Varying linguistic purposes of emoji in (Twitter) context

Noa Na'aman, Hannah Provenza, Orion Montoya
Brandeis University
{nnaaman,hprovenza,obm}@brandeis.edu

Abstract

Research into emoji in textual communication has, thus far, focused on high-frequency usages and the ambiguity of interpretations. Investigation of emoji uses across a wide range of uses can divide them into different linguistic functions: function and content words, or multimodal affective markers. We report on an annotation task on English Twitter data with the goal of classifying emoji usage by these categories, and on the effectiveness of a classifier trained on these annotations. We find that it is reasonably easy to train a classifier to tell the difference between content words and multimodal markers, but that sub-classification of these multimodal emoji needs more data and a more feature engineering.

1 Background

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

Emoji are sometimes said to function as way to make written text a form of multimodal communication, 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.

Early work on Twitter emoticons (Schnoebelen, 2012) pre-dated the wide spread of Unicode emoji on mobile and desktop devices. Schnoebelen studied the Recent work (Miller et al., 2016) has explored the cross-platform ambiguity of emoji renderings, finding that readers interpret emoji in drastically different ways depending on what emoji font is being used to render them.

We felt that a lexical semantics of emoji characters is implied in these studies without being directly addressed. Words are not used randomly, and neither are emoji, but emoji are used for different purposes than words. We believe that work on emoji would be better informed if there were an explicit typology of the linguistic functions that emoji can serve in expressive text.

Our aim was 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. We would then use this corpus to make predictions of what emoji characters are doing in novel contexts.

2 Annotation task

2.1 Guidelines

Although recognizing the presence of emoji characters is trivial and unambiguous, the linguistic distinctions we sought to annotate were ambiguous and seemed prone to disagreement. Therefore in the guidelines we structured the annotation process in such a way as to minimize cognitive load and lead the annotators to intuitive and natural decisions. Our task was motivated partially by the observation that emoji are used in contexts that make them graphical replacements that map to the existing lexicon, and are therefore straightforward to interpret. The guidelines build on this observation and dictate a process that aims to take advantage of such uses. The process was a flow that presented annotators with a few simple questions

at each step, to determine whether to assign a label or to move on to the next step.

Our guidelines gave a cursory background about emoji and their uses in social media, assuming no particular familiarity with the range of creative uses of emoji. In hindsight we realized that we assumed that the annotators would have a fair degree of familiarity with Twitter. The short-message social platform has many distinctive cultural and communicative codes of its own, not to mention subcultures, and continuously evolving trends combined with a long memory. As two of the authors are active and engaged users of Twitter, we unfortunately took it for granted that our annotators would be able to decipher emoji in contexts that required knowledge of InterNet language and Twitter norms.

The task comprised:

- Identifying each emoji in the tweet
- Deciding whether multiple contiguous emoji should be considered separately or as a group
- Choosing the right tag for the emoji (or sequence)
- Providing a translation or interpretation for each tagged span.

Eliciting an interpretation serves two main goals. First, as a coercive prompt for the user to bias them toward a linguistic interpretation. A replaceable phrase that fits with the grammar of the sentence is a different proposition than a marker that amounts to a standalone utterance such as "I am laughing" or "I am sad". Secondly, one of the eventual applications of annotated corpus may be emoji-sense disambiguation (ESD), and mapping to a lexicalized expression would be useful grounding for future ESD tasks. The text field was very helpful during the adjudication process, clarifying the annotators' judgments and understanding of the task (when done correctly).

For each tweet, annotators were asked to do a first read without annotating anything, to get a sense of the general message of the tweet and to think about the relationship between the emoji and the text. After the first reading, they are asked to determine whether the emoji is serving as punctuation or a function word; then if it is a content word; and if it is neither of those, then to examine it as a multimodal emoji. A key test, in our opinion, was asking annotators to simulate reading the message of the tweet aloud to another person. If a listeners comprension of the core message seemed to require a word or phrase to be spoken in place of an emoji, then that would be a compelling sign that it should be tagged as function or content.

Multimodal was the category assigned to uses that failed the first two tests. We provided some guidance for deciding between 'topic', 'attitude' or 'gesture' as sub-types of the multimodal category.

2.1.1 Later amendments to the guidelines

After the first batch of annotation work, in response to concerns and suggestions of our annotators, we made three major changes to the guidelines:

- Added the option of tagging a tweet as "out of scope," since some uses did not fit in our task description. One major example was an entire movie plot retold as a sequence of emoji characters. This is out of scope because here the emoji do not appear in a context that puts them in relation to other text, which is the phenomenon we sought to examine.
- Added co-reference tagging for topic-marker emoji that had overt co-referring words in the tweet. These are the only cases where we tagged non-emoji character spans.
- Added the option of tagging sequences of emoji that were split up by non-emoji text but should still be treated as one unit. This wasn't used at all in the following annotation batches, because the tweet format that made it necessary did not appear in the tweet sets for those batches.

2.2 Data collection and filtering

Tweets were pulled from the public Twitter streaming API using the `tweepy` Python package. The collected tweets were automatically filtered to include only tweets with characters from the Emoji Unicode ranges (i.e. generally `U+1FXXX`, `U+26XX`–`U+27BF`); only tweets labeled as being in English; to exclude tweets with embedded images or links (more below). Redundant/duplicate tweets were filtered by comparing tweet texts after removal of hashtags and @mentions; this left only a small number of cloned duplicates. After that, tweets were hand-selected to get a wide variety of emojis and context in a small

sample size — but, therefore, our corpus does not reflect the true distribution of emoji uses or context types.

Tweets with links and images were excluded from consideration to reduce time investment and cognitive load for annotators. Our early explorations found frequent cases where emoji were tweeted to show a reaction to an attached image or linked page (especially a blog post or news story) and that these tended toward ambiguous interpretations akin to those found by Miller et al. A given tweet's `U+1F62D 'loudly crying face'` might be showing true sympathy, or sarcastically saying "cry my a river." The amount of annotator effort necessary for an annotator to determine this in context would require an understanding of the tweeter's past opinions and their stance on the parties involved in the story they linked. These are very interesting questions for future research, but we determined them to be out of scope for the present research, which focuses on uses of emoji within predominantly textual expression.

2.3 Inter-annotator agreement

Our annotators were given 567 tweets with 878 total occurrences of emoji characters; in the gold standard these amounted to 775 tagged emoji spans. In weeks 1–2, each tweet was marked by four annotators; in weeks 3 and four we split them into two groups and had only two annotators per tweet.

There are two separate aspects of annotation for which IAA was relevant; the first, and less interesting, was the marking of the extent of emoji spans. Since emoji are unambiguously visible, we anticipated strong agreement. The one confounding aspect was that annotators were encouraged to group multiple emoji in a single span if they were a semantic/functional unit. This exposed a few differences of opinion, with one annotator tagging a pair of characters together, another separately. The overall Krippendorff α for extent markings was around 0.9.

The more significant place to look at IAA is the labeling of the emoji's functions. Because we were categorizing tokens, and because these categories are not ordered and we presented more than two labels, we used Fleiss's κ. But Fleiss's κ requires that annotators have annotated the same things, an in some cases annotators did not complete the dataset or missed an individual emoji character in a tweet. In order to calculate the statistics on actual agreement, rather than impute disagreement in the case of an 'abstention', we removed from our IAA-calculation counts any spans that were not marked by all annotators. There are many of these in the first dataset, and progressively fewer in each subsequent dataset as the annotators become more experienced. We used a Python script with an XML processor and the NLTK metrics package to calculate our scores, which are shown in Table 1.

2.4 Agreement/disagreement analysis

Content words. Part-of-speech identification is a skill familiar to most of our annotators from grammar school, so we were not surprised to see excellent levels of agreement among words tagged for part of speech —especially after we were able to review our guidelines with annotators and revise after the first week. These content words, however, were a very small proportion of the data — 51 out of 775 emoji spans — which may be problematically small. For datasets 3B and 4B, annotators were in perfect agreement.

Multimodal. Agreement on multimodal sublabels was much lower, and did not improve as annotation progressed. This may be because the categories could have been better defined; see "Adjudication" below.

Worst overall cross-label agreement scores were for week one, but all of the following datasets improved on that baseline after the annotation guidelines were refined.

2.5 Adjudication and gold standard

Given the low agreement in Week 1, adjudication required a lot of work, including translation from our 1.0 DTD version to the revised 1.0.1 used in weeks 2–4.

During adjudication it became obvious that we left the multimodal subtypes too open to interpretation, and when annotators disagreed, we had trouble deciding which of two disagreeing annotators was 'right' about choosing one type over another. For example, a smiley-face `U+1F60A` might be interpreted as a `gesture` (a smile), an *attitude* (joy), or a `topic` (for example, if the tweet is about what a good day the author is having) — and any of these would be a valid interpretation of a single tweet. In face-to-face human interaction, why do we use gestures? To clarify

Dataset	# taters	span rem	total	mm	content
Week 1	4	78	0.2071	0.4251	0.1311
Week 2	4	49	0.8743	0.7158	0.8531
Week 3A	2	11	0.9096	0.4616	0.792
Week 3B	2	6	0.7436	0.3905	1.0
Week 4A	2	3	0.8789	0.4838	0.7435
Week 4B	2	1	0.3954	0.5078	1.0
Total/mean	4	150	0.6681	0.4974	0.7533

Table 1: Fleiss's κ scores and other annotation numbers

Label	count
Multi-modal (mm)	total 686
attitude	407
topic	184
gesture	93
other	2
Content (cont)	total 51
noun	40
adj	6
verb	4
adv	1
Functional (func)	total 38
punct	34
aux	2
dt	1
other	1
emoji spans	total 775
words	6174
punctuation	668

Table 2: Label counts and subtypes in gold-standard data

attitudes or stances, or to indicate topics. So multimodal emoji may be inherently ambiguous, and we need a labeling system that can account for this. A clearer typology of multimodal emojis, and, if possible, a more deterministic procedure for labeling emoji with these subtypes, may be one approach.

3 Machine-learning task

Our experiment was to train a sequence tagger to assign the correct linguistic-function label to an emoji character. Our annotators had assigned labels and subtypes, but due to the low agreement on multimodal (`mm`) labels, and the small number of `cont` and `func` labels assigned, we narrowed the focus of our classification task to simply categorizing things correctly as either `mm` or `cont/func`. After one iteration, we saw that the low number of `func` tokens was preventing us from finding any `func` emoji, so we combined the `cont` and `func` tokens into a single label of `cont`. Therefore our sequence tagger needed simply to decide whether a token was serving as a substitute for a textual word, or was a multimodal marker.

3.1 Data sparseness

For reasons described above, we had a small and arbitrary sample of emoji usage available to study. Although we tagged 775 spans in 567 tweets, we only saw 300 distinct emoji, and 135 of them occurred only once. This drove us away from using an HMM tagger, because from the first attempt we could see that we would need to consider complex features independently. So we went with Conditional Random Fields (Lafferty et al., 2001).

3.2 Feature engineering

We used CRFSuite (Okazaki, 2007) and, after experimenting with the different algorithms available, found that the averaged perceptron algorithm (Collins,) gave the best results. Results for several iterations of features are given in Table 3, generally in order of increasing improvement until "prev +emo_class (best?)". The baseline feature was, of course, the emoji span itself, here called "character" although it may also be a sequence of emoji. "emo?" is a binary feature of either `emo` or `txt` — i.e. whether the token contains emoji characters, or is purely word characters.

The "POS" feature was a part-of-speech tag obtained by running the tweet text through `nltk.pos_tag`, which did apply part-of-speech labels to some emoji characters, and sometimes even correct ones. "position" was a set of three positional features: an integer 0–9 indicating a token's position in tenths of the way through the tweet; a three-class `BEGIN/MID/END` to indicate tokens at the beginning or end of a tweet (slightly different from the 0–9 feature in that multiple tokens may get 0 or 9, but only one token will get `BEGIN` or `END`); and the number of characters in the token. The "contexty" feature is another set of three features, this time related to con-

feature	F1 word	F1 mm	P cont	R cont	F1 cont	Macro-avg F1
character	0.9721	0.7481	0.3571	0.3333	0.3448	0.8441
prev +emo?	0.9914	0.8649	0.4286	0.4000	0.4000	0.8783
prev +POS	0.9914	0.8784	0.5000	0.4667	0.4828	0.8921
prev +position	0.9914	0.8844	0.4667	0.4667	0.4667	0.9028
prev +contexty	0.9914	0.8831	0.6250	0.3333	0.4348	0.8848
prev +emo_class (best?)	0.9914	**0.8933**	**0.7273**	0.5333	0.6154	**0.9168**
best − character	0.9906	0.8514	0.6429	**0.6000**	**0.6207**	0.9090
best − contexty	0.9922	0.8750	0.4706	0.5333	0.5000	0.8945
emo?+POS+emo_class	0.9914	0.8421	0.6000	0.4000	0.4800	0.8855

Table 3: Performance of feature iterations. Only the F1 score is given for word and mm labels because precision and recall were pretty consistent. cont labels are broken down by precision, recall and F1 because they varied in interesting ways.

text. A boolean preceded_by_determiner aimed to catch noun emoji; then two features to record the pairing of the preceding and following part of speech with the present token type (i.e. emo/txt).

A very useful feature is one that currently inheres in the ordering of emoji characters in Unicode blocks. Thus far, emoji have been added in semantically-related groups that tend to be contiguous. So there is a block of smiley faces and other 'emoticons'; a block of transport images; blocks of food, sports, animals, clothing; a whole block of hearts of different colors and elaborations; office-related, clocks, weather, hands, plants, and celebratory characters. These provide a very inexpensive proxy to semantics, and the "emo_class" feature yielded a marked improvement in both precision and recall on content words, although the small number of cases in the test data make it hard to be sure of their exact improvement.

We did a few other experiments to explore our features. "best − character" checked whether the features we had might be stronger without looking at the characters themselves. This showed that ignoring the character actually improved recall on content words, at the expense of precision. "best − contexty" removed the "contexty" feature, since it had actually slightly worsened several metrics, but removing it from the final "(best?)" feature set also worsened several metrics.

This was why we used CRFs: so we could see how different features, considered independently, might strengthen our labeling. A standard HMM would have treated any observations of the 135 nonce emoji as entirely unique and incomparable. The best thing we could likely have done with an HMM is shown on the final line of Table 3: just using a tuple of < is_emoji, POS, emoji_class >, and ignoring the individual token, yields performance that is competitive with the early iterations of the CRF feature set—not too shabby, but certainly creating a desire for improvement on content words.

3.3 Full-feature performance

The results in Table 3 show what we could reliably label with coarse-grained labels given the small size of our data set. But given that we annotated with finer-grained labels as well, it is worth looking at the performance on that task so far; results are shown it Table 3.3. Our test set had only two of each of the verbal content words — content_verb and func_aux — and didn't catch either of them, nor label anything else with either label. In fact, the only two func_aux in our dataset were in the test set, so they never actually got trained on. We got fairly reasonable recall on the mm_topic and mm_attitude labels, but given that those are the most frequent labels in the entire data set, it is more relevant that our precision was low.

4 Future directions

We clearly need more data. 89 examples of content and functional uses of emoji is not enough to reliably model the behavior of these characters. Since our classification is mostly binary. More annotation may yield much richer models of the variety, and will help get a better handle on the range of emoji polysemy. Clustering of contents based on observed features may help induce more em-

feature	TP	labeled	true	precision	recall	F1
mm_topic	38	53	44	0.7170	0.8636	0.7835
mm_attitude	11	26	16	0.4231	0.6875	0.5238
content_noun	6	11	11	0.5455	0.5455	0.5455
mm_gesture	2	2	8	1.0000	0.2500	0.4000
content_verb	0	0	2	0.0000	0.0000	0.0000
func_aux	0	0	2	0.0000	0.0000	0.0000

Table 4: performance of best model on subtype labels

pirically valid subtypes than the ones defined by our specification.

This project was motivated by a curiosity, and a hypothesis, about emoji semantics. Anglophone Twitter users use emoji in their tweets for a wide range of purposes, and a given emoji character means different things in different contexts. It seems inevitable that emoji will become subject to lexicographical description. Modern lexicography has evolved from millennia of attention to words and their combinations. Every emoji linguist notes the fascinating range of pragmatic and multi-modal effects that emoji can have in electronic communication. If these effects are to be given lexicographical treatment and categorization, they must also be organized into functional and pragmatic categories that are not part of the typical range of classes used to talk about words.

We have mentioned the notion of emoji-sense disambiguation (ESD). ESD would require an empirical inventory of emoji senses, presumably from an empirical lexicon of this sort.

Even our small sample has shown a number of characters that are genuinely used both as content words and as topical or gestural cues. Sometimes emoji of the "Earth globe" (U+1F30D–U+1F30F) are used to mean 'world', and sometimes they simply decorate a tweet that is expressing an environmental message. Sometimes these three globes—one showing Earth centered on Europe/Africa, one showing Earth centered on the Americas, and one centered on Asia/Australia—are used to denote those specific regions, or sometimes they are concatenated, either for emphasis ('the whole entire world') or to indicate the passing of time. There are a number of flower emoji, and sometimes they are used to decorate a message about flowers themselves, and sometimes they add sentiment to a message—and, just as in culture away from keyboards, a rose U+1F339 conveys a different sentiment than a sunflower U+1F33B.

There can be little question that different people use emoji differently, and this will certainly confound the study of emoji semantics in the immediate term. Using emoji as a functional word seems intuitively rare, and it is also rare in our (skewed) data set. Many people might send a multimodal smiley face, but far fewer will hunt their mobile keyboards for the KEYCAP DIGIT 4 to embellish the functional preposition 'for' when a plain number 4 will do just as well. The study of community dialects will be essential to emoji semantics, and there is certain also to be strong variation on the level of idiolect. The categorizations may need refinement, but the phenomenon is undeniably worthy of further study.

References

Michael Collins. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully happy" or "ready to fight": Varying interpretations of emoji. In *Proceedings of the Tenth International Conference on Web and Social Media, ICWSM 2016, Cologne, Germany, May 17–20, 2016*. Association for the Advancement of Artificial Intelligence, May.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Tyler Schnoebelen. 2012. Do you smile with your nose? Stylistic variation in Twitter emoticons. In *University of Pennsylvania Working Papers in Linguistics*, volume 18, pages 117–125. University of Pennsylvania.