



HHS Public Access

Author manuscript

Proc Int Workshop Soc Inform. Author manuscript; available in PMC 2017 July 20.

Published in final edited form as:

Proc Int Workshop Soc Inform. 2016 November ; 10046: 527–541. doi:10.1007/978-3-319-47880-7_33.

EmojiNet: Building a Machine Readable Sense Inventory for Emoji

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran

Kno.e.sis Center, Wright State University, Dayton, Ohio, USA, <http://www.knoesis.org>

Abstract

Emoji are a contemporary and extremely popular way to enhance electronic communication. Without rigid semantics attached to them, emoji symbols take on different meanings based on the context of a message. Thus, like the word sense disambiguation task in natural language processing, machines also need to disambiguate the meaning or ‘sense’ of an emoji. In a first step toward achieving this goal, this paper presents EmojiNet, the first machine readable sense inventory for emoji. EmojiNet is a resource enabling systems to link emoji with their context-specific meaning. It is automatically constructed by integrating multiple emoji resources with BabelNet, which is the most comprehensive multilingual sense inventory available to date. The paper discusses its construction, evaluates the automatic resource creation process, and presents a use case where EmojiNet disambiguates emoji usage in tweets. EmojiNet is available online for use at <http://emojinet.knoesis.org>.

Keywords

EmojiNet; Emoji Analysis; Emoji Sense Disambiguation

1 Introduction

Pictographs commonly referred to as ‘emoji’ have grown from their introduction in the late 1990’s by Japanese cell phone manufacturers to an incredibly popular form of computer mediated communication (CMC). Instagram reported that as of April 2015, 40% of all messages posted on Instagram consist of emoji [6]. From a 1% random sample of all tweets published from July 2013 to July 2016, the service Emojitracker reported its processing of over 15.6 billion tweets with emoji¹. Creators of the SwiftKey Keyboard for mobile devices also report that 6 billion messages per day contain emoji [15]. Even authorities on language use have acknowledged emoji; the American Dialect Society selected ‘eggplant’ 🍆 as the most notable emoji of the year², and The Oxford Dictionary recently awarded ‘face with tears of joy’ 😄 as the word of the year in 2015³. All these reports suggest that emoji are now an undeniable part of the world’s electronic communication vernacular.

¹<http://www.emojitracker.com/api/stats>

²<http://www.americandialect.org/2015-word-of-the-year-is-singular-they>

³<http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/>

People use emoji to add color and whimsiness to their messages [7] and to articulate hard to describe emotions [1]. Perhaps by design, emoji were defined with no rigid semantics attached to them⁴, allowing people to develop their own use and interpretation. Thus, similar to words, emoji can take on different meanings depending on context and part-of-speech [8]. For example, consider the three emoji 🤔, 🤔, and 🤔 and their use in multiple tweets in Figure 1. Depending on context, we see that each of these emoji can take on wildly different meanings. People use the 🤔 emoji to mean laughter, happiness, and humor; the 🤔 emoji to discuss killings, shootings or anger; and the 🤔 emoji to express that something is expensive, working hard to earn money or simply to refer to money.

Knowing the meaning of an emoji can significantly enhance applications that study, analyze, and summarize electronic communications. For example, rather than stripping away emoji in a preprocessing step, sentiment analysis application reported in [11] uses emoji to improve its sentiment score. However, knowing the meaning of an emoji could further improve the sentiment score. A good example for this scenario would be the 😊 emoji, where people use it to describe both happiness (using senses such as laugh, joy) and sadness (using senses such as cry, tear). Knowing under which sense the emoji is being used could help to understand its sentiment better. But to enable this, a system needs to understand the particular meaning or *sense* of the emoji in a particular instance. However, no resources have been made available for this task [8]. This calls for the need of a machine readable *sense inventory for emoji* that can provide information such as: (i) the plausible part-of-speech tags (PoS tags) for a particular use of emoji; (ii) the definition of an emoji and the senses it is used in; (iii) example uses of emoji for each sense; and (iv) links of emoji senses to other inventories or knowledge bases such as BabelNet or Wikipedia. Current research on emoji analysis has been limited to emoji-based sentiment analysis [11], emoji-based emotion analysis [17], and Twitter profile classification [2, 18] etc. However, we believe introduction of an emoji sense inventory can open up new research directions on emoji sense disambiguation, emoji similarity analysis, and emoji understanding.

This paper introduces **EmojiNet**, the first machine readable sense inventory for emoji. EmojiNet links emoji represented as Unicode with their English language meanings extracted from the Web. To achieve this linkage, EmojiNet integrates multiple emoji lexicographic resources found on the Web along with BabelNet [10], a comprehensive machine readable sense inventory for words, to infer sense definitions. Our contributions in this work are threefold:

1. We integrate four openly available emoji resources into a single, query-able dictionary of emoji definitions and interpretations;
2. We use word sense disambiguation techniques to assign senses to emoji;
3. We integrate the disambiguated senses in an open Web resource, EmojiNet, which is presently available for systems to query.

The paper also discusses the architecture and construction of EmojiNet and presents an evaluation of the process to populate its sense inventory.

⁴http://www.unicode.org/faq/emoji_dingbats.html#4.0.1

This paper is organized as follows. Section 2 discusses the related literature and frames how this work differs from and furthers existing research. Section 3 discusses our approach and explains the techniques we use to integrate different resources to build EmojiNet. Section 4 reports on the evaluation of the proposed approach and the evaluation results in detail. Section 5 offers concluding remarks and plans for future work.

2 Related Work

Emoji was first introduced in the late 1990s but did not become a Unicode standard until 2009 [5]. Following standardization, emoji usage experienced major growth in 2011 when the Apple iPhone added an emoji keyboard to iOS, and again in 2013 when the Android mobile platform began emoji support [6]. In an experiment conducted using 1.6 million tweets, Novak *et al.* report that 4% of them contained at least one emoji [11]. Their recent popularity explains why research about their use is not as extensive as the research conducted on emoticons [8], which are the predecessor to emoji [11] that used to represent facial expression, emotion or to mimic nonverbal cues in verbal speech [13] in CMC.

Early research on emoji focuses on understanding its role in computer-aided textual communications. From interviews of 20 participants in close personal relationships, Kelly *et al.* reported that people use emoji to maintain conversational connections in a playful manner [7]. Pavalanathan *et al.* studied how emoji compete with emoticons to communicate paralinguistic content on social media [12]. They report that emoji were gaining popularity while emoticons were declining in Twitter communications. Miller *et al.* studied whether different emoji renderings would give rise to diverse interpretations [8], finding disagreements based on the rendering. This finding underscores the need for tools to help machines disambiguate the meaning and interpretation of emoji.

The Emoji Dictionary⁵ is a promising Web resource for emoji sense disambiguation. It is a crowdsourced emoji dictionary that provides emoji definitions with user defined sense labels, which are `word(POS tag)` pairs such as `laugh(noun)`. However, it cannot be utilized by a machine for several reasons. First, it does not list the Unicode or short code names for emoji, which are common ways to programmatically identify emoji characters in text. Secondly, it does not list sense definitions and example sentences along with different sense labels for emoji. Typically, when using machine readable dictionaries, machines use such sense definitions and example sentences to generate contextually relevant words for each sense in the dictionary. Thirdly, the reliability of the sense labels is unclear as no validation of the sense labels submitted by the crowd is performed. With EmojiNet, we address these limitations by linking The Emoji Dictionary with other rich emoji resources found on the Web. This allows sense labels to be linked with their Unicode and short code name representations and discards human-entered sense labels for emoji that are not agreed upon by the resources. EmojiNet also links sense labels with BabelNet to provide definitions and example usages for different senses of an emoji.

⁵<http://emojidictionary.emojifoundation.com/home.php?learn>

3 Building EmojiNet

We formally define EmojiNet as a collection of octuples representing the senses of an emoji. Let E be the set of all emoji in EmojiNet. For each $e_j \in E$, EmojiNet records the octuple $e_j = (u_j, c_j, d_j, K_j, I_j, R_j, H_j, S_j)$, where u_j is the Unicode representation of e_j , c_j is the short code name of e_j , d_j is a description of e_j , K_j is the set of keywords that describe basic meanings attached to e_j , I_j is the set of images that are used in different rendering platforms such as the iPhone and Android, R_j is the set of related emoji for e_j , H_j is the set of categories that e_j belongs to, and S_j is the set of different senses in which e_j can be used within a sentence. An example for an octuple notation is shown as part of Figure 2. Each element in the octuple provides essential information for sense disambiguation. EmojiNet uses unicode u_j and short code name c_j of an emoji $e_j \in E$ to uniquely identify e_j , and hence, to search EmojiNet. d_j is needed to understand what is represented by the emoji. It can also help to understand how an emoji should be used. K_j is essential to understand different human-verified senses that an emoji could be used for. I_j is needed to understand the rendering differences in each emoji based on different platforms. Images in I_j can also help to conduct similar studies as [8], where the focus is to disambiguate the different representations of the same emoji on different rendering platforms. R_j and H_j could be helpful in tasks such as calculating emoji similarity and emoji sense disambiguation. Finally, S_j is the key enabler of EmojiNet as a tool to support emoji sense disambiguation as S_j holds all sense labels and their definitions for e_j based on crowd and lexicographic knowledge. Next, we describe the open information EmojiNet extracts and integrates from the Web to construct the octuples.

3.1 Open Resources

Several emoji-related open resources are available on the Web, each carrying their own strengths and weaknesses. Some have overlapping information, but none has all of the elements required for a machine readable sense inventory. Thus, EmojiNet collects information across multiple open resources, linking them together to build the sense inventory. We describe the resources EmojiNet utilizes below.

Unicode Consortium—Unicode is a text encoding standard enforcing a uniform interpretation of text byte code by computers⁶. The consortium maintains a complete list of the standardized Unicodes for each emoji⁷ along with manually curated keywords and images of emoji. Let the set of all emoji available in the Unicode emoji list be E_U . For each emoji $e_u \in E_U$, we extract the Unicode character u_j of e_u , the set of all images I_{e_u} associated with e_u that are used to display e_u on different platforms, and the set of keywords $K_{U_{e_u}} \subset K_{e_u}$ associated with e_u , where K_{e_u} is the set of all manually-assigned keywords available for the emoji e_u .

Emojipedia—Emojipedia is a human-created emoji reference site⁸. It provides Unicode representations for emoji, images for each emoji based on different rendering platforms, short code names, and other emoji manually-asserted to be related. Emojipedia organizes

⁶<http://www.unicode.org/>

⁷<http://www.unicode.org/emoji/charts/full-emoji-list.html>

⁸<https://en.wikipedia.org/wiki/Emojipedia>

emoji into a pre-defined set of categories based on how similar the concepts are represented by each emoji, i.e., Smileys & People, Animals & Nature, or Food & Drink. Let the set of all emoji available in Emojipedia be E_P . For each emoji $e_p \in E_P$, we extract the Unicode representation u_p , the short code name c_p , and the emoji definition d_p of e_p , the set of related emoji R_{e_p} , and its category set H_{e_p} .

iEmoji—iEmoji⁹ is a service tailored toward understanding how emoji are being used in social media posts. For each emoji, it provides a human-generated description, its Unicode character representation, short code name, images across platforms, keywords describing the emoji, its category within a manually-built hierarchy, and examples of its use in social media (Twitter) posts. Let the set of all emoji available in iEmoji be E_{IE} . For each emoji $e_{ie} \in E_{IE}$, we collect the Unicode representation u_{ie} of e_{ie} and the set of keywords $K_{IEe_{ie}} \subset K_{e_{ie}}$ associated with e_{ie} , where $K_{e_{ie}}$ is the set of all keywords available for e_{ie} .

The Emoji Dictionary—The Emoji Dictionary¹⁰ is a crowdsourced site providing emoji definitions with sense labels based on how they could be used in sentences. It organizes meanings for emoji under three part-of-speech tags, namely, nouns, verbs, and adjectives. It also lists an image of the emoji and its definition with example uses spanning multiples sense labels. Let the set of all emoji available in The Emoji Dictionary be E_{ED} . For each emoji $e_{ed} \in E_{ED}$, we extract its image $i_{ed} \in I_{ED}$, where I_{ED} is the set of all images of all emoji in E_{ED} and the set of crowd-generated sense labels $S_{e_{ed}}$.

BabelNet—BabelNet is the most comprehensive multilingual machine readable semantic network available to date [10]. It is a dictionary with lexicographic and encyclopedic coverage of words tied to a semantic network that connects concepts in Wikipedia to the words in the dictionary. It is built automatically by merging lexicographic data in WordNet with the corresponding encyclopedic knowledge extracted from Wikipedia¹¹. BabelNet has been shown effective in many research areas including word sense disambiguation [10], semantic similarity, and sense clustering [4]. For the set of all sense labels $S_{e_{ed}}$ in each $e_{ed} \in E_{ED}$, we extract the sense definitions and examples (if available) for each sense label $s_{e_{ed}} \in S_{e_{ed}}$ from BabelNet.

Table 1 summarizes the data about an emoji available across the four open resources. A ‘✓’ denotes the availability of the information in the resource where ‘X’ denotes the non-availability. It is important to note that unique crowds of people deposit information about emoji into each resource, making it important to integrate the same type of data across many resources. For example, the set of keywords K_i , the set of related emoji R_i , and the set of categories H_i for an emoji e_i are defined by the crowds qualitatively, making it necessary to compare and scrutinize them to determine the elements that should be considered by EmojiNet. Data types that are ‘fixed’, e.g. the Unicode u_i of an emoji e_i , will also be useful to link data about the same emoji across the resources. We also note that The Emoji Dictionary uniquely holds the sense labels of an emoji, yet does not store its Unicode u_i .

⁹<http://www.iemoji.com/>

¹⁰<http://emojidictionary.emojifoundation.com/home.php?learn>

¹¹<http://babelnet.org/about>

This requires EmojiNet to link to this resource via emoji images, as we discuss further in the next section.

3.2 Integrating Emoji Resources

We now describe how EmojiNet integrates the open resources described above. The integration, illustrated in Figure 2, starts with the Unicode's emoji characters list as it is the official list of 1,791 emoji accepted by the Unicode Consortium for support in standardized software products. Using Unicode character representation in the emoji list, we link these emoji along with the information extracted from Emojipedia and the iEmoji websites. Specifically, for each emoji $e_u \in E_U$, we compare u_u with all Unicode representations of the emoji in E_P and E_{IE} . If there is an emoji $e_u \in E_U$ such that $u_u = u_p = u_{ie}$, we merge the three corresponding emoji $e_u \in E_U$, $e_p \in E_P$, and $e_{ie} \in E_{IE}$ under a single emoji representation $e_j \in E$. In other words, we merge all emoji where they share the same Unicode representation. We store all the information extracted from the merged resources under each emoji e_j as the octuple described in Section 3.

Linking to The Emoji Dictionary—Unfortunately, The Emoji dictionary does not store the Unicode of an emoji. Thus, we merge this resource into EmojiNet by considering emoji *images*. We created an index of multiple images of the 1,791 Unicode defined emoji in the Unicode Consortium website. We have downloaded a total of 13,387 images for the 1,791 emoji. These images are referred to as our example image dataset I_x . We additionally downloaded images of all emoji listed on The Emoji Dictionary website, which resulted in a total of 1,074 images. We refer to this set of images as the test image dataset I_t .

To align the two datasets, we implement a nearest neighborhood-based image matching algorithm [14] that matches each image in I_t with the images in I_x . Because images are of different resolutions, we normalize them into a 300x300px space and then divide them along a lattice of 25 non-overlapping regions of size 25x25px. We then find an average color intensity of each region by averaging its *R*, *G* and *B* pixel color values. To calculate the dissimilarity between two images, we sum the L_2 distance of the average color intensities of the corresponding regions. The final accumulated value that we receive for a pair of images will be a measure of the dissimilarity of the two images. For each image in I_t , the least dissimilar image from I_x is chosen and the corresponding emoji octuple information is merged.

Emoji sense and part-of-speech filtering—With The Emoji Dictionary linked to the rest of the open resources via images, EmojiNet can now integrate its sense and part-of-speech information (sense labels). However, as mentioned in Section 2, The Emoji Dictionary does not validate the sense labels collected from the crowd. Thus, EmojiNet must pre-process the sense labels from The Emoji Dictionary to verify its reliability. This is done in a three step process and it is elaborated in Figure 3. First, we use the set of keywords K_j of emoji e_j collected from the Unicode Consortium and iEmoji as seed words to identify reliable sense labels. These keywords are human-generated and represent the meanings in which an emoji can be used. For example, the 😊 emoji has been tagged with the keywords *face*, *joy*, *laugh*, and *tear* in the Unicode emoji list and *tear*, *cry*, *joy*, and *happy* in the iEmoji

website. Taking the union of these lists as a set of seed words, we filter the crowdsourced sense labels of an emoji from The Emoji Dictionary. For each keyword $k_i \in K_i$, we extract crowdsourced sense labels. For example, for the emoji 🤔, The Emoji Dictionary lists three sense labels for the sense *laugh* as *laugh(noun)*, *laugh(verb)* and *laugh(adjective)*. However, the word *laugh* cannot be used as an adjective in the English language. Therefore, in the second step, we cross-check if the sense labels extracted from The Emoji Dictionary are valid using the information available in BabelNet. In this step, BabelNet reveals that *laugh* cannot be used as an adjective, so we discard *laugh(adjective)* and use *laugh(noun)*, *laugh(verb)* in EmojiNet. We do this for all seed keywords we obtain from the Unicode emoji list and iEmoji websites. In the final step, for any sense label in The Emoji Dictionary that is not a seed word but was submitted by more than one human (commonly agreed senses in Figure 3), EmojiNet validates these sense labels using BabelNet. For example, the sense label *funny(adjective)* has been added by more than one user to The Emoji Dictionary as a possible sense for 🤔 emoji. This was not in our seed set; however, since there is common agreement on the sense label *funny(adjective)* and the word *funny* can be used as an adjective in the English language, EmojiNet extracts *funny(adjective)* from The Emoji Dictionary and adds it to its sense inventory under 🤔 emoji. Note that EmojiNet does not assign BabelNet sense IDs (or sense definitions) to the extracted sense labels yet. That process will require a word sense disambiguation step, which we will discuss in the next section.

3.3 Linking Emoji Resources with BabelNet

Having access to sense labels extracted from The Emoji Dictionary for each emoji, EmojiNet can now link these sense labels with BabelNet. This linking allows EmojiNet to interpret each sense label on how it can be used in a sentence. For example, the current version of BabelNet lists 6 different sense definitions for the sense label *laugh(noun)*. Thus, EmojiNet must select the most appropriate sense definition out of the six. As we described in Section 2, The Emoji Dictionary does not link its sense labels with example sentences. Therefore, we cannot directly perform WSD on the sense labels or example sentences available in The Emoji Dictionary. Thus, to align the two resources, we use the MASC¹² corpus with a most frequent sense (MFS) baseline for WSD. MASC corpus is a balanced dataset that represents different text categories such as tweets, blogs, emails, letters, essays, and speech; words in the MASC corpus are already disambiguated using BabelNet [9]. We use these disambiguated words to calculate MFS for each word in the MASC corpus. Once the MFS of each word is calculated, for every sense label in EmojiNet, we assign its definition as the MFS of that same sense label retrieved from MASC corpus. We use an MFS-based WSD baseline due to the fact that MFS is a very strong, hard-to-beat baseline model for WSD tasks [3]. Figure 4 depicts the steps followed in our WSD approach.

EmojiNet has a total of 3,206 sense labels that need to be sense disambiguated using BabelNet. However, not all sense labels in EmojiNet were assigned BabelNet senses in the above WSD task. There were sense labels in EmojiNet which were not present in MASC

¹²[https://en.wikipedia.org/wiki/Manually_Annotated_Sub-Corpus_\(MASC\)](https://en.wikipedia.org/wiki/Manually_Annotated_Sub-Corpus_(MASC))

corpus, hence they were not disambiguated. To disambiguate such sense labels which were left out, we define a second WSD task. We calculate the most popular sense (MPS) for each BabelNet sense, which we define as follows. For each BabelNet sense label B_s , we take the count of all sense definitions BabelNet lists for B_s . The MPS of B_s is the BabelNet sense ID that has the highest number of definitions for B_s . If there are more than one MPS available, a sense ID will be picked randomly out of the set of MPS sense IDs as the MPS. Once the MPS is calculated, those will be assigned to their corresponding sense labels in EmojiNet which were left out in the first WSD task. Note that BabelNet holds multiple definitions that come from multiple resources such as WordNet, VerbNet, Wikipedia, etc. which are integrated into it. Hence, MPS of B_s gives an indication of the popularity of B_s . With this step, we complete the integration of open resources to create EmojiNet.

3.4 EmojiNet Web Application

We expose EmojiNet as a web application at <http://emojinet.knoesis.org/>. The current version of EmojiNet supports searching for emoji based on Unicode character and short code name. It also lets the user search emoji by specifying a part-of-speech tagged sense and returns a list of emoji that are tagged with the searched sense. Table 2 lists statistics for EmojiNet. It currently holds a total of 1,074 emoji. It has a total of 3,206 valid sense definitions that are shared among 875 emoji, with an average of 4 senses per emoji. The resource is freely available to the public for research use¹³.

4 Evaluation

Note that the construction of EmojiNet is based on linking multiple open resources together in an automated fashion. We thus evaluate the automatic creation of EmojiNet. In particular, we evaluate the nearest neighborhood-based image processing algorithm that we used to integrate emoji resources and the most frequent sense-based and most popular sense-based word sense disambiguation algorithms that we used to assign meanings to emoji sense labels. Note that we do not evaluate the usability of EmojiNet based on its performance on a selected task or a benchmark dataset. While sense inventories such as BabelNet have been evaluated on benchmark datasets for WSD or word similarity calculation performance, emoji sense disambiguation and finding emoji similarity are two research problems on their own that have not been explored yet [8]. The focus of this paper is not to study or solve those problems. Evaluating the usefulness of EmojiNet should, and will, be addressed once emoji similarity tasks and emoji sense disambiguation tasks are defined with baseline datasets. In lieu of task evaluation, we demonstrate the usefulness of EmojiNet with a use case of how it can be used to address the emoji sense disambiguation problem.

4.1 Evaluating Image Processing Algorithm

We next evaluate how well the nearest neighborhood-based image processing algorithm could match each image in I_t with the images in I_x . I_x could contain multiple images for a given emoji (7 images per emoji on average), based on different rendering platforms on which the emoji could appear. The set of all different images $I_j \in I_x$ that belongs to e_j are

¹³<http://emojinet.knoesis.org/>

tagged with the Unicode representation u_i , which is the Unicode representation of e_i . For us to find a match between I_t and I_x , we only require one of the multiple images that represents an emoji from I_x match with any image from I_t . Once the matching process is done, we pick the top ranked match based on the dissimilarity of the two matched images and manually evaluate them for equality. While the image processing algorithm we used is naive, it works well in our use case due to several reasons. First, the images of emoji are not complex as they represent a single object (e.g. eggplant 🍆) or face (e.g. smiling face 😊). Second, the emoji images do not contain very complex color combinations as in textures and they are small in size. The image processing algorithm combines color (spectral) information with spatial (position/distribution) information and tends to represent those features well when the images are simple. Third, Euclidean distance ($L2$ distance) prefers many medium disagreements to one large disagreement as in $L1$ distance. Therefore, this nearest neighborhood-based image processing algorithm fits well for our problem.

Manual evaluation of the algorithm revealed that it achieves 98.42% accuracy in aligning images in I_t with I_x . Out of the 1,074 image instances we checked, our algorithm could correctly find matching images for 1,057 images in I_t and it could not find correct matches for 17 images. We checked the 17 incorrect alignments manually and found that eight were clock emoji that express different times of the day. Those images were very similar in color despite the fact that the two arms in the clocks were at different positions. There were three incorrect alignments involving people characters present in the emoji pictures. Those images had minimal differences, which the image processing algorithm could not identify correctly. There were two instances where flags of countries were aligned incorrectly. Again, those flags were very similar in color (e.g. Flag of Russia and Flag of Slovenia). In our error analysis, we identified that the image processing algorithm does not perform correctly when the images are very similar in color but have slight variations in the object(s) it renders. Since the color of the image plays a huge role in this algorithm, the same picture taken in different lighting conditions (i.e. changes in the background color, while the image color stays the same) could decrease the accuracy of the program. However, that does not apply in our case as all the images we considered have a transparent background.

4.2 Evaluating Word Sense Disambiguation Algorithm

Here we discuss how we evaluate the most frequent sense-based and most popular sense-based word sense disambiguation algorithms that we used to link Emoji senses with BabelNet sense IDs. We use a manual evaluation approach based on human judges to validate whether a BabelNet sense ID assigned to an emoji sense is valid. We sought the help of two human judges in this task and our judges were graduate students who had either worked on or taken a class on natural language processing. We provided them with all the emoji included in EmojiNet, listing all the valid sense labels extracted from The Emoji Dictionary and their corresponding BabelNet senses (BabelNet sense IDs with definitions) extracted from each WSD approach. The human judges were asked to mark whether they thought that the suggested BabelNet sense ID was the correct sense ID for the emoji sense label listed. If so, they would mark the sense ID prediction as correct, otherwise they would mark it as incorrect. We calculated the agreement between the two judges for this task using

Cohen's kappa coefficient¹⁴ and obtained an agreement value of 0.7355, which is considered to be a good agreement.

Out of the 3,206 sense labels to disambiguate, the MFS-based method could disambiguate a total of 2,293 sense labels. Our judges analysed these sense labels manually and marked 2,031 of them as correct, with an accuracy of 88.57% for the MFS-based WSD task. There were 262 cases where the emoji sense was not correctly captured. The correctly disambiguated sense labels belong to 835 emoji. The 913 sense labels which were not disambiguated in the MFS-based WSD task were considered in a second WSD task, based on the MPS. Our evaluation revealed that the MPS-based WSD task could correctly disambiguate 700 sense labels, with an accuracy of 76.67%. There were 213 cases where our MPS-based approach failed to correctly disambiguate the sense label. The correctly disambiguated sense labels belong to 446 emoji.

Table 3 integrates the results obtained by both word sense disambiguation algorithms for different part-of-speech tags. The results shows the two WSD approaches we used have performed reasonably well in disambiguating the sense labels in EmojiNet. They have sense-disambiguated with a combined accuracy of 85.18%. These two methods combined have assigned BabelNet sense IDs to a total of 875 emoji out of the 1,074 emoji we extracted from The Emoji Dictionary website. It shows that our WSD approaches combined have disambiguated senses for 81.47% of the total number of emoji present in The Emoji Dictionary. However, we do not report on the total number of valid sense labels that we did not extract in our data extraction process since The Emoji Dictionary had 16,392 unique sense labels, which were too big for one to manually evaluate.

4.3 EmojiNet at Work

We also provide an illustration of EmojiNet in action with a disambiguation of the sense of the 🙏 emoji as it is used in two example tweets. We choose this emoji since it is reported as one of the most misused emoji on social media¹⁵. The tweets we consider are:

T_1 : Pray for my family 🙏 God gained an angel today.

T_2 : Hard to win, but we did it man 🙏 Lets celebrate!

EmojiNet lists `high five(noun)` and `pray(verb)` as valid senses for the above emoji. For `high five(noun)`, EmojiNet lists three definitions and for `pray(verb)`, it lists two definitions. We take all the words that appear in their corresponding definitions as possible context words that can appear when the corresponding sense is being used in a sentence (tweet in this case). For each sense, EmojiNet extracts the following sets of words:

`pray(verb)`: { *worship, thanksgiving, saint, pray, higher, god, confession* }

`highfive(noun)`: { *palm, high, hand, slide, celebrate, raise, person, head, five* }

To calculate the sense of the 🙏 emoji in each tweet, we calculate the overlap between the words which appear in the tweet with words appearing with each emoji sense listed above.

¹⁴https://en.wikipedia.org/wiki/Cohen's_kappa

¹⁵<http://www.goodhousekeeping.com/life/g3601/surprising-emoji-meanings/>

This method is called the Simplified Lesk Algorithm [16]. The sense with the highest word overlap is assigned to the emoji at the end of a successful run of the algorithm. We can see that 🙏 emoji in T_1 will be assigned `pray(verb)` based on the overlap of words {god, pray} with words retrieved from the sense definition of `pray(verb)` and the same emoji in T_2 will be assigned `high five(noun)` based on the overlap of word {celebrate} with words retrieved from the sense definition of `high five(noun)`. In the above example, we have only shown the minimal set of words that one could extract from EmojiNet. Since we link EmojiNet senses with their corresponding BabelNet senses using BabelNet sense IDs, one could easily utilize other resources available in BabelNet such as related WordNet senses, VerbNet senses, Wikipedia, etc. to collect an improved set of context words for emoji sense disambiguation tasks. It should be emphasized that this example was taken only to show the usefulness of the resource for research directions.

5 Conclusion and Future Work

This paper presented the construction of EmojiNet, the first ever machine readable sense inventory to understand the meanings of emoji. It integrates four different emoji resources from the Web to extract emoji senses and align those senses with BabelNet. We evaluated the automatic creation of EmojiNet by evaluating (i) the nearest neighborhood-based image processing algorithm used to align different emoji resources and (ii) the most frequent sense-based and the most popular sense-based word sense disambiguation algorithms used to align different emoji senses extracted from the Web with BabelNet. We plan to extend our work in the future by expanding the sense definitions extracted from BabelNet with words extracted from tweets, using a word embedding model trained on tweets that contain emoji. We also plan to evaluate the usability of EmojiNet by first defining the emoji sense disambiguation and emoji similarity finding problems, and then applying EmojiNet to disambiguate emoji senses based on different contexts in which they appear. We are working on applying EmojiNet to improve sentiment analysis and exposing EmojiNet as a web service.

Acknowledgments

We are grateful to Sujan Perera for thought-provoking discussions on the topic. We acknowledge partial support from the National Institute on Drug Abuse (NIDA) Grant No. 5R01DA039454-02: “Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use”, National Institutes of Health (NIH) award: MH105384-01A1: “Modeling Social Behavior for Healthcare Utilization in Depression”, and Grant No. 2014-PS-PSN-00006 awarded by the Bureau of Justice Assistance. The Bureau of Justice Assistance is a component of the U.S. Department of Justice’s Office of Justice Programs, which also includes the Bureau of Justice Statistics, the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, the Office for Victims of Crime, and the SMART Office. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice, NIH or NIDA.

References

1. Emogi research team - 2015 emoji report (2015)
2. Balasuriya, L., Wijeratne, S., Doran, D., Sheth, A. Finding street gang members on twitter. The 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2016); San Francisco, CA, USA. 08 2016; p. 685-692.
3. Basile P, Caputo A, Semeraro G. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. COLING. 2014:1591–1600.

4. Camacho-Collados J, Pilehvar MT, Navigli R. Nasari: a novel approach to a semantically-aware representation of items. *Proceedings of NAACL*. 2015:567–577.
5. Davis M, Edberg P. Unicode emoji - unicode technical report #51. *Technical Report*. 2016; 51(3)
6. Dimson T. Emojineering part 1: Machine learning for emoji trends. *Instagram Engineering Blog*. 2015
7. Kelly R, Watts L. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*. 2015
8. Miller H, Thebault-Spieker J, Chang S, Johnson I, Terveen L, Hecht B. “blissfully happy” or “ready to fight”: Varying interpretations of emoji. *ICWSM’16*. 2016
9. Moro A, Navigli R, Tucci FM, Passonneau RJ. Annotating the masc corpus with babelnet. *LREC*. 2014:4214–4219.
10. Navigli R, Ponzetto SP. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*. 2012; 193:217–250.
11. Novak PK, Smailovi J, Sluban B, Mozeti I. Sentiment of emojis. *PloS one*. 2015; 10(12):e0144296. [PubMed: 26641093]
12. Pavalanathan U, Eisenstein J. Emoticons vs. emojis on twitter: A causal inference approach. 2015 arXiv preprint arXiv:1510.08480.
13. Rezabek L, Cochenour J. Visual cues in computer-mediated communication: Supplementing text with emoticons. *Journal of Visual Literacy*. 1998; 18(2):201–215.
14. Santos, R. Java image processing cookbook. 2010. <http://www.lac.inpe.br/JIPCookbook>
15. SwiftKey, P. Most-used emoji revealed: Americans love skulls brazilians love cats the french love hearts [blog]. 2015. <http://bit.ly/2c5biPU>
16. Vasilescu, F., Langlais, P., Lapalme, G. Lrec. 2004. Evaluating variants of the lesk approach for disambiguating words.
17. Wang, W., Chen, L., Thirunarayan, K., Sheth, AP. Harnessing twitter “big data” for automatic emotion identification. *Privacy, Security, Risk and Trust (PAS-SAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*; IEEE; 2012. p. 587-592.
18. Wijeratne, S., Balasuriya, L., Doran, D., Sheth, A. *IJCAI Workshop on Semantic Machine Learning (SML 2016)*. CEUR-WS; New York City, NY: Jul. 2016 Word embeddings to enhance twitter gang member profile identification; p. 18-24.




					
Sense	Example	Sense	Example	Sense	Example
Laugh (noun)	I can't stop laughing 😂	Kill (verb)	He tried to kill one of my brothers last year. 🖱️🖱️	Costly (Adjective)	Can't buy class la 💰
Happy (noun)	Got all A's but 1 😂😁	Shot (noun)	Ooooooh shots fired! 🖱️🖱️	Work hard (noun)	Up early on the grind 💰
Funny (Adjective)	Central Intelligence was damn hilarious! 😂	Anger (noun)	Why this the only emotion I know to show anger? 🖱️	Money (noun)	Earn money when one register /w ur link 💰

Fig. 1.
Emoji Usage in Social Media with Multiple Senses.

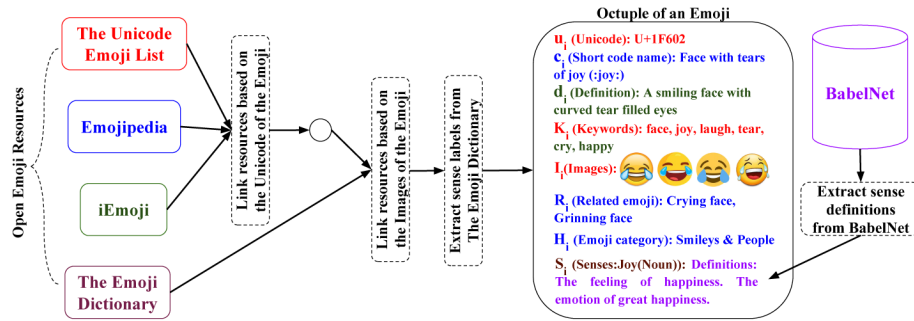


Fig. 2. Building EmojiNet by Integrating Multiple Open Resources.

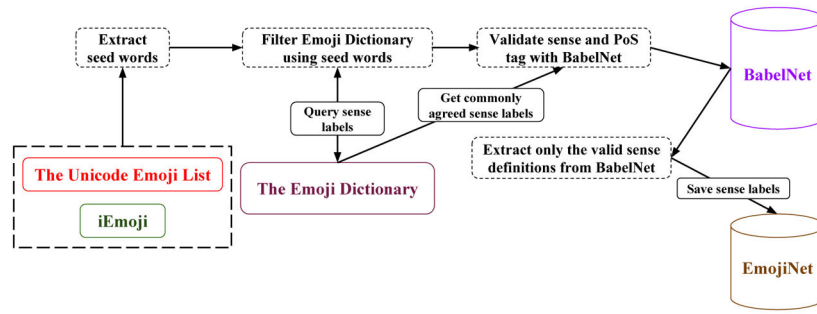


Fig. 3.
Emoji Sense and Part-of-Speech Filtering.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

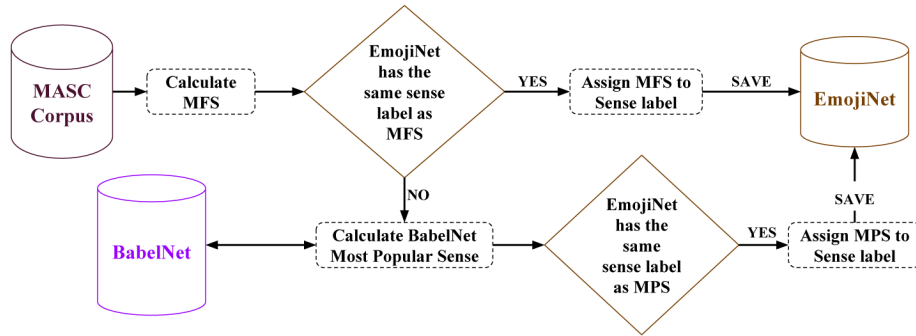


Fig. 4.
Using BabelNet to Assign Sense Definitions.

Table 1

Emoji Data Available in Open Resources

Emoji Resource	u	c	d	K	I	R	H	S
Unicode Consortium	✓	✓	X	✓	✓	X	X	X
EmojiJpedia	✓	✓	✓	X	✓	✓	✓	X
iEmoji	✓	✓	✓	✓	✓	X	✓	X
The Emoji Dictionary	X	X	X	X	✓	X	X	✓

Table 2

EmojiNet Statistics

Emoji Statistic	<i>u</i>	<i>c</i>	<i>d</i>	<i>K</i>	<i>I</i>	<i>R</i>	<i>H</i>	<i>S</i>
Number of emoji with each feature	1,074	845	1,074	1,074	1,074	1,002	705	875
Amount of data stored for each feature	1,074	845	1,074	8,069	28,370	9,743	8	3,206

Table 3

Word Sense Disambiguation Statistics

	Correct	Incorrect	Total
Noun	1,271 (83.28%)	255 (16.71%)	1,526
Verb	735 (84.00%)	140 (16.00%)	875
Adjective	725 (90.06%)	80 (9.93%)	805
Total	2,731 (85.18%)	475 (14.81%)	3,206

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript