Named Entity Recognition for Food Logging System

Hyun Gi Ahn (2014-31117)
Bio & Health Informatics Lab, Dept. of Computer Science & Engineering, Seoul National University
puppybit@gmail.com

Abstract

In Natural Language Processing, named-entity recognition and classification (NERC) is a task of information extraction that seeks to locate and classify elements in text into pre-defined categories. In the project we are going to implement Food Logging system using NER algorithm for Korean. Food Logging system consists of following elements. At first, the system must categorize whether the intent of user utterance is for food logging or not. Secondly, the system recognize food related words from classifed sentence and extract these words. Finally extracted words are classifed into 3 kinds of group - Food, Value of Unit, Unit. We'll use word vectorization algorithm for feature extraction method. In real world it is difficult to get lots of good annotated data. So, in order to annotate training data, we try to use unsupervised learning based on seed data which are related with food, value of unit, unit.

1 Introduction

People who want to live longer and healthier are more interested in diet. There are two factors that need to manage weight: exercise and diet. There are many smartphone apps that help people track their workouts and meals. However, in general, the food input method is very inconvenient. At first the meal time must be entered. Then search for food, insert the quantity, and then repeat the selection of food. In this project, we want to develop a text based food input system. The main idea of food logging system is Named Entity Recognition and Classification. The term "Named Entity", now widely used in Natural Language Processing, was coined for the Sixth Message Understanding Conference (MUC-6) (R. Grishman & Sundheim 1996). At that time, MUC was focusing on Information Extraction (IE) tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In defining the task, people noticed that it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Identifying references to these entities in text was recognized as one of the important sub-tasks of IE and was called "Named Entity Recognition and Classification (NERC)".

2 Overview of project idea

2.1 Preprocessing of potential data sets for food logging system

In order to extract food-related words from sentences, special notations are required for each food-related word in each sentence. However, it is very difficult to hand-write a lot of data. So, we use unsupervised learning algorithm to tag food related words. After clustering similar words based on pre-prepared food-related DBs and we try to tag all the words that included in the cluster as [Food]. In case of [Unit] and [Value of Unit] the same method applies. Then finally [B][[O][I] is tagged for all words based on [Food][Unit][Value of Unit] tags. The data needed for clustering will be obtained from namu wiki, twitter, and blog for recipe. Especially because food names have many proper nouns, the data from the cooking blog is very important.

2.2 Intention of food logging

Even if the name of the food and related words are extracted, the food logging system can not be

operated unless the intention of the utterance is food logging. Therefore, it is necessary to classify whether the utterance is a storage or recording intention. After generating a variety of sentences containing the intent of saving or recording as regular expressions, we try to find similar sentences from the training data by using clustering algorithm. A classifier for intention of food loging is constructed by using this classified data. The reason for not using directly regular expression data is to create a General model while avoiding over fitting.

2.3 The method for calulating similarity in unsupervised learning

In order to calculate similarity among words with L2 distance etc., we'll use word embedding model. Sentences are used as inputs for learning algorithm. Representation of words in the sentence is via the form of embeddings. Hence the features for learning algorithm are sentences a.k.a sequence of words a.k.a sequence of embeddings. Each unique word should have certain number of features, these are called embeddings or also vectors. These are the input features to the neural architecture we are using. Word2vec was originally conceived by Tomas Mikolov' team at Google, it provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. Word2vec is a neural network which learns word's meaning by running deep learning algorithms, by feeding huge training dataset, it can make highly accurate guesses about the words' meaning, and clusters words by meaning. It outputs uniform length vector, as the representation of the input word, so it also perfectly solved the "different length" problem.

2.4 Named entity recognition algorithm

NER tools and frameworks implement a broad spectrum of approaches, which can be subdivided into three main categories: dictionary-based, rule-based and machinelearning approaches. The first systems for NER implemented dictionary-based approaches, which relied on a list of named entities (NEs) and tried to identify these in text. Following work then showed that these approaches did not perform well for NER tasks such as recognizing proper names. Thus, rule-based approaches were introduced. These approaches rely on handcrafted rules to recognize NEs. Most rule-based approaches combine dictionary and rule-based algorithms to extend the list of known entities. Nowadays, hand-crafted rules for recognizing NEs are usually implemented when no training examples are available for the domain or language to process. When training examples are available, the methods of choice are borrowed from supervised machine learning. Approaches such as Hidden Markov Models, Maximum Entropy Models and Conditional Random Fields have been applied to the NER task. Due to scarcity of large training corpora as necessitated by supervised machine learning approaches, the semi-supervised and unsupervised machine learning paradigms have also been used for extracting NER from text. Recently a system was presented that combines with stacking and voting classifiers which were trained with several languages, for language-independent NER. Many different classifier types have been used to perform machine-learned NER, with conditional random fields being a typical choice. The current state of the art approaches are Multi-task learning based models and Residual stack bidirectional LSTM with conditional random fields.

2.5 Challenge point

In order to extract the food name, the amount of food consumed and its unit, the following should be considered.

- "오늘 아침에 사과 바나나를 2개씩 먹었어" : We should understand that user ate two apples and two bananas each.
- "점심에 딸기 우유 1잔을 마셨어" : We have to distinguish between "strawberry and milk" or "strawberry milk".
- "저녁에 쌀밥 대신 잡곡밥 반 그릇 먹었어" : It is necessary to understand the negation expression to enable accurate tagging.

We will use the parts - of - speech information of each word in proposed algorithm and design it considering various edge cases.

2.6 For demo system

We are goint to implement Food Logging Chatbot with Facebook messenger. If user enter the utterance with Facebook messenger, chatbot reply its parsed output quickly and store its data into server.

3 Plan of activities

1. 2017-10-22 : Secure training data with web scraping
2. 2017-11-05 : preprocessing training data with clustering
3. 2017-11-24 : Implementing NERC algorithm and Demo system
4. 2017-11-30 : Performance Evaluation and Documentation