Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 47 (2018) 16–27

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Question classification in Persian using word vectors and frequencies

Action editor: Ali Minai

Mohammad Razzaghnoori [a], Hedieh Sajedi [a,*], Iman Khani Jazani [b]

[a] Dept. of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Iran
[b] Dept. of Computer Engineering and Information Technology, AmirKabir University of Technology, Iran

Received 2 January 2017; received in revised form 20 May 2017; accepted 16 July 2017
Available online 22 July 2017

Abstract

The necessity of the existence of Question Answering (QA) systems becomes evident by considering the fact that the enormous amount of unstructured data created by humans nowadays, results in ineffectiveness of search engines to provide the exact solution for a given question. However, an outstanding question answering system requires an outstanding Question Classification (QC) system. Question classifier is a system that assigns a label to each question. There exist different ways of solving this problem such as rule-based, machine learning, and hybrid approaches. This paper provides a better solution for QC using machine-learning approaches. Three methods of feature extraction are proposed in this paper. The First method uses clustering algorithms to partition vocabulary into clusters and acquires feature vector corresponding to each question using clustering information. The second one suggests a method of extracting features from questions to dispose of using recurrent neural networks and to use feedforward neural networks, which have the advantage of learning faster and less need for data, instead. Each question is converted to a feature vector, which is obtained by the Word2vec method and weighted by tf-idf coefficients. The results of question classification using Support Vector Machine and Neural Network classifiers indicate the effectiveness of this type of feature vector and based on that, high performance of the proposed QC system. Finally, the third approach keeps the innovation behind first approach, but it also keeps the fact that we are dealing with a sequence based type of data into consideration. Eventually, it would be concluded that even with a limited amount of data it is reasonable to take Recurrent Neural Networks into consideration.

Keywords: Question classification; Word2vec; Tf-idf; Feedforward neural networks; Recurrent Neural Networks (RNN); LSTM

1. Introduction

Nowadays, search engines receive keywords and return some relevant and irrelevant pages to the users, whereas QA systems were designed to get a natural language question or query and retrieve more probable and appropriate answers. An appropriate answer has some characteristics such as being concise, comprehensible, and correct. These answers can be reported as a word, sentence, paragraph, image, audio fragment, or an entire document (Kolomiyets & Moens, 2011). The focus of current QA systems is on types of questions generally asked by users.

One of the most important modules of a QA system is question classification, which is the task of assigning a label to an input question that has been addressed before, using rule-based, machine learning and hybrid techniques. Rule-based approaches are implemented by designer defining some rules (Huang, Thint, & Qin, 2008) often based on the assumption that natural language can be controlled by finite definitive rules. However, this assumption is almost always not completely true (Metzler & Bruce Croft, 2005); in addition, the process of defining these rules

* Corresponding author.
E-mail addresses: m.razzaghnoori@ut.ac.ir (M. Razzaghnoori), hhsajedi@ut.ac.ir (H. Sajedi), imankhanijazani@aut.ac.ir (I.K. Jazani).

http://dx.doi.org/10.1016/j.cogsys.2017.07.002
1389-0417/

are often time-consuming (Sherkat & Farhoodi, 2014) and does not worth the effort due to the poor performance of the final system in comparison with machine-learning-based systems (Mikolov, Kombrink, Burget, Černocký, & Khudanpur, 2011). On the other hand, machine learning suggests a better way by learning the data without making any assumption about the structure of the language (Zhang & Lee, 2003). Therefore, the whole system can be changed easily to be ported to another language.

Previously, a popular method of extracting features from questions was bag-of-words (Laokulrat, 2016), yet this method has serious downsides. For example, if this kind of systems receive a training question containing the word "cat" and then asked to classify the very same question with the word "cat" being replaced by "kitten" misclassification would be probable. To overcome this problem we need a mathematical representation of words like function f to assign each word x to a vector $f(x)$ such that if x and y are semantically or syntactically close then $f(x)$ and $f(y)$ are close vectors. A novel representation of words with aforementioned feature was introduced by Mikolov et al. named Word2vec (Mikolov & Dean, 2013) which is applied in this paper in feature extraction phase. The idea of Word2vec method is that words which are semantically or syntactically close occur in the same contexts with high probability. Therefore, if words w_1 and w_2 were seen in the same context, then their vectors should get a little bit closer to each other.

First method introduced in this paper uses clustering algorithms to cluster words in the vocabulary and convert each question into a sparse vector.

Second method considers each feature vector as a linear combination of vectors of the words of the question. For extracting vector of words, the Word2vec method is employed. Afterwards, we set the coefficients of previously mentioned linear combination using the tf-idf method.

Then for classification, we use Multi Layered Perception (MLP) and Support Vector Machines (SVM). By applying these two proposed method, an average accuracy of 72.46% using MLP and 72% using SVM over three question datasets was achieved.

Our third proposed method converts each question to a matrix in which each row represents word2vec representation of a word. Later, a LSTM network is used for the purpose of classification which in our experiments led to an average accuracy of 81.77% over three question datasets.

Furthermore, in this paper, we introduce University of Tehran Question Dataset 2016 (UTQD.2016) which was gathered from some sort of jeopardy game hosted by Iran's official television.

The rest of the paper is organized as follows: In Section 2 we briefly review related works that have been done in the field of question classification both in Persian and in other languages. Section 3 describes the Word2vec method. In Section 4, the proposed methods will be introduced. In this section, we describe our feature extraction methods, learning procedure and details of the databases. Further, the experimental results will be reported. Finally, the conclusion section ends the paper.

2. Related works

A typical pipeline Question Answering System consists of three distinct phases: Question Processing, Document Processing, and Answer Processing. Question Processing phase classifies user questions (also termed as question classification), derives expected answer types, extracts keywords, and reformulates a question into semantically equivalent multiple questions. Reformulation of a query into similar meaning queries is also known as query expansion and it boosts up the recall of the information retrieval system. The Document Processing phase retrieves documents containing keywords in the original as well as expanded questions, applies ranking algorithms on the retrieved document set and returns the top ranked documents. In Answer Processing phase, the system identifies the candidate answer sentences, validates the correctness of the answers, ranks them and finally presents the answers to the user using information extraction techniques (Ray, Singh, & Joshi, 2010).

Question classification as mentioned can be done using three main approaches: rule-based (manual), machine learning, and hybrid approaches. Manual question classification (Hermjakob, 2001) tries to match questions with handcrafted rules to identify question's answer type. Aside from the fact that writing these rules are tedious and the final system often suffers from being too specific (Metzler & Bruce Croft, 2005), the obvious reason why this type of system is rare is that the overall performance is not even close to machine learning based methods (Mikolov et al., 2011). For a discussion about machine learning methods outperforming manual methods see (Li, 2002). Machine learning on the other hand, suggests a relatively easier method to classify questions; instead of manually writing the rules, we can learn them from the data and as a result, this kind of systems can be adapted to new situations with minimum effort. Because of the fact that hybrid methods are new and not quite common, in the following a single hybrid approach will be reviewed.

Machine learning methods are quite diverse; some of them try to represent questions as a tree. For instance, Zhang and Lee (2003) proposed to use a kernel function named tree kernel to enable SVM to take advantage of syntactical structures of questions. Furthermore, another kernel named HDAG is proposed by Suzuki, Taira, Sasaki, and Maeda (2003) which directly accepts structured natural language data, such as several levels of chunks and their relations.

Blunsom, Kocik, and Curran (2006) used a maximum entropy model with a rich set of syntactic and semantic features. They also have observed that hierarchical classifier performs better in practice.

A two-layered taxonomy has been proposed by Li and Roth (2002), which contains six coarse-grained and 50