3C-02

Dish-Ingredient Relationship Extraction From Japanese Reviews

Weichang Chen† Katsuhiko Kaji† Kei Hiroi‡ Nobuo Kawaguchi‡
Graduate School of Engineering, Nagoya University†
Institute of Innovation for Future Society, Nagoya University‡

1 Introduction

Following name entity recognition (NER), entities relation extraction is considered as another valuable topic in text mining. It aims to find the definite semantic relation among named entities. Linked Data represent a new norm that the resource is organized in form of entities linking with semantic relation rather than original hyperlink. Building semantic network of named entities based on relation extraction can not only help people to find more latent interest spots but also augment existed Knowledge Base by entity linking.

2 Related Works

For determining candidates from the results of NER, text feature selection methods can be referred. Prateek[1] elaborated partonomic relation from linguistics and proposeed a strict definition of "part-of" relation. They acquired and filtered candidate set by applying "WikiPageLink" property from Wikipedia. All of the entities which were used to determine "part-of" method were coming from well-defined LOD clouds. So, this method could be weak in handling the problems of noise from restaurant reviews. In order to transform unstructured text into Linked Open data, Exner[2] proposed the pipeline methods for generating lined data from unstructured text by using off-the-shelf tools on Natural language processing (NLP) and semantic web. Many methods about transforming unstructured text into RDF mainly used existed ontology to link entities. These methods are referred in our method to infer new relation which is not defined in knowledge base.

3 The proposed method

In previous research, we have proposed a method to recognize dish name entities from users' reviews[3]. Integrating the results of NER, we present a solution by employing a pipeline method to extract Dish-Ingredient (D-I) relation.

3.1 Subject List Building

In order to remove noise from previous results, subject list related with food domain can be better way. We use a small subject seeds set to bootstrap all related subjects in DBpedia in food domain. We collect 83 subjects distributed in 11 categories as seed set manually. Ontology "skos:broader" is used to search hypernymy and hyponymy layers of each seed. After gotten extended list, we compute the confidence for each subject in it.

$$P(Subj|Food) = \frac{Count(WikiPageLink_{Subj \in SS})}{Count(WikiPageLink_{All})} \quad (1)$$

where *WikiPageLink* is ontology property of "dbpedia-owl:wikiPageWikiLink" corresponding to out-degree of linked pages from a subject in Wikipedia. The numerator denotes the number of linked pages whose subjects are contained in seed set from "Subj". The denominator denotes the number of all linked pages from from "Subj". By means of subject extending, the subject list increases from 83 to 353 with 97% precision.

3.2 Subject Register

Having acquired subject list, a register operation is executed for each entity in input set whose subject is annotated by "dcterms:subject". In this step, only those entities whose subjects could be included in subject list and the number of included subjects are larger than two can be reserved into candidate set.

3.3 Coreference Resolution

That two or more different expressions of named entities in documents substituting the same real world entities are called "Coreference Resolution". In Japanese, multiple written expressions of the same entity always causes more complex in text processing. For those entities whose multiple expressions are derived from being written in different alphabets, we address this problem by transforming all types of Japanese alphabets into "Hiragana " which is the basic spelling character. According to Aliases and Abbreviation in which the transforming can not distinguish the entities who have same semanteme with unrelated writing, we utilize property "dbpedia-owl:wikiPageRedirects" in DBpedia to infer the coreference chain for the entities.

3.4 Preliminary Paths Generation

After coreference resolution, the process could be continued to discovery preliminary paths in the range from dish root node to ingredients leaf nodes passing by some

sub-dish intermediate nodes. The paths are generated by reusing "dbpedia-owl:wikiPageWikiLink". Here, we make an assumption that all entities which can be retrieved by "dbpedia-owl:wikiPageWikiLink" from target dish and have been existed in candidate set have strong association with target dish. Consequently, we acquire a set of preliminary paths for D-I tree.

3.5 Feature Selection

Having acquired preliminary paths, the next step will be operated to locate most probable ingredients which can constitute D-I relation with target dish. Here, users' reviews of "Tabelog" are used as document corpus. Using χ^2 statistic, we compute score for every candidate in preliminary path. Here, the number of each candidate is counted by the norm that the number of occurrence of an entity equals the sum of occurrence of all expressions of it in coreference chain. We assume the reviews including target dish name can be considered as topic documents. Firstly, we compute score for each child node which have been generated by root node. Secondly, we apply same method on those child nodes which are generated by second layer nodes. Iteratively, all the nodes in D-I tree are assigned the scores. Under this operation, some ingredients with same name might get different scores, because they might be linked from different parent nodes. For example, "山椒" (Japanese Pepper) gets two scores because of being linked from both "タレ" (Tare Sauce) and "薬味" (Yakumi) in D-I tree of "ひつまぶし" (Sea eel rice). Next, the scores of "山椒" will be compared to determine that "山椒" more likely is an ingredient of which parent node.

4 Evaluation Experiment

In this experiment, we use the same data source from "Tabelog" with dish names recognition. Here, eight famous foods in Nagoya region are chosen as target dishes. In order to keep the computation tractable, we assume that most probable ingredients usually locate on first following sentence. In Table 1, the second column means the number of extracted paths and the third column means accuracy. The forth column is the number of ingredients for each dish. The accuracy of relation extraction can achieve 86% for all dishes. Although the knowledge base does not provide any D-I relation ontology, our method can get high accuracy by using ontology inference in LOD cloud. Currently, there are few researches focusing on D-I extraction in Japanese. Most of them paid attention to structure analysis of recipes where the ingredients had been given out by users[4]. So, the results of our method could be accepted to some extent. Here, the loss of accuracy is mainly due to the mistake of feature selection and the disorder in relation paths.

Table 1: Results of relation extraction

Name	χ^2		
	D-I Paths		#Ingre.
	#Num	Accu.%	
Sauce dressing spaghetti	18	94.44	18
Curry Udon	25	84.00	25
Flat Japanese noodle	18	94.44	21
Sea eel rice	16	75.00	18
Morning Service	44	90.91	50
Nagoya cochin	13	76.92	17
Miso dressing meet	36	83.33	37
Miso soup udon	36	86.11	37

5 Conclusion and Future Work

In this paper, we present a Dish-Ingredient relation extraction methods by combining NLP and semantic web techniques. After acquired the results of dish name recognition, we realize D-I relation extraction through using existed ontology to infer probable paths in the range of dish name to ingredients and employing text feature selection to get positive ingredients for target dish. The experimental results show that our method gets the better effect of D-I relation extraction. In the future, we will continue the research of relation extraction on elevating recall of ingredients discovering and transforming D-I pairs from unstructured text into RDF.

Reference

[1] Prateek Jain, Pascal Hitzler, Kunal Verma, Peter Z Yeh, and Amit P Sheth. Moving beyond sameas with plato: Partonomy detection for linked data. In *HT12*, pp. 33–42, 2012.

[2] Peter Exner and Pierre Nugues. Entity extraction: From unstructured text to dbpedia rdf triples. In *WoLE12*, 2012.

[3] Weichang Chen, Katsuhiko Kaji, Nobuo Kawaguchi, and Kei Hiroi. Non-local dictionary based japanese dish names recognition using multi-feature crf from online reviews. In *WIMS14*, p. 14, 2014.

[4] Yoko Yamakata, Shinji Imahori, Yuichi Sugiyama, Shinsuke Mori, and Katsumi Tanaka. Feature extraction and summarization of recipes using flow graph. In *Social Informatics*, pp. 241–254, 2013.