up to the mapping tables we had already constructed to obtain the mapping information, we need another way to find new mappings from potential attribute pairs that do not have any connection yet. Therefore, we decided to use instance-based method introduced in [2] to find new alignments between such attributes. The steps are as follows.

1. Let S be a set of article pairs $P_l$-$P_{l'}$ where $l$ is the Source language and $l'$ is the Target language and each $P_l$ contains an infobox with template T
2. Let A be the set of attributes from all $P_l$ and A′ be a set of attributes from all $P_{l'}$ where each element in A does not exist in the mapping table. For each attribute pair ($a_l$, $a_{l'}$), we compute $\text{sim}_\text{a}$

$$sim_a(a_l, a_{l'}) = \frac{\sum_{s \in S} sim_{instance}(a_l, a_{l'})}{|S|}. \quad (1)$$

   The algorithm that we used to calculate the similarity between the attributes is further explained in detail in Sect. 3.5
3. All ($a_l$, $a_{l'}$) whose $sim_a < \alpha$ will be discarded
4. For each $a_l$, find ($a_\text{l}$, $a_{l'}$) with the maximum value and add to matching set $\text{M}_\text{a}$
5. Add $\text{M}_\text{a}$ to the mapping table

We decided to use an instance-based method due to the format-loose nature of infobox values. Wikipedia does not provide a specific convention that must be followed to define infobox value (e.g. birth_date must be in YYYYMMDD or DDMMYYYY). Instead, it let authors use their own style. We found it difficult to use other similarity measures to compare two different strings that are semantically similar but are written in different forms. Moreover, infobox values are often composed of several elements other than texts. Therefore we use an instance-based method as a heuristic to measure the similarity of two strings of infobox value.

### 3.4 Infobox Generation and Population

We could use all information about templates and attributes alignments to align and complete the cross language infoboxes. First of all, we had to determine whether an infobox exists in the Target article. If the Target infobox I′ does not exist, it means that we have to create a new infobox in the Target article by translating all information available from the Source infobox into the Target language. The process is called the infobox generation process. Otherwise, we had to compare both infoboxes from a pair of article, which talks about a same topic (later we refer to it as article pair) to determine whether new attribute-value pairs should be added to the Target infobox.

The infobox population process adds new potential attribute-value pairs which are not yet available to the Target infobox. Basically, both processes consists of three steps.

1. Template assignment
   As mentioned before, to create new infobox we need to define all the components. This step finds a mapping of T by looking at the mapping table and assigns it the corresponding template of T as the template of the new infobox. This step is particularly important in the generation process while in the population process we could skip it because the Target infobox usually already has its own template.
2. Attribute translation and insertion
   This step maps a set of attributes A from the Source infobox to its corresponding mapping in the Target language by looking at the attribute mapping table and assigns them as attributes for the new infobox. In the generation process, we inserted all translated attributes to the new infobox while in the population process we omit attributes that already exist in the Target infobox and insert the new ones.
3. Value translation
   For each attribute, we translated its value to the Target language as the new values by using translator API[3]. For the values that contain links, we substitute them with the corresponding link in the Target language by utilizing Wikipedia interlanguage links. For example, 1990년 03월 09일 <미국> will be translated as 09-03-1990 <United States> where the brackets denote a link.

### 3.5 Similarity Measure

To get $sim_{instance}$ of an attribute pair we break down the value into four parts: text, number, date, and links, and then calculate a similarity score for each part. We then aggregate the results to find the final similarity score. We adapted the methods in [2] to calculate the similarity score for each part and aggregate them in Table 2.

To determine the overall instance similarity between a pair of attribute values, [2] took into account the portion of respective components in the original attribute value string. $f_{s1}$ and $f_{s2}$ are the fraction of string values of both attribute values, while $f_n$ represents fraction of number, and $f_d$ is fraction of date. $len_1$ and $len_2$ are the length of the original attribute values. Then, $f_s$, $f_n$, and $f_d$ can be defined as follow.

$$f_s = \frac{len_1 . f_{s1} + len_2 . f_{s2}}{len_1 + len_2} \quad f_n = \frac{len_1 . f_{n1} + len_2 . f_{n2}}{len_1 + len_2}$$

$$f_d = \frac{len_1 . f_{d1} + len_2 . f_{d2}}{len_1 + len_2}$$

After the similarity scores from each value components were obtained, similarities of texts, numbers, and dates were calculated and then weighted together with similarities of links to produce the overall instance similarity value.

[3] https://www.microsoft.com/en-us/translator/translatorapi.aspx.

**Table 2.** Similarity measures for number, date, link, and text extracted from attribute values

| | |
|---|---|
| Number similarity | $sim_{num}(n_1, n_2) = \begin{cases} 1, & if\ n_1 = n_2 \\ 0.5.\dfrac{\min\{\lvert n_1\rvert, \lvert n_2\rvert\}}{\max\{\lvert n_1\rvert, \lvert n_2\rvert\}}, & otherwise \end{cases}$ |
| | $sim_{numset}(N_1, N_2) = \dfrac{\sum_{<n_1, n_2> \in M_n} sim_{num}(n_1, n_2)}{\max\{\lvert N_1\rvert, \lvert N_2\rvert\}}$ |
| Date similarity | $sim_{date}(d_1, d_2) = 1 - \dfrac{\lvert d_1 - d_2\rvert}{maxDate - minDate}$ |
| | $sim_{dateset}(D_1, D_2) = \dfrac{\sum_{<d_1, d_2> \in M_n} sim_{date}(d_1, d_2)}{\max\{\lvert D_1\rvert, \lvert D_2\rvert\}}$ |
| Link similarity | $sim_{wikilinks}(w_{l1}, w_{l2}) = \dfrac{2.\lvert w_{l1} \cap w_{l2}\rvert}{\lvert w_{l1}\rvert + \lvert w_{l2}\rvert}$ |
| | $sim_{exlinks}(e_{l1}, e_{l2}) = \dfrac{2.\lvert e_{l1} \cap e_{l2}\rvert}{\lvert e_{l1}\rvert + \lvert e_{l2}\rvert}$ |
| Text similarity | $sim_{str}(s_1, s_2) = \dfrac{\lvert T_1 \cap T_2\rvert}{\lvert T_1 \cup T_2\rvert}$ |

$$sim_{val}(a_1, a_2) = \frac{w_s \,.f_s \,.\, sim_{str} + w_n \,.f_n \,.\, sim_{numset} + w_s f_s \,.\, sim_{dateset}}{w_s \,.f_s + w_n \,.f_n + w_d \,.f_d} \tag{2}$$

$$sim_{instance}(a_1, a_2) = \frac{w_v.sim_{val}(a1, a2) + w_w.sim_{wikilinks} + w_e.sim_{exlinks}}{w_v + w_w + w_e} \tag{3}$$

According to [2], the weights for each data type portion and links were largely determined empirically. The weight used in the experiment are $w_s = 0.11$, $w_n = 0.44$, $w_d = 0.44$, $w_v = 0.3$, $w_w = 0.6$, and $w_e = 0.1$. Note that the links similarity could be calculated if both set have at least one member. Otherwise, the weight, either $w_w$ or $w_e$ will become 0.

## 4 Experiment

In our experiment, we used Korean Wikipedia and English Wikipedia articles dump[4] and DBpedia mappings[5]. In the data pre-processing step, we extracted infobox data by using infobox2rdf [25][6], which robustly extracts the infoboxes from xml files, cleanses them, and transforms them into RDF triples. To test our model, we picked five different infobox templates from Korean Wikipedia that became the Source infoboxes; 군인 (Military person), 학교 (School), 왕 (Monarch), 회사 (Company), and 대학 (University). We then extracted Korean-English article pairs from each template as our dataset in the experiment by using interlanguage links.

[4] https://dumps.wikimedia.org/.

[5] DBpedia mapping (http://mappings.dbpedia.org/) version 5 March 2016.

[6] https://github.com/thomlee/infobox2rdf.

### 4.1 Infobox Alignment by Instance-Based Method

As the number of Wikipedia attributes is huge, it is almost impossible to map all existing Wikipedia attributes into the DBpedia properties. We wanted to find other attributes in Source language that have potential to be mapped to the Target language to expand our mapping tables. Therefore, we used instance-based approach to find new alignments.

First, we did the data pre-processing step to filter and cleanse the attributes. It is important to note that we only focused on the attributes whose values are text, number, date, or links, so we omitted any attributes whose value is related to pictures, logos, captions, or signatures. Then, we matched each attribute against the candidate attributes, which are the attributes that already exist in the mapping table. We set a threshold 0.6 to filter the similarity scores of each attribute pair. The score was ranged from 0 to 1. The higher score means the higher probability of two attributes being similar. Each attribute pairs with the score lower than the threshold would be ignored. We only considered attributes that occur $\geq 10$ times in the whole articles for the same template. We added another constraint to only accept the alignments whose number of the matching article pairs $\geq 5$. For the pairs that did not pass, we depended on human judgment to determine whether they are acceptable. There are two criteria we used to determine whether the matching is acceptable or not; the types of their values (e.g. currency, location, organization, etc.) and the values themselves. We compared the values of each pair with other values from the same attribute. For example, for 관할관청–district, we looked at all possible values of 관할관청 attribute in the Korean Wikipedia and all possible values of district in the English Wikipedia. If both attributes share the same values from the same type for at least five different articles, we defined it as acceptable.

After we aligned all possible attribute pairs, we could generate total 41 new mappings for attributes in our dataset. The discovery of the new mappings might also help in the generation process to add more attributes that exist in the Source article but do not exist in the Target article.

### 4.2 Infobox Attribute-Value Pairs Generation

After the Source and Target infoboxes were aligned, we applied our generation technique to generate new infobox tuples to solve error Type I and II in the infoboxes. Figure 3 shows an example of the original attribute-pairs for infobox and the new attributes that have been generated by our approach.

We tested our approach to generate new attribute-value pairs for all articles from five templates that we had chosen. Table 3 and Fig. 4 show the comparison between the number of the existing tuples before alignment and after alignment.

### 4.3 Evaluation

We evaluated the accuracy of our method in two ways. First, we compared similarity between the newly generated values in English with their original tuples in Korean to see the number of attributes that had been correctly translated. Second, we evaluated


| | |
|---|---|
| battles | <Korean War> <Vietnam War> |
| birth_date/death_date | 1924\|10\|18 ~ |
| branch | < Combined forces command headquarters > < Joint Chiefs of Staff (Republic of Korea) > combined forces command headquarters, the Joint Chiefs of staff |
| children/offspring | That 4 m |
| commands | < Combined forces command headquarters > < Korea joint chiefs chairman > combined forces command headquarters of the Joint Chief of staff |
| country | <Republic of Korea Army> |
| laterwork | < Korea land development corporation > |
| nickname | (號) wheel (夏 崙) |
| rank | < South Korea > < South Korea army > South Korea Army Chief < Four-star rank > |
| serviceyears | <1945> <1981> |
| spouse | Yang Jung-Hee (; born radio DJ) |

**Fig. 3.** Attribute-value pairs for article Lew Byung Hyun (류병현) from the original infobox and new attribute-value pairs generated by our method

**Table 3.** Statistics about the number of attribute-value pairs before and after generation process

| Template | Total article pairs | Existing tuples | | Tuples after alignment | | Expanded (%) |
|---|---|---|---|---|---|---|
| | | Ko | En | DBpedia | DBpedia + IB | |
| 군인/Military person | 457 | 5249 | 6669 | 7478 | 8444 | 21.02 |
| 학교/School | 219 | 3000 | 2426 | 3846 | 3940 | 38.43 |
| 왕/Monarch | 584 | 5640 | 6654 | 7333 | 8273 | 19.57 |
| 회사/Company | 1568 | 21016 | 18831 | 27466 | 27827 | 32.33 |
| 대학/University | 879 | 9523 | 14991 | 18694 | 18788 | 20.21 |

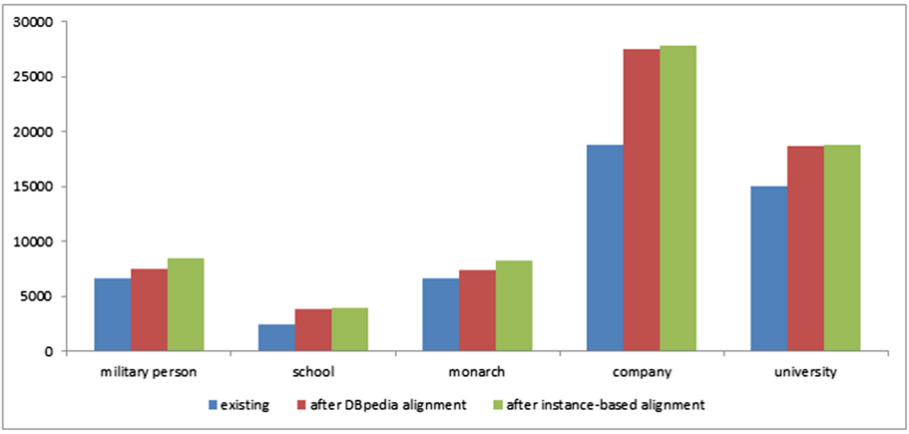


**Fig. 4.** Comparison of the number of attribute-value pairs before and after generation process

the overall accuracy of the method by re-generating the existing English attribute-value pairs using our method and comparing the results with the original ones. We took 20% of the generated pairs as our sample and used human evaluator to do the task. The results show that 73% of the new generated pairs were translated correctly while they also show that our method has overall accuracy of 61%.

It is hard to compare our result with the existing approaches since in our experiment we only used data from five infobox templates from while the other approaches used the data from all infobox templates. If we compared our results with [4, 15], our method has better performance in expanding the existing pairs (tuples) and accuracy. According to [15], their approach could generate 27% new tuples from the existing Dutch Wikipedia while our method could generate up to 38% new tuples. Meanwhile, we compared our accuracy with [4] since [15] did not state the accuracy of their result. The method proposed by [4] was able to match cross language template-attribute pairs with 60% accuracy by using the most frequent tuples, while our method has the slightly higher accuracy, which is 61%. It is important to note that we have not tested the performance against the whole infobox data. Therefore, the number might still be changed. We will leave the evaluation for the future works.

While performing evaluation, we also found common errors that occurred in the results. They happened due to the translation errors, API errors, link errors, and inconsistencies. We found that this approach might be useful to detect value inconsistencies for the same attribute. However, we did not do any validation to resolve the problem and left it for future work.

## 5 Conclusion and Future Work

The purpose of the present study is to fix information gap between cross language Wikipedia articles. We have proposed an approach that takes advantages from the existing DBpedia mappings to align two attributes in different languages that are semantically similar by constructing the mapping tables, which were derived from the extracted DBpedia mapping files that contain the existing mapping of Wikipedia infobox attributes to DBpedia properties. Two attributes that were mapped to the same property were aligned. In addition, we also attempted to expand the number of the existing alignments by using instance-based method to align attributes that do not exist in the mapping table. Our approach was able to expand up to 38% of the existing attribute-value pairs from our dataset.

Previous studies have attempted to complete missing information in infoboxes by using various techniques (e.g. [4, 5]). Even though those approaches show good performances on aligning cross language infoboxes whose languages came from the same root, e.g. Indo-European, they have not been tested for languages that came from the different root, e.g. English and Korean. Our approach is able to generate cross-lingual infoboxes regardless of their root, alphabetical system, or grammar structure, as long as the language is available on Wikipedia and DBpedia. Since Wikipedia covers more than 200 different languages and DBpedia mappings are also available in 40 languages, our approach could be used in broader range in terms of languages. Moreover, to the best of our knowledge, a study that examined the contribution of the existing DBpedia

mapping to the infobox completion process among Wikipedia pages has not been conducted yet. Thus, our study represents the first step into this new direction, thereby making a new contribution to this research field.

Our approach needs to be further refined. Given that our approach relies on human intervention, we would like to do some improvements to reduce the human effort, such as using a robust XML parser to construct the mapping tables. We also would like add a validation component to resolve inconsistencies in the aligned attribute values. Finally, we are planning to expand our dataset to the whole Korean Wikipedia and to other language versions and evaluate our approach to get a holistic view about its performance.

**Acknowledgments.** This work was supported by the Industrial Strategic Technology Development Program, 10052955, Experiential Knowledge Platform Development Research for the Acquisition and Utilization of Field Expert Knowledge, funded by the Ministry of Trade, Industry & Energy (MI, Korea).

## References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyaniak, R., Hellmann, S.: DBpedia - a crystallization point for the web of data. Web Sem. Sci. Serv. Agents World Wide Web **7**(3), 154–165 (2009)
2. Rinser, D., Lange, D., Naumann, F.: Cross-lingual entity matching and infobox alignment in Wikipedia. Inf. Syst. **38**(6), 887–907 (2013)
3. Palmero Aprosio, A., Giuliano, C., Lavelli, A.: Towards an automatic creation of localized versions of DBpedia. In: Alani, H., et al. (eds.) ISWC 2013, Part I. LNCS, vol. 8218, pp. 494–509. Springer, Heidelberg (2013)
4. Adar, E., Skinner, M., Weld, D.S.: Information arbitrage across multi-lingual Wikipedia. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM (2009)
5. Wu, F., Weld, D.S.: Autonomously semantifying Wikipedia. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. ACM (2007)
6. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. VLDB J. **10**(4), 334–350 (2001)
7. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In: Proceedings of 18th International Conference on Data Engineering. IEEE (2002)
8. Li, W.-S., Clifton, C.: SEMINT: a tool for identifying attribute correspondences in heterogeneous databases using neural networks. Data Knowl. Eng. **33**(1), 49–84 (2000)
9. Nottelmann, H., Straccia, U.: Information retrieval and machine learning for probabilistic schema matching. Inf. Process. Manag. **43**(3), 552–576 (2007)
10. Kohonen, T.: Adaptive, associative, and self-organizing functions in neural computing. Appl. Opt. **26**(23), 4910–4918 (1987)
11. Fuhr, N.: Probabilistic datalog—a logic for powerful retrieval methods. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (1995)

12. Wang, H., et al.: Identifying indirect attribute correspondences in multilingual schemas. In: 17th International Workshop on Database and Expert Systems Applications, 2006. DEXA 2006. IEEE (2006)
13. Fu, B., Brennan, R., O'Sullivan, D.: Cross-lingual ontology mapping – an investigation of the impact of machine translation. In: Gómez-Pérez, A., Yu, Y., Ding, Y. (eds.) ASWC 2009. LNCS, vol. 5926, pp. 1–15. Springer, Heidelberg (2009)
14. Dos Santos, C.T., Quaresma, P., Vieira, R.: An API for multilingual ontology matching. In: Proceedings of 7th Conference on Language Resources and Evaluation Conference (LREC). No commercial editor (2010)
15. 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. Association for Computational Linguistics (2009)
16. Nguyen, T., et al.: Multilingual schema matching for Wikipedia infoboxes. Proc. VLDB Endow. **5**(2), 133–144 (2011)
17. Cojan, J., Cabrio, E., Gandon, F.: Filling the gaps among DBpedia multilingual chapters for question answering. In: Proceedings of the 5th Annual ACM Web Science Conference. ACM (2013)
18. Kim, E.-K., Choi, K.-S.: Cross-lingual property alignment for DBpedia ontology using triple conceptualization (2014)
19. Palmero Aprosio, A., Giuliano, C., Lavelli, A.: Automatic expansion of DBpedia exploiting Wikipedia cross-language information. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 397–411. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38288-8_27
20. Kim, E.K., et al.: An approach for supplementing the Korean Wikipedia based on DBpedia. Liliana Cabral (Open University, UK) Tania Tudorache (Stanford University, USA), p. 7 (2010)
21. 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)
22. Tacchini, E., Schultz, A., Bizer, C.: Experiments with Wikipedia cross-language data fusion. In: Workshop on Scripting and Development (2009)
23. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: Aroyo, L., et al. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 665–680. Springer, Heidelberg (2011)
24. Salhi, A., Camacho, H.: A string metric based on a one-to-one greedy matching algorithm. Res. Comput. Sci. **19**, 171–182 (2006)
25. Lee, T.Y., et al.: Automating relational database schema design for very large semantic datasets. Technical report, Department of Computer Science, University of Hong Kong (2013)
26. Lehmann, J., et al.: DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015)