Non-Local Dictionary Based Japanese Dish Names Recognition Using Multi-Feature CRF from Online Reviews

Weichang Chen
chen@ucl.nuee.nagoya-u.ac.jp

Katsuhiko Kaji
kaji@ucl.nuee.nagoya-u.ac.jp

Kei Hiroi
k.hiroi@ucl.nuee.nagoya-u.ac.jp

Nobuo Kawaguchi
Kawaguti@nagoya-u.jp

Graduate School of Engineering,Nagoya University
Furo-Cho, Chikusa-ku, Nagoya,Aichi, Japan

ABSTRACT

In cuisine recommender service, online user review is an important data source avoiding a cold-start problem. Cuisine-domain named entity recognition(NER) can be used as an entrance to comprehend the semantic information of reviews. This paper describes a supervised approach recognizing Japanese dish name entity (DNE) from online reviews of Japanese cuisine website. In the first stage, this work adopts tweets as the data source to construct the dictionary of dish name elements through semantic rules and use Bayesian posterior to remove noise. Next stage, we maps first-stage dictionary as a non-local feature into Conditional Random Field (CRF) to recognize the dish name. This method can automatically add new dish name elements into the non-local dictionary by iteration during the recognition proceeding. By using 10-fold validation, experimental results show our method can reach 84.38% in F1 score and outperform the two baselines using the dictionary or CRF with term feature separately.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.7 [**Natural Language Processing**]: Text analysis

General Terms

Algorithm, Design, Experimentation

Keywords

Text Extraction, Twitter, Conditional Random Field, Non-Local Dictionary

1. INTRODUCTION

WIMS'14, June 2-4, 2014 Thessaloniki, Greece

In text extracting field, named entity recognition is a fundamental method to comprehend natural language. Nowadays many natural language processing (NLP) tasks are built upon NER, such as Information Extracting, Question and Answering, Knowledge Discovery, etc. It aims to locate and classify identical names 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 rules-based knowledge base which usually can not 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) as shown in figure 1. Hence, people can not get detailed descriptions for making choice. It is necessary to extend POI by means of obtaining dishes names from restaurant reviews automatically.

```
<Item>
  <Rcd>23002562</Rcd>
  <RestaurantName>ステーキハウス スエ</RestaurantName>
  <TabelogUrl>http://tabelog.com/aichi/A2301/A230101/23002562/</TabelogUrl>
  <TabelogMobileUrl>http://m.tabelog.com/aichi/A2301/A230101/23002562/</TabelogMobileUrl>
  <TotalScore>3.83</TotalScore>
  <TasteScore>3.94</TasteScore>
  <ServiceScore>3.94</ServiceScore>
  <MoodScore>3.9</MoodScore>
  <Situation>友人・同僚と</Situation>
  <DinnerPrice>￥10,000～￥14,999</DinnerPrice>
  <LunchPrice>￥4,000～￥4,999</LunchPrice>
  <Category>ステーキ、洋食・欧風料理（その他）、居酒屋</Category>
  <Station>国際センター、名鉄名古屋、伏見</Station>
  <Address>名古屋市中村区名駅南1-8-18</Address>
  <Tel>052-551-5815</Tel>
  <BusinessHours>昼12：00～14：00夜17：00～21：00（土・祝は夜のみ）</BusinessHours>
  <Holiday>日曜</Holiday>
  <Latitude>35.1668574879958</Latitude>
  <Longitude>136.890344625265</Longitude>
</Item>
```

Figure 1: An example illustrating POI of the restaurant

There are some distinctive characteristics in Japanese dish names. For traditional Japanese dish names, almost all of them are made up by term units like "目玉焼き" (fried eggs) where "目玉" means eggs and "焼き" means fried. In some other phrases, such as "塩焼きそば" (salty fried noodle) and "特製大盛つけ麺" (premium big size sauce noodle), the elements are combined freely with each other. In view of this, we extract traditional Japanese DNEs by ignoring bound-

ary rather than by adopting BILOU or BIO chunks patterns [14]. For foreign dish names, they are usually expressed in the form of KATAKANA (Japanese syllabary) which can be used to denote transliteration. For instance, "パスタ" means pasta and "バタートースト" means butter toast. Unlike in Chinese dish names, there are hardly any abstract names in Japanese like Chinese "佛跳牆" which is cooked with the pork and chicken (buddha jumps over the wall) [20].

In this paper, we propose a supervised approach to recognize and extract named entities which belong to a special domain. First of all, we use Japanese twitter[1] as the data source to construct a dictionary of dish elements by semantic rules. As a popular self-media tool, twitter is used to publish real-time messages named tweets by millions of persons in their daily life without content limitation. Many users are prone to share cuisine menus with their followers. Meanwhile, because of tweets' short-text nature, we can acquire relatively purer dish name entities without unnecessary description. However, only relying on constructing dish name dictionary from tweets can not meet the requirement of covering and recognizing all DNEs. Therefore, we import the dictionary as a non-local feature into the Conditional Random Field. By combining the dictionary with machine learning algorithm, our method can achieve high precision and recall with the multiple iterations, and it can find and draw the new dish name elements back to the dictionary continuously.

The contributions of this research can be concluded as: we build a dictionary of dish name elements collected from tweets on the base of Japanese semantic rules. Then, we lead the dictionary as feature into CRF learning algorithm to recognize the whole Japanese dish names from online cuisine reviews. In this paper, the iterative proceeding of recognizing can add new dish elements into dictionary.

The remainder of this paper is organized as follows. In section 2, we introduce the related work. Section 3 provides some detailed descriptions about the methods and section 4 presents and evaluates the experimental results. In section 5, we show an application scenario on dish names extraction. Conclusion and the future work are described in the last section.

2. RELATED WORK

Most of NER tasks focus on conventional direction, such as person, organization, location, date and so on [13]. CoNLL 2003 shared task [19] also listed four name entity classes in person, organization, location and miscellaneous. In this section, we introduce some related works in three aspects.

2.1 NER on Dish Name Domain

There have been some existing studies on dish name recognition that belongs to a single semantic class. Tsai and Chou [20] proposed an unsupervised method which can acquire high recall to identify dish name sequences that appear more than twice as candidates. Then, they used CRF to validate those candidates in their experiment which added some auxiliary features in CRF training proceeding like quotation marks, font, color, hyperlink and image proximity. Shin and Peng [17] presented an application of restaurant recommendation. They adopted blog posts as a data source and used the supervised method to divide blogs into two classes(restaurant and unrestaurant). Then their study used mutual information to extract dish units and serialized n-gram grouping & weighting to score dish name sequences. Vechtomova [21] gave out a semi-supervised approach which used a small set of seed words representing the dish name class, and applied the distributional similarity measure to rank all single words in the corpus. Also this research can get most probable dish name relations by calculating NP-Score. This method can achieve high precision and it could be utilized to remove noise by other researches. In dictionary construction, Shinzto [18] described an automatic method for NER in specific domains like restaurant guides and showed many named entity classes related restaurants such as area, time, and station. Their dictionary construction algorithm used the annotated corpus as seeds which are expanded by using a large number of HTML documents. However, this method needs to obtain a large number of web pages and costs an amount of human labour to annotate named entities.

[1]https://twitter.com/

2.2 NER on Different Data Source

Considering those researches in NER from different data sources, we divide them into two categories. In a formal data source like news and science documents, Zhang [23] proposed a stochastic model to tackle the problem of Chinese NER on People's daily and MET-2. Zhang and Yoshida [24] adopted a new method named enhanced mutual information and collocation optimization to extract multi-word expressions on science reference documents. Yoshida and Tsujii [22] used multi-feature set such as label feature, word-based feature and chunk-based feature to extract biomedical names in Japanese medicine articles. Jimeno [6] used the semantic method to recognize disease names on a corpus of annotated sentences. As an informal data source, twitter is a typical example. Liu [10] conducted an experiment on tweets in which they mentioned and analysed some challenges about insufficient information in tweets and the unavailability of training data. In that work, a semi-supervised learning framework based on K-nearest neighbor 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. Ritter [15] approached an experimental study to address high noise and informal problems of tweets by re-building the NLP pipeline from Part-of-Speech (POS) tagging and chunking to Named entity recognition.

2.3 Non-Local Feature in CRF

Named Entity recognition is always limited by the low recall as a result of asymmetric data distribution in which the NONE class dominate entity classes [11]. Chieu and Teow [1] applied a dual decomposition approach to combine local sentential model and non-local label consistency model in NER proceeding. This method can simplify complicated sequences into two CRF models(sentence CRF and word CRF) and can acquire remarkable results. Krishman and Manning [8] showed a two-stage approach to handle non-local dependencies in NER. In this research, they defined three pairs of features and extracted them from the output of the first stage CRF, then combined them with the second CRF stage. Based on Krishman's research, Mao [11] pre-

sented an approach using four types of non-local features to improve NER recall on MSRA and CityU datasets. But all of non-local features came from the training set. So there is no universality in this method when using change to training set. It is necessary to guarantee enough non-local features for a large training data.

3. METHOD

Following, we show some challenges in Japanese dish name entity recognition, and present detailed methods to address these challenges. Figure 2 shows research architecture and data flow including dictionary construction and multi-feature CRF framework.

3.1 Challenges

Main challenges consist of three aspects:

Challenge 1: There exist varieties of studies about NER relying on external corpora or crowdsourcing such as Wikipedia [3] [4] [7]. However, according to dish name domain, no large-scale corpora can be utilized in Japanese NER like WordNet[2]. How to construct non-local dictionary and combine with machine learning for boosting precision and recall would be first challenge.

Challenge 2: In Japanese, there are large number of elements being used in composing dish names. And these elements also appear anywhere independent with dish names. Therefore, the second challenge is how to define and classify correlation of elements among dish names.

Challenge 3: The data source used to recognize dish names comes from online restaurant reviews where there are with a mass of colloquial expression, likes "サンド" (sand) representing "サンドイッチ" (sandwich). How to discovery and recognize the colloquial expression is third challenge.

3.2 Overview

In order to cope with three challenges, a hybrid model is adopted by means of combining machine leaning method with external dictionary constructed by collecting dish name elements from tweets using sementic rules.

For challenge 1, the primary problem is running out of external corpora in special-domain NER. As a result, we focus on constructing our own cuisine corpora. The informal and real-time natures of tweets can satisfy our requirement for recognizing from online restaurant reviews.

Next step, for challenge 2, we find that four kinds of POS are distinctive which always are used as elements of dish names. We decompose tweets message into term vectors, and annotate POS for every terms employing off-the-shelf Japanese morphology analyser. Finally, only terms belonging to four kinds of POS are put into the dictionary.

For colloquial expression of online restaurant reviews in challenge 3, we find tweets sharing the same written structure with online reviews. Therefore, some frequently used colloquial expression could be collected into the dictionary. Although we also encounter limitation in complete coverage of rule-based dictionary, Markov process based machine learning method can recognize the terms which do not exist in the dictionary by sequence model. Our algorithms are given as follows.

[2]http://nlpwww.nict.go.jp/wn-ja/index.en.html

Algorithm 1 Non-Local Dictionary Construction

Require: Tweets input stream tis;
Require: General noun gn; Proper noun in;
Require: Independent verb idv; Suffix noun sn;
Require: Term set ts; Dictionary D;
  Initialize: $\vec{V}_{tis} \leftarrow \vec{0}$, $\vec{V}_{gn} \leftarrow \vec{0}$, $\vec{V}_{in} \leftarrow \vec{0}$, $\vec{V}_{idv} \leftarrow \vec{0}$;
  Initialize: $\vec{V}_{sn} \leftarrow \vec{0}$, $\vec{V}_{ts} \leftarrow \vec{0}$, $\vec{V}_{D} \leftarrow \vec{0}$;
  while tis.contain(「Menu is」)=ture **do**
    get $\vec{V}_{tis}$.add(tis);
  end while
  for each $tis_i \leftarrow$ Projection($\vec{V}_{tis}$) in dimension j **do**
    get $\vec{V}_{term} \leftarrow$ MeCab(tis_i) with {$\vec{V}_{term}$| $\vec{V}_{gn}$,$\vec{V}_{in}$,$\vec{V}_{idv}$,$\vec{V}_{sn}$};
    if Num of $\vec{V}_{term}$ more than threshold **then**
      $\vec{V}_{ts} \leftarrow \vec{V}_{term}$;
    end if
  end for
  $\vec{V}_{D} \leftarrow$ Bayes($\vec{V}_{ts}$);
  return D;

Algorithm 2 Non-Local Dictionary Based Dish NER Using CRF

Require: Non-local dictionary D; Training set trs;
Require: Training set tes; Labeling result lr;
Require: New term set nt;
  Initialize: $\vec{V}_{trs} \leftarrow \vec{0}$, $\vec{V}_{tes} \leftarrow \vec{0}$, $\vec{V}_{D} \leftarrow \vec{0}$;
  Initialize: $\vec{V}_{nt} \leftarrow \vec{0}$;
  for each ds $\neq$ null with {ds| trs,tes,D} **do**
    $\vec{V}_{i} \leftarrow$ MeCab(ds);
    get $\vec{V}_{i}$ with {ds| $\vec{V}_{trs}$,$\vec{V}_{tes}$,$\vec{V}_{D}$};
  end for
  $\vec{V}_{nt} \leftarrow \vec{V}_{D}$;
  while $\vec{V}_{nt} \neq \vec{0}$ **do**
    for each $t_j \leftarrow$ Projection($\vec{V}_{D}$) in dimention j **do**
      if $t_j \in \vec{V}_{trs}$ **then**
        $F_D(t_j) \leftarrow$ Y;
        if Form of t_j = 「KATAKANA」 **then**
          $F_{ka}(t_j) \leftarrow$ K;
        end if
      end if
      $F_{term}(t_j) \leftarrow t_j$;
    end for
    model $\leftarrow$ Train($\vec{V}_{trs}$) with F_{term}, F_{ka} and F_{pos};
    lr $\leftarrow$ Test($\vec{V}_{tes}$) with model;
    $\vec{V}_{nt} \leftarrow$ Compare(lr,D);
    $\vec{V}_{nt}$.add($\vec{V}_{nt}$);
  end while
  return lr;

3.3 Dictionary Construction

It is necessary to make clear the composition of Japanese DNEs. The research of Shin [17] on Chinese DNEs gives us some enlightenments that they divided composition of dish names into meaningful elements. We also split DNEs from different aspects as follows:

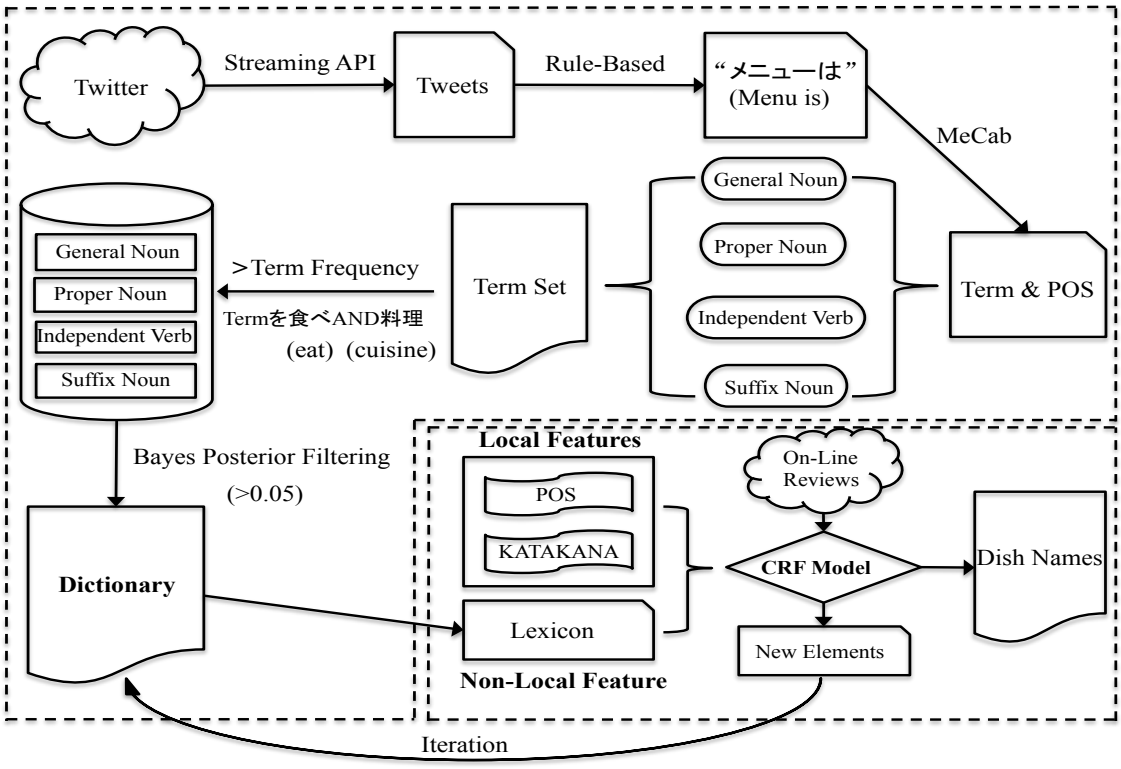

Figure 2: Research architecture and data flow

- **Ingredients:** Main ingredients of dish name, such as “アボカドとパストラミのバター醤油” (Butter soy sauce dressing on avocado and pastrami).
- **Culinary methods:** How the dish being cooked or seasoned, such as “鶏肉と厚揚のトマト煮” (Boiled chicken, fried bean curd and tomato) and “鴨南蛮” (soba noodles served with a cooked duck on top).
- **Cooking tools:** Tools used on cooking dish, such as “ひつまぶし” (sea eel rice holding with small box).
- **Origin:** Where the dish being from. Like “台湾ラーメン” (Taiwan noodle).
- **Transliteration:** Transliteration of foreign dish name. Most of foreign dish names are expressed by transliteration in the form of “KATAKANA”.

Four types of POS 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. Algorithm 1 illustrates the detailed procedure for the dictionary construction. Data stream *tis* is from twitter search API in which we set a semantic rule for all Japanese tweets including “メニューは” that means “menu is”. Many users like publishing their lunch or diner lists written in this expression by their accounts. The sentence are always expressed in form of “メニューはXXX,XXX,XXX”. Only the second half behind “メニューは” is employed as import of our experiment. For segmentation and taggering, we adopt off-the-shelf Japanese morphology tool named “MeCab”[3].

[3]https://code.google.com/p/mecab/

After obtaining taggering result, four types of POS are stored into term sets $\vec{V}_{term}$ respectively. However, an amount of noise would be brought into the term sets simultaneously that belongs to the four types of POS but is not related with cuisine. We remove part of noise using Japanese search engine “InfoSeek[4]” by another semantic rule with “XXXを食べ & 料理” meaning “eat XXX & cuisine” and obtain cuisine term sets $\vec{V}_{ts}$. The thresholds are denoted as 90-500-150-100. It is still not enough to lead the cuisine term sets into the dictionary, because some terms always occur in cuisine documents but are not as components of dish names. So, we utilize Bayesian posterior to rank all terms of $\vec{V}_{ts}$, then add the terms whose probability is larger than 5% into the dictionary D . The formula of Bayesian posterior is as follows:

$$P(Cuisine|Term) = \frac{P(Cuisine, Term)}{P(Term)}$$

Here, we consider “Tweets” is a proper data source with the find format. For one thing, because menu related tweets are needed to locate accurately, we only use those tweets which are expressed explicitly by semantic rules. For another, by getting lots of menu related tweets, we can ensure the coverage of dish name elements in some degree. Thus, the tweets with implicit expression are ignored.

3.4 Dish Name Recognition Using CRF

Conditional Random Field [9] is a sort of discriminative undirected graphical model where vertices are denoted as random variables and edges are denoted as probabilistic de-

[4]http://www.infoseek.co.jp/

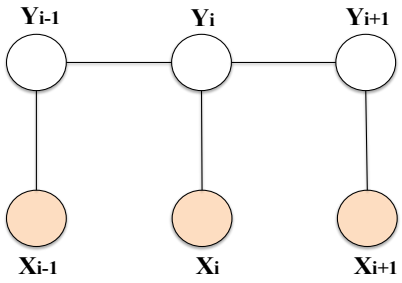

Figure 3: Graphical structure of linear-chain CRF

pendency between variables. CRF has been employed successfully on some territories of natural language processing and biological sequence prediction, such as named entity recognition [12], shallow parsing [16] and gene prediction [2]. CRF is defined as below:

Definition: [9] Supposing a undirected graph $G = (V, E)$ as $Y = \{Y_v | \mathrm{v} \in V\}$, so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variables Y_v obey the Markov property with respect to the graph: $p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G. Figure 3 shows a graphical structure of linear-chain CRF.

In linear-chain CRF model, give two sets of observation sequence $\{\vec{x}_i|\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n\}$ and random variable $\{\vec{y}_i|\vec{y}_1, \vec{y}_2, \cdots, \vec{y}_n\}$. based on fundamental theorem of random fields [5]:

$$P(y|x, \lambda) \propto exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i))$$

$t_j(y_{i-1}, y_i, x, i)$ is transmission eigenfunction corresponding observation sequence between i and $i-1$. And $s_k(y_i, x, i)$ is eigenfunction of observation sequence. Unify two eigenfunction as $f_i(y_{i-1}, y_i, x, i)$:

$$P(y|x, \lambda) = \frac{1}{Z(x)} exp(\sum_i^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i))$$

$$Z(x) = \sum_j exp(\sum_i^n \sum_j \lambda_j f_j(y_{i-1}, y_i, x, i))$$

The maximum probability of label sequence for given x is,

$$Y = argmax P(y|x, \lambda) \quad (1)$$

In algorithm 2, we segment and tagger trs, tes, D using MeCab, then annotate each term in four dimensions as part-of-speech, dictionary, KATAKANA and term-self. The model is built upon CRF on training set trs which is annotated by handicraft. By means of prediction on testing set tes , we can obtain a labelling result set lr that has been annotated automatically. Comparing F_{dic} of lr and the dictionary D, we put new terms $\vec{V}_{nt}$ back to D which is annotated as dish names but do not exist in the dictionary D. The system iterates the whole proceeding by repeating previous method until no new terms appear. Although this method may bring some noise into the dictionary, it is worth of improving recall with small loss of precision.

Here, four features are shown as follows:

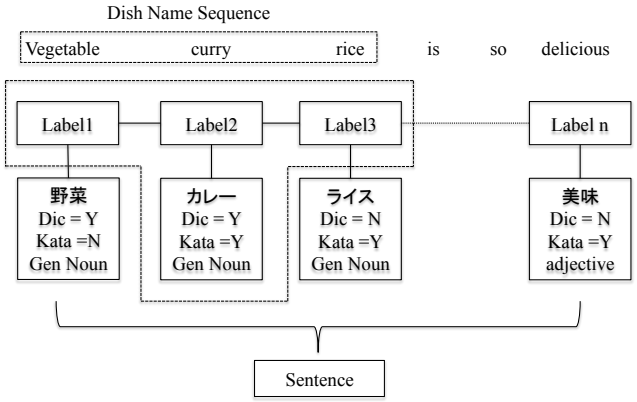

Figure 4: Labelling sample of non-local dictionary based dish name recognition using CRF

- **Term:** In the experiment, we denote five-term-window as template to train. The trainer needs to train all the rules on each five-term-sequence it passes by. The reason of choosing five-term-window is that numbers of dish names are composed within five words.
- **Part-of-speech:** POS is an important feature for DNE recognition. As description above, most of dish name elements are included in these four types of POS. The combination of them can change results of recognition deeply.
- **Dictionary:** It is the core feature for our method. This feature can provide a high confidence for training proceeding. In dataset, large number of terms with "$F_{dic} = Y$ " would be taggered as dish names.
- **KATAKANA:** From testing results, we find that some terms written in form of KATAKANA often are used as single-term dish name and a part of them are not collected into the dictionary yet. Markov chain based machine learning algorithm is skilled in recognizing term sequence. This algorithm is not suitable for some foreign dish names that often appear in form of single word. Therefore, we add KATAKANA feature as observation intendedly into CRF to overcome this shortcoming .

A sample is shown in figure 4. In this sample (three-term-window), the window locates "Vegetable curry rice ", and the current term is curry. In equation (1), we can get observations $\{x_i\}$ corresponding to four features and the terms that are in former or latter position than current term. By observations $\{x_i\}$ and parameter λ from training, CRF can calculate the maximum probability of Y to predict the result that "Vegetable curry rice" is a dish name.

4. EXPERIMENT AND EVALUATION

We will show experimental results and evaluation from some aspects in this section.

4.1 Data Set

According to the training and testing set, we utilize web crawler to snatch reviews from Tabelog[5] (Japanese cuisine

[5]http://tabelog.com/

Table 1: Results for CRF+Features

	Pre.(%)	Rec.(%)	F1(%)	S(%)
DICTIONARY	70.70	42.70	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 2: Results for Combining with Bayes Posterior

	Pre.(%)	Rec.(%)	F1(%)	S(%)
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

classification website). The data set includes 27997 online registered restaurants where 185494 reviews are written. We randomly select 500 reviews as experimental sample which 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. In the dictionary construction, twitter authority API is used to obtain tweets by semantic rules, and 37443 cuisines related messages are collected within 24days. We can see the statistics in table 3.

Table 3: Statistics of data set

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

4.2 Evaluation Metrics

Precision, recall and F1 score are used in evaluating results. F1 score can achieve harmonic mean between precision and recall. In NER system, we not only should regard the precision of recognition results, but also consider recall, especially for recommender service. The formulas of precision, recall and F1 score are given as follows:

$$\textbf{Precison} = \frac{\#(correct\ DNEs\ recognized)}{\#(all\ DNEs\ recognized)}$$

$$\textbf{Recall} = \frac{\#(correct\ DNEs\ recognized)}{\#(all\ ture\ DNEs)}$$

$$\textbf{F1 score} = (1+\beta^2)\frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

Our experiment assumes $\beta = 1$.

4.3 Results

In the experiment, we consider two baselines. One is dictionary looking up, and another is CRF+Term. In table 1, 2, 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 features and non-local feature into system step by step. Following, we will evaluate the results of baseline and our method.

4.4 Baseline

Table 1 shows the low precision and recall gotten from dictionary looking up baseline. The reason is that many terms in the dictionary can not be as dish name independently, and many dish name elements are not still collected into the dictionary. So, it is not available for only using the dictionary as NER method. For another baseline, 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 which often achieve high precision accompanied with low recall.

4.5 Effect on Local Features

We discuss the contribution of the features on the whole recognition procedure. In the first step, term as unique feature is used to test second baseline which 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 1, 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 can not ignore it. There exist a mass of foreign dish names in Japanese, especially in ethnic cuisine. So, we experiment on another data set which 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.

4.6 Effect on Non-local Dictionary

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

Table 4: Results for adding KATAKANA

	Pre.(%)	Rec.(%)	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**

4 to 6 in table 1, a small increase can be observed in proceeding of adding D, K, and I without Bayes posterior. In table 2, an evident increase can be achieved by the dictionary filtered with Bayes posterior compared with table 1, and there is a little loss of precision as the compensation. High recall is beneficial for getting enough information from cuisine documents.

5. APPLICATION SCENARIO

This research is the part of "Smart Station Nagoya" which aims to provide comprehensive services concluding voice navigation and recommendation based on indoor location in Nagoya station. As the method described above, we present a preliminary application scenario of dish name extraction. There are tens of Nagoya local famous foods that can be enjoyed in restaurants surrounding Nagoya station. When the user reaches Nagoya station area and wants to find a restaurant by mobile phone, a popular or famous dishes list would be displayed depending on the user's location by the way of inquiring user's target dinner genre such as instant food, Japanese food or the exotic flavour. After clicking on the item, user can obtain detailed information about the dish by text and image like figure 6 and 7 showing. Hence, the user could make a definite decision based on preference. Here, we may encounter a problem that someone gives out the rogue reviews on the specific restaurant or dish name. In this work, we adopt the real online reviews as data souce which are refereed as crowd-sourcing. This crowd-based method can avoid noise leaded into by the rogue behaviours. We collect 14 sorts of most famous dishes which are popular among the tourists and residents, and rank them by three criterion as Mean, Variance and Weighting Bayes Averaging that are given as follows. In view of the uneven distribution of user reviews, we adopt a weighted average method to compensate less amount of dish reviews in all samples.

$$\mathbf{WBA} = \frac{\#(Review_i)}{\#(Review_i)+\#(LeastOfReviews)} Mean_i + \frac{\#(LeastOfReviws)}{\#(Review_i)+\#(LeastOfReviews)} Mean_{all}$$

Here, $\#(Review_i)$ denotes the number of reviews for dish i and $\#(LeastOfReviews)$ denotes the number of reviews for least the quantity of the dish. $Mean_i$ and $Mean_{all}$ represent the mean of user ratings of dish i and all dishes respectively. All the user ratings are obtained from the reviews referring to all 14 dish names in Tabelog website. We can see an example of user review shown as figure 5. In table 5, we give out the statistic ranking for 14 famous dishes. Here, the mean denotes average score of every dish that is given out by all users(the score is from 0 to 5).


```
<Item>
  <NickName>flowing3stellate3</NickName>
  <VisitDate>13/01</VisitDate>
  <ReviewDate>13/01/22</ReviewDate>
  <UseType>Night</UseType>
  <Situations>Friends and Fellowship</Situations>
  <TatalScore>3.17</TatalScore>
  <TasteScore>3.0</TasteScore>
  <ServiceScore>3.5</ServiceScore>
  <MoodScore>3.0</MoodScore>
  <DinnerPrice>￥1,000～￥1,999</DinnerPrice>
  <LunchPrice>" "</LunchPrice>
  <Title>山本屋味噌煮込みうどん 食べ比べ</Title>
  <Comment>
  名古屋駅隣接JRセントラルタワーズにあります。普通煮込みうどん(1008円)を頼みます。こちらは打ち粉にそば粉を使用しているので、注文時にそばアレルギーが無いかどうか聞かれます。紙エプロンが…
  </Comment>
  <PcSiteUrl>
    http://tabelog.com/aichi/A2301/A230101/23000114/dtlrvwlst/4879199/
  </PcSiteUrl>
  <MobileSiteUrl>
    http://m.tabelog.com/aichi/A2301/A230101/23000114/dtlrvwlst/4879199/
  </MobileSiteUrl>
  <Latitude>35.171416586007</Latitude>
  <Longitude>136.882526175785</Longitude>
</Item>
```


Figure 5: An example of user review

After computing popularity rank of user rating, a simple recommender list is presented. Considering individual flavour for every user, it is necessary to show more detailed information or images for the users. By means of our research, we can extract related food names from online restaurant reviews of the most famous dishes automatically and can provide detailed list concluding ingredient and set meal for the tourists and residents. Two samples are shown as follows:

〈薬味(Flavorant), 135〉; 〈わさび(Horseradish), 111〉; 〈タレ(Sauce), 94〉; 〈ネギ(Green onion), 87〉; 〈鰻(Sea eel), 157〉; 〈海苔(See weed), 85〉; 〈大葉(Perilla leaf), 49〉; 〈ご飯(Rice), 35〉; 〈出汁(Soup stock), 26〉; 〈山椒(Pepper), 26〉; 〈吸い物(Japanese soup), 21〉; 〈茶碗蒸し(Yellow cake), 18〉; 〈お茶漬け(Rice soaked in tea), 14〉; 〈緑茶(Green tea), 12〉; 〈蒲焼(Grilled fish), 11〉; 〈山葵(Horseradish), 10〉; 〈刺身(Sashimi), 9〉

Figure 6: Detail description and image of Sea eel rice

Table 5: Statistic ranking for 14 famous dishes in Nagoya

Name	Mean	Variance	#(Restaurants)	#(Reviews)	WBA
ひつまぶし(Sea eel rice)	3.8425	0.6410	17	702	3.7636
名古屋コーチン(Nagoya cochin)	3.6421	0.6961	439	1516	3.6303
カレーうどん(Curry noodle)	3.6197	0.7040	693	1950	3.6130
天むす(Fried prawn on rice roll)	3.6065	0.6710	295	841	3.5966
味噌煮込みうどん(Noodle with Miso soup)	3.5874	0.7122	722	2164	3.5850
味噌おでん(Boiled vegetable with Miso soup)	3.6023	0.6875	162	282	3.5846
どて煮(Boiled haslet with Miso soup)	3.5878	0.7084	405	661	3.5816
台湾ラーメン(Taiwan noodle)	3.5788	0.7195	809	2033	3.5774
手羽先(Baked chicken wings)	3.5683	0.7261	1061	3076	3.5682
味噌カツ(Breaded pork chops with Miso dressing)	3.5307	0.7258	960	2134	3.5349
エビフライ(Fried prawn)	3.5249	0.7224	981	1927	3.5303
きしめん(Flat Japanese noodle)	3.4659	0.7479	1165	3076	3.4743
モーニング(Morning)	3.4009	0.7555	3040	6617	3.4077
あんかけスパ(Sauce dressing noodle)	3.3533	0.7835	443	1267	3.3921

〈パン(Bread), 1750〉; 〈トースト(Toasted bread), 1676〉; 〈珈琲(Coffee), 1095〉; 〈スープ(Soup), 848〉; 〈サラダ(Salad), 2462〉; 〈紅茶(Black tea), 706〉; 〈デザート(Dessert), 683〉; 〈サンドイッチ(Sandwich), 525〉; 〈フルーツ(Fruits), 434〉; 〈ヨーグルト(Yogurt), 410〉; 〈パスタ(Pasta), 350〉; 〈玉子(Omelette), 313〉; 〈バター(Butter), 295〉; 〈ハム(Ham), 249〉; 〈ポテトサラダ(Potato salad), 243〉; 〈ドレッシング(Dressing), 234〉; 〈ジャム(Jam), 230〉; 〈バナナ(Banana), 185〉

Figure 7: Detail description and image of Morning(a kind of breakfast)

In figure 6 and figure 7, the first example is the No.1 famous traditional Japanese food in Nagoya named "ひつまぶし" (Sea eel rice), and the second example is the western-style breakfast named "モーニング" (Morning). There are 702 reviews of "Sea eel rice" in 17 restaurants and 6617 reviews of "Morning" in 3040 restaurants in statistics results. Japanese characters showed in textboxes represent ingredient of dish or some foods which are often eaten with these two dishes together. The number in bracket stands for mutual frequency between them. Results show our method can extract detailed dish name entities fundamentally. By the images and key words of dishes, tourists or residents can choose their preferring flavour.

6. CONCLUSION AND FUTURE WORK

In this paper, we present a non-local dictionary based named entity recognition method by using supervised leaning algorithm CRF. Meanwhile, we examine results validation and compare them with baseline system. Experimental results show our method acquire high F1 score with balance precision and recall. Finally, we evaluate and discuss advantages of our method. In all, our approach can contribute in some aspects: the non-local dictionary is constructed by semantic rules with tweets as the data source, and the small volume of the dictionary as bootstrap can avoid complexity work for building or using large-scale knowledge base; Using machine learning method on sequence model can elevate recall radically; High precision of learning proceeding can guarantee constructing the dictionary with less noise in iteration proceeding.

In the future, we will build an ample Japanese cuisine database using semantic web technologies and extract sentiment from online reviews to judge users' opinion by semantic method. Finally we will realize location and time based cuisine recommender service for users.

7. ACKNOWLEDGEMENT

A part of this work is supported by Microsoft Research Asia and SCOPE (Strategic Information and Communications R&D Promotion Programme) number 132306007 operated by Ministry of Internal Affairs and Communications of JAPAN.

8. REFERENCES

[1] H. L. Chieu and L.-N. Teow. Combining local and non-local information with dual decomposition for named entity recognition from text. In *Proceedings of 15th International Conference on Information Fusion*, 2012.

[2] D. DeCaprio, J. P. Vinson, M. D. Pearson, P. Montgomery, M. Doherty, and J. E. Galagan. Conrad: Gene prediction using conditional random fields. *Genome Res*, 17:1389–1398, 2007.

[3] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's*

Mechanical Turk, pages 80–88, Los Angeles, California, 1992.

[4] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, Nantes, France, August 23-28 1992.

[5] J.Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society.B*, 36(2):192–236, 1974.

[6] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinform*, 9(Suppl 3), 2008.

[7] J. Kazama and K. Torisawa. Exploiting wikipedia as external knowledge for named entity recognition. In *EMNLP*, 2007.

[8] V. Krishnan and C. D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1121–1128, Sydney, Australia, July 17-18 2006.

[9] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001.

[10] 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.

[11] X. Mao, W. Xu, Y. Dong, S. He, and H. Wang. Using non-local features to improve named entity recognition recall. In *The 21st Pacific Asia Conference on Language, Information and Computation*, Seoul, Korea, 2007.

[12] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields. In *feature induction and web-enhanced lexicons, Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191, Edmonton, Canada, 2003.

[13] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, pages 30:3–26, 2007.

[14] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *CoNLL*, pages 147–155, 2009.

[15] A. Ritter, M. S. Clark, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1524–1534, Edinburgh, Scotland, 2011.

[16] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 134–141, Edmonton, Canada, 2003.

[17] C. C. Shih, T. C. Peng, and W. S. Lai. Mining the blogosphere to generate local cuisine hotspots for mobile map service. In *Fourth International Conference on Digital Information Management*, page 151–158, 2009.

[18] K. Shinzato, S. Sekine, N. Yoshinaga, and K. Torisawa. Constructing dictionaries for named entity recognition on specific domains from the web. In *Web Content Mining with Human Language Technologies Workshop on the 5th International Semantic Web*, 2006.

[19] K. S. E. F. Tjong and F. D. Meulder. Introduction to the conll- 2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2003*, page 142–147. W. Daelemans and M. Osborne, Eds. Edmonton, Canada, 2003.

[20] 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.

[21] O. Vechtomova. A semi-supervised approach to extracting multiword entity names from user reviews. In *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search*, pages 1–6, 2012.

[22] K. Yoshida and J. Tsujii. Reranking for biomedical named-entity recognition. In *Workshop: Biological translational and clinical language processing*, page 209–216, 2007.

[23] H. Zhang, Q. Liu, H. Yu, X. Cheng, and S. Bai. Chinese named entity recognition using role model. *the International Journal of Computational Linguistics and Chinese Language Processing*, 8(2):29–60, 2003.

[24] W. Zhang, T. Yoshida, X. Tang, and T.-B. Ho. Improving effectiveness of mutual information for substantival multiword expression extraction. *Expert Systems with Applications: An International Journal*, 36(8):10919–10930, October 2009.