Multi-class Multilingual Classification of Wikipedia Articles Using Extended Named Entity Tag Set

Hassan S. Shavarani[*]
School of Computing Science
Simon Fraser University, BC, Canada
sshavara@sfu.ca

Satoshi Sekine
AIP Center for Advanced Intelligence
Riken, Tokyo, Japan
satoshi.sekine@riken.jp

Abstract

Wikipedia is a great source of general world knowledge which can guide NLP models better understand their motivation to make predictions. Structuring Wikipedia is the initial step towards this goal which can facilitate fine-grain classification of articles. In this work, we introduce the *Shinra 5-Language Categorization Dataset* (SHINRA-5LDS), a large multi-lingual and multi-labeled set of annotated Wikipedia articles in Japanese, English, French, German, and Farsi using *Extended Named Entity* (ENE) tag set. We evaluate the dataset using the best models provided for ENE label set classification and show that the currently available classification models struggle with large datasets using fine-grained tag sets.

1. Introduction

Major progress has been made in different tasks in Natural Language Processing, yet our models are still not able to describe why they make their decisions when summarizing an article, translating a sentence, or answering a question. Lack of meta information (e.g. general world knowledge regarding the task) is one important obstacle in the construction of language understanding models capable of reasoning about their considerations when making decisions (predictions).

Wikipedia is a great resource of world knowledge for human beings, but lacks the proper structure to be useful for the models. To address this issue and make a more structured knowledge-base, Sekine et al. (2018b) try to structure Wikipedia. Their final goal is to have, for each Wikipedia article, known entities and sets of attributes, with each attribute linking to other entities wherever possible. The initial step towards this goal would be to classify the entities into predefined categories and verify the results using human annotators[1].

Throughout the past years, many have tried classifying Wikipedia articles into different category sets mostly containing between 3 to 15 class types (Toral and Munoz, 2006; Watanabe et al., 2007; Dakka and Cucerzan, 2008; Chang et al., 2009; Tardif et al., 2009). Such categorization type sets are not much helpful when the classified articles are being used as the training data for question answering systems, since the extracted knowledge-base does not provide detailed enough information to the model.

On the other hand, much larger categorization type sets such as Cyc-Taxonomy (Lenat, 1995), Yago-Taxonomy (Suchanek et al., 2007), or Wikipedia's own taxonomy of categories (Schönhofen, 2009) are not suitable for classifying Wikipedia articles since the tags are not verifiable for annotators[2]. In addition, taxonomies are not designed in a tree format, so some categories might have multiple super-categories and this would make the verification process much harder for articles discussing multiple topics.

Considering the mentioned problem requirements, we believe *Extended Named Entities Hierarchy* (Sekine et al., 2002), containing 200 fine-grained categories tailored for Wikipedia articles, is the best fitting tag set.

Higashinaka et al. (2012) were the first to use this extended tag set as output labels when categorizing Wikipedia pages. Their model was trained using a hand-extracted feature set that converted the pages into model compatible input vectors. Following their work, Suzuki et al. (2016) augmented the extracted input features with trained vectors modelling the links between different Wikipedia pages. They proposed a more complex model for learning the mapping between the converted articles and the labels. Although providing useful insights, neither have considered exploring the multi-lingual nature of many Wikipedia articles.

---

[*] The author was an intern at AIP Center for Advanced Intelligence, during this project.

[1] Please note that the verification process plays an important role in the knowledge-base construction process since it leads to what is represented to our models as world facts.

[2] They need to keep 200K+ classes in mind to find the most suitable ones for the article at hand or verify the classifier category prediction for it.

language	average size in folds				total classes	average count		max annotations
	train	dev	test	total		article/class	annot./article	
ja	96,321.8	12,004.9	12,006.3	120,333	141	853.426	1.0359	5
en	42,652.8	5,301.1	5,301.1	53,228	127	419.331	1.0359	5
fr	27,750.5	3,425.7	3,424.8	34,601	113	306.204	1.0347	5
de	23,969.8	2,958.8	2,959.4	29,888	108	276.741	1.0309	5
fa	11,329.4	1,388	1,386.6	14,104	80	176.3	1.0342	5

Table 1: Statistics about *Shinra 5-Language Categorization Dataset* as well as the suggested average train/dev/test size of the data sectors used in the benchmark experiments.

We base this work on Sekine et al. (2018a)'s work in which they have hired linguists as annotators and educated them on the Extended Named Entities (ENE) tag set to annotate each article with up to 6 different ENE classes. We exploit the Wikipedia language links in the annotated articles to create our multi-lingual Wikipedia classification dataset. Section 2 details our dataset creation process.

We then use the models suggested by Higashinaka et al. (2012) and Suzuki et al. (2016), the only works close enough to our task at hand (to the best of our knowledge), to benchmark the created dataset. Section 3 provides more details about our multi-lingual feature selection method and the models. Section 4 presents our experimental setup and the classification results.

2. Dataset Collection and Annotation

Recently, Sekine et al. (2018a) created an annotated dataset containing 782,517 Japanese Wikipedia articles in different areas, covering 175 out of 200 ENE labels[3]. The articles are selected from Japanese Wikipedia with the condition of being hyperlinked at least 5 times from other articles in Wikipedia. They had instructed annotators[4] to label the collected articles using at most 6 labels[5] from the 200 suggested ENE labels[6].

We considered a subset of the annotated articles which have been hyperlinked at least 100 times (as Suzuki et al. (2016) suggest) that led to a 120,333 Japanese

Wikipedia articles (annotated with 141 out of 200 ENE labels and maximum 5 annotations per article). We collected the content of the same article titles in English, French, German, and Farsi Wikipedia sections[7], relying only on Wikipedia language links. Language links connect the articles representing the exact same topic from one language to another. We used the labels assigned to Japanese version of the articles to all the articles in other four languages (in case any existed), since ENEs are language agnostic and the pages offered the same content.

To perform the language link exploration, we first created the graph of language links for all the ("wikipedia id", "language") pairs linking one article in one of the five languages to another article in another language. We also took into account the Wikipedia redirect links in our exploration process, since sometimes language links connect articles to redirect pages in other languages. Using the language links graph, we formed "Entities" grouping all different ("wikipedia id", "language") pairs representing the same subject and then applied the ENE labels to the articles in different languages.

We call this multi-lingual multi-labeled collection of Wikipedia articles, the "*Shinra 5-Language Categorization Dataset*" (SHINRA-5LDS)[8], and we release the dataset alongside this paper to enable the other researchers to perform the benchmark on multi-labeled Japanese, English, French, German, and Farsi Wikipedia categorization using their suggested methods. Table 1. contains the total number of annotated articles in each of the languages as well as the total number of ENE classes with at lease one article annotated in that class, the average number of articles collected in each class, and the average number of annotations assigned to each article by the annotators.

---

[3] The rest of the categories were not covered since they did not find any articles under the category which could meet the selection criteria at the time.

[4] Majority of the annotators were post secondary degree holders in linguistics.

[5] They report no inter annotator agreement data, but report that 200 samples from the data have been randomly selected and passed to skilled annotators to validate/verify the quality of the annotations.

[6] The data is provided and maintained for SHINRA2020-ML classification task. The latest version of it is available via http://shinra-project.info/shinra2020ml/

[7] The wikidump data used for extracting the articles' content was the May 20, 2018 snapshot of Wikipedia in all five languages.

[8] The data (available at shinra-project.info/download/) will only contain the ("wikipedia id", "language") pairs and can be combined with the actual articles (in wiki-dumps) using wikipedia_id references.

3. Feature Selection and Models

To perform the benchmark, we surveyed the available suggested models for multi-class categorization of Wikipedia articles and selected the models suggested by Higashinaka et al. (2012) and Suzuki et al. (2016), since both have suggested classifying Wikipedia articles using ENEs. We also decided to study the usefulness of the hierarchy in the process of training the classifiers using ENEs. Hence, we also selected the models suggested by Wehrmann et al. (2018) as our third set of models. The following sections describe our feature selection procedure and briefly explain each of the models.

3.1. Feature Selection

A fair comparison between the models on the dataset is not possible unless we can guarantee the same input to each of them. With that in mind, we went through the feature selection methods suggested in (Wang and Manning, 2012), (Higashinaka et al., 2012) and (Suzuki et al., 2016) and created a union of their suggestions.

However, we had to remove some of the features such as '*Last one/two/three characters in the headings or titles*" or "*Last character type (Hiragana/Katakana/Kanji/Other)*" from the union due to the multi-lingual nature of our task.

Figure 1 summarizes the final unified schema for categorization of the Wikipedia articles in SHINRA-5LDS.

3.2. Binary Logistic Regression

Higashinaka et al. (2012) suggested learning a set of separate *Binary Logistic Regression Classifier Models* to learn the contribution of the extracted features towards the final selected class. We employ this model to indicate the classification difficulty level of our dataset using a simple model.

3.3. Joint-NN and Joint-NN++

Suzuki et al. (2016) suggested that combining all the separate Logistic Regression Classifier Models into a *2-Layer Perceptron Neural Network* may result in capturing more information for better confidence in assigning ENE classes to the articles. They call their suggested model *Joint-NN* and conclude that their model is better in learning the correlation of the extracted features with the output ENE labels than a separate set of logistic regression models or even a separate set of 2-Layer Perceptron Networks each of which trying to predict one of the labels. We employ their suggested *Joint-NN* model and also try augmenting it with another additional layer (we call the augmented model *Joint-NN++*) in our benchmark experiments.

Content-Based Features

- token uni/bigrams; char uni/bigrams; and token part-of-speech uni/bigrams of the title
- token uni/bigrams of the first sentence
- token uni/bigrams of the category titles
- token unigrams of the wiki-link anchors
- token unigrams of the titles of outgoing linked wiki-pages
- token unigrams of the heading lines
- "_" merged template name tokens concatenated with each key name in the template
- last token part-of-speech tagged as noun in the title / the first sentence

Article Vector Features

- D dimensional dense vector embedding of the wiki-links representing each article in other wikipedia pages; created with Word2Vec skip-gram model exactly as mentioned in (Suzuki et al., 2016)

Figure 1: Features extracted from each article

3.4. Hierarchical Multi-Label Classification Networks

To examine the extent of information lying in the Hierarchy of ENEs, we propose using *Hierarchical Multi-Label Classification Networks (HMCN)*. Wehrmann et al. (2018) suggest two different settings for the HMCNs both of which perform the prediction of the label hierarchy in a top-down manner. The first setting, *HMCN Feed-forward (HMCN-F)*, uses a separate explicit part of the network for predicting each level of the hierarchy. On the other hand, *HMCN Recurrent (HMCN-R)* learns of the hierarchy by recurrently feeding the prediction of the previous top layer to the next lower level in hierarchy. We suggest to employ *HMCN-R* in addition to *HMCN-F* to examine the effect of model compression on learning to predict the hierarchy of ENEs at test time.

3.5. Training and Evaluation

To preform the multi-label classification, we suggest passing all the model predicted membership distributions through a Sigmoid layer and assign the label to the article if the predicted probability after passing

Model	dev					test				
	ja	en	de	fr	fa	ja	en	de	fr	fa
Binary Logistic Regression	71.25	76.24	69.56	69.74	79.70	71.18	72.69	69.27	65.83	66.45
Joint-NN †	**80.19**	78.43	**81.58**	81.23	79.71	77.31	78.18	**81.41**	78.85	76.34
Joint-NN++	77.73	**81.13**	79.88	**83.53**	**85.25**	**77.40**	**80.80**	79.88	**83.43**	**79.78**
HMCNF	72.07	73.59	71.43	73.54	76.07	71.25	73.31	69.71	70.22	75.83
HMCNR	61.63	64.28	64.66	64.80	70.45	61.38	63.04	61.70	64.65	70.20

Table 2: The classification accuracy of the predicted labels. Partially correct labels have also contributed partially to the scores.

† Despite our endeavor to keep the settings comparable to the original model, comparison between our results and theirs would not be fair, since the size of datasets used in our experiments and also the number of classes are different than theirs.

through Sigmoid is above 0.5.

The evaluation measure would then be the micro-averaged precision (Sorower, 2010) of the predicted labels. In addition, to prevent the domination of more frequent classes on the training procedure, we suggest weighted gradient back-propagation. The back-propagation weight of each article would be calculated using $w = \frac{N}{\sum_{n=1}^{N} f(l_n)}$ where N is the number of labels assigned to the article (with a maximum of 6) and $f(l_n)$ counts the total train-set articles to which label l_n has been assigned. The loss function used for training all the models has been *Binary Cross Entropy Loss* averaged over all the possible classes.

4. Experiments and Results

We implemented all the models suggested in §3. using the PyTorch framework[9]. For part-of-speech tagging the title and first sentences of the articles mentioned in the feature selection schema (Figure 1) and also normalization and tokenization of the articles, we used Hazm Toolkit[10] for Farsi, Mecab Toolkit (Kudo, 2006) for Japanese, and TreeTagger Toolkit[11] for English, French, and German.

In all of our experiments, we have used Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e^{-3}$ and have performed gradient clipping (Pascanu et al., 2013) of 5.0. We have initialized all of the network parameters with random values between $(-0.1, 0.1)$. We have done training on mini-batches of size 32, and to have a fair comparison, all the experiments have been conducted with 30,000 steps (batches) of randomly shuffled training instances to train the model parameters. The hidden layer size of

all the models in each layer has also been set to 384[12].

We have performed the evaluation in a 10 fold cross validation manner in each fold of which 80% of the data has been used for training, 10% for validation and model selection, and 10% for testing. In addition, classes with a frequency less than 20 in the dataset have been ignored in the train/test procedure.

Table 2 depicts the benchmarked micro-averaged precision of classification prediction of the articles in our dataset. The results initially demonstrate that the dataset is not an easy one as the Binary Logistic Regression model is not achieving very high accuracy scores. Besides, the lower scores for Japanese in comparison to the other languages demonstrate the higher difficulty level of classification for a larger category set size for all the models.

On the other hand, the consistency of the superior results of non-hierarchical models to the hierarchical models shows that the leaf-node ENEs contain all the necessary information to perform the classification over them, and the hierarchy may only add more confusion to model decisions.

Last but not least, the overall precision scores depict that the currently available models struggle with larger more complex annotated sets of Wikipedia articles.

In our future studies, we will focus on providing more complex models which can capture more information from the articles (leading to better classification scores) and we will also focus on using the results of our classifier to create a bigger structured knowledge-base to augment the currently available NLP models.

---

[9]https://pytorch.org - v0.4.1
[10]https://github.com/sobhe/hazm
[11]https://github.com/miotto/treetagger-python

[12]We have also tried larger sizes of hidden layers for simpler models but the results did not vary much, so we removed the probability of difference in learning capability of the models in different parameter set sizes from our experiment result analysis.

5. Bibliographical References

Chang, J., Tsai, R. T.-H., and Chang, J. S. (2009). Wikisense: Supersense tagging of wikipedia named entities based wordnet. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, volume 1.

Dakka, W. and Cucerzan, S. (2008). Augmenting wikipedia with named entity tags. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Higashinaka, R., Sadamitsu, K., Saito, K., Makino, T., and Matsuo, Y. (2012). Creating an extended named entity dictionary from Wikipedia. In *Proceedings of COLING 2012*, pages 1163–1178, Mumbai, India, December. The COLING 2012 Organizing Committee.

Kingma, D. P. and Ba, L. (2015). Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

Kudo, T. (2006). Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318.

Schönhofen, P. (2009). Identifying document topics using the wikipedia category network. *Web Intelligence and Agent Systems: An International Journal*, 7(2):195–207.

Sekine, S., Sudo, K., and Nobata, C. (2002). Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Sekine, S., Ando, M., Kobayashi, A., Matsuda, K., Suzuki, M., Nguyen, D., and Inui, K. (2018a). Wikipedia categorization data based on extended named entity (in japanese). *The 24th Annual conference of Association for Natural Language Processing, Japan*, pages 504–507.

Sekine, S., Kobayashi, A., and Nakayama, K. (2018b). Shinra: Structuring wikipedia by collaborative contribution. *Automated Knowledge Base Construction*.

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A Core of Semantic Knowledge. In *16th International Conference on the World Wide Web*, pages 697–706.

Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., and Inui, K. (2016). Neural joint learning for classifying wikipedia articles into fine-grained named entity types. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, pages 535–544, Seoul, South Korea, October.

Tardif, S., Curran, J. R., and Murphy, T. (2009). Improved text categorisation for wikipedia named entities. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 104–108.

Toral, A. and Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics.

Watanabe, Y., Asahara, M., and Matsumoto, Y. (2007). A graph-based approach to named entity categorization in wikipedia using conditional random fields. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Wehrmann, J., Cerri, R., and Barros, R. (2018). Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5225–5234.