MojiSem:
Varying linguistic purposes of emoji in (Twitter) context

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

Abstract

Early research into emoji in textual communication has focused largely on high-frequency usages and ambiguity of interpretations. Investigation of a wide range of emoji usage shows these glyphs serving at least two very different purposes: as content and function words, or as multimodal affective markers. Identifying where an emoji is replacing textual content allows NLP tools the possibility of parsing them as any other word or phrase. Recognizing the import of non-content emoji can be a a significant part of understanding a message as well.

We report on an annotation task on English Twitter data with the goal of classifying emoji uses by these categories, and on the effectiveness of a classifier trained on these annotations. We find that it is possible to train a classifier to tell the difference between those emoji used as linguistic content words and those used as paralinguistic or affective multimodal markers even with a small amount of training data, but that accurate sub-classification of these multimodal emoji into specific classes like attitude, topic, or gesture will require more data and more feature engineering.

1 Background

Emoji characters were first offered on Japanese mobile phones around the turn of the 21st century. These pictographic elements reached global language communities after being added to Unicode 6.0 in 2010, and then being offered within software keyboards on smartphones. In the ensuing half-decade, digitally-mediated language users have evolved diverse and novel linguistic uses for emoji.

The expressive richness of emoji communication would, on its own, be sufficient reason to seek a nuanced understanding of its usage. But our initial survey of emoji on Twitter reveals many cases where emoji serve direct semantic functions in a tweet or where 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. Recent work (Miller et al., 2016) has explored the cross-platform ambiguity of emoji renderings; (Eisner et al., 2016) created word embeddings that performed competitively on emoji analogy tasks; (Ljubešic and Fišer, 2016) mapped global emoji distributions by frequency; (Barbieri et al., 2017) used LSTMs to predict them in context.

We feel 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 even when they replace a word, 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. The current project offered annotators a framework and heuristics to classify uses of emoji by linguistic and discursive function. We then used a model based on this corpus to predict the grammatical function of emoji characters in novel contexts.

2 Annotation task

Although recognizing the presence of emoji characters is trivial, the linguistic distinctions we sought to annotate were ambiguous and seemed prone to disagreement. Therefore in our annotation guidelines we structured the process to minimize cognitive load and lead the annotators to in-

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 136–141

https://doi.org/10.18653/v1/P17-3022

tuitive decisions. This was aided somewhat by the observation that emoji are often used in contexts that make them graphical replacements for existing lexical units, and that such uses are therefore straightforward to interpret. Taking advantage of such uses, our flow presented annotators with a few simple questions at each step, to determine whether to assign a label or to move on to the next category.

2.1 Categories and subtypes

The high-level labels we defined for emoji uses were:

- **Function** (`func`): stand-ins for a function word in an utterance. These had a `type` attribute with values `prep`, `aux`, `conj`, `dt`, `punc`, `other`. An example from our data: "I 🍩 like u".
- **Content** (`cont`): stand-ins for lexical words or phrases that are part of the main informative content of the sentence. These have natural parts of speech, which annotators could subtype as: `noun`, `verb`, `adj`, `adv`, `other`. "The 🔑 to success is 🍕"
- **Multimodal** (`mm`): characters that enrich a grammatically-complete text with markers of affect or stance, whether to express an attitude ("Let my work disrespect me one more time... 🙃"), to echo the topic with an iconic repetition ("Mean girls 🎬", or to express a gesture that might have accompanied the utterance in face-to-face speech ("Omg why is my mom screaming so early 🙄"). Subtypes: `attitude`, `topic`, `gesture`, `other`.

The POS tags we chose were deliberately coarse-grained and did not include distinctions such as noun sub-types. We wanted to capture important differences while knowing that we would have fewer instances for the function and content labels. For all three labels, annotators were asked to provide a `replacement`: a word or phrase that could replace the emoji. For `func` and `cont`, replacements were a criterion for choosing the label; for `mm` there was room for interpretation.

2.2 Data Collection

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. We excluded tweets with embedded images or links. 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 — therefore, our corpus does not reflect the true distribution of emoji uses or context types.

2.3 Guidelines

Our guidelines gave annotators 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 noticed our assumption that annotators would have a fair degree of familiarity with modes of discourse on 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 nuanced knowledge of InterNet language and Twitter norms. This left annotators occasionally bewildered: by random users begging celebrities to follow them, by dialogue-formatted tweets, and by other epigrammatic subgenres of the short-text form.

The analytical steps we prescribed were:

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

Eliciting an interpretation serves two 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.

For each tweet, annotators 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. On subsequent readings, 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 listener's comprehension 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.

For each step we provided examples of tweets and emoji uses that clearly belong in each category. These examples were not included in the data set. Uses that failed the first two tests were assigned the multimodal category. We provided guidance and examples for deciding between 'topic', 'attitude' or 'gesture' as subtypes of the multimodal category.

2.4 Inter-annotator agreement

Four annotators, all computational linguistics grad students, were given 567 tweets with 878 total occurrences of emoji characters; in the gold standard these amounted to 775 tagged emoji spans. For the first 200 tweets annotated ('Set 1' and 'Set 2' in Table 1), each was marked by four annotators. After establishing some facility with the task we divided annotators into two groups and had only two annotators per tweet for the remaining 367.

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. The overall Krippendorff's 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, and 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. A total of 150 spans were excluded from Fleiss' kappa calculations for this reason.

2.5 Agreement/disagreement analysis

Content words. Part-of-speech identification is a skill familiar to most of our annotators, so we were not surprised to see excellent levels of agreement among words tagged for part of speech. 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. Multimodal emoji may be inherently ambiguous, and we need a labeling system that can account for this. A smiley face 😊 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. A clearer typology of multimodal emojis, and, if possible, a more deterministic procedure for labeling emoji with these subtypes, may be one approach.

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

3 Tag prediction experiment

We trained 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` la-

Dataset	# taters	span rem	total	mm	content
Set 1	4	78	0.2071	0.4251	0.1311
Set 2	4	49	0.8743	0.7158	0.8531
Set 3A	2	11	0.9096	0.4616	0.792
Set 3B	2	6	0.7436	0.3905	1.0
Set 4A	2	3	0.8789	0.4838	0.7435
Set 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/agreement variables

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

bels 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 Feature engineering

For reasons described above, we had a small and arbitrary sample of emoji usage available to study. After annotating 775 spans in 567 tweets, we had tagged 300 distinct emoji, 135 of which occurred only once. Given that our task is sequence tagging and our features are complex and independent, Conditional Random Fields seemed a good choice for our task. We used CRFSuite (Okazaki, 2007) and, after experimenting with the training algorithms available, found that training with averaged perceptron (Collins, 2002) yielded the best predictive results. Results for several iterations of features are given in Table 3, generally in order of increasing improvement until "prev +emo_class".

- The emoji span itself, here called 'character' although it may span multiple characters.
- 'emo?' is a binary feature indicating whether the token contains emoji characters (`emo`), or is purely word characters (`txt`).
- 'POS', a part-of-speech tag assigned by `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 (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 context: A boolean `preceded_by_determiner` aimed at catching noun emoji; and two features to record the pairing of the preceding and following part of speech with the present token type (i.e. `emo/txt`);
- Unicode blocks, which inhere in the ordering of emoji characters. 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

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.

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 true contribution.

We did a few other experiments to explore our features. 'best − character' 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.

3.2 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: 511 training tweets, 56 test tweets. 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 in Table 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 saw 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

89 examples of content and functional uses of emoji is not enough to reliably model the behavior of these categories. More annotation may yield much richer models of the variety of purposes of emoji, and will help get a better handle on the range of emoji polysemy. Clustering of contexts based on observed features may induce more empirically valid subtypes than the ones defined by our specification.

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. Every emoji linguist notes the fascinating range of pragmatic and multimodal 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 printed words.

We have mentioned the notion of emoji-sense disambiguation (ESD). ESD in the model of traditional WSD would seem to require an empirical inventory of emoji senses. Even our small sample has shown a number of characters that are used both as content words and as topical or gestural cues. Our data included "Mean girls 🎬", i.e. 'I am watching the movie Mean Girls', which has no propositional content in common with (untattested in our data set) "Mean girls 😡", i.e. 'girls who are mean upset me'. There are a number of flower emoji: 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 🌹 is always marked as conveying a message of 'love', while a cherry blossom 🌸 is consistently associated with 'beauty'.

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

There can be little question that individuals use emoji differently, and this will certainly confound the study of emoji semantics in the immediate term. 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

Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? *EACL 2017*, page 105.

Michael Collins. 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Conference on Empirical Methods in Natural Language Processing*, page 48.

Nikola Ljubešic and Darja Fišer. 2016. A global analysis of emoji usage. *ACL 2016*, page 82.

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.