A Simple Text-based Relevant Location Prediction Method using Knowledge Base

Mei Sasaki **Shumpei Okura** **Shingo Ono**
{mesasaki, sokura, shiono}@yahoo-corp.jp
Yahoo Japan Corporation
Tokyo, Japan

Abstract

In this paper, we propose a simple method to predict salient locations from news article text using a knowledge base (KB). The proposed method uses a dictionary of locations created from the KB to identify occurrences of locations in the text and uses the hierarchical information between entities in the KB for assigning appropriate saliency scores to regions. It allows prediction at arbitrary region units and has only a few hyperparameters that need to be tuned. We show using manually annotated news articles that the proposed method improves the f-measure by > 0.12 compared to multiple baselines.

1 Introduction

Predicting relevant locations from news articles can result in numerous useful applications. For example, it enables the delivery of news related to a specific city that is of user interest, or facilitates the prediction of a disease outbreak in a specific region when used with event detection techniques.

In this paper, we focus on predicting relevant locations from news articles. The goal of this task is to identify locations that are salient to the article, not those that simply appeared in the article. For example, consider the following excerpt: "The Aoi Festival is one of the three major festivals in Kyoto. It originated as a series of rites to calm down angry gods. A visitor from Australia said..." In this example, the phrase "Kyoto" is highly relevant to the article, but "Australia" is not.

Traditional methods to predict locations require specific data, such as a training dataset or phrase distribution, that match the application domain and the granularity of the prediction. However, it is costly to prepare such data for individual applications.

We propose a knowledge base (KB)-based method that only requires a general-purpose KB instead of a labeled dataset for training. It propagates phrase-level importance to region entities following their relationship in the KB. It can theoretically be applied to predictions at an arbitrary level of granularity (e.g. countries, prefectures, cities) without dedicated training data.

In this study, we focus on Japanese news articles and report the performance of predictions at the Japanese prefecture-level. We provide practical tips to tackle un-tokenized language like Japanese.

2 Related Works

Depending on the objective, geolocation prediction tasks from texts are roughly divided into two types. One type is for detecting and identifying mentions of points-of-interest (POIs) in the text, well known by entity linking (Nadeau and Sekine, 2007; Shen et al., 2015). This task focuses on extracting all mentions regardless of their saliency. The other type is for estimating the author's current location or home town from his or her posts (Huang and Carley, 2019). It is mainly performed to complement user profiles in services that deal with user-generated content (e.g. SNS). In this case, in addition to text, various user metadata such as the relationship between authors is available (Backstrom et al., 2010). Our work is similar to the former type in terms of purpose, but we focus on identifying salient regions rather than extracting all of the individual POIs.

There are two main approaches to predicting location. One is the dictionary-based approach (Berggren et al., 2015; Han et al., 2012; Li et al., 2014), where dictionaries of location indicative words are created in advance and used for the prediction. In addition to explicit region names, the choice of which words to add to the dictionary is a hot topic of discussion (Han et al., 2014). The other is the machine learning(ML)-based ap-

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 116–121

proach (Zhou and Luo, 2012; Miyazaki et al., 2018). Methods based on this approach usually perform well if sufficient training data is available. However, in practice, it is difficult to prepare data whose granularity matches the requirements of the application. In particular, estimating regions that are rarely found in the training data is one of the weaknesses of machine learning.

3 Task Setting and Baseline

3.1 Task

Let A be the set of articles and R be the set of candidate regions, e.g., Japanese prefectures. Our goal is to construct a function $\mathcal{F} : A \rightarrow \mathfrak{P}(R)$ such that $r \in \mathcal{F}(a)$ if and only if there are mentions of region r in the article a and region r is salient to the text, where $\mathfrak{P}(R)$ denotes the power set of R. Note that even when there are mentions of r in a, if r is not a main topic in a, $\mathcal{F}(a)$ does not contain r. Similarly, when a is not a location-aware article, then $\mathcal{F}(a)$ is $\emptyset$. In this paper, we assume A to be a set of Japanese news articles and R to be a set of Japanese prefectures.

3.2 Gazetteer baseline

The baseline method we adopted is similar to the baseline used in (Berggren et al., 2015) and consists of the following three steps:

1) Create a gazetteer from an external data source. 2) Identify the strings contained in the gazetteer from the given text. 3) Aggregate the results and return the relevant locations.

In practice, there are several choices regarding step 3). For example, we could return all the regions mentioned in the text, return the region mentioned most frequently in the text, or return only the region that appears earliest in the text. We decided to return locations that appear in the first 20% of the text.

4 KB-based Methods

4.1 Knowledge base

A KB consists of information about entities expressed in a structured, machine-readable graph format. YAGO and DBpedia are examples of KBs (Hoffart et al., 2011; Lehmann et al., 2015). Entities are assigned class(es) that represent what kind of entity they are. Examples of possible entity classes include person, place, and company. Mount Fuji is an example of a place entity.

A KB can be regarded as an edge-labeled directed graph. Entities correspond to nodes in the graph, and the relationships between entities correspond to edges in the graph. These relationships are given relation-type labels called *predicates*.

We focus on the subgraph of KB that is useful in terms of location prediction. Let us reduce the graph by keeping only *Place* class entities, and vertexes connected to such entities with inclusive relations such as *containedBy* predicates and notation relations such as *name, alsoKnownAs* predicates, and name the resulting graph $G = (V, E)$. The notation relations in the graph will be used in §4.3 to create a gazetteer, while the inclusive relations will be used in §4.5 and §4.6 to determine the set of corresponding entities for each mention and calculate the corresponding score for different candidate regions, respectively.

4.2 Overview of the proposed method

The overview of our proposed method is shown in Fig. 1. It consists of four steps, three of which correspond to those in §3.2 and the rest is the entity linking step performed between 2) and 3). As shown in the following sections, we add the efficient use of KB information at each step.

4.3 Create gazetteer

Create a dictionary $\mathcal{D}$ with location names as keys and corresponding entities as values using the notation relations in the KB. Note that multiple entities may have the same name, so $\mathcal{D}(m)$ is a set of entities that belong to V for each key m.

However, in practice, if we use all of the notation relations for $\mathcal{D}$, it may adversely affect the prediction. For example, "the park" can be an alias for all parks in the world, but due to some inconsistencies in KB entries, some parks have such an alias in the KB and others do not. Therefore, by using inclusive relations between entities in the KB, we systematically extract phrases that appear as the prefix/postfix of entity names in wildly distant multiple regions and create a blacklist of phrases by manually reviewing them. In addition, we manually added the names of central ministries to the blacklist, since occurrences of such names rarely indicate the locality of the news. The blacklist currently consists of 151 words.

4.4 Phrase identification

When given an article a, identify phrases that serve as clues by the following three steps:

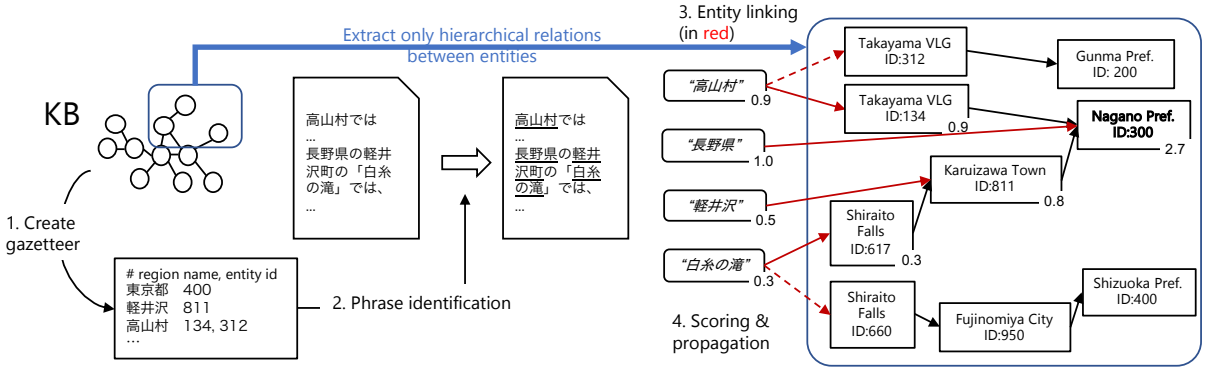

Figure 1: Overview of the proposed method.

1. Tokenize text of the article into morphemes.

2. Chunk the list of morphemes so that it results in as long and as many matches for keys in $\mathcal{D}$.

3. Perform named entity recognition (NER) and only keep phrases that at least partially overlap with named entities whose IREX[1] category is *LOCATION*, *ARTIFACT*, or *ORGANIZATION*.

The additional NER step is necessary to avoid confusion between the names of persons and places. There are many family names in Japanese that have similar characters to region names.

The reason we do not limit the phrases to those categorized as *LOCATION* is that some *ORGANIZATION* or *ARTIFACT* entities may contain region names in their names, e.g., small local businesses. The sets of phrases identified from article a_i in this step are represented by $\{m_i\}$ hereafter.

4.5 Entity linking

Entity linking is the task of mapping entity mentions to the corresponding entities in a KB. The purpose of this step is to reduce the candidate entities for m_i from $\mathcal{D}(m_i)$ using the contexts of article a. $\mathcal{D}_a(m_i)$ denotes the candidates remaining after the following steps.

1. If $|\mathcal{D}(m_i)| = 1$, we have nothing to do.

2. If $\mathcal{D}(m_i)$ contains e s.t. $\exists m_j$ in a, $\mathcal{D}(m_j) = \{e\}$, we remove other candidates for m_i.

3. If $\mathcal{D}(m_i)$ doesn't satisfy the above but contains e s.t. $\exists m_j$ in a, $\mathcal{D}(m_j) = \{e'\}$, e and e' are both contained by the same region $r \in R$, we remove other candidates for m_i.

In short, we give preference to entities when there is relevant evidence elsewhere in the article.

Unlike in traditional entity linking tasks, for our purposes, if the procedure fails to resolve the phrase to a single entity, but finds a list of candidates, it is still quite useful in terms of location prediction. As discussed later in the paper, such phrases and corresponding candidate entities will be taken properly into account in the later steps.

4.6 Scoring and propagation

In this step, we score each phrase occurrence m_i and propagate the score to corresponding entities.

First, we define phrase scores ϕ_a as:

$$\phi_a(m_i) = \frac{\text{length}(m_i)}{\log(\text{pos}(m_i) + C)},$$

where $\text{pos}(m_i)$ means the number of words that precede m_i in the article and C is a positive constant.

Next, we calculate entity scores ψ_a as:

$$\psi_a(e_i) = \sum_{(e_j, e_i) \in E} \frac{\psi_a(e_j)}{|\{(e_j, \cdot) \in E\}|} + \sum_{e_i \in \mathcal{D}_a(m_j)} \frac{\phi_a(m_j)}{|\mathcal{D}_a(m_j)|},$$

where E has a DAG structure because it is composed of inclusive relations and the calculation order is naturally determined.

Finally, return regions that satisfy certain criteria as $\mathcal{F}(a)$. There are several possibilities for the actual criteria to determine which regions to return, such as

$$\begin{aligned}
\mathcal{F}(a) &= \{r \in R | \psi_a(r) > T\} \text{ (absolute)}, \\
\mathcal{F}(a) &= \{r \in R | \frac{\psi_a(r)}{\max_{r'}(\psi_a(r'))} > \alpha\} \text{ (relative)}, \\
\mathcal{F}(a) &= \{r \in R | \text{rank}(\psi_a(r)) \leq N\} \text{ (rank)}.
\end{aligned}$$

[1] https://nlp.cs.nyu.edu/irex/NE/df990214.txt

# of articles	1,711
# of candidate prefs	47
# of salient prefs / article	1.29
articles with no salient prefs	28.4%

Table 1: Statistics on dataset.

After experiments, we decided to adopt the intersection of the above three criteria with parameters $T = 0.5, \alpha = 0.7$, and $N = 2$.

5 Experiments

5.1 Knowledge base resource

We implemented the method proposed in §4 using an in-house KB of Yahoo Japan Corporation (Yamazaki et al., 2019) and in-house morphological analysis/NER tools for the following experiments. In short, the KB consists of data integrated from various open data, data purchased from our suppliers, and information extracted from web crawling. When open data is available in multiple languages (e.g., Wikipedia), a Japanese data dump is used to construct the KB. For historical reasons, the entities that correspond to regions in the KB were little used, and there were problems regarding the quality of data in this domain. Therefore, we incorporated various official data sources containing lists of regions, regional codes, and zip codes into the KB, as the accuracy/completeness of regions and the inclusive relations between them play a crucial role in location prediction.

5.2 Dataset

Since there is no publicly available Japanese corpus of salient locations, we asked a team of professional annotators to label a total of 1,711 news articles with relevant prefectures. There are 47 prefectures in Japan. The details of this dataset are shown in Table 1. The team consists of five annotators independent of us. Although each article is labeled by one annotator, the annotation team created an annotation guideline in an iterative way as follows to ensure consistency of the annotation:

First, create a temporary annotation guideline and annotate a relatively small number of articles. Then, share the annotated results and discuss whether an annotation guideline needs to be updated. This iteration was repeated until a reasonable annotation guideline is fixed. Note that the annotation guideline was finalized before the development of the proposed method started.

Although our method enables prediction with finer granularity, we evaluate only at the prefecture-level in this first research due to the cost of annotation.

5.3 Metrics

We evaluate the prediction performance by micro-averaged precision (p_m), micro-averaged recall (r_m), and article-averaged f-measure (f_A) calculated as:

$$p_\mathrm{m} = \frac{\sum_{a \in A} |R_a \cap \mathcal{F}(a)|}{\sum_{a \in A} |\mathcal{F}(a)|}, f_A = \frac{1}{|A|} \sum_{a \in A} \frac{2 p_a r_a}{p_a + r_a},$$
$$r_\mathrm{m} = \frac{\sum_{a \in A} |R_a \cap \mathcal{F}(a)|}{\sum_{a \in A} |R_a|},$$

where R_a is the set of salient prefectures for article a in the ground truth and $p_a = |R_a \cap \mathcal{F}(a)|/|\mathcal{F}(a)|$, $r_a = |R_a \cap \mathcal{F}(a)|/|R_a|$ are article-level precision and recall, respectively. We consider $p_a = 1.0$ if $\mathcal{F}(a) = \emptyset$, and $r_a = 1.0$ if $R_a = \emptyset$. Hence, when a is not a location-aware article, the harmonic average of the two is 1.0 if and only if the method returns an empty set, and 0.0 otherwise.

5.4 Baseline methods

We adopted two different baseline methods to demonstrate the validity of the proposed method, the baseline method that relies on gazetteer described in 3.2 and ML-based method.

The second ML-based baseline method treats location prediction as a multi-label classification problem (i.e., an article can have multiple subject regions). In this setting, the classifier assigns different labels that correspond to Japanese prefectures to each article. We used fastText (Joulin et al., 2017) library for this task and tokenized the text for each article using the same in-house morphological tool described in 5.1.

5.5 Comparison with baselines

The results for the proposed and baseline methods are listed in Table 2. As shown, the proposed method outperformed the baseline methods in all performance metrics. Note that the evaluation metrics for fastText baseline are calculated in a slightly different way from other methods and are meant as approximate reference values. It was calculated by taking an average of models obtained

methods	p_{m}	r_{m}	f_A
gazetteer baseline	0.501	0.476	0.708
fastText baseline	0.660*	0.420*	0.430*
proposed	0.824	**0.515**	0.830
proposed + BL	**0.856**	**0.515**	**0.852**

Table 2: Results of proposed and baseline methods.
* The average over nested 4-fold cross-validation.

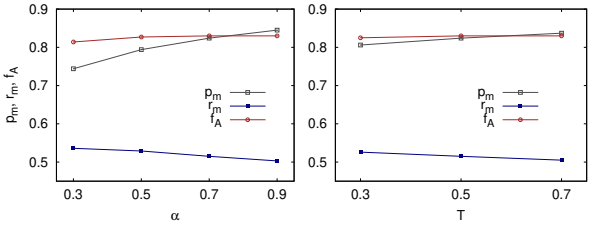

Figure 2: Effect of varying α (left) and T (right).

by nested 4-fold cross-validation over the evaluation dataset. We tuned the hyperparameters to optimize article-averaged f-measure (f_A) in each inner loop of cross-validation. For the proposed method and the gazetteer baseline that require no dataset for training, the evaluation metrics are calculated using the entire dataset.

The gazetteer baseline suffers from low precision. We give the following example to demonstrate how the proposed method's output improved over that of the gazetteer baseline. Yokohama most often represented the well-known city in Kanagawa Pref. but on rare occasions represented the small town with a similar name in Aomori Pref. The gazetteer baseline is not able to prioritize between them and returns both locations. The proposed method considered the other entity mentions to resolve Yokohama into the correct region.

The fastText baseline performs differently for different kinds of articles. While most of the articles in the evaluation dataset are labeled one or two prefectures, some articles contain phrases that collectively refer to multiple Japanese prefectures[2] and are labeled a large number of prefectures. Since such phrases have only a limited number of variations and appear in the dataset repeatedly, it is relatively easy for the ML-based approach to learn such expressions. Therefore it performs relatively well in terms of micro-averaged metrics heavily weighted to articles with a high number of relevant prefectures. However, the article-averaged metric f_A is incredibly low compared to other methods. This is because the knowledge of names of individual prefectures or cities is essential in order to make correct predictions for the rest of the articles. We found that fastText classifier often fails to predict locations for such articles even when names of prefectures are explicitly stated in the article. We conclude that it is practically impossible to learn all the necessary region names from a few thousand articles and that utilizing external resources such as the KB is the critical element in achieving good performance.

As another example, there is "Tokyo Disneyland" that is actually located in Chiba Pref., not in Tokyo Pref. It is crucial to treat it as an entity and not be confused by their apparent region names.

When we added the blacklist created in §4.3 to the proposed method, there was a huge improvement in precision. This highlights the incompleteness of the aliases in the KB and indicates that care must be taken when applying entries in KB to a real service.

5.6 Impact of hyperparameters

The hyperparameters that govern the performance of our proposed method are T, α and N (introduced in §4.6). We can see the effect of varying these hyperparameters in Fig. 2. These results demonstrate that the precision/recall tradeoff can be adjusted by varying hyperparameters.

6 Conclusion

In this paper, we presented a simple KB-based method to predict relevant locations from articles. The proposed method requires no training data or maintenance of a dictionary thanks to a freshly generated KB, and it can be used to make predictions at an arbitrary level of granularity, as long as the corresponding data is present in the KB. We demonstrated the effectiveness of this method at predicting salient Japanese prefectures using manually annotated articles. In future work, we plan to make location predictions at the city-level and evaluate its performance.

Acknowledgments

We thank our teammates for help in developing and deploying our location prediction system. We are grateful to the annotation team for providing us with an evaluation dataset. We wish to thank the anonymous referees for helpful comments.

[2]Examples include "Tōhoku region" that refers to 6 prefectures and "Western Japan" that refers to $>$ 20 prefectures.

References

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70.

Max Berggren, Jussi Karlgren, Robert Östling, and Mikael Parkvall. 2015. Inferring the location of authors from words in their texts. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 211–218.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49(1):451–500.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard De Melo, and Gerhard Weikum. 2011. Yago2: exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference Companion on World Wide Web*, pages 229–232.

Binxuan Huang and Kathleen Carley. 2019. A hierarchical location prediction neural network for twitter user geolocation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4732–4742.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Guoliang Li, Jun Hu, Jianhua Feng, and Kian-lee Tan. 2014. Effective location identification from microblogs. In *2014 IEEE 30th International Conference on Data Engineering*, pages 880–891.

Taro Miyazaki, Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Twitter geolocation using knowledge-based methods. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 7–16.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Tomoya Yamazaki, Kentaro Nishi, Takuya Makabe, Mei Sasaki, Chihiro Nishimoto, Hiroki Iwasawa, Masaki Noguchi, and Yukihiro Tagami. 2019. A scalable and plug-in based system to construct a production-level knowledge base. In *DI2KG@ KDD*.

Youjie Zhou and Jiebo Luo. 2012. Geo-location inference on news articles via multimodal plsa. In *Proceedings of the 20th ACM International Conference on Multimedia*, pages 741–744.