Requirement analysis

Emoji Prediction

Description

The advent of social media has brought along a novel way of communication where meaning is composed by combining short text messages and visual enhancements, the so-called emojis. This visual language is as of now a de-facto standard for online communication, available not only in Twitter, but also in other large online platforms such as Facebook, Whatsapp, or Instagram.

Despite its status as language form, emojis have been so far scarcely studied from a Natural Language Processing (NLP) standpoint. Notable exceptions include studies focused on emojis' semantics and usage or sentiment. However, the interplay between text-based messages and emojis remains virtually unexplored. We aim to fill this gap by investigating the relation between words and emojis, studying the problem of predicting which emojis are evoked by textbased tweet messages.

Domain

We will describe the most used use case scenarios that will occur in the development of this project and after the development is done.

Interests

The extraction module: This module will select from a wide range of tweets (a few millions) and will remove hyperlinks from each tweet and lowercase all textual content in order to reduce noise and sparsity. From the dataset, we select the one which include one and only one of the 20 most frequent emojis, resulting in a final dataset.

The word representation module: Using recurrent neuronal networks we will model our tweets based on bi-directional Long Short-term Memory Networks and normalize it afterwards. Thus, we will have to compute the probability distribution of emojis given a message. We genereate word embeddings which are learned together with the updates to the model. We stochastically replace (with p = 0.5) each word that occurs only once in the training data with a fixed represenation (outof-vocabulary words vector). When we use pretrained word embeddings, these are concatenated with the learned vector representations obtaining a final representation for each word type.

The experiment and evaluation module: We divide each dataset in three parts, training (80%), development (10%) and testing (10%). The three subsets are selected in sequence starting from the oldest tweets and from the training set since automatic systems are usually trained on past tweets, and need to be robust to future topic variations. For evaluation, the classic Precision and Recall metrics over each emoji are used. The official results will be based on Macro F-score, as the fundamental idea of this

task is to encourage systems to perform well overall, which would inherently mean a better sensitivity to the use of emojis in general, rather than for instance overfitting a model to do well in the three or four most common emojis of the test data. Macro F-score can be defined as simply the average of the individual label-wise F-scores. We will also report Micro F-score for informative purposes.

Use case scenarios

1. **Finding out the most used emoji in a certain period of time, or a specific day**
   Scenario/Steps:
   1. The user will select a date in the calendar (the range is prepoluated based on the tweets retrieved at extraction)
   2. A signal is received that the user have made a new request
   3. The date is analyzed and using the algorithm it will retrieve the most used emoji in that period of time
   4. A response is generated and sent to the user.

2. **Predicting an emoji based on input text**
   Scenario/Steps:
   1. The user will type some text in the bar.
   2. A signal is received that the user have made a new request.
   3. Based on the algorithm and the scores that this text returns, and using the training data, an emoji is generated.
   4. A response is generated and sent to the user.