Construction of Japanese Dish Names Database Using Multi-Feature CRF from On-Line Reviews

Weichang Chen† Katsuhiko Kaji† Nobuo Kawaguchi†
Graduate School of Engineering, Nagoya University†

1.Introduction

In text extracting field, named entity recognition (NER) is a fundamental method to comprehend natural language. Nowadays many natural language processing tasks are built upon NER, such as Information Extracting, Question and Answering, Knowledge Discovery, etc. It aims to locate and classify identical name of person, location, organization, time, measurement unit, dish name and so on. NER can achieve high accuracy with the automatic extracting method rather than with the rule-based knowledge base that usually cannot cover knowledge completely.

As an important application in local recommender services, restaurant recommending is more likely to encounter the cold-start problem. We can only acquire some limited information from POI (Point of Interest). Hence, users cannot get detailed descriptions for making choice. It is necessary to extend POI by means of obtaining dishes names from restaurant reviews automatically.

2.Related Work

Most of NER tasks focus on conventional direction, such as person, organization, location, date and so on. CoNLL 2003 shared task also listed four name entity classes in person, organization, location and miscellaneous.

There have been existing studies on dish name recognition that belong to a single semantic classification. Tsai and Chou[1] proposed an unsupervised method which can acquire high recall to identify dish name sequences that appear more than twice as candidates. Then, they used Condition Random Field (CRF) to validate those candidates in their experiment that added some auxiliary features in CRF training proceeding like quotation marks, font, color, hyperlink and image proximity.

For an informal data source, twitter is a typical example. Liu[2] conducted an experiment on tweets in which they mentioned and analyzed some challenges about insufficient information in tweets and the unavailability of training data. Then a semi-supervised learning framework basing on KNN classifier and linear CRF model was used to recognize named entities. They also added non-local gazetteer into CRF in order to elevate recall. Because of no general rules in informal data, Liu did not consider special local features in CRF proceeding, such as capitalization, quotation and so on.

3.Proposal Method

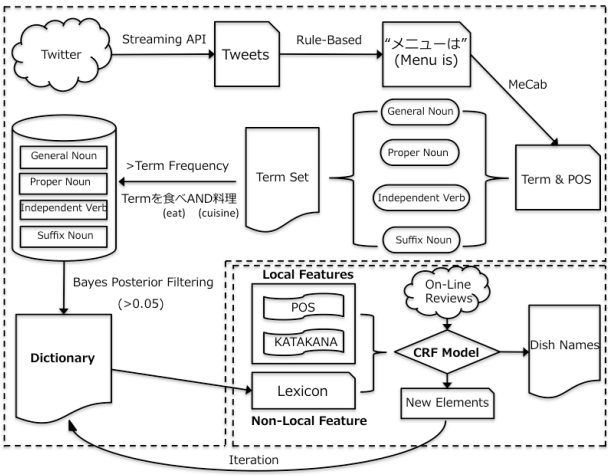

Figure 1: Research architecture and data flow

In this paper we propose a supervised approach to recognize and extract named entities that belong to a special domain (dish name domain). Figure 1 shows research architecture and data flow including dictionary construction and multi-feature CRF framework. First of all, we use Japanese twitter¶ as a data source to construct a dictionary of dish elements by semantic rules and use Bayesian posterior to remove noise. In this algorithm, four types of POS (part of speech) are always used as dish name elements corresponding to General Noun, Proper Noun, Independent Verb and Suffix Noun respectively. So only terms within these four types of POS would be stored into the dictionary. Secondly, we map the dictionary as a non-local feature into the Conditional Random Field. In this stage, we also refer to other three local features (Term, POS and KATAKANA) as observations in CRF. By combining the dictionary with machine learning, our method can achieve high precision and recall with the multiple iteration, and it can find and draw the new dish name elements back to the dictionary continuously. For segmentation and tagger, we adopt off-the-shelf Japanese morphology tool named MeCab‡. A Japanese dish name database of Aich prefecture is constructed from Tabelog* (Japanese cuisine classification website) in our experiment.

¶ http://twitter.com/
*http://Tabelog.com/
‡ https://code.google.com/p/mecab/

4.Evaluation experiment

For constructing dictionary of dish elements, we use tweets as the data source in which 37443 cuisines related messages are collected within 24days by twitter authority API. Meanwhile we utilize web crawler to snatch reviews from Tabelog as training set and testing set for CRF learning proceeding. The data set includes 27997 on-line registered restaurants where 185494 reviews are written. We randomly select 500 reviews as experimental sample that is annotated by one person within 3 days. We adopt 10-fold cross validation to evaluate the experimental results with 90% for training and 10% for testing. We can see the statistics in table 1.

Table 1: Statistics of data set

Data Set	Statistics	
Tabelog	#(reviews):	185494
	#(Restaurants):	27997
Experiment Data	#(reviews):	500
	Training Set:	90%
	Testing Set:	10%
Dictionary	#(Tweets):	37443

In the experiment, we consider two baselines. One is dictionary looking up, and another is CRF+Term. In table 2, 3, 4, symbol T, P, D, K, B and I denote term, part-of-speech, dictionary, KATAKANA and iteration respectively. The forth column F1 score is calculated by the second and the third columns. The fifth column is standard deviation of F1 score. The proceeding of experiment is completed through adding each local feature and non-local feature into system step by step. Following, we will evaluate the results of baseline and our method.

Table 2 shows low precision and recall of only using dictionary. The reason is many terms in the dictionary cannot be as dish name independently, and many dish name elements are not collected into the dictionary. So, it is not available for only using the dictionary as NER method. For using CRF and T, although high precision (91.14%) can be acquired, a low recall (54.39%) also appears simultaneously. That is a common problem of the NER systems that often achieve high precision accompanied with low recall.

Following, we add the other features into CRF progressively. In the first step, term as unique feature is used to test second baseline that can acquire high precision. Next, we add the POS into the proceeding, and it brings huge increase on recall and F1 score. As mentioned above, most of elements consisting of Japanese dish name belong to these four types of part-of-speech. Therefore, they could be considered as powerful observations for sequence recognition. As rows 5 and 6 in table 2, KATAKANA hardly contributes anything to final results. By observation on training data, we find KATAKANA accounts for a small proportion in data set. But we cannot ignore it. There exist large mount foreign dish names in Japanese, especially in ethnic cuisine. So, we experiment on another data set that is extracted with a large number of foreign dish names by handicraft. In table 4, we can see that KATAKANA plays an important role in elevating recall.

Table 2: Results for CRF+Features

	Pre.%	Re.%	F1%	SD%
DICTIONARY	70.7	42.7	53.25	
CRF+T	**91.14**	54.39	68.06	3.32
CRF+T+P	88.66	71.47	79.08	2.11
CRF+T+P+D	88.37	74.39	80.68	2.75
CRF+T+P+D+K	88.19	74.44	80.65	2.43
CRF+T+P+D+K+I	88.76	**75.52**	81.11	2.42

Table 3: Results for Combining with Bayes Posterior

	Pre. %	Re.%	F1%	SD%
CRF+T+P+D+B	88.25	79.41	83.54	2.94
CRF+T+P+D+K+B	88.41	78.99	83.37	2.89
CRF+T+P+D+K+B+I	87.49	**81.61**	**84.38**	2.48

Table 4: Results for adding KATAKANA

	Pre. %	Re.%	F1%
CRF+T+P+D+B+I	87.12	73.72	79.86
CRF+T+P+D+B+K+I	**87.23**	**78.84**	**82.82**

According to the non-local dictionary feature, we divide the experiment into two parts - the first part uses whole dictionary without dealing with Bayes posterior and another uses the dictionary processed by Bayes posterior. From row 4 to 6 in table 2, a small increase can be observed in proceeding of adding D, K, and I without Bayes posterior. In table 3, an evident increase can be achieved by the dictionary filtered with Bayes posterior compared with table 2, and there is a little loss of precision as the compensation. High recall is beneficial for getting enough information in cuisine documents.

5.Future work

In the future, we will build Japanese cuisine database using probabilistic graphical model and extract sentiment from on-line reviews for judging users' opinion. Finally we will provide location and time based cuisine recommender service for users.

Reference

[1] R. Tsai and C. Chou. Extracting dish names from Chinese blog reviews using suffix arrays and a multi-modal CRF model. In First International Workshop on Entity-Oriented Search. ACM SIGIR, 2011.

[2] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, June 19-24 2011.