Sentence2Emoji - Project Proposal

Chaolong Tang, Kai Ye, Chang Zeng, Xuming Deng

October 18, 2021

1 Introduction

Emoji has been developed for over two decades and being popular ever since Apple added emoji keyboard on iPhone OS 5 in 2011.[1] Moreover, emoji started to become one of the most popular languages around the world. The "laugh out loud" 😂 face is officially the world's popular emoji, according to researchers from Adobe who surveyed 7,000 users all over the world. People can understand the emotion behind the message via emoji and the description of a particular word, which bring people from all over the world with different languages background much closer.dd

One idea starts to get to our mind - 'what if we can translate English to emoji?' Will this helping different language speakers easier to understand each other than a plain text message? Over 3000 emoji that has been released to the public, and by 2022 the total number of emoji would be 3460.[2] 3460 is a vast number. One emoji can be translated into a couple of synonym words and un-synonym words. For example, 👍 could be translated to "good", "agree with you" or "good job",etc.

If there is a relation between people's word or tweets and their emoji choice, we can build a model by learning those sentences. We could find out the best emoji choice to represent a word or the whole sentences. The final goal of our group project is to translate a English sentences to one or a sequence of emoji.

2 Applications or Examples

1. Sample with no emoji:

   Do you like Apollo How are beautiful. => ❤️ 🚀 ❓ 🥰

2. Sample with one emoji:

   7pm what E.A.T or W. A.T 🤔 => 🕖 🏞️ 🤔

3. Sample with multiple same emoji:

   Another day to remind you how LX 570s are the most beautiful beasts ever made by man 😍😍😍=> 🌟🏎️😍

4. Sample with multiple different emoji:

   Ted Virtue? 🤣😋😋😋😋😂😂😂😂😂 => 🧑💼 😂

3 Previous Paper

3.1 Emoji2vec

The author of the first paper believe that emojis usage is prevalent and most of them inherently have emotions build in. Emojis can give a lot of insight into sentiment of online text, which is beneficial for us to analyse tweets or sentences.

The author was inspired by the fact that Zwidge separated one emoji into two emojis. [3] For example, the Zwidge separated 👩‍🚀 to 👩 🚀, which led the author to emoji2vec. The author used

a method for normalization called the skip-gram method, which is similar to n-gram, but instead of using consecutive words, the method allows word pairs with any word in the sentence. After the normalization, the author asked people to score each labeled data set of emoji pairs manually. Another group of people has been asked differently by the author. Their task adds a description to each emoji. After both tasks, the author was embedding from 300-dimensional space into 2-dimensional space by using t-SNE. The author concluded that as emojis continue using on the internet, emoji embedding will be advanced.

Base on this paper, we have few ideas about how to begin our project by analysing the relationship between sentences and emoji. Also, since the author also mentioned that jupyter notebook could not display the latest emoji, but could print the emojis separately, we may need to come up with a new way to analyse those 'combined' emoji.

3.2 EmojiNet: An Open Service and API for Emoji Sense Discovery

The second paper presents a general idea of what is EmojiNet, and how EmojiNet functioning. [4] EmojiNet is a machine-readable emoji-English representation inventory, which has a store of over 2389 emoji with a 12,904 sense label. Similar to the previous paper, EmojiNet was inferred through word embedding models and trained over Google News corpus and Twitter message corpus to define each emoji sense, which inspired us to use both corpora as our vocabulary. During the author's research, they have found out that google emoji and iPhone emoji look different. Therefore, they have collected the emoji from 3 different open resources - The Unicode Emoji List, Emojipedia, and The Emoji Dictionary. Since the image of the same emoji could be different, they need to link different resources based on the Images of the Emoji. The modeling of EmojiNet is each emoji have their corresponding nonuple. Let E be all emoji in EmojiNet and $\mathbf{e_i} \in \mathbf{E}$ the corresponding nonuple for their model.

$$\mathbf{e_i} = (\mathbf{u_i}, \mathbf{n_i}, \mathbf{c_i}, \mathbf{d_i}, \mathbf{K_i}, \mathbf{I_i}, \mathbf{R_i}, \mathbf{H_i}, \mathbf{S_i}).$$

Symbol	Meaning
$\mathbf{u_i}$	Unicode of the Emoji
$\mathbf{n_i}$	Name of the Emoji
$\mathbf{c_i}$	Other name of the Emoji
$\mathbf{d_i}$	Definition of the Emoji
$\mathbf{I_i}$	Images of the Emoji
$\mathbf{R_i}$	Related
$\mathbf{H_i}$	Category for that Emoji
$\mathbf{S_i}$	Senses behind the Emoji

We think that the resources and modeling from this paper are beneficial for our project. Those resources could help us define our emoji corpus and give us suggestions on how to model.

4 Dataset

4.1 Requirement

The data sources should contain multiple different categories of sentences so that we can have a variety of data set to train and test the model. Since we want to train the model to translate from a sentence to a sequence of emojis, we want a sentence with some sentiment included. If we have a neutral sentiment sentence, such as 'There is a dining common down the road.', it is tough to extract proper emotion -> emoji from it.

4.2 Data Source

Fetching data from social media, such as Twitter, would be good practice for us. There are millions of users lively share their sentiments and there are lots of sentence already containing emoji. Thus we

will use sentences on Twitter as our dataset. There are lots of different public tweets dataset online, or we can use Twitter API to get tweets and build our own dataset if none of the online dataset meet our requirement.

4.3 Possible Dataset

On the first step, we plan to use the dataset provided by Sentiment 140. This dataset contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment. This dataset is also available on kaggle.

4.4 Human Annotation

We do not plan to annotate the data by human at this point.

5 Preliminary Experiment

5.1 Baseline Algorithm

Every emoji has its own English explanation, so we can translate a word into emoji by finding the emoji which the explanation is most similar to that word. We can translate a sentence into a emoji sequence by selecting all the emoji which match at least one word in this sentence. We can also use

5.2 Models

We plan to mostly use pre-existing software or pre-trained model to create our Sentence2Emoji model. Pre-trained model includes BERT, GPT-J-6B, GPT-3, Word2Vec, and Emoji2vec .

Our first idea is using autoregressive language model like EleutherAI/gpt-j-6B or GPT-3 to do the translation task by prompting. We prompt the model by some examples of sentence to emoji translation. Select the first consecutive emoji from the text generated by model as our translation result.

References

[1] Daniel Hånberg Alonso. *Emoji Timeline*. https://emojitimeline.com, E-Book Version, 2021.

[2] Katharina Buchholz. Emoji count: can you guess how many there will be by 2022? https://www.weforum.org/agenda/2021/06/emoji-count-is-increasing-diversity, June 2021. Accessed on 2021-10-12.

[3] Caroline Vanacore. Emoji2vec. https://medium.com/@vanacorec/emoji2vec-dc78f0b9e2ca, May 2019. Accessed on 2021-10-16.

[4] Balasuriya L. Sheth A. Doran D. Wijeratne, S. Emojinet: An open service and api for emoji sense discovery. https://ojs.aaai.org/index.php/ICWSM/article/view/14857, May 2017.