HUKB at SHINRA2020-ML task

Masaharu Yoshioka
Faculty of Information Science and Technology, Hokkaido University N-14 W-9, Kita-ku, Sapporo 060-0814, Japan.
Graduate School of Information Science and Technology, Hokkaido University
Global Station for Big Date and Cybersecurity, Global Institution for Collaborative Research and Education, Hokkaido University.
Center for Advance Intelligence Project, RIKEN
Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi,Chuo-ku, Tokyo 103-0027, Japan
yoshioka@ist.hokudai.ac.jp

Yoshiaki Koitabashi
Graduate School of Information Science and Technology, Hokkaido University
yk-wwm-aa@eis.hokudai.ac.jp

ABSTRACT

The HUKB team participated in the SHINRA 2020 ML task of the NTCIR-15. This paper introduces our approach to solve the problem and discuss the official results.

TEAM NAME

HUKB

SUBTASKS

SHINRA 2020-ML (30 languages)

1 INTRODUCTION

SHINRA2020-ML task [1] is a task to classify the Wikipedia entities in 30 languages into fine-grained categories provided by SHINRA project[1].

For this task, original categorized data was created by using Japanese Wikipedia pages as training data and this data is used for training data of other languages using inter-language links. The participants are asked to categorize Wikipedia entities of one or more target language(s) using the training data.

HUKB participate this task to discuss the method to categorize Wikipedia entities using Wikipedia category only. Since Wikipedia category was designed to categorize Wikipedia pages from multiple aspects, it may be possible to make matching between Wikipedia categories and fine-grained categories provided by SHINRA project.

In this paper, we introduce a method to make such matching lists between Wikipedia categories and SHINRA fine-grained categories. This method is used for making submission list of SHINRA 2020 tasks and we also discuss the characteristics of the method using the evaluation results.

2 METHODS

SHINRA is a project for making a resource to structure the knowledge in Wikipedia. Estimation of the categories of Wikipedia entities is the first step to make such knowledge from the Wikipedia. In SHINRA, there are 219 categories in Extended Named Entity [2] and they also construct data that classify Japanese Wikipedia entities into these categories.

Wikipedia categories are information associated with Wikipedia pages to classify the pages. However, since there are multiple aspects of information in the Wikipedia pages, there are categories that is not so meaningful to classify the pages. As we have discussed in the Wikipedia category ontology[2], there are four types of categories; topics (e.g., "Japan"), set (e.g., "Cities"), constrained topics (e.g., "2019 in Japan") and constrained set (e.g., "Cities in Japan"). Since SHINRA fine-grained categories is for classify the named entity types, most of the categories related to topics are not useful for the category estimation of the Wikipedia page.

2.1 Matching lists between Wikipedia categories and SHINRA categories

Therefore, we make matching lists between Wikipedia categories and SHINRA fine-grained categories based on the information of the Wikipedia category information for the training data.

Followings are our basic approach to make the matching list.

(1) Make a simple matching list using precision
Since it is difficult to estimate the matching between Wikipedia category and SHINRA category when there are two or more candidate categories for the entities, we use set of pages that have only one candidate category in the training data (P_{all}). We calculate precision for Wikipedia category ($wcat$) for SHINRA category ($scat$) using training data for all Wikipedia categories that have a page that belongs to $scat$ for the category using following equation.

$$prec_{wcat,scat} = \frac{|P_{wcat} \cup P_{scat}|}{|P_{wcat}|} \quad (1)$$

In this equation, $P_{wcat}(\subset P_{all})$ and $P_{scat}(\subset P_{all})$ are set of pages that have $wcat$ in the category list and

[1] https://shinra-project.info/

[2] http://ene-project.info/

pages that are categorized for *scat* in the training data, respectively.
We select pair of $(wcat, scat)$ whose $prec_{wcat,scat}$ is greater or equal to $precmin$ and $|P_{wcat} \cup P_{scat}|$ is greater or equal to $minfreq$ for the candidate pairs. We use $precmin = 0.85$ and $minfreq = min(5, |P_{scat}|/5$ in this experiment. $prec_{wcat,scat}$ is used for *original-based score*.

(2) Make a generalized matching list using SHINRA category hierarchy.
Due to the mismatch of the Wikipedia categories and SHINRA categories, there are several cases that pages for one Wikipedia categories belongs to two or more Wikipedia categories. For example, "Columbia Record artists" have pages for the artists that are classified as "1.1"(Person) for solo artists and "1.4.2" (Show organization) for a group of artist; e.g., music band. In such cases, when they share the abstract category ("1" for this case), we calculate the precision for the abstract category using the set of page $P_{scat'}$ defined by equation 2 for the abstract SHINRA category $scat'$ instead of P_{scat} in equation 1.

$$P_{scat'} = \bigcup_{i=1}^{i<4} P_{scat_i} \tag{2}$$

where $scat_i$ is a i-th highly frequent SHINRA categories in P_{wcat} and share the abstract category with $scat_1$. We also select pair of $(wcat, scat')$ whose $prec_{wcat,scat'}$ is greater or equal to $precmin$ and $|P_{wcat} \cup P_{scat}|$ is greater or equal to $minfreq$ for the candidate pairs. Since this task asks the participants to classify the category in the detailed level, $prec_{wcat,scat_i}$ is used for *generalized-based score*.

(3) Expansion of candidate pairs by using Wikipedia category hierarchy.
Due to the limitation of the training data, matching lists generated by the previous procedures are not good enough to cover wide varieties of the Wikipedia pages. We expand the category list using Wikipedia category hierarchy based on the string pattern matching.
In the Wikipedia, there is a policy for diffusion to make the big category into small categories with different constraints [2], subcategories for such a big category use same word sequence (substring) to represent the diffused relationships. For example, "Cities in Japan" have subcategories such as "Cities in Aichi Prefecture", "Cities in Aomori Prefecture", and so on. categories that share such common word sequence (substring) are mostly diffused categories of the parent category. Based on this understandings, we expand the list using this relationships by following procedures.

(a) Selection of a parent category for each category in the candidate pair. (e.g., select "Cities in Japan" from "Cities in Aomori Prefecture").

(b) Make a list of subcategories that shares word sequence (substring) from the start or the end.
For the language that uses space for the word separator (other than Japanese and Chinese), we split the word sequence by space. For Japanese and Chinese, we split the category name as a sequence of character. We check the number of categories who share sequence from the start or the end. For example, from parent category "Cities in Japan", we can extract sub-sequence such as "Cities" and "Cities in" from the start and "Japan", "in Japan", and "Prefecture" from the end. In order to select the meaningful sub-sequence for category classification, we set minimum length for the sequence as $minlength$. In addition, since we would like to select most common diffused category from the relationships, we select categories that satisfy following conditions.

$$\frac{|subcategoriesthatsharesubstring|}{|allsubcategories|} \leq minratio \tag{3}$$

We use $minlength = 2$ and $minratio = 0.7$ in this experiment. Using this parameter, substring such as "Cities", "Japan", and "Prefecture" are excluded by $minlength$ and "in Japan" is excluded by $minratio$. "Cities in" is selected for candidate of the expansion.

(c) Calculate precision using expanded categories.
For this case, the precision for each expanded category is calculated by using the set of page P_{wcat_g} defined by equation 4 for the generalized Wikipedia category $wcat_g$ instead of P_{wcat} in equation 1.

$$P_{wcat_g} = \bigcup_{s_i \in S_{g,substring}} P_{wcat_{s_i}} \tag{4}$$

where $S_{g,substring}$ is a set of subcategories of Wikipedia category g and share $substring$. If $prec_{wcat_g,scat}$ is greater or equal to $precmin$, we select all pairs of $(wcat_{s_i}, scat)$ for the candidate pairs. For the value of $prec(wcat_{s_i}, scat)$, we use $prec_{wcat_g,scat}$ for the *expand-based score*, if $(wcat_{s_i}, scat)$ is not selected as a candidate for the first step.

(d) Check ancestor categories
When $wcat_g$ is selected for the expansion, we use $substring$ to expand category for the parent category of $wcat_g$ recursively. Back to the step 3b for expansion.

We also make lists of candidate pairs using simplified version of the expansion used Wikipedia category hierarchy.

(1) Expansion of candidate pairs by using bigrams.
From the result of expansion used for the expansion of candidate pairs by using Wikipedia category hierarchy, there are many common $substring$ used for expansion for a particular SHINRA category (e.g., "Cities in" for 1.5.1.1 (city)). In addition there are several categories that contains number for representing years. In order to normalize such patterns we replace the terms that contains numeric character (0-9) only for "¡num¿" and sequence of numeric character (0-9) +

"s" for "¡num¿s". For this case, the precision for the categories that contain particular word (character) bigram is calculated by using the set of page $P_{wcat_{bigram}}$ instead of P_{wcat} in equation 1.

$$P_{wcat_{bigram}} = \bigcup_{s_i \in S_{bigram}} P_{wcat_{s_i}} \quad (5)$$

where S_{bigram} is a set of Wikipedia categories that have *bigram* in its text. If $prec_{wcat_{bigram},scat}$ is greater or equal to *precmin* and $|P_{wcat_{bigram}} \cup P_{scat}|$ is greater or equal to $minfreq$ for the candidate pairs. After making the matching list for the bigrams, we check the consistency of estimated SHINRA category for the all Wikipedia categories. When there are two or more SHINRA categories are assigned for one Wikipedia category, this category is excluded from the matching list. After excluding such list we use highest value for the corresponding $prec_{wcat_{bigram},scat}$ as $prec(wcat_{s_i}, scat)$ for the *bigram-based matching list score*.

(1) Expansion of candidate pairs by set (English only).
As we analyzed in the Wikipedia category ontology, there are many cases that constrained set "Geography of Tokyo" have set related category "Geography by city" and constraint (topic) related category (Tokyo) as parent categories. Since "by ..." is a style for showing a criteria for the diffusion in Wikipedia category, this category has "Geography" (category without a criteria for diffusion) as an ancestor category. By using this information, we can split the data using these categories. First example is using "Geography" ("Geography" + "of Tokyo") and the other case is using "Tokyo" ("Geography of" + " Tokyo"). We check all categories for making such splitting data. Based on the comparison between the split data with Wikipedia category ontology for Japanese Wikipedia data [2], most of the terms for representing set appears at the beginning with preposition (e.g., "in", "from", ...) or at the end that ends "s" for representing the plural. By using this patterns, we extract terms for representing set and also make a list of categories that contains set at the beginning with preposition or at the end S_{set}. The precision for each set is calculated by using the set of page $P_{wcat_{set}}$ instead of P_{wcat} in equation 1.

$$P_{wcat_{set}} = \bigcup_{s_i \in S_{set}} P_{wcat_{s_i}} \quad (6)$$

We use $prec_{wcat_{set},scat}$ for the *set-based score*.

(2) Expansion of candidate pairs by substring (Other language).
Based on the analysis of the English Wikipedia category analysis data, we suppose common keywords from the start or the end may represent set keywords for the target language. We make a common keyword list from start and end. For each keyword, we make a set of categories that start or end with given keywords $S_{substring}$. The precision for each substring is calculated by using the set of page $P_{wcat_{set}}$ instead of P_{wcat} in equation 1.

$$P_{wcat_{set}} = \bigcup_{s_i \in S_{substring}} P_{wcat_{s_i}} \quad (7)$$

We use $prec_{wcat_{substring},scat}$ for the *substring-based score*.

Another approach is using language links. Wikipedia categories of the target language that have language links from Japanese or English Wikipedia categories for the candidate pairs are also used for the candidate pairs. $prec_{wcat,scat}$ of original language (Japanese or English) is used for $prec_{wcat,scat}$ of the *language-links-based matching score*.

2.2 Estimation of the SHINRA category using Matching lists

Estimation of the SHINRA category for the Wikipedia pages are conducted using matching lists. General idea for the category estimation is calculating score using $prec_{wcat,scat}$ for each SHINRA category of all Wikipedia categories of the page using matching lists. All scores are summed up for each SHINRA category and category with highest score is selected as estimated category. If there is no appropriate SHINRA category for the page (the page does not have Wikipedia category listed in the matching list), it was selected as "0" (CONCEPT).

Followings are details of the algorithms.[3]

There are five types of matching list with different scores (original, expand, generalize, set, bigram, and language-links).

(1) Estimation of score for each Wikipedia category of the target page.

(a) Calculation of the score using *unique categories*.
Since there are categories that are useful to identify the SHINRA category without considering other Wikipedia categories (e.g., "Japanese footballers" for 1.1 ("Person")), we treat category pairs whose $prec_{wcat,scat}$ are larger than 0.98 as *unique categories* . Because of the usefulness of such categories, we calculate scores using *uniqueratio* as a parameter to represent its importance (*uniqueratio* × $prec_{wcat,scat}$). However, There are several bigram and substrings for representing topic related keywords that may not be good for estimating the category with high $prec_{wcat,scat}$, we exclude category pairs generated by bigram and substring approach for selecting *unique categories* .[4]

(b) Calculation of the score for other pairs.
For other cases, $prec_{wcat,scat}$ is used for the score for the category. However the quality of the simple substring based method is not so good compared with other ones. We discount substring-based matching score with *subratio*. In addition, since we would like

[3]Due to the bugs of the submitted code, weighting schema of the submitted results are not same as the one used for submission. However, based on the analysis using Leaderboard, there is no significant difference between the results (it is slightly better than submitted one).
[4]Submitted version selects *unique categories* from bigram instead of others.

to normalize the score from one Wikipedia category, we divide $prec_{wcat,scat}$ with number of matching list (e.g., original, bi-gram, ...) for the target category. For the substring-based score, $subratio \times prec_{wcat,scat}$ is used for the score.

(2) Selection of SHINRA category.
After calculating the score for each Wikipedia category and sum of the scores for all Wikipedia categories, SHINRA category with highest score is selected for the candidate category. If there is no corresponding category for the page, "0" (Concept) is selected for the candidates.

3 EVALUATION OF THE SUBMITTED RESULTS

Since pages with two or more SHINRA categories may have Wikipedia categories that correspond for each SHINRA category and it is difficult to decide the corresponding SHINRA category for each Wikipedia category, we exclude such page from the training examples.

After excluding the training data, we select Wikipedia category information for each Wikipedia page from the cirrus dump. From the Wikipedia category, category belongs to hidden categories are excluded because those categories are mostly used for Wikipedia maintenance purpose.

However for el (Greek), there is no cirrus dump provided from the organizers, we add Wikipedia category information from the recent Wikipedia dump (downloaded at 2020-08-01 version) and add category using "pageid" as a reference. However, due to the modification of the pages after the data of the official runs, there are many cases that the system fails to make a list of Wikipedia categories for such pages.

We submit following three runs.

HUKB1 uses all options (original, generalize, set, bigram and language-links).

HUKB2 uses original, generalize, set, bigram (except language-links).

HUKB3 uses original, generalize, set, language-links (except bigram).

Followings are evaluation results of the submitted runs.

From this results, evaluation of HUKB-1 and HUKB-2 are almost similar, but those results are significantly better than HUKB-3.

This big difference suggests that categories pair made by bigram approach contributes a lot to select appropriate SHINRA categories. We also confirmed that bigram approach works well mostly in all languages except zh (Chinese). Since there is no clear word boundary in Chinese, we use character bigram instead of word bigram. It may be better to utilize different methods for such texts.

Worst performance of the el (Greek) may comes from the lack of information about the Wikipedia category information.

Based on the analysis using the training data for evaluation, we confirmed that original approach may works well. However, this approach did not work well for the outside data. One of the reason for the problem is the limited variation of Wikipedia categories exist in the training data. Since outside data contains large number of pages that do not have links from Japanese Wikipedia pages, we suppose such outside data do not share many Wikipedia categories in the training data.

Table 1: Evaluations results of the submitted runs for 30 languages

language	HUKB-1	HUKB-2	HUKB-3
ar	30.98311817	30.98311817	13.50546177
bg	61.05577689	60.85657371	28.0876494
ca	42.5384234	42.34010907	16.26177491
cs	52.60545906	52.60545906	18.85856079
da	49.00793651	49.00793651	13.98809524
de	53.81961557	53.72104485	26.81123706
el	7.509881423	7.509881423	7.509881423
en	45.10680576	45.10680576	11.92250373
es	49.10891089	49.20792079	19.5049505
fa	45.58969277	45.58969277	15.65906838
fi	53.44571145	53.14823996	17.05503223
fr	43.73757455	43.83697813	11.23260437
he	60.04962779	59.95037221	15.78163772
hi	39.50617284	39.7037037	22.02469136
hu	69.44444444	69.1468254	26.09126984
id	44.46650124	44.06947891	16.27791563
it	45.55335968	45.55335968	12.05533597
ko	63.97608371	63.67713004	13.9511709
nl	42.45423058	42.35526967	17.12023751
no	34.5752608	34.5752608	11.32637854
pl	63.51418616	62.71777003	32.55350921
pt	42.62295082	42.32488823	16.09538003
ro	57.69612711	57.59682224	28.50049652
ru	42.24095191	42.04263758	11.30391671
sv	50.6215813	50.32322228	21.97911487
th	40.37605146	39.98020782	24.04750124
tr	62.47524752	61.88118812	16.73267327
uk	60.18839861	60.28755578	22.50867625
vi	60.47666336	60.37735849	22.1449851
zh	21.42152024	21.22408687	17.5715696

In this system, when there is no corresponding category for the page, "0" (Concept) is selected for the candidates. When there are limited numbers of candidates categories for estimating the category, number of page selected as "0" (Concept) increases. Table 2 shows ratio of pages selected as concepts by default selection rule (no appropriate categories for estimating the class). As we can see from the big difference between HUKB-3 and others, candidate pairs based on the bigram are widely used compared with others pairs. In addition, for the language whose number of the page that can not be estimated without bigram is small, the evaluation results of such language is comparatively smaller than others. For example, the results for "ar" (0.000), "fr" (0.002), "el"(0.015), "en"(0.016)) are smaller than 0.5. In addition,

Table 2: Number of pages that are selected as concepts by default selection for 30 languages

language	HUKB-1	HUKB-2	HUKB-3
ar	0.093	0.093	1.000
bg	0.285	0.288	0.783
ca	0.231	0.234	0.904
cs	0.302	0.305	0.904
da	0.354	0.360	0.934
de	0.414	0.418	0.841
el	0.985	0.986	0.985
en	0.209	0.209	0.984
es	0.221	0.224	0.892
fa	0.162	0.162	0.974
fi	0.252	0.255	0.832
fr	0.161	0.161	0.998
he	0.226	0.229	0.941
hi	0.518	0.524	0.865
hu	0.173	0.181	0.804
id	0.127	0.130	0.944
it	0.232	0.234	0.904
ko	0.233	0.235	0.914
nl	0.682	0.685	0.933
no	0.152	0.154	0.973
pl	0.251	0.257	0.738
pt	0.299	0.301	0.925
ro	0.151	0.153	0.738
ru	0.055	0.056	0.955
sv	0.079	0.080	0.961
th	0.626	0.636	0.878
tr	0.261	0.265	0.885
uk	0.245	0.247	0.840
vi	0.548	0.548	0.898
zh	0.876	0.880	0.919

when the value for HUKB-1 is large (e.g., "de" (0.414), "el" (0.985), "hi" (0.518), "nl" (0.682), "th" (0.626), "vi" (0.548), "zh" (0.876)), evaluation results are smaller than 0.5 except "vi". For those cases, it is necessary to find out appropriate methods to increase the number of candidate pair for the estimation. For the "vi" case, precision of the estimated results using candidate pairs may be very good. If we can expand the candidate pairs for default estimation case, there is a good room to improve the performance of the "vi" results.

4 DISCUSSION

Since our system uses only Wikipedia category information, our system does not works well compared to the system using text contents. However, the failure analysis of the training data suggests that there are several inconsistency found in the training data.

One of the example is "Meridians (geography)". Most of the pages belongs to the category are classified as "3.11" (Latitude_Longitude), but 41 pages are categorized as "1.5.0"(Location Other). The other example is inconsistency between "1.1" (Person) for solo artists and "1.4.2" (Show organization) for a group of artist. For example "Icona Pop" and "The Peanuts" are vocal groups but categorized as "1.1". Those categories are classified by automatic approach. These inconsistency may affect our system to make appropriate pairs and it is better to take into account the quality of automatic data. We plan to conduct such analysis for understanding the characteristics of our approach.

5 CONCLUSIONS

In this paper, we introduce our approach to estimate SHINRA categories for the Wikipedia pages using Wikipedia category. Even though our system does not perform well compared to the approach using text contents, our system may have a characteristics to check the quality of the data. We would like to discuss this issue for the future works.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Number 18H03338.

REFERENCES

[1] Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. 2020. Overview of SHINRA2020-ML Task. In *Proceedings of the 15th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Quesiton Answering, And Cross-Lingual Information Access.*

[2] Masaharu Yoshioka, Myungha Jang, James Allan, and Noriko Kando. 2020. Wikipedia Category Ontology: A Framework for Utilization of the Wikipedia Category Structure by Knowledge Engineers. In *Proceedings of the ISWC 2020 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas) co-located with 19th International Semantic Web Conference (ISWC 2020).*