Ruchi: Rating Individual Food Items in Restaurant Reviews

Burusothman Ahiladas, Paraneetharan Saravanaperumal, Sanjith Balachandran, Thamayanthy Sripalan and **Surangika Ranathunga**
Department of Computer Science and Engineering
University of Moratuwa, Katubedda 10400, Sri Lanka
{brusoth.10,parane.10,sanjith.10,thamayanthy.10,surangika}@cse.mrt.ac.lk

Abstract

Restaurant recommendation systems are capable of recommending restaurants based on various aspects such as location, facilities and price range. There exists some research that implements restaurant recommendation systems, as well as some famous online recommendation systems such as Yelp. However, automatically rating individual food items of a restaurant based on online customer reviews is an area that has not received much attention. This paper presents Ruchi, a system capable of rating individual food items in restaurants. Ruchi makes use of Named Entity Recognition (NER) techniques to identify food names in restaurant reviews. Typed dependency technique is used to identify opinions associated with different food names in a single sentence, thus it was possible to carry out entity-level sentiment analysis to rate individual food items instead of sentence-level sentiment analysis as done by previous research.

1 Introduction

Today, many factors affect a person's selection of a particular restaurant to dine in. People can find information on factors such as price, wifi and service from a restaurant's website and/or brochures. However, information about some important factors is not directly available. Ratings of individual dishes in restaurants is one such factor that is not directly available. Therefore it is common nowadays for people to rely on the reviews and ratings in restaurant review sites given by other customers. However, reading each customer review on restaurant review sites is time consuming, boring and exhaustive. This becomes more complex if someone wishes to search for a particular food item that he is interested in.

Existing restaurant recommendation systems such as Yelp do not have the facility to rate the individual food items of a restaurant. Moreover, as far as we are aware, existing research on restaurant recommendation systems has not focused on rating individual food items of restaurants, except for the work of Trevisiol et al. (2014).

This paper presents Ruchi (Ruchi means taste in Sinhala and Tamil, the two local languages in Sri Lanka), which is a system for rating individual food items (both food and beverages) in restaurants by automatically analyzing customer reviews. It combines the techniques of machine learning, natural language processing and information retrieval.

In order to rate and recommend individual food items, it was necessary to identify food names in customer reviews, and the customer opinions associated with them. Food names in reviews are identified using a trained NER model. For this purpose, a corpus[1] created from online customer reviews for restaurants was automatically annotated with food names extracted from various sources. As far as we are aware, this is the only freely available comprehensive corpus annotated with food names. Opinions related to these identified food items are extracted using a typed dependency parser. We then perform entity-level sentiment analysis to find the polarity of these opinions. Finally individual food items are rated based on the polarities of all the opinions received for each of these food items.

The rest of the paper is organized as follows. Next section discusses related work. Section three gives an overview of sentiment analysis. Section four discusses the data collection process. Section five discusses research and development work. Section six contains evaluation and discussion, and finally section seven concludes the paper.

[1] https://raw.githubusercontent.com/brusoth09/Ruchi/master/res/review_train

2 Related Work

Recommendation systems are available for many domains, including online business, specific products, restaurants and movies. As for restaurant recommendation systems, a very popular recommendation system is Yelp[2]. In Yelp, restaurant profiles are rated using a large data set with customer ratings and reviews. Yelp is capable of recommending the best restaurants, but not individual food items. Tripadvisor[3] also provides restaurant recommendations. It recommends a set of restaurants in a country that have famous food items, however these food items are not rated.

Snyder and Barzilay (2007) used the good grief algorithm to rate multiple aspects in restaurants. In their approach, each review is given a rating of 1 to 5 for five different aspects in a restaurant review: food, service, ambiance, value, and overall experience. But they did not rate the individual food items. Gupta et al. (2015) also discussed about sentiment based summarizing of restaurant reviews based on three aspects: food, ambiance and service.

As far as we are aware, the work by Trevisiol et al. (2014) is the only research that focused on rating individual food items using customer reviews. Their BuonAppetito system is capable of recommending personalized menus in a restaurant. In this system, a menu is considered to be comprised of food items. Food items are rated based on the customer opinions in the customer text reviews. The main difference between their approach and our approach is that they have carried out sentence-level sentiment analysis. In contrast, we carry out entity-level sentiment analysis. This is because one sentence of a review can contain more than one food item and associated opinions. For example, consider the sentence "Pizza was tasty but pasta was terrible". Here two food items are mentioned in one sentence - one has a positive opinion and the other one has a negative opinion. Therefore, if sentence-level sentiment analysis was performed, the overall review will be neutral. Despite the fact that we used a data set different to what was used by Trevisiol et al., we note that our recommendation system achieved a better precision and F1 measure than what was received by Trevisiol et al.

[2] http://www.yelp.com/

[3] http://www.tripadvisor.com

3 Sentiment Analysis

Sentiment analysis is a natural language processing technique that involves collecting and categorizing opinions (Liu, 2010). Sentiment analysis (or classification) can be done at different levels.

In document-level sentiment analysis, a whole opinion document is classified as a positive or negative sentiment (Liu, 2010; Liu, 2012). Document-level sentiment analysis assumes that each document expresses opinion on only one entity. More fine-grained analysis can be done using sentence-level sentiment analysis. In this level, each sentence is classified as positive, negative, or neutral. However, sentence-level sentiment analysis is not capable of handling cases where a single sentence contains opinions on multiple entities. Thus entity and aspect-level analysis can be done to obtain better insight to customer opinions. At aspect-level, opinions on multiple aspects (e.g. food, service, ambiance, value and overall experience aspects of the restaurant entity, or the size and color aspects of a mobile phone entity) are analyzed.

In this research, different food items could be considered as different aspects of the food entity in restaurants. Note that food is considered an entity here, rather than an aspect of a restaurant, as considered by Snyder and Barzilay (2007), and Gupta et al. (2015). However, we see that using the term aspect-level sentiment analysis is slightly misleading in our context, because in its true sense, a food item is not really an aspect of food, as opposed to color being an aspect of a mobile phone. Rather, in our context, we see a food item as a sub-entity of the food entity. Therefore in this paper, the term entity-level sentiment analysis is used.

4 Data Collection

Data collection had two aspects - collecting restaurant reviews and collecting food names.

Multiple sources for review collection were identified. These are Yelp, CityGrid[4] and tasty.lk[5]. One of the biggest issues with customer review system is opinion spam (Ott et al., 2013). Fake reviews can lead to false conclusions. We only used Yelp data source in our final system because Yelp has its very own spam filtering mechanism.

Food names were collected from the A-Z of Food and Drink dictionary published by Oxford

[4] http://www.citygrid.com/

[5] tasty.lk

University (food Dictionary, 2015), food timeline (food list, 2015) and the Oregon State Glossary of food items (state food list, 2015).

5 Rating Individual Food Items in Restaurant Reviews

Figure 1 shows our overall approach for rating food items based on customer reviews. This has four main steps: extracting food names from customer reviews, associating opinions with each food name in a review, calculating the sentiment value for the given opinion, and finally rating the food item using all the sentiment scores recorded for it.

A NER system proceeded by a pre-processing step was used to extract food names from customer reviews. This NER system is particularly trained for food domain using our food list. Opinion word associated with each food item is determined using a typed dependency parser, and phrases that contain only one subject (i.e. a food item) are created. Sentiment analysis tool in the StanfordNLP toolkit is used for sentiment analysis. We used a modified version of Suresh et al. (2014)'s ranking algorithm for rating food items using the sentiment scores. Calculated ratings for the individual food items are finally saved in a persistent storage.

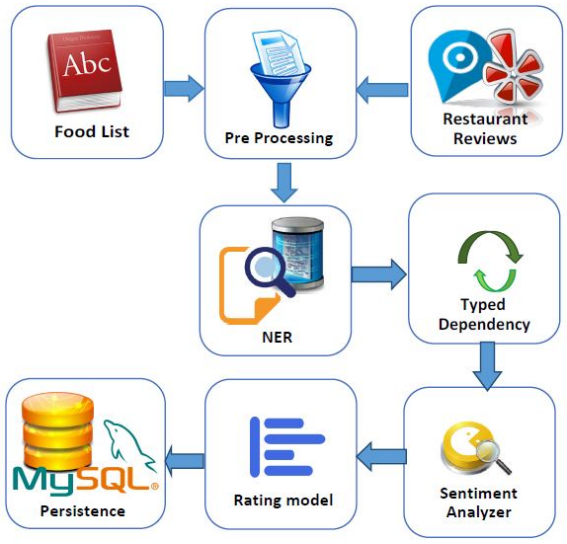

Figure 1: Overall system architecture

5.1 Pre-processing

In the pre-processing module, we focused on getting various data sources to a usable format for the NER module. Stemming, language detection, and symbol removing are these pre-processing steps.

5.2 Food Name Extraction

Named Entity Recognition (NER) was used for food name extraction. An NER model trained for the food name domain can be used to extract food names from sentences without explicitly searching for word tokens. POS tagging can be used to get the noun phrase from sentences, so that while training the NER, we can search only for noun phrases to tag food names.

Apache OpenNLP machine learning toolkit [6] was used for NER purpose. OpenNLP toolkit has NameFinder API for NER. It uses Maximum Entropy principle to classify entities using a pre-trained data model. Tokenized sentence should be given as input to NameFinder to predict categorized entities.

To make use of NER to identify food names in restaurant reviews, a corpus annotated with food names was required. Since such corpus was not available, we had to create one. We automated the process of creating a corpus by using the food names we collected from various sources. Figure 2 shows the process of creating this annotated corpus. We picked 150,000 reviews to create this corpus. From these reviews, about 300,000 sentences contained food names and thus got automatically annotated.

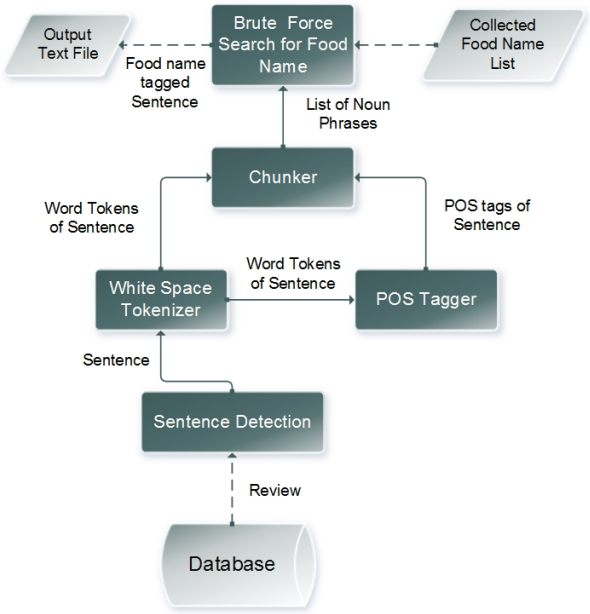

Figure 2: Process of creating the annotated corpus

First, each review was broken into its constituent sentences. Sentence detector in the

[6] https://opennlp.apache.org/

OpenNLP toolkit was used to break reviews into sentences using punctuation characters. Then, each of these sentences was tokenized and POS tagged. The whitespace tokenizer and the POS tagger (respectively) in the OpenNLP toolkit were used for these tasks. The POS annotated tokens were then sent to the chunker, which combines these tokens into syntactically correlated parts of words, such as noun groups and verb groups. This was required since some food items have more than one token. Finally, the noun groups are checked against the food name list we have prepared, and matching noun phrases are tagged with a unique tag ($< START : food > food\_name < END >$) to identify food item names.

Now this annotated corpus could be used for NER. However, sometimes the output was just a part of the actual food name (e.g.: pepperoni pizza was identified when the actual name was Italian pepperoni pizza). This was because the dish names of most of the restaurants are not just the standard food names included in our food list.

A post-processing technique was used to eliminate this problem. This post-processing step is based on the observation that the common features (food names) come as noun phrases. So we first picked up the noun phrases from each sentence using POS tags and checked each phrase with our NER predicted food names. This process considers noun phrases around any food item as part of that food name. If a phrase contains the predicted name, then it will be considered as a food name. It is equally possible to apply this post-processing technique while annotating the corpus. However, we decided against it in order to make our corpus as general as possible, in order to use it for food name detection in a different type of application, say identifying food names in a supermarket context.

5.3 Sentiment Analysis

In our system, we are rating individual food items, therefore sentiment extraction is done at entity-level. Sentences may contain several subjects with different opinions. Stanford typed dependency representation (De Marneffe and Manning, 2008) is used to find the opinion associated with each food item in a sentence, and to create phrases that contain only one subject (i.e., a food name).

Many researchers have mentioned that opinion words are usually adjective or adverb. Gupta et al. (2015) have used two grammatical relations - amod and nsubj, to determine the noun that an adjective modifies. amod, short for adjectival modifier is any adjectival phrase that serves to modify the meaning of the noun phrase. nsubj, short for nominal subject is a noun phrase, which is the syntactic subject of a clause. In nsubj relation, there is a possibility that both words involved are nouns so we have to check for the presence of adjective in the relation. Other than these two grammatical relations, we also used advmod, which is the short form for adjectival modifier. This is because adverbs can also refer to opinion words.

First, opinion words are identified. Then other words that have grammatical relationship with food and opinion words are identified. After identifying all the words, we create opinion phrases containing only one subject.

Once the opinion phrases are identified, their sentiment orientation is determined by the sentiment analysis tool. Sentiment analysis tool in the StanfordNLP toolkit was used for this purpose. Sentiment analysis tool in the StanfordNLP toolkit uses deep learning technique. This technique uses a recursive neural sensor network to compute compositional vector representations for phrases of variable length and syntactic type. These representations will then be used as features to classify each phrase.

Other than the deep learning method in the StanfordNLP sentiment analysis tool, we also experimented with few other machine learning techniques (multilayer perceptron neural network, Support Vector Machine (SVM), PART, REPtree, Random Forest and J48) to determine the sentiment polarity of phrases. Since our system is for restaurants, restaurant reviews were used to train these algorithms. 1150 reviews were manually tagged according to Stanford Sentiment Treebank and were used to train the model.

StanfordNLP sentiment scores are: 4 - Very Positive, 3 - Positive, 2 - Neutral, 1 - Negative, 0 - Very Negative.

In the Treebank representation, training data structure is a binary tree. In the StanfordNLP sentiment analysis process, first the individual words are assigned a sentiment score. Then two words are combined and a single score is assigned to the combined words. This process continues combining word with word, phrase with word and phrase

with phrase. Finally a sentiment score for the complete opinion phrase can be obtained.

5.4 Rating System

Rating system aims at scoring the food items in a restaurant based on customer reviews, using their sentiment weight. In order to rate a particular food item, sentiment weights assigned to it across all of the reviews should be considered.

In order to rate the food items using weak and strong positive and negative words, we first built a subjectivity lexicon. A subjectivity lexicon is a list of positive or negative opinion words. We created a master list that contains this subjectivity lexicon (each word in the lexicon has a sentiment weight), an intensifier word list (really, very, too, such etc.), and a negation word list (no, not etc.). We prepared the subjectivity lexicon from AFINN-111 word list[7] and the restaurant reviews that were not used to test the system. The POS tagged reviews are fed into our rating algorithm.

Our rating algorithm is an extension of Suresh et al. (2014)'s opinion score assignment algorithm. Suresh's algorithm focused on word-level scoring. However, in our approach, we focused on sentence-level scoring since word-level scoring mostly relies on sentiment weight of individual words and it fails to calculate rate for complex sentences. For example, if you consider the sentence "pizza was good, not service", this sentence has 2 opinion phrase and word-level score gives a wrong rating for pizza.

The rating algorithm takes polarity tagged phrases as input and provides a scoring value depending on the polarity of the phrase. If the word is POS tagged as an adverb or an adjective, it is considered as an opinion word. If the word appeared in a master list where all possible polarity words are classified according to sentiment weight, the score is calculated according to the sentiment weight of the word.

In the next step, if the opinion word is POS-tagged as a superlative sentiment, the score is increased or decreased by 2. If the opinion word is POS-tagged as comparative sentiment, the score is increased or decreased by 1. Words that modify the polarity (using negation word e.g. no) and intensifiers (e.g. too, very) are also considered for scoring the opinion word. Final score converges to a value between 1 to 5 according to the sentiment score.

[7] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

6 Results

Experiments were carried out to validate our (1) corpus creation process, (2) food name extraction, (3) sentiment analysis, and (4) the overall process.

Corpus creation process: In order to validate our corpus creation, a sample set of reviews containing 1000 sentences was randomly selected from the tagged corpus. Then these sentences were manually inspected to see how the tagging process has performed. This manual investigation identified 1898 occurrences of food names in these 1000 sentences. Out of these 1898 occurrences, 71.97% food names have been correctly tagged. It was noticed that sometimes only a part of a long food name was tagged.

Food name extraction: We used 1219 review sentences with manually tagged food names to evaluate the NER approaches. We achieved 63.2% precision, and 83.5% recall by using Maximum Entropy model technique of OpenNLP toolkit.

Sentiment analysis: We used 1150 sentiment sentences hand selected from restaurant reviews that included food names, and sentiments of each of these sentences were manually tagged. Our sentiment evaluation results obtained for different machine learning techniques are summarized in Table 1. However, after training the sentiment model for the restaurant context using the deep learning technique of StanfordNLP toolkit, we achieved 85.74109% accuracy, 82.9684% recall, 98.2708% precision and 89.9736% F1-measure. Deep learning technique is very promising. Therefore we used deep learning in our final system. For this experiment, our sentiment classification demonstrated an improvement of 6.7% over StanfordNLP baseline. We were able to achieve this because we used a trained model containing food item names.

Overall process: For the overall evaluation of the system, reviews were manually tagged with ratings. For each sentence in a review, all the food names and the corresponding opinion were tagged by human annotators. Whether the opinion rate is 1 to 5 (very negative to very positive) was also identified.

It is easy to judge whether an opinion of the sentence is positive or negative. However, deciding the opinion rate (score) can be somewhat subjec-

Table 1: Machine learning algorithm result for sentiment classification

Algorithm	precision	recall	F1
NeuralNetwork	0.8072	0.7867	0.7665
SVM(SMO)	0.7798	0.7205	0.6634
PART	0.8047	0.8014	0.7914
DecisionTable	0.8312	0.8161	0.8031
J48	0.8014	0.8047	0.7914
REPTree	0.8068	0.7941	0.7783
RandomForest	0.7520	0.7573	0.7492
RandomTree	0.7523	0.7573	0.7527
Naive Bayes	0.8222	0.8235	0.8226

tive. In order to validate our human tagged rating for sentence opinion rate, we carried out an inter-rater reliability (IRR) test. Each sentence was given to a primary human tagger (participant) and the secondary tagger (one of the authors). Final rate value was calculated using joint probability of agreement, and we received a joint probability of agreement value of 75%. Finally, all the results generated by our system are compared with the manually tagged result. Result of our final system is evaluated with respect to precision and F1 measure using 10-fold cross validation. The average precision of our recommendation system was 0.4177 and F1 measure was 0.4518.

7 Conclusion

This paper presented Ruchi, a system capable of rating individual food items. Ruchi makes use of NER for extracting the food item names from reviews, and typed dependency representation to identify the customer opinions. A corpus created from restaurant reviews was automatically tagged to be used in NER, using a list of food names compiled from various resources. This automated approach proved to be effective in creating a large corpus tagged with food names, as opposed to corpora manually tagged (Yasavur et al., 2013).

As for future work, we are planning to modify our system to be able to carry out time-based food rating. This feature will give the rating based on the reviews that were written within a preferred time period and avoid giving false rating to food items based on very old reviews.

References

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Oxford food Dictionary. 2015. Oxford food dictionary. http://www.oxfordreference.com/view/10.1093/acref/9780192803511.001.0001/acref-9780192803511. Accessed: 2015-01-04.

Foodtimeline food list. 2015. Foodtimeline food list. http://www.foodtimeline.org/foodfaqindex.html. Accessed: 2015-01-04.

Abhishek Gupta, Tejaswi Tenneti, and Ankit Gupta. 2015. /sentiment based summarization of restaurant reviews. http://nlp.stanford.edu/courses/cs224n/2009/fp/9.pdf. Accessed: 2015-01-30.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference*, pages 497–501.

Benjamin Snyder and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference*, pages 300–307.

Oregon state food list. 2015. Oregon state food list. http://health.oregonstate.edu/food/glossary/index.html. Accessed: 2015-01-04.

Vaishak Suresh, Syeda Roohi, Magdalini Eirinaki, and Iraklis Varlamis. 2014. Using social data for personalizing review rankings. In *Proceedings of the 6th Workshop on Recommender Systems and the Social Web*.

Michele Trevisiol, Luca Chiarandini, and Ricardo Baeza-Yates. 2014. Buon appetito: recommending personalized menus. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 327–329. ACM.

Ugan Yasavur, Reza Amini, Christine L Lisetti, and Naphtali Rishe. 2013. Ontology-based named entity recognizer for behavioral health. In *27th International Conference of the Florida Artificial Intelligence Research Society (2013)*. AAAI press.