Restaurant Information Extraction (Including Opinion Mining Elements) for the Recommendation System

Ekaterina Pronoza, Elena Yagunova, Svetlana Volskaya, and Andrey Lyashin

Saint-Petersburg State University, Saint-Petersburg, Russian Federation
{katpronoza,iagounova.elena,svetlana.volskaya}@gmail.com
Scicon Ltd, Saint-Petersburg, Russian Federation
andrey.lyashin@scicon.com

Abstract. In this paper information extraction method for the restaurant recommendation system is proposed. We aim at the development of an information extraction (IE) system which is intended to be a module of the recommendation system. The IE system is to gather information about different aspects of restaurants from online reviews, structure it and feed the recommendation module with the obtained data. The analyzed frames include service and food quality, cuisine, price level, noise level, etc. In this paper service quality, cuisine type and food quality are considered. As part of corpus preprocessing phase, a method for Russian reviews corpus analysis (as part of information extraction) is proposed. Its importance is shown at the experimental phase, when the application of machine learning techniques to aspects extraction is analyzed. It is shown that the ideas obtained at the corpus preprocessing stage can help to improve machine learning models performance.

Keywords: corpus analysis, restaurant reviews, information extraction, recommendation system, machine learning.

1 Introduction

In this paper information extraction (IE) method for the Russian restaurant recommendation system is proposed. It is based on the application of linguistic information gathered from corpus analysis and can be used for similar domains and under-resourced languages. Our information extraction framework is a part of the project which aims at implementing restaurants recommendation system, and in this paper we consider two tasks: reviews corpus analysis and the application of machine learning techniques to the problem in question. During the latter task we use the information obtained at the corpus analysis phase. Our approach includes opinion mining since restaurant characteristics are both objective and subjective.

Our corpus analysis method is based on non-contiguous bigrams and part of speech (POS) distribution analysis. Trigger words dictionaries are learnt using the bootstrapping method.
The frames to be extracted include service quality, food quality, cuisine type, price level, noise level, etc. Each frame has its own set of aspects. We suppose that the most important characteristics of a restaurant are service and food quality and cuisine type and therefore we only consider these three frames and focus on the extraction of their aspects. Such an assumption is proved by the distribution of the aspects in the data.

We also suppose that the proposed IE system can be highly effective despite the difficulties imposed by the structure of a typical Russian restaurant review. Although the key information about restaurant characteristics does not always lie on the surface, tuning machine learning models according to the results of corpus analysis can help to improve the performance of an IE system.

2 Related Work

Information extraction (IE) task as part of recommendation system development is discussed in [21]. The authors propose a rule-based approach to the extraction of key words from user’s email. These keywords are put into a car recommendation system which is using content-based filtering approach with Jaccard Coefficient method.

In [16] a tag extraction algorithm for web pages entering a recommendation system is presented. The algorithm is based on semi-supervised document classification.

An approach to key words extraction for a learning recommendation system is described in [11]. The algorithm employs TF-IDF measure and a combination of collaborative filtering (see [27], [31]) and content-based filtering strategies (see [26], [27], [34]).

The bootstrapping approach has been successfully applied in web page classification [17], text classification ([13], [18]), named entity classification ([7], [15], [19], [22], parsing ([28], [36]) and IE ([6], [12], [29], [37]).

As some of restaurant aspects are quite subjective and represent users’ opinion or sentiment, our task becomes similar to sentiment classification. In our research we experiment with such classification models as Naive Bayes, Logistic Regression and Linear Support Vector Machines. These models are commonly used both in sentiment analysis and opinion mining ([2], [4], [10], [14], [20], [23], [30], [35], [39]), and typical tasks usually include classifying movie or product reviews or attitudes toward a political figure/event (in a blog, tweet, etc.) into positive and negative ones.

When it comes to applying machine learning techniques to sentiment classification, it is crucially important to identify relevant features, and we pay special attention to it. In sentiment classification n-grams (contiguous ones) are certainly the most commonly used features ([2], [10], [23], [38], [39]). Sometimes words in n-grams are substituted with their semantic classes [10].