MojiSem □😲💻

Montoya • Naaman • Provenza

Background and general goal:

Emoji characters were first offered on JPmobile phones around the turn of the 21st century. But these pictographic elements reached 🌎🌍 language communities after being added to Unicode 6.0 in 2010, and then offered as software keyboards on →📱. In the ensuing half-decade, communities of language users have quickly developed many linguistic uses for emoji.

Emoji are sometimes said to function as discourse markers, and this alone would be a good motivation to gain a nuanced understanding of their application. But our initial survey of emoji usage on Twitter reveals many cases where emoji serve direct semantic functions in a tweet or they are used as a grammatical function such as a preposition or punctuation.

Emoji now play an undeniably central role in the expression of meaning in the media where they are heavily used (web and chat), but they remain under-studied and under-analyzed. Some interesting patterns may be found automatically. Instagram Engineering blog used word2vec to cluster emoji from 50 million posts by their distributional similarity:
http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji
yielding this very enriching map of semantic clusters:
https://s3.amazonaws.com/instagram-static/engineering-blog/emoji-hashtags/tsne_map_tight.png
Tyler Schnoebelen did a fairly shallow and summary investigation of emoji usage in tweets to make certain generalizations about high-frequency emoji in 500,000 tweets:
http://time.com/2993508/emoji-rules-tweets/
But there are no results for 'emoji' in the ACL Anthology.

We propose to go deeper, as far as a semester project may allow. We aim to offer annotators a framework and heuristics to classify uses of emoji by linguistic and discursive function, in order to reach a better understanding of the ways that different emoji are used and their relationship to any surrounding text. This corpus and annotation schema will be used to collect human judgment of what an emoji is doing in a given context, and to make predictions of what it is doing in novel contexts.

At the highest level, we have already observed emoji serving the following 3 major functions. These are the minimal categories which annotators will be identifying, though the range of tags within these categories is still to be determined:

1. **Discourse markers**
   Additions that allow a written text to become multimodal in the way that speech often is. This includes: attitude, physical gestures, topic markers and more.

2. **Semantic function/content words**
   Cases where the emoji replaces a content word in the sentence, usually based on the meaning represented in the image, or a culturally defined meaning (could be compositional). We hope that emojis in this category could be classified correctly by existing POS taggers (even if some minor changes will be needed).
3. **Grammatical functions**
   This category has been discussed less in connection to emoji than others, since it is widely thought to be what prevents emojis from becoming a separate language. The inability to express auxiliaries, prepositions, determiners and other functional words makes it harder to express things like tense and aspect. However, in our research, we have already come across several emojis that are used in such roles. Some are a natural continuation of shorthand conventions such as using the number 4 instead of the word “for”, and others are graphical representations of punctuation or mathematical symbols. Further exploration is needed to see if a closed set of options exists, or if novel uses can be generated by users.

Corpus: Twitter data in English

For this project we will be looking at tweets. As a social network, Twitter is a communication platform where emojis have become a significant part of the language. Its users vary in age, dialect, register and interests. The limited length of each tweet (140 characters) provides a convenient scope for determining the function of an emoji character in relation to the rest of the post. Tweets include both standalone utterances and conversations between users, which yields a variety of discourse styles that cannot be found in chat data or blog posts in isolation.
From a technical point of view, large quantities of tweets are accessible and easy to analyze. Corpus size will be determined in a later date, and will be based on further research into the distribution of emojis in Twitter data.

We will focus on tweets in English, both for interpretability, and because we expect that different languages adopt different uses and meanings for emoji characters, based on syntax or similar-sounding words.