Building a Knowledge Base for QA System by Linking Korean Vocabulary and Wikipedia

Yongbae Lee[1], Pyung Kim[1] & Jungsik Yang[2]

[1] Department of Computer Education, Jeonju National University of Education, 50 Seohak-ro, Wansan-gu, Jeonju-si, Jeonbuk, Korea, 55101

[2] IWAZ Cooperation, 73 Jukdong-ro, Yuseong-gu, Daejeon-si, Korea, 34127

Correspondence: Pyung Kim, Department of Computer Education, Jeonju National University of Education, 50 Seohak-ro, Wansan-gu, Jeonju-si, Jeonbuk, Korea, 55101. E-mail: pyung@jnue.kr

Received: April 4, 2018 Accepted: September 28, 2018 Online Published: May 21, 2019

doi:10.5430/ijhe.v8n3p103 URL: https://doi.org/10.5430/ijhe.v8n3p103

Abstract

For a QA system it's very important to have a notion of vocabulary used in questions and for building correct answers. Especially, when a word represents a concept, one can use related lexical instances to understand it and further extend the knowledge by using the associated information. In this work, we suggest a process of building a knowledge base for such concepts as people, organizations, and places, and linking their instances to Wikipedia articles. We also develop a workbench for KB building. This workbench should efficiently support all features needed to collect necessary data and build the knowledge base. We have created 150,941 links to Korean Wikipedia for 2,394 instances of Korean vocabulary. This KB can be used in QA systems to extend questions, while the workbench can be used to build the KB itself.

Keywords: knowledgebase construction, workbench for knowledgebase construction, question and answering system

1. Introduction

For QA system it's very important to understand questions and the meaning of words used to build the correct answers. In order to understand word meaning, it's useful to have the word definition, identify taxonomic relations among the words, and establish instances that correspond to the concepts. Wikipedia is the most representative body of knowledge in the general domain. One can use semi structured Wikipedia articles to extract appropriate features and connect them to the vocabulary, which can help extend the knowledge and refine correct answers in a QA system.

Although knowledge bases are useful for knowledge extension, their construction is a lengthy and expensive process. It can be facilitated, however, with a workbench specially designed for the purpose. In this work, we limit target vocabulary to people, organizations, and places, and suggest a process and a workbench for linking vocabulary instances to Wikipedia articles. We start with drawing up a guideline and defining processes for KB building. After that, we develop an appropriate workbench, which we will use in the process of building our KB that links target vocabulary to Wikipedia articles. We handle some 390,000 Wikipedia articles and about 580,000 Korean words for concepts classified into people, organizations, and places. We build then the corresponding concept vocabulary and search for Wikipedia articles that match the best the selected lexical core. The discovered instances are linked to the vocabulary and finally verified.

In Section 2 of this article we briefly refer to the related works on the applicability of knowledge bases and workbench development. Section 3 outlines in detail the used data and the process of KB building. Section 4 reviews the workbench features and shows how it works at each stage of the process. Section 5 examines the results of KB building. In Section 6 we draw the conclusion and make some considerations for future research.

2. Related Works

The knowledge base is useful for understanding the vocabulary or the expansion of the information possessed in the QA system or the intelligent services (Bao et al, 2014; Zhang et al, 2016; Park et al, 2016).

Bao, et al (2014) proposed a translation-based KB-QA method that integrates semantic parsing and QA in one unified framework and showed better results on a general domain evaluation set. Zhang et al, (2016) adopt a heterogeneous network embedding method, termed as TransR, to extract items' structural representations by considering the heterogeneity of both nodes and relationships. They proposed Collaborative Knowledge Base Embedding (CKE) to jointly learn the latent representations in collaborative filtering as well as items' semantic representations from the knowledge base. Park (Park et al, 2016; Zesch et al, 2007; Lehmannm et al, 2015; Rebele et al, 2016; Ponzetto and Strube, 2013; Wang and Kim, 2017; Tezcan Kardas and Sadik, 2018; Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017) proposed a method to automatically generate the object name recognition corpus using knowledge base. Two methods are applied according to the type of knowledge base. The first method is to create a learning corpus by attaching an object name tag to a sentence of Wikipedia text based on Wikipedia. The second method generates a learning corpus by collecting various types of sentences from the Internet and attaching an object name tag using a pre-base which holds the relation between various objects in the database.

Wikipedia is a useful resource for building knowledge bases and is actively used in many areas (Zesch et al, 2007; Lehmannm et al, 2015; Rebele et al, 2016; Ponzetto and Strube, 2013; Mokhtar, 2017; Khan, Hassan, & Marimuthu, 2017; Garaeva and Ahmetzyanov, 2018; Kamau., Mwania and Njue, 2018; Aina and Ayodele, 2018; Audu, 2018; Promsri, 2018; Wang and Yang, 2018; Hassan and Kommers, 2018; Agbabiaka-Mustapha and Adebola, 2018). Zesch et al, (2007) developed a general purpose, high performance Java-based Wikipedia API to use Wikipedia as a lexical semantic resource in large-scale NLP tasks. DBpedia project (Lehmannm et al, 2015; Rebele et al, 2016; Ponzetto and Strube, 2013; Wang and Kim, 2017; Tezcan Kardas and Sadik, 2018; Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018) extracts knowledge from 111 different language editions of Wikipedia. The largest DBpedia knowledge base which is extracted from the English edition of Wikipedia consists of over 400 million facts that describe 3.7 million things. The DBpedia knowledge bases that are extracted from the other 110 Wikipedia editions together consist of 1.46 billion facts and describe 10 million additional things. Yago (Rebele et al, 2016; Ponzetto and Strube, 2013; Wang and Kim, 2017; Tezcan Kardas and Sadik, 2018; Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018; Yildirim and Çoban, 2018) is a large knowledge base that is built automatically from Wikipedia, WordNet and GeoNames. This project combines information from Wikipedias in 10 different languages, thus giving the knowledge a multilingual dimension. Wikitaxonomy (Ponzetto and Strube, 2013; Wang and Kim, 2017; Tezcan Kardas and Sadik, 2018; Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018; Yildirim and Çoban, 2018; Lee et al, 2017) is a taxonomy automatically generated from the system of categories in Wikipedia. Categories in the resource are identified as either classes or instances and included in a large subsumption. Knowledge base is used as language resources in various research fields including search and classification fields (Wang and Kim, 2017), (Tezcan Kardas and Sadik, 2018).

The workbench is used in various studies to build knowledge base (Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018). Rybina (Vafa et al, 2018; Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018; Yildirim and Çoban, 2018; Lee et al, 2017; Rybina et al, 2017) suggested knowledge acquisition processes that use technologic knowledge base of intelligent planner of AT-TECHNOLOGY workbench and special program tools. This work is focused on models and methods of distributed knowledge acquisition from databases as additional knowledge sources and automation of the process via intelligent program environment. Choi (Wadmany and Melamed, 2018; Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018; Yildirim and Çoban, 2018; Lee et al, 2017; Rybina et al, 2017; Choi et al, 2012) suggested SINDI-WALKS, an integrated workbench that can extract and systematically manage technical knowledge inherent in scientific and technical literature such as academic papers and patents. SINDI-WALKS basically includes a technology knowledge extraction engine that identifies the PLOT, ie, names, names, institutions, and technical terms in text and extracts semantic relationships between them, and a testbed function for monitoring and error analysis of these engines. do. It also supports the ability to build test collections to efficiently build a learning set that can be utilized by a technology knowledge extraction engine. A workbench was developed and used to support all the processes needed to build a terminology dictionary in the defence field (Wyatt et al, 2018; Yang et al, 2017; Yildirim, 2018; Yildirim and Çoban, 2018; Lee et al, 2017; Rybina et al, 2017; Choi et al, 2012; Choi et al, 2012). The workbench is composed of terminology dictionary construction process and organization structure, definition of headwords, selection of target document for extracting terminology candidate, extraction of terminology candidate, creation of terminology candidate group, dictionary construction, verification of dictionary.

3. Knowledge Base Construction: Data and Process

The process of knowledge base construction starts with data selection and goes through the number of steps to final verification of the KB. For the purpose of this study, we categorize target vocabulary into people, organizations, and places and link their instances to the corresponding Wikipedia articles. In this section we examine the data used for KB building and the construction process.

A. Data for Knowledge Base Construction

For KB construction we use Korean vocabulary and Korean Wikipedia. We limit target vocabulary used for KB construction to people, organizations, and places. Accordingly, we pick up 81,272 words, which makes out 14% out of 585,039 vocabulary corpora. We use 396,335 articles from Korean Wikipedia as a baseline for September 2017, which we collect for KB construction purposes. There are all together 226,601 Wikipedia articles on people, organizations, and places, which make 57% of the total.

Table 1 shows the distribution of articles in Korean Wikipedia for each category: people, organizations, and locations. Vocabulary attributes include word definition, hypernyms and hyponyms, word type and other information. Wikipedia articles include the body text and category. For some Wikipedia articles management template is further available.

Table 1. Number of Vocabulary and Wikipedia

Kind	**Vocabulary**		**Wikipedia**		**Description**
	#	**%**	**#**	**%**	
Person	36,806	6,2	109,644	27.7	Person, Group, Job Title
Organization	21,274	3.6	46,427	11.7	Team, Cooperation, Organization
Location	23,912	4.1	70,530	17.8	Place, Building, Country
Etc.	510,318	87.2	169,524	42.8	Other vocabulary except Person, Organization, Location
Total	585,039	100	396,335	100	

B. Process of Knowledge Base Construction

In order to link concepts and vocabulary instances to Korean Wikipedia, we need to collect concept vocabulary, look up for related Wikipedia articles, and provide verification.

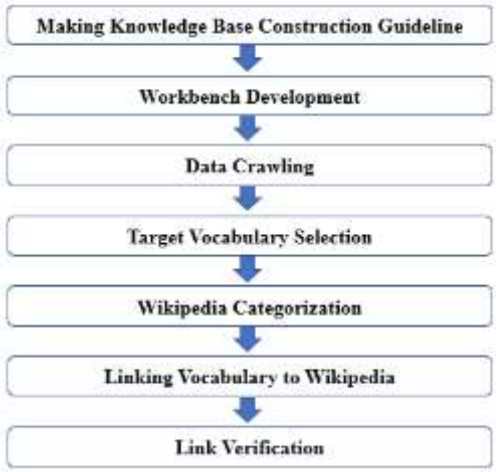
Making Knowledge Base Construction Guideline
Workbench Development
Data Crawling
Target Vocabulary Selection
Wikipedia Categorization
Linking Vocabulary to Wikipedia
Link Verification

Figure 1. Process of Knowledge Base Construction

The process of KB development is shown in Figure 1 and consists of 7 steps from making the guideline to final verification.

Making Knowledge Base Construction Guideline: At this step, we determine the scope of data, the working processes, user roles and permissions, the method of creating related data, and verification. In general, we set up exactly what and how we will do at each following step. The target Korean vocabulary is limited to people, organizations, and places for which we should link concepts to Wikipedia articles. We also use only Wikipedia articles that fall into people, organizations, and places categories, although it's possible to consider articles from other categories, as well. Two operators work on vocabulary selection, Wikipedia categorization, and links creation. Two supervisors check work results and perform verification and approval.

Workbench Development: In the process of building a knowledge base, several users should be able to efficiently create links for multiple vocabulary entries and Wikipedia articles. For the purpose of this study, we develop the workbench first, and then use it on the following stages. At this stage, we design Workbench features and UI for collecting and screening the vocabulary, Wikipedia categorization, linking selected instances to Wikipedia, performing verification and user management. Finally, we develop the program.

Data Crawling: At the of step gathering and storing target vocabulary and Korean Wikipedia articles, users initiate data collection, and depending on the progress can suspend or terminate the task.

Target Vocabulary Selection: First, vocabulary related to people, organizations, and places is selected. Then, the vocabulary is confined to the concepts that can be linked to Wikipedia. The workbench automatically classifies target vocabulary into people, organizations, and places. Users can edit the results, determine whether a specific concept is needed at all, and specify exceptions. If two operators select different classification and processing options for the specific vocabulary, the supervisor checks the result and makes the final decision.

Wikipedia Categorization: Wikipedia articles are also initially categorized into people, organizations, places, and "other" based on collected article properties and template information. If two operators make different categorization for the specific Wikipedia article, the supervisor checks the result and makes the final decision.

Linking Vocabulary to Wikipedia: Two operators use the target vocabulary to perform Wikipedia searches. At this step, it's possible to further use hypernyms, hyponyms and synonyms for the selected word.

Link Verification: The supervisor finally checks how vocabulary is linked to the Wikipedia, and can edit, approve or reject the instance.

4. Knowledge Base Construction Workbench

The workbench should efficiently support all features needed to collect necessary data and build the knowledge base. Also, it should be possible to save and restore job results for multiple users working with the program at the same time. In the course of this study, we developed a workbench that supports necessary features specified in the Knowledge Base building process and used it to build the KB.

In order to reduce the number of KB errors, the workbench facilitates the procedure where two operators concurrently perform vocabulary selection, Wikipedia categorization, and vocabulary linking, and one supervisor verifies work results. Accordingly, the workbench further provides the possibility to assign tasks to individual workers, monitor activity progress, and verify and approve work results.

A. User Registration and Rights Management

The workbench supports such user roles as administrator, operator, and supervisor, and limits functionality available to the user depending on her role. The administrator assigns user roles when the user is registered in the system. Table 2 shows user functions depending on the role.

Table 2. Users Rights of Workbench

Type	Rights
Administrator	- Register new users, change user roles - Data collection and monitoring - Monitoring the status of operator's and supervisor's work - KB construction, verification, recovery, editing and approving jobs, etc.
Operator	- Select vocabulary candidates for the task - Select candidate Wikipedia articles and categorize them - Select link candidates
Supervisor	- Mange user rights - Edit and select task vocabulary - Edit and approve Wikipedia categories - Edit and approve links among vocabulary and Wikipedia

The administrator has full access to all workbench functions and can register users, monitor job progress and results, as well as edit, approve, and reject the results. The operator can select vocabulary candidates, Wikipedia articles and categories, and edit links. The supervisor can further edit and approve the selected entries.

B. Data Crawling

For KB construction we need Korean vocabulary and data from Korean Wikipedia. Data collection feature facilitates selection of entries from Korean vocabulary database, collecting and storing necessary values, and collecting and storing articles from Korean Wikipedia.

The administrator can use this feature to specify data attributes, whose values should be collected from Korean vocabulary, check Korean Wikipedia statistics, and start, suspend, resume, and terminate data collection tasks. Also, it's possible to monitor data collection progress and individually check collected entries stored in the database.

C. Selection of Target Conceptual Vocabulary

Lexical classification feature is used to automatically classify the vocabulary into people, organizations, places, and "other" based on attributes retrieved from Korean vocabulary database, and lets operators search for selected words and edit classification. Because vocabulary attributes contain information about the word class, preliminary classification can be made automatically. Where automatic preliminary classification is not possible due to the lack of the corresponding attribute or where classification results are not correct, the operator can edit the entry manually and approve the changes she has made.

During vocabulary classification it is necessary to discriminate concept and non-concept words and to exclude too general concepts and relative terms. Table 3 below shows how vocabulary is classified into target concepts, non-target concepts, and non-concepts for people, organizations, and places.

The functions supported by the workbench for selection of target conceptual vocabulary are:

- Classify Korean Wikipedia to people, organization, location
- Search vocabulary: forward, middle, backward search and nearby, within, digits search
- Search by adding search condition directly to title and body of vocabulary and Wikipedia
- Save and edit all or selectively classification information of Wikipedia
- Store and manage work done by multiple workers

When handling the vocabulary, operators use word meaning, word type, hypernyms and hyponyms, etc. The workbench, accordingly, supports necessary features and further makes it possible to use Excel to upload vocabulary classification results. It's also possible to process hypernyms and hyponyms for target vocabulary in batch. Two operators handle vocabulary classification, and two supervisors check the results and make final decisions.

Table 3. Target Vocabulary and Non-Target Vocabulary

Type	Descriptions
Conceptual Vocabulary (Target)	- Terms related to people, organizations, and places - People: occupation, job title, activity, team, nationality, etc. - Organization: name, group, affiliation, etc. - Place: administrative district, country, city, building, etc.
Conceptual Vocabulary (Non-Target)	- Concept is too general or relative - Too general: household, mountain district, riverside, etc. - Relative rich, poor, modern building, cool place, etc.
Non-Conceptual Vocabulary (Non-Target)	- Non-concept proper names - Proper names: Napoleon, Seoul, Namdaemun, etc.

D. Linking Vocabulary to Wikipedia

Linking Wikipedia articles to concept instances from the vocabulary is the most time-consuming and the important task in building a relevant knowledge base.

As shown in Table 4, we use target vocabulary to search across the Wikipedia. The retrieved Wikipedia articles are clustered, and link candidates are suggested for possible vocabulary entries arranged in order of frequency. In this study we suggest links using vocabulary-based Wikipedia search, which makes the entire process more transparent and increases the accuracy of work.

Table 4. Linking Method Pros and Cons

Linking Method	Considerations
Vocabulary-based Wikipedia search	- The operator searches Wikipedia based on her knowledge about the vocabulary - As vocabulary descriptions are very brief, automatic vocabulary expansion can be difficult - When limited to search results and operator's knowledge only, a lot of link candidates remain uncertain - During linking, the list of missing Wikipedia articles is created - There are many time-consuming tasks associated with Wikipedia checks - It's possible to enhance productivity of Wikipedia linking with tools and make it more transparent and accurate
Wikipedia clustering for vocabulary search	- Accurate clustering and creating cluster definitions can take a lot of time - In case no frequency vocabulary can be displayed for a specific cluster, it's necessary to repeat the search process again - Synonyms, North Korean variants, rare words, archaisms, etc. are off the cluster - Cluster precision is of great importance, and it takes effort to keep it accurate and consistent - Not all information necessary for Wikipedia linking is available in the Wikipedia documentation - In case the clusters are not accurate, the operator should be able to categorize and link articles manually

The functions supported by the workbench for linking vocabulary to Wikipedia are:

- Register vocabulary work results stored in Excel with batch process
- Replicate vocabulary and Wikipedia connection information to other vocabularies
- Search Wikipedia using vocabulary
- Search for the upper and lower vocabularies of the target vocabulary, and to link the all or selectively vocabularies with the Wikipedia
- Return previously processed vocabulary to a previous step
- Search Wikipedia using title, body, and object name
- Provide vocabulary information with work status, and separately stores and manages linked information of multiple workers
- Search work period, type, area of linked information

Figure 2 shows the program interface for Wikipedia search. In the ① area, the operator selects the vocabulary. It also displays vocabulary classification, job status, word number, etc. In the ② area, vocabulary definition is displayed along with information whether this is a terminal node and vocabulary classification results. In the ③ area, you can see the taxonomically related words. The ④ area is where the results of conditional search for

Wikipedia titles and content are displayed. You can further filter Wikipedia categories here. In the ⑤ area, you can confirm Wikipedia search results and categorization. In the ⑥ area, you can check the article content.

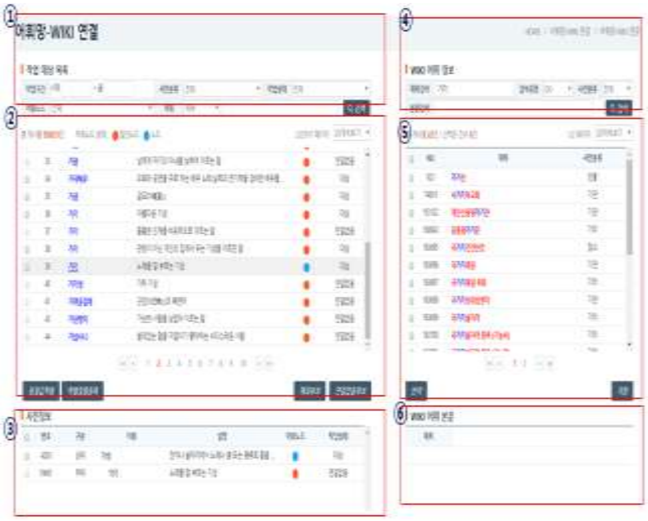
어휘망-WIKI 연결

Figure 2. Interface for Linking Vocabulary to Wikipedia

The operator uses vocabulary definitions and her knowledge to search Wikipedia for different conditions. After he checks search results, she can select link candidates. In such a manner, the operator can use different Wikipedia searches to extend vocabulary linking and further connecting hypernyms, hyponyms, and synonyms.

This enhances linking efficiency because the established connections can be further cloned for North Korean variants, synonyms, archaisms, etc.

E. Link Verification

The list of Wikipedia links for the vocabulary can vary a lot depending on operator's understanding of the vocabulary and search queries she uses. Therefore, candidate links are finally validation at the next step by the supervisor who does not establish the links himself.

Figure 3 shows the verification interface where two supervisors check linking results. The ① area displays the vocabulary list, lexical classification, work progress, word number, and the number of operators. The ② area displays word definitions, work progress by operator, and operator's confirmation status. The ③ area features final approval of Wikipedia links. The ④ area lets confirm link candidates created by each operator.

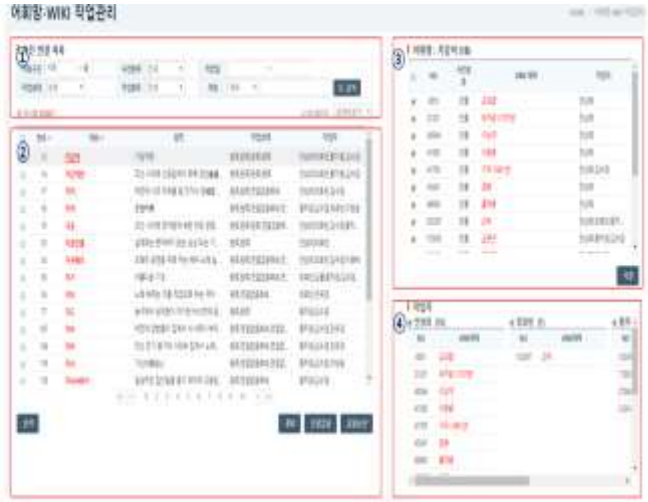
어휘망-WIKI 작업관리

Figure 3. Interface for Link Verification

Two supervisors check links created concurrently by two operators. If no faults are detected, links from individual operators are consolidated to create a combined linkage information. The supervisor can approve or reject link candidates or mark a vocabulary instance as link missing.

F. Monitoring Work Progress

Because many operators and supervisors can work concurrently with the same tool, multiuser statistics on work results is required along with possibility to edit the results where necessary. Depending on user rights, the workbench makes it possible to check statistics by task type, results and edit the results as may be required.

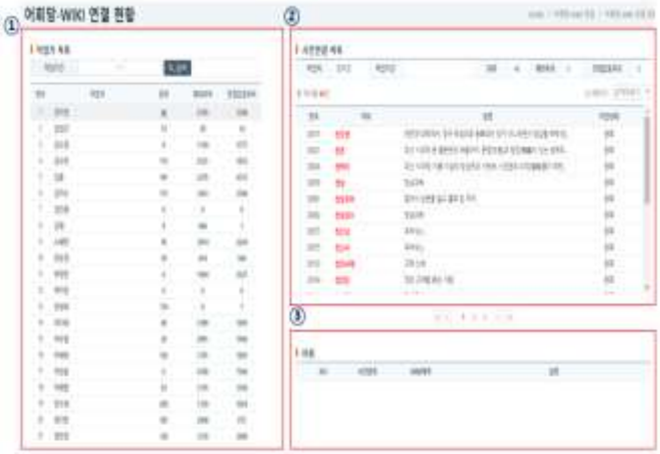

Figure 4. Interface for Monitoring Work Status

Operators can check their work results for vocabulary selection, Wikipedia categorization, and link generation made on specific date. Supervisors can also check and edit vocabulary selection, Wikipedia categorization, and link generation made by each operator on a specific day. Supervisors can further monitor the progress of operators' work and edit the results for each stage.

Figure 4 shows the program interface for monitoring work status of operators in the process for linking vocabulary to Wikipedia. The administrator can see a list of worker 's connection work status in the area ①, when the work kind of a specific worker is selected, the vocabulary list can be displayed in the area ②, if a vocabulary is selected in area 2 then the vocabulary information will be shown in the area ③.

G. DB Schema for Workbench

The DB table consists of a table for storing Wikipedia information, three tables for storing vocabulary information, a table for storing vocabulary and Wikipedia connection information, and two tables for user and code management.

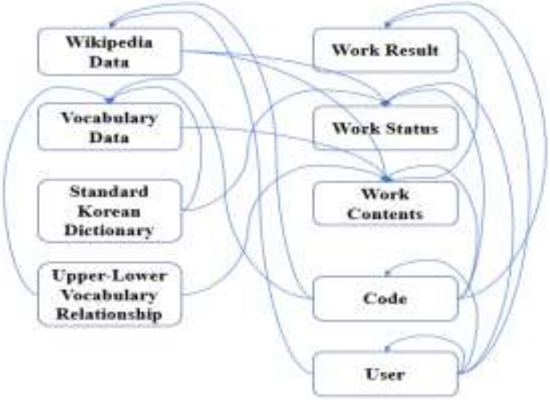

Figure 5. Relationships between Tables

The roles of the 9 tables used in the workbench are as follows.

1) Wikipedia: The Wikipedia table stores the data collected from Wikipedia and the classification data generated during the collecting and storing process.

2) Vocabulary: The vocabulary table stores vocabulary information for vocabulary and Wikipedia connection. It contains vocabulary information in the standard Korean dictionary, search and classification data of vocabulary, data for linking to Wikipedia with workbench.

3) Standard Korean Dictionary: The standard Korean dictionary table distinguishes vocabulary according to the sense of the vocabulary and assign the description and the unique number according to the sense of the vocabulary.

4) Upper-Lower Vocabulary Relationship: According to the concept of vocabulary, it can be divided into upper vocabulary and lower vocabulary. The upper-lower vocabulary relation table is for storing relationship of vocabularies according to the concept.

5) Work Result: The work result table contains ranking information about vocabulary connection to Wikipedia.

6) Work Status: The work state table stores user-created vocabulary connection information. The work status has 6 kinds of status information corresponding to completion, approval, exclusion candidate, exclusion, no connection candidate, and no connection.

7) Work Contents: The work content table is a table for storing all the work contents that are processed through the workbench, and includes a vocabulary number, a Wikipedia number, a work content, and information about the worker.

8) Code: The code table is a table for managing the code used in the workbench. It contains information for creating and modifying the code, checking whether the code is used or not, as well as information for the upper and lower code structure.

9) User: The user table is a table for managing information on users who use the workbench, and includes not only the user ID and password, but also the role of the worker and recent login information.

Table 5 shows the table and attribute values of the database used in Workbench.

Table 5. Tables for Workbench

Table	**Attributes**
Wikipedia	wikipedia document id, title, category tag, weight, redirect, contents, check, category, update date, update user id
Vocabulary	sequence, vocabulary id, ontology id, ontology position tag, original vocabulary id, search title, title, analogy vocabulary id, word sense id, description, category tag, Upper-Lower sequence
Standard Korean Dictionary	vocabulary id, ontology id, original vocabulary id, title, word sense, description
Upper-Lower Vocabulary Relationship	sequence, ontology id, vocabulary id, lower vocabulary id,
Work Result	sequence, vocabulary id, wikipedia id, user id, rank, insert date, insert user id, update date, update user id, delete date, delete user id
Work Status	sequence, vocabulary id, code id, user id, rank
Work Contents	sequence, vocabulary id, wikipedia id, code id, user id, rank, insert date, insert user id, update date, update user id, delete date, delete user id
Code	code id, group ig, parent id, level, code name, code description, check, sequence, insert date, insert user id, update date, update user id
User	user id, user name, user type, password, password fail count, department name, tell number, mobile phone number, last login date, last login ip, statue, update date, update user id

5. Results of Knowledge Base Construction

Twenty operators and ten supervisors had been working on KB construction for the period of five months, starting from September 2017.

Vocabularies can be divided into leaf nodes and non-leaf nodes according to the concept, and Table 2 shows the number of leaf nodes and non-leaf nodes for each vocabulary type. In this study, the task of linking vocabulary as a class and Wikipedia as an instance is performed, whether the leaf node of vocabularies is also considered in the target vocabulary selection task. Table 6 shows the number of vocabularies belonging to non-leaf node and leaf node. 9,441 vocabularies belonging to people, organization, and location belong to non-leaf node, and this information was also taken into consideration in the process of determining the vocabulary to be connected.

Table 6. Number of Leaf Node Vocabularies

Type	Non-Leaf Node	Leaf Node	Total
Person	4,477	20,199	24,676
Organization	2,332	11,468	13,800
Location	2,632	13,182	15,814
Total	9,441	44,849	54,290

As a result, among 81,272 words from target vocabulary 2,394 words for people, organizations, and places 150,941 Wikipedia links have been created as shown in Table 7. For about a half of vocabulary entries for people, organizations, and places linking have failed for the absence of relevant Wikipedia articles or because the specific vocabulary was too general or relative. The remaining non-linked vocabulary are proper names that do not represent concepts. As many Wikipedia articles are linked to more than one vocabulary item, 150,941 links correspond to all together 84,852 Wikipedia articles linked to the target vocabulary.

Table 7. Number of Vocabulary Links to Wikipedia

Type	# of Linked Vocabulary	# of Linked Wikipedia (with redundancy)	# of Linked Wikipedia (without redundancy)
Person	875	54,123	29,058
Organization	829	57,913	27,705
Location	757	38,905	28,089
Total	2,394	150,941	84,852

Because one vocabulary is linked to several Wikipedia articles, the vocabulary according to the number of articles of connected Wikipedia is as shown in Table 8. There are 1,332 vocabularies linked to less than 10 Wikipedia articles, accounting for more than 50% of the total. There are 633 vocabularies linked to 11 ~ 50 Wikipedia articles, accounting for 26.4% and 21 vocabularies linked to more than 1000 Wikipedia articles.

Table 8. Number of Linked Wikipedia

# of Linked Wikipedia	# of Vocabulary	%
1~10	1,332	55.6
11~50	633	26.4
51~100	148	6.2
101~1000	260	10.9
1001 ~	21	0.9
Total	2,394	100

Table 9 shows examples of Wikipedia linked by vocabulary. There are vocabularies that can easily be linked to Wikipedia articles such as general, school, and museum, but there are some vocabularies that are difficult to find Wikipedia articles that need to be linked like archaeologists, open schools, breeding grounds.

Table 9. Examples of Linked Vocabulary

Vocabulary		Wikipedia
Person	General	Sunshin Lee, Kamchan Kang, Munduk Eulgi, Yongwoo Kim,…
	Musician	Eddi Kim, Roi Kim, C Kim, Kunmo Kim, Yeon Park, …
Organization	School	Korea University, Seoul National University, Daejeon Middle School, …
	Cooperative	Seoul Milk Cooperative, National Federation of Fisheries Cooperatives, Worker Cooperative, …
Location	Museum	Gail Museum, Kansong Museum, Kyungwun Museum, Korea University Museum, …
	National Park	Raeryong Mountain, Dukyu Mountain, Sokri Mountain, Joowang Mountain, …

6. Conclusion

This study proposes a process and a workbench for building a knowledge base and uses them for creating a KB that links Korean vocabulary instances to Korean Wikipedia articles. The work continued for five months with the involvement of twenty operators and ten supervisors who created Wikipedia links for people, organization, and places concepts. In the process of KB creation, 150,941 Wikipedia links have been created for 2,394 words for people, organizations, and places among 81,272 words from the target vocabulary.

In the process of Wikipedia categorization, vocabulary selection for the task, and generating linking data we used vocabulary and Wikipedia attributes for automatic processing, and then verified the results in manual mode. To ensure the accuracy of the KB, two operators worked separately in parallel, and one supervisor checked and edited work results where necessary. The obtained KB can help improve understanding questions in a QA system, and further extend subject knowledge by using structured collection of documents associated with a specific vocabulary instance.

In order to link vocabulary to Wikipedia articles, the operator should understand vocabulary concepts first. Thus, in spite of the ambiguity of Wikipedia search results, although the process takes a long time, the quality of the entire work is high. When direct vocabulary search for Wikipedia yields no results, however, operators may opt to similar words, which may result in data that depend on operator's preferences. On the other hand, vocabulary-based Wikipedia search suggests that primary Wikipedia clusters are created first. Upon that, a representative cluster vocabulary is selected, which operators can use in their work. Operators are supposed to understand well cluster characteristics. If clusters are built accurately enough, operators can efficiently exclude or edit the articles in question. Yet another problem is how to link similar vocabulary that is not available in Wikipedia.

In other words, both searching and linking Wikipedia for vocabulary entries and search vocabulary based on representative vocabulary from Wikipedia clusters have their pros and cons. Accordingly, for the purpose of this study we use vocabulary-based Wikipedia search, which makes it possible to enlarge the connected domain and enhance links quality.

In order to enhance the quality of KB links and to ensure efficiency and usability of the workbench, however, the used tool needs some more improvements. Also, in order to improve the usability of the knowledge base, it would be helpful to expand the vocabulary beyond people, organizations, and places, and create Wikipedia links for these categories. Further improvements to link building should include the possibility to take advantage of both methods: vocabulary-based Wikipedia search, and vocabulary search based on Wikipedia cluster, as well as making it possible for the operator to check and remedy missing links. There is also a need to extend Wikipedia and vocabulary linking around similar vocabulary.

References

Agbabiaka-Mustapha, M., & Adebola, K. S. (2018). Exploring Curriculum Innovation as a Tool Towards Attainment of Self Reliance of NCE Graduates of Islamic Studies. *International Journal of Emerging Trends in Social Sciences, 2*(1), 21-27. https://doi.org/10.20448/2001.21.21.27

Aina, J. K., & Ayodele, M. O. (2018). The Decline in Science Students' Enrolment in Nigerian Colleges of Education: Causes and Remedies. *International Journal of Education and Practice, 6*(4), 167-178. https://doi.org/10.18488/journal.61.2018.64.167.178

Audu, T. A. (2018). Effects of Teaching Methods on Basic Science Achievement and Spatial Ability of Basic Nine Boys and Girls in Kogi State, Nigeria. *Humanities and Social Sciences Letters, 6*(4), 149-155. https://doi.org/10.18488/journal.73.2018.64.149.155

B.G. Lee, D.H. Lim and J.S. Kim. (2017). Performance Improvement of Wave Information Retrieval Algorithm Using Noise Reduction. *Journal of Information and Communication Convergence Engineering, 15*(3), 175-181.

F. Zhang, N. J. Yuan, D. Lian, X. Xie, W. M. (2016). Collaborative knowledge base embedding for recommender systems, In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 353-362. https://doi.org/10.1145/2939672.2939673

G. V. Rybina, Y. M. Blokhin, E. S. Sergienko. (2017). Distributed knowledge acquisition basing on integration of Data Mining and Text Mining methods and their usage with AT-TECHNOLOGY workbench, In Future Internet of Things and Cloud Workshops, 1-6. https://doi.org/10.1109/FiCloudW.2017.77

Garaeva, A. K., & Ahmetzyanov, I. G. (2018). Awareness of Historical Background as One of the Factors of Better Language Acquisition. *International Journal of English Language and Literature Studies, 7*(1), 15-21.

Hassan, M. I. A., & Kommers, P. (2018). A Review on Effect of Social Media on Education in Sudan. *International Journal of Educational Technology and Learning, 3*(1), 30-34. https://doi.org/10.20448/2003.31.30.34

J. Bao, N. Duan, M. Zhou, T. Zhao. (2014). Knowledge-based question answering as machine translation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 967-976. https://doi.org/10.3115/v1/P14-1091

J. Lehmannm R. Islel, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. can Kleef, S. Auer, C. Bizer. (2015). DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web, 6*(2), 167-195.

J. W Choi, J. H. Park, K. S. Kim, P. Kim. (2012). Science and Technology Terminology Dictionary Building Process and Workbench Development in Defense Area. *JOURNAL OF THE KOREA CONTENTS ASSOCIATION, 12*(8), 420-428. https://doi.org/10.5392/JKCA.2012.12.08.420

Kamau, L. M., Mwania, J., & Njue, A. K. (2018). Technology resources for teaching secondary mathematics: lessons from early and late adopters of technology in Kenya. *Asian Journal of Contemporary Education, 2*(1), 43-52.

Khan, H., Hassan, R., & Marimuthu, M. (2017). Diversity on corporate boards and firm performance: An empirical evidence from Malaysia. *American Journal of Social Sciences and Humanities, 2*(1), 1-8. https://doi.org/10.20448/801.21.1.8

Mokhtar, S. B. (2017). Teaching-Learning Model of Islamic Education at Madrasah Based on Mosque in Singapore. *International Journal of Asian Social Science, 7*(3), 218-225. https://doi.org/10.18488/journal.1/2017.7.3/1.3.218.225

Promsri, C. (2018). The Influence of External Locus of Control on Life Stress: Evidence from Graduate Students in Thailand. *International Journal of Social Sciences Perspectives, 3*(1), 38-41. https://doi.org/10.33094/7.2017.2018.31.38.41

S. P. Choi, H. W. Chun, C. H. Jeong, H. M. Jung. (2012). SINDI-WALKS : A Workbench for Scientific Intelligence Discovery. *Journal of KIISE : Computing Practices and Letters, 18*(12), 906-910.

S. P. Ponzetto, M. Strube. (2013). WikiTaxonomy: A Large Scale Knowledge Resource. *In ECAI, 178*, 751-752.

T. Rebele, F. Suchanek, J. Hoffart, J, Biega, E. Kuzey, G. Weikum. (2016). YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. *The Semantic Web – ISWC*, 177-185. https://doi.org/10.1007/978-3-319-46547-0_19

T. Zesch, I. Gurevych, M. Mühlhäuser. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, 197205.

Tezcan Kardas, N., & Sadik, R. (2018). An Analysis of the Effect of Educational Game Training on Some Physical Parameters and Social Skills of the Children with Autism Spectrum Disorders. *Asian Journal of Education and Training, 4*(4), 319-325.

Vafa, S., Sappington, K., & Coombs-Richardson, R. (2018). Using Augmented Reality to Increase Interaction in Online Courses. *International Journal of Educational Technology and Learning, 3*(2), 65-68. https://doi.org/10.20448/2003.32.65.68

Wadmany, R., & Melamed, O. (2018). “New Media in Education” MOOC: Improving Peer Assessments of Students' Plans and Their Innovativeness. *Journal of Education and e-Learning Research, 5*(2), 122-130. https://doi.org/10.20448/journal.509.2018.52.122.130

Wang, K., & Yang, Z. (2018). The Research on Teaching of Mathematical Understanding in China. *American Journal of Education and Learning, 3*(2), 93-99. https://doi.org/10.20448/804.3.2.93.99

Wyatt, Z., Hoban, E., & Macfarlane, S. (2017). Trauma-Informed Education Practice in Cambodia. *International Journal of Asian Social Science, 8*(2), 62-76.

X. Wang & H.C. Kim. (2017). New Feature Selection Method for Text Categorization. *Journal of Information and Communication Convergence Engineering, 15*(1), 53-6.

Y. M. Park, Y. J. Kim, S. W. Kang, J. Y. Seo. (2016). Automatic Training Corpus Generation Method of Named Entity Recognition Using Knowledge-Bases. *KOREAN JOURNAL OF COGNITIVE SCIENCE, 27*(1), 27-41. https://doi.org/10.19066/cogsci.2016.27.1.002

Yang, D. C., Chang, M. C., & Sianturi, I. A. (2017). The Study of Addition and Subtraction for Two Digit Numbers in Grade One Between Singapore and Taiwan. *Learning, 2*(1), 75-82. https://doi.org/10.20448/804.2.1.75.82

Yildirim, M. (2018). Investigation of Physical Activity Levels of Physical Education and Sports School Students. *Asian Journal of Education and Training, 4*(4), 347-355. https://doi.org/10.20448/journal.522.2018.44.380.390

Yildirim, M., & Çoban, O. (2018). Examination of the Aggression Levels of Physical Education and Sport School Students. *Asian Journal of Education and Training*, *4*(4), 380-390. https://doi.org/10.20448/journal.522.2018.44.380.390