Abstract

With the explosive growth of Internet and computing technology, human beings are confronted by a great amount of unstructured text data. The need to extract useful knowledge from the data also grows. Researchers in the natural language processing community have delivered many marvellous technologies for various applications, such as information retrieval, machine translation, sentiment analysis, etc. Traditional methods usually rely on rigid language assumptions and require great efforts and time to be devoted to feature engineering. The research goal of this thesis is to develop machine learning models that can automatically learn semantic representations from texts with few or no human interventions. The models proposed in this thesis can induce effective representations for sentences or documents which are used to solve high-level language understanding tasks. The models are shown in four main chapters in this thesis according to the tasks they are addressing. The first task is document summarization which is addressed by two new approaches; after that, another two innovative algorithms are proposed for sentiment analysis and sentence modeling respectively; at last, one model is developed for human demography prediction. However, the models are never limited to these applications but can easily generalized to diverse natural language understanding tasks. The core of all the models lies in learning good semantic representations.

Document summarization is aimed at generating a brief summary for a long document or a set of documents. In this thesis, the task is transformed into a regression problem which ranks sentences by saliency scores. Methods are explored to represent sentences as vectors so as to obtain scores of sentences by a regressor. The first model leverages on word embedding to represent sentences so as to avoid the intensive labor of feature engineering. A new technique, termed window-based sentence representation, is proposed and achieves satisfactory summarization performance compared with baseline methods. However, the representation power is still weak because of its simple structure. To improve the representation capability, we employ deep learning algorithms and develop an innovative variant of the convolutional neural network, namely multi-view convolutional neural network which can obtain the features of sentences and rank sentences jointly. The performance of the new model is evaluated on five benchmark datasets and demonstrates better performance than the state-of-art approaches.

The second natural language understanding task addressed in this thesis is sentiment analysis which has been applied to recommender systems, business intelligence and automated trading, etc. A new architecture termed comprehensive attention recurrent model is developed to access comprehensive information contained in sentences. The model employs the recurrent neural network to capture the past and future context information and the convolutional neural network to access local information of words in a sentence. Empirical results on large-scale datasets demonstrate that the new architecture effectively improves the prediction performance compared with standard recurrent methods.

The sentence modelling problem is at the core of many natural language processing tasks whose main objective is to learn good representations for sentences. Actually the objective of the thesis is to learn good semantic representations for texts. Therefore, this task lies at core and is the foundation of the other three tasks addressed in this thesis. One innovative model combining the bidirectional long-term short memory and convolutional structures is developed for the problem. A new pooling scheme for the convolutional neural networks, which better retains significant information than the popular max pooling method, is proposed by leveraging on attention mechanism. The model achieves state-of-art performance on seven benchmark datasets for text classification.

At last, a simple but effective document representation approach is designed for predicting demographic attributes of web users based on the browsing history. I put this task at the last position because The task is a practical application of natural language understanding. The new representation approach exploits word embedding and term frequency and inverse document frequency weighting scheme and proves to be more powerful than other feature representation methods for this task.