The ELCo Dataset: Bridging Emoji and Lexical Composition

Zi Yun Yang, Ziqing Zhang, Yisong Miao
Web IR / NLP Group (WING), National University of Singapore
{ziyun_yang, zhangziqing}@u.nus.edu, yisong@comp.nus.edu.sg

Abstract

Can emojis be composed to convey intricate meanings like English phrases? As a pioneering study, we present the **E**moji-**L**exical **Co**mposition (ELCo) dataset, a new resource that offers parallel annotations of emoji sequences corresponding to English phrases. Our dataset contains 1,655 instances, spanning 209 diverse concepts from tangible ones like "right man" (✔️👨) to abstract ones such as "full attention" (🧐✍️, illustrating a metaphoric composition of a focusing face and writing hand). ELCo enables the analysis of the patterns shared between emoji and lexical composition. Through a corpus study, we discovered that simple strategies like direct representation and reduplication are sufficient for conveying certain concepts, but a richer, metaphorical strategy is essential for expressing more abstract ideas. We further introduce an evaluative task, Emoji-based Textual Entailment (EmoTE), to assess the proficiency of NLP models in comprehending emoji compositions. Our findings reveals the challenge of understanding emoji composition in a zero-shot setting for current models, including ChatGPT. Our analysis indicates that the intricacy of metaphorical compositions contributes to this challenge. Encouragingly, models show marked improvement when fine-tuned on the ELCo dataset, with larger models excelling in deciphering nuanced metaphorical compositions.

Keywords: Corpus (Creation, Annotation, etc.), Lexicon, Lexical Database, Semantics, Textual Entailment and Paraphrasing

1. Introduction

Emojis, with their rich visual and emotional lexicon, traditionally enhance verbal languages like English, for example, 👍 signifies agreement. Even though without a standardized grammar, users have intuitively developed conventions for crafting longer emoji sequences: 👍👍👍 signifies strong recommendation, while 🎊🥳🍾🎁🎈🎉 symbolises celebration. This prompts the question: Can the composition of emojis convey complex meanings on their own?

Words merge to craft meanings: "baby oil" is perceived as oil *made for* babies, not *extracted from* them. Similarly, consider the pairing of emojis like ❄️ and 🐻, the common interpretation leans towards bear *living in* snowy habitats, rather than a bear *consuming* snow. This leads us to explore the underlying strategies that guide emoji composition.

To the best of our knowledge, no existing dataset offers systematic annotations for emoji composition. Most prior studies incorporate emojis only as supplementary features to text comments (Na'aman et al., 2017; Barbieri et al., 2018). Exceptions to this trend are primarily confined to theoretical analysis (Cohn et al., 2019; Wicke and Bolognesi, 2020) or are restricted in scale due to data limitations (Shoeb and de Melo, 2021). To address this research gap, our study makes an initial exploration into lexical-level emoji compositions, which we believe form the foundation for tackling more complex sentence-level tasks. To this end, we introduce the ELCo dataset, offering parallel annotations of English phrases and corresponding emoji sequences. Our analysis of ELCo reveals that emoji composition follows certain rules. Simple strategies, such as direct mapping, are employed for tangible concepts (✔️👨 ≡ "*right man*"). Perhaps more interestingly, more intricate strategies are preferred when expressing metaphorical meanings (🧐✍️ entails "*full attention*" due to the combination of a focusing face and a writing hand) or using a series of emojis to suggest a concept (😎👨‍🔬👩‍⚕️🧑‍💻 entails "*bright future*" as the emojis together suggest a joyful state and prosperous professions). We also observe a slight correlation that the more metaphorical the phrase, the greater the diversity in emoji choices. These strategies not only present unique challenges to emoji composition, but also make it an intriguing subject of study.

To benchmark NLP models' ability to comprehend emoji composition, we introduce an Emoji-based Textual Entailment (**EmoTE**) task. This task assesses a model's capability to determine if an emoji sequence entail an English phrase, demanding a deep comprehension of emoji composition. Our task draws inspiration from Lyu et al. (2022) who studied complex lexical compositions. Our findings indicate that EmoTE is challenging for BERT, RoBERTa, and BART models, as they only modestly outperform a random baseline (achieving 60% accuracy compared to 50%), and their per-

LREC-COLING 2024, pages 15899–15909

formance significantly falls relative to their performance on their standard training dataset (MNLI for textual entailment). Next, we explore if fine-tuning on the ELCo dataset can enhance these models' performance on the EmoTE task. Post fine-tuning, we observe an uptick in accuracy across all models. Notably, metaphorical compositions appear to be the most challenging to interpret, although larger models exhibit a superior capacity to learn these representations. Finally, although not replicable, we evaluate the performance of ChatGPT on the EmoTE task out of interest. ChatGPT faces similar challenges to BERT models under a zero-shot setting (achieving 60% accuracy), highlighting the unique challenges inherent in emoji composition.

Our contributions can be summarized as follows:

- We present ELCo dataset as a pioneering resource to analyze the parallels and disparities between the composition of emojis and traditional languages.[1]

- Leveraging the EmoTE task, we validate comprehending emoji composition as a challenging task, with initial model accuracy around 60%, and highlight ELCo's role in boosting this performance to > 80%, thus aiding the learning of human methods for composing emoji sequences (§5-6);

- Through an in-depth corpus study, we observe that metaphorical representations are both abundant and intriguing. Humans tend to subtly favor diverse emoji choices in such contexts, underscoring the inherent complexity and challenge of this representation (§4).

2. Related Work and Background

2.1. Lexical Compositions

Lexical composition involves deriving the meaning of larger linguistic units by combining individual lexical items in a grammatical manner (Reddy et al., 2011; Yu and Ettinger, 2020). Modeling lexical composition is challenging because the compound's meaning goes beyond a simple sum of its constituent words. The NLP community has identified challenges: (1) **Meaning shift** occurs when the phrase's meaning deviates from that of its individual words. For example, the English verb particle construction "give up" conveys a different sentiment than the standalone verb "give" and the word "up" (Tu and Roth, 2012). (2) **Implicit meaning** often requires world knowledge to uncover, as seen in "hot debate" referring to emotional intensity, rather than temperature (Hartung, 2015). BERT models still struggle to comprehend both meaning shift and implicit meaning (Shwartz and Dagan, 2019).

Motivated by these challenges, we investigate whether emoji compositions in ELCo face similar difficulties as lexical compositions.

2.2. Emoji Corpora and Representations

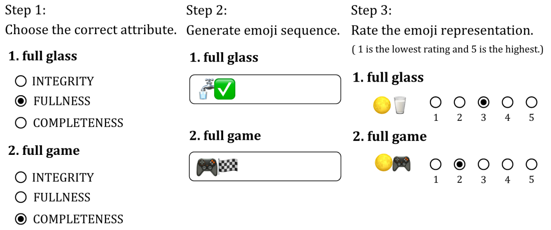
Emoji corpora primarily focus on the relationships between individual emojis, or between emojis and textual content, with scant attention given to emoji composition: EmojiNet (Wijeratne et al., 2017b) privides a WordNet (Miller, 1995) alike emoji sense inventory that links Unicode emoji representations to their English meanings extracted from the Web. (Shoeb and de Melo, 2020) introduced the EmoTag1200 corpus to study the association between Emoji and emotions. (Wijeratne et al., 2017a) contribute a EmoSim508 corpus to facilitate a semantic measure of emoji relatedness. The ambiguity of emojis are also recognized in context (Miller et al., 2017) or out-of-context (Częstochowska et al., 2022). Na'aman et al. (2017) annotated various linguistic purposes of emojis, encompassing functional, content, and multimodal usages. Kirk et al. (2022) presented a benchmark for emoji-based hate speech detection. Yet, a clear gap remains: none of these resources center on emoji compositions. A notable outlier is the Unicode Emoji Zero Width Joiner (ZWJ emojis) (Davis and Edberg, 2015). It melds existing emojis, such as "polar bear" 🐻‍❄️ = 🐻 + ❄️, to aid users on older platforms. Nevertheless, after filtering for skin-tone or gender modifiers, only 33 distinct ZWJ emojis remain in Unicode Version 14.0[2], making it inadequate for computational analysis.

Emoji representations, like Emoji2vec (Eisner et al., 2016) and embeddings by Barbieri et al. (2016), enhance various downstream tasks by providing expressive emoji embeddings. Although recent studies evaluate emoji usage in NLP systems (Shoeb and de Melo, 2021), none delve into the nuanced meanings of emoji compositions.

Using these representations, various applications have been developed, including English-to-emoji translation (Day et al., 2020), sentence-ending emoji prediction (Barbieri et al., 2018), sentiment analysis (Felbo et al., 2017), and cross-lingual learning (Chen et al., 2019). However, most over-

[1] ELCo dataset and code are publicly available at `http://github.com/WING-NUS/ELCo`

[2] `http://www.unicode.org/Public/emoji/14.0/emoji-zwj-sequences.txt`, accessed on 23th March 2024.



Step 1:
Choose the correct attribute.

1. full glass

○ INTEGRITY
● FULLNESS
○ COMPLETENESS

2. full game

○ INTEGRITY
○ FULLNESS
● COMPLETENESS

Step 2:
Generate emoji sequence.

1. full glass

🚰✅

2. full game

🎮🏁

Step 3:
Rate the emoji representation.
(1 is the lowest rating and 5 is the highest.)

1. full glass

🌕🥛 ○ 1 ○ 2 ● 3 ○ 4 ○ 5

2. full game

🌕🎮 ○ 1 ● 2 ○ 3 ○ 4 ○ 5

Figure 1: ELCo's annotation process consists of three steps: (1) select the attribute of the phrase, (2) execute the annotation, and (3) rate the output from a rule-based system, Emojinating.

look the nuanced compositionality of emojis. Emojinating (Cunha et al., 2018), a rule-based system, generates fixed two-emoji sequences using dictionary lookups from ConceptNet (Speer et al., 2017) and EmojiNet (Wijeratne et al., 2017b), but fails to address deeper emoji compositionality. Cohn's Grammar (Cohn et al., 2019) explores a potential grammatical structure for emoji use and introduces a taxonomy of grammars for emoji sequences, but it does not offer parallel annotations akin to ELCo.

3. Emoji–Lexical Composition (ELCo)

Given the constrained size of the ZWJ emojis, we propose a systematic workflow to annotate emoji composition given English phrases.

In our current study, we focus solely on Adjective–Noun (AN) compounds as one simplistic type of lexical composition, as motivated by Fyshe et al. (2019). To construct our dataset, we extract AN compounds from the HeiPLAS dataset (Hartung, 2015), which has been extensively studied (Shwartz and Dagan, 2019). The HeiPLAS dataset comprises 1,598 AN compounds, with each compound's adjective labeled with its corresponding attribute meaning. For instance, the adjective *full* in the phrase *"full glass"* is associated with the attribute *fullness*, while in the phrase *"full game"*, it corresponds to the attribute *completeness*. We sort the 1,598 AN compounds in descending order based on the number of attribute choices and the frequency of the adjectives. From this ranking, we select the top 209 AN compounds, which encompass 45 adjectives and 77 attributes.

3.1. Annotation Guideline

We collect emoji representation of AN compounds from human annotators. We design a three-step annotation workflow (Figure 1). **First,** annotators are tasked with selecting the correct attribute meaning from several possible choices for each provided phrase. The objective of this stage is understanding reinforcement — we ensure the annotators have an accurate understanding of the phrase's meaning by thinking deeply about it. **Second,** annotators construct an emoji sequence that most effectively represents the given AN compound. We emphasize that the order of the emojis in these sequences is critical. Moreover, we have not placed any restrictions on the length of the emoji sequences. The emojis should reveal the implicit or concealed meaning within the concept. For instance, "*full glass*" could be represented by 🚰✅, signifying *fullness*, while "*full game*" might be denoted by 🎮🏁, illustrating *completeness*. To facilitate their annotation process, annotators can use `emojicopy.com`, a web-based copy-and-paste interface for emoji and emoji search by keyword. **Third,** annotators assess the output from Emojinating, which, to the best of our knowledge, stands as the sole rule-based system specifically crafted to produce emoji sequences from distinct concepts. Annotators rate Emojinating's performance on a 5-point Likert scale, where 1 signifies the lowest possible rating, and 5 the highest. The intent of this stage is to discern the limitations of a rule-based system when contrasted with human judgment.

3.2. Annotation Details

We recruited a group of 40 university students, aged between 18 and 30, after obtaining approval from the Institutional Review Board (IRB). On average, each participant annotated around 41 emoji sequences corresponding to unique English phrases. The annotators were compensated fairly based on the number of samples they annotated. We collected a total of 1,655 responses for 209 English phrases, with each phrase receiving 7 or 8 annotations. Since we received multiple emoji sequence annotations for a single phrase, **each annotation is treated as an independent valid entry**.

Validating the ELCo Dataset: Given the creative nature of emoji composition, establishing a definitive gold standard for annotation validity is challenging. As a solution, two of our authors manually checked all 1,655 instances to ensure their coherence and relevance. This was conducted in conjunction with our corpus study (§4). Their primary objective was the determination of compositional strategies, but in the process, they also validated the data. Upon thorough examination, **both authors identified no malicious annotations**, affirming the validity of all instances.

English Phrase	Attribute	ELCo's Annotations (length = 2)	ELCo's Annotations (length > 2)	Average length of ELCo's Annotation	Emojinating's Output	Emojinating Rating
full attention	INTEGRITY	💯❗ and 🧐✍️	🙋‍♂️👂🏫	2.43	🌕🛎️	3.1
full glass	FULLNESS	🥛🈵 and 🚰✅	🥛➕🚰🚰🚰🚰🚰🈵	2.71	🌕🥛	3.0
full game	COMPLETENESS	🎮🏁 and 🎮💯	💯🆓🎮	2.29	🌕🎮	2.7
full auditorium	FULLNESS	🈵🏫	👩‍💻👨‍💻🏫 and 🎫👱👨🧑‍🦱👩‍🦱🈵	4.14	🌕⛹️	2.0
full life	FULLNESS	😊❤️	❤️👨‍👩‍👧💼💰💰💰💯 and 😆🥰💏	4.00	🌕😔	1.1

Table 1: **ELCo Dataset Samples:** ELCo's human-annotated emoji sequences showcase diverse, equally valid representations. The Emojinating system, deemed literal and less coherent, often receives low ratings. To ensure a fair comparison with Emojinating, which produces outputs limited to a length of 2, we separately showcase ELCo samples of length 2 or greater.

Comparing ELCo with Emojinating: To underscore the need for ELCo, we compare its instances with the Emojinating system, the sole prior study for generating emoji sequences. As shown in Table 1, the Emojinating system presents average ratings ranging from 1.1 to 3.1 (of 5), indicating low coherence when assessed by humans. The reason is that Emojinating typically opts for word-to-word translations by matching emojis with the closest literal meanings, resulting in a lack of coherence. While the average emoji sequence length for the entire ELCo dataset stands at 2.59, the Emojinating system is limited to a fixed length of 2. This distinction requires us to make a separate comparison using ELCo samples of similar length. For the phrase "full glass", Emojinating chooses emojis that are semantically closest to the words "full" and "glass", producing the sequence 🌕 and 🥛. This sequence is not intuitive and introduces ambiguity since 🌕 is also associated with unrelated concepts like nighttime or lunar cycles. On the other hand, our human annotations convey a richer meaning. When the length is 2, ELCo's sequence uses 🚰 and ✅, aptly representing a glass filled to its brim and the completion of the filling action. For sequences longer than 2, the progression 🥛➕🚰🚰🚰🚰🚰🈵 illustrates the process of pouring water into a container until it reaches 🈵 (symbolizing "no vacancy" in Japanese), further emphasizing the nuanced, dynamic representation of the action. This discrepancy underscores the importance of our human annotation in capturing nuanced human approaches to emoji composition.

4. Corpus Study

Leveraging the ELCo dataset, we delve into the potential rules and tactics our annotators employed to assemble emojis into coherent sequences. Following the taxonomy of grammar introduced by Cohn et al. (2019), we discern several predominant compositional strategies within our ELCo dataset:

(1) Direct representation is a straightforward approach to map each word in English phrase to one or a few emojis (e.g. the direct mapping in Ex. 1 ✔️👨). **(2) Metaphorical Representation:** This approach involves a sequence of emojis that together embody the metaphorical meaning of an English phrase. Such a strategy includes: Representing a sequence of events to convey a concept (e.g., falling in love 💏 and raising children 👨‍👩‍👦‍👦 in Ex. 2 signifies the concept of the "right man"), composing a metaphorical phrase that encapsulates the event (e.g., a teacher giving a 100 mark to an explanation in Ex. 3 depicts clear explanation), and creating a scene that hints at the event (e.g., the depiction of a stove, clock, bread, and hot spring in Ex. 4 suggests freshly baked bread). **(3) Semantic List:** A list of related emojis is used to imply the concept's meaning (e.g., the assembly of occupation-related emojis in Ex. 5 implies a promising future). **(4) Reduplication** as a method to accentuate the intensity of an adjective, and **(5) Single Emoji Representation**, where a solitary emoji captures the essence of the concept.

	Compositional Strategy	EN Phrase	Emoji Sequence
Ex. 1	Direct	right man	✔️👨
Ex. 2	Metaphorical	right man	💏👨‍👩‍👦‍👦
Ex. 3	Metaphorical	clear explanation	👩‍🏫💯🗣️
Ex. 4	Metaphorical	fresh bread	🎛️⏰🍞♨️
Ex. 5	Semantic list	bright future	😎👨‍🔬👩‍⚕️👨‍💻
Ex. 6	Reduplication	big group	👬👬👬
Ex. 7	Single	right thing	✔️

Table 2: Examples for compositional strategies.

Annotation: Two authors annotated the ELCo dataset, checking the quality of the data and assigning one of five compositional strategies to each pair of emoji sequence and English phrase. They began with a pilot of 119 instances covering 15 phrases, resolving initial disagreements through discussion. They then annotated the remaining dataset, with a subset of 229 overlapping instances (15%). Agreement was achieved on 210 pairs, with an agreement rate of 91.7%.

From the annotated compositional strategies, we

ID: EN PHRASE	Human Annotation Samples	Main Compositional Strategy	Jaccard Similarity
EN_1: WRONG MEDICINE	❌💊, 🙅‍♀️💊, ❌💊🤒, 🙅‍♂️💊	Direct	0.57
EN_2: WRONG ROAD	❌🛣️, 🙅‍♀️🛣️, 🙅🛣️, 🙅‍♂️🛣️	Direct	0.51
EN_{-2}: FAR SIDE	🧍🌍🧍, ↙️🛣️, 😪👉, 🚫🔜📏	Metaphorical	0.0
EN_{-1}: IMMEDIATE INFLUENCE	🧑‍🤝‍🧑↔️❗, 💃🕺🙋🙋‍♀️🙋‍♂️, ⏩👩‍⚖️👨‍⚖️	Metaphorical	0.0

Table 3: **Top 2 and Bottom 2 Examples Based on Jaccard Similarity:** Phrases with higher similarity scores usually relate to tangible concepts, whereas those with lower scores tend to be more abstract.

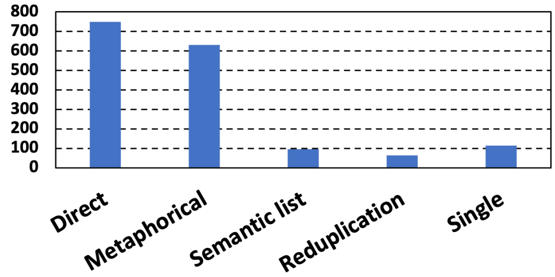


Figure 2: Number of compositional structures identified in our corpus study (1,655 samples in total).

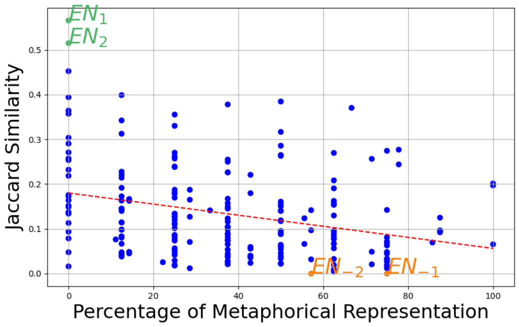


Figure 3: Impact of a phrase's metaphorical representation percentage on its Jaccard similarity score. Linear regression is shown as the dashed line. A low score indicates diverse emoji choices.

derived two key insights: **Insight 1: Metaphorical strategy is widely employed.** Figure 2 illustrates that, although direct representation is the preferred choice (700+ instances), the metaphorical strategy is also common (600+ instances), indicating a propensity for indirect expressions when direct mappings are challenging. **Insight 2: Metaphorical strategies often adopt more diverse emoji choices.** In Figure 3, the X-axis depicts the percentage of metaphorical representation for EN phrases. The left end highlights literal concepts, typically associating a tangible noun (e.g., medicine, water) with an emoji-compatible adjective (e.g., wrong, hot). Conversely, the right end features concepts such as business or flavor, which are frequently conveyed metaphorically.

Dataset Diversity: Contrary to studies that prioritize **agreement** among annotations, our emphasis is on gauging the **diversity** within our dataset. This choice is rooted in the nature of our task, which is fundamentally a creative generation endeavor. Our corpus study also finds no malicious annotations. We use the pairwise Jaccard similarity (Skjærholt, 2014) to measure the overlap between emoji sequences annotated by distinct annotators. Specifically, we define $J_{avg} = \sum_{i \neq j} J(A_i, A_j) / \frac{n*(n-1)}{2}$, where n is the total number of emoji sequences. Here, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ represents the Jaccard similarity computation. A lower Jaccard similarity reflects a broader variety of emoji selections.

Figure 3 plots the Jaccard similarity against the metaphorical percentage of English phrase annotations, revealing a slight correlation ($k = -0.123$ by linear regression fit). This indicates that more metaphorical phrases often uses more diverse emoji choices. Table 3 shows the top and bottom instances by similarity, highlighting that tangible concepts often use common emojis like 💊 and 🛣️. In contrast, metaphorical concepts tend to have diverse emoji representations. This aligns with Wicke and Bolognesi (2020), suggesting that intangible concepts yield varied representations. Our dataset's average Jaccard similarity of 0.13 emphasizes diverse emoji interpretations, suggesting multiple valid representations for a single concept.

5. Task Formalization

We adopt an Emoji-based Textual Entailment (**EmoTE**) task to examine the capacity of models to comprehend the composition of emojis. Lyu et al. (2022) also employs textual entailment to study model's understanding of nuanced meanings in lexical composition.

Emoji-based Textual Entailment (EmoTE) determines if a sequence of emojis *EM* implies a English phrase *EN*; formally,

Premise: **P em**$_1$, **em**$_2$, ... **em**$_n$.
Hypothesis: **P en**$_1$, **en**$_2$, ... **en**$_n$.
Labels: `Entailment` | `Non-entailment`.

where **P** is a sentence prefix like *This is*. Emoji tokens **em**$_1$, **em**$_2$, ... **em**$_n$ compose an emoji sequence EM. Similarly, English tokens **en**$_1$, **en**$_2$, ... **en**$_n$ form an English phrase EN.

Input	Golden Label
Premise: This is 🧐✍️. Hypothesis: This is full attention.	Entailment
Premise: This is 🚰✅. Hypothesis: This is full attention.	Non-Entailment

Table 4: Examples for EmoTE.

We posit that an English phrase EN is entailed by an emoji sequence EM if the sequence captures the phrase's meaning. For instance, as seen in Table 4, the emoji sequence comprising a focusing face and a writing hand (🧐✍️) metaphorically entails "full attention". Conversely, the sequence 🚰✅ does not entail "full attention" because it expresses an irrelevant topic (glass of water). We follow Lyu et al. (2022) to simplify the three-way classification to a binary task of predicting entailment or its absence.

Given the complex nature of emoji sequences, we have not yet structured our task in a generative setting. As our corpus study demonstrates, a single English phrase can be validly represented by various emoji sequences, employing different compositional strategies. This inherent diversity, evidenced by the low Jaccard similarity, complicates both the generation and evaluation of emoji compositions. Consequently, our current focus is on understanding emoji composition, with plans to explore generation in future research.

6. Experiments

We conduct our experiments in pursuit of the following research questions:

- RQ1: Can existing models comprehend emoji compositions, and can they be improved by fine-tuning on ELCo ?
- RQ2: Is ELCo sufficient in size?
- RQ3: What are the specific instances of success and failure in model performance?
- RQ4: How does ChatGPT perform on ELCo?

6.1. Experiment Setup

Sampling: ELCo offers annotations for 209 English phrases, with each phrase having multiple emoji representations, totaling 1,655 annotations. We regard each of these emoji sequences associated with an English phrase as a separate and equally valid representation. Inspired by the evaluation protocol in Shwartz and Dagan (2019), we generate negative samples through a noun-flipping technique. For a given AN phrase, we select an emoji sequence from another phrase, AN', where the sole distinction between AN' and the original AN is the noun. This results in the pairing (AN, EM') acting as a negative sample, given that EM' doesn't encapsulate AN. Using this strategy, we produce one negative sample for each dataset entry, leading to an aggregate of 3,310 instances.

Dataset Split: We distribute the dataset in a rough 70:15:15 ratio for training, validation, and testing, respectively. To avoid shortcut learning and to evaluate genuine compositional skills, we ensure no adjective overlap exists between the training set and the validation or testing datasets, which is also employed in (Shwartz and Dagan, 2019). This results in 2,398 training instances, 394 validation instances, and 518 testing instances, with a balanced positive to negative ratio of 1:1.

Models: As per Lyu et al. (2022), we utilize BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2020) models, each pre-fine-tuned on the MNLI dataset (Williams et al., 2018) for textual entailment capabilities. In line with Lyu et al. (2022), we use available model checkpoints on HuggingFace, given the extensive size of the MNLI dataset, and consider the sum of `neutral` and `contradiction` scores as indication of `non-entailment`. To address these models' lack of emoji tokenization, we tokenize emoji descriptions, thereby preserving semantics and focusing on our study of compositions.

Training Details: Based on models fine-tuned on MNLI, we continue fine-tuning them on ELCo using Adam with a learning rate of 0.0001. Typically, training converges within ten epochs and takes approximately 30 minutes on two Titan RTX GPUs with 24GB RAM. We perform multi-run with 5 different random seeds and report the average performance.

Evaluation Metric: We primarily measure models' performance by accuracy score.

6.2. Evaluation and Fine-tuning (RQ1)

Initially, we evaluate the overall performance of models pre-fine-tuned on MNLI against our EmoTE task. These models attain an accuracy of approximately 60% on EmoTE (Figure 4), in contrast to their 80+% accuracy on the standard textual entailment benchmark, MNLI. This significant gap underscores the complexity introduced by emoji

	w/o fine-tuning on ELCo					Fine-tuned on ELCo				
	BERT$_{base}$	RoBERTa$_{base}$	RoBERTa$_{large}$	BART$_{large}$	Avg	BERT$_{base}$	RoBERTa$_{base}$	RoBERTa$_{large}$	BART$_{large}$	Avg
Direct	34.7	35.6	41.5	51.7	40.9	$85.1_{2.3}$	$89_{3.3}$	$91.9_{2.2}$	$88.5_{2.8}$	88.6
Metaphorical	19.4	24.7	34.4	36.6	28.8	$68.4_{3.9}$	$73.1_{2.6}$	$80.4_{1.9}$	$82.8_{4.3}$	76.2
Semantic list	33.3	41.7	50.0	58.3	45.8	$86.7_{7.5}$	$78.3_{4.6}$	$85.0_{3.8}$	91.7_{0}	85.4
Reduplication	13.3	0	6.7	0	5.0	$65.4_{3.0}$	$52.0_{7.3}$	$62.7_{6.0}$	$88.0_{8.7}$	67.0
Single	19.0	19.0	19.0	52.4	27.4	$66.7_{3.4}$	$88.6_{2.6}$	$85.7_{4.8}$	$83.8_{2.6}$	81.2
Negative	83.4	83.0	90.3	82.2	84.7	$84.3_{4.1}$	$87.2_{1.9}$	$85.2_{0.7}$	$84.8_{1.2}$	85.4
Overall	55.0	55.8	62.9	62.9	59.2	$80.4_{1.5}$	$84.0_{0.8}$	$85.2_{0.9}$	$85.5_{0.9}$	83.8

Table 5: **Fine-grained Analysis (RQ1):** Accuracy of models, pre and post-fine-tuning of models across various compositional strategies in the EmoTE task. Models are fine-tuned over 5 runs with different seeds, with performance reported as a mean and standard deviation (as $_{\text{subscript}}$). We also present the average scores across the four models, and underline the three most challenging categories: metaphorical, reduplication, and single emojis.

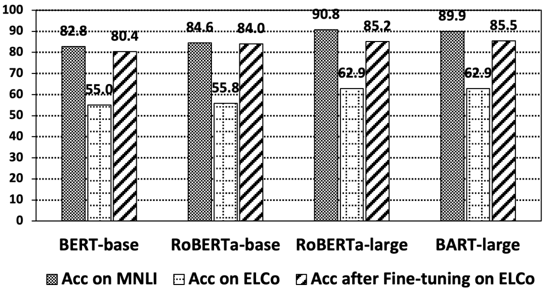


Figure 4: **Overall Performance (RQ1):** A comparison of model accuracy on the standardly evaluated MNLI dataset (1st bar) and ELCo (2nd bar), highlighting a significant drop between them. The accuracy improves following fine-tuning (3rd bar).

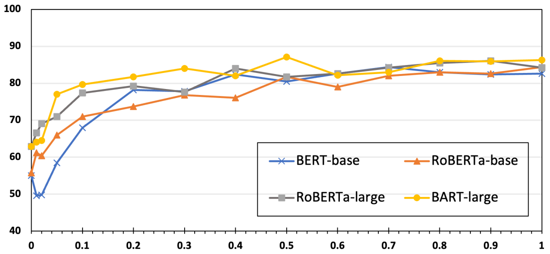


Figure 5: **Scaling Experiment (RQ2):** Accuracy trends across varying proportions of training data, showcasing a sharp rise up to 0.1 and reaching near convergence post 0.5.

composition, highlighting the limitations of current datasets in fully capturing this aspect. However, after fine-tuning on ELCo, we observe a substantial improvement in performance, which implies that models can effectively acquire and utilize emoji composition skills.

We then proceed to a more granular examination of model performance in relation to distinct composition strategies, guided by our corpus analysis. We find that: (1) Before fine-tuning on ELCo, models exhibit limited understanding of metaphorical compositions (accuracy only 28.8%) and struggle with reduplication (5.0% Acc). (2) All models display substantial improvement after fine-tuning on ELCo, particularly in the reduplication strategy, where accuracy leaps to 67.0. This substantial gain, observed even for a straightforward strategy, signals the efficacy of ELCo in facilitating learning of compositional patterns. (3) Despite the prevalence of metaphorical representation in our dataset (constituting 38% of the data), models only reach an accuracy of 76.0% in this category post-fine-tuning. This marks this category — along with reduplications and single emojis — as consistently challenging. (4) We observe a consistent trend of larger models performing better in metaphorical representation, with two of the larger models achieving accuracies of 80.4 and 82.8 post-fine-tuning, suggesting their enhanced capability to grasp nuanced metaphorical compositions.

6.3. Validating Data Sufficiency (RQ2)

It is of interest to determine whether our dataset is sufficient for effective model learning. We replicate the fine-tuning experiments by progressively adding incremental portions of the training data. Our observations include: (1) All models converge and plateau after utilizing half of the training data. This implies that our dataset is sufficiently large for current model learning paradigms. (2) Larger models perform better, even when provisioned with smaller data volumes. For instance, we observe a substantial increase in BART$_{large}$'s performance, even with just 2–% of ELCo, while BERT$_{base}$ takes more, requiring 20% of the training data to attain comparable performance.

6.4. Result Analysis (RQ3)

English	Emoji	Pre	Post
1 big group	👩‍👩‍👦‍👦 👩‍👩‍👦‍👦 👩‍👩‍👦‍👦 👩‍👩‍👦‍👦 👩‍👩‍👦‍👦	×	✓
2 big city	🌇 🌁 🌆 🏙️ 🌃 🗽	×	✓
3 hot forehead	🥴 🤒 👨‍⚕️	×	✓
4 thin soup	💦 💦 🥣	×	✓
5 big city	🌃 🌃 🌃 🌃 🌃	×	×
6 ineffectual ruler	👨‍💻 👩‍💻 👨‍💻 👎 👎	×	×
7 full attention	🧐 ✍️	×	×
8 full life	🥰 ☺️ 🏡	×	×

Table 6: **(RQ3)** Case studies depicting the accuracy of predictions, pre- and post-fine-tuning.

We showcase $\text{BART}_{\text{large}}$, the best-performing model, by elucidating its successful and failed cases. Cases 1-4 in Table 6 reflect correct predictions post-ELCo finetuning. Notably, the model effectively learns Reduplication (Case 1) and Semantic list strategies (Case 2) to represent "big" in "*big group*" and "*big city*". Furthermore, it demonstrates a non-trivial grasp of metaphors; for instance, it associates a thermometer-bearing person visiting a doctor with a hot forehead (Case 3), and a droplet symbolizes thin soup (Case 4).

Cases 5-8 illustrate instances of unsuccessful predictions even after ELCo finetuning. Case 5 necessitates the **visual information** encapsulated in 🌃 (city night view), which is absent from its text description ("night-with-stars"). This implies the potential benefit of enhanced emoji representation with visual features. Case 6 calls for **commonsense knowledge** to understand that a 👨‍💻 ("technologist") could be an administrator (i.e., ruler). Cases 7 and 8 involve the use of emojis that are significantly **distant** from the phrase's meaning, demanding additional metaphorical reasoning.

6.5. ChatGPT's Performance on ELCo (RQ4)

Given the recent prominence of ChatGPT, assessing its competence in interpreting emoji compositions is of interest. We adopt a zero-shot evaluation approach in anticipation of user freedom in emoji composition. Our experiments employ the ChatGPT-3.5 version from May 24th, 2023, using the subsequent prompt template.

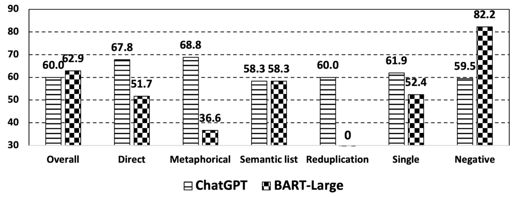


Figure 6: **(RQ4)** Accuracy scores of ChatGPT and $\text{BART}_{\text{large}}$, broken down by composition strategy.

> Please predict whether Sent1 entails Sent2 and provide a binary response (1 for true, 0 for false). Note that Sent1 contains emoji compositions, they can combine together to represent a meaning. Please just return a single response (1 or 0) for each line, indicating the entailment relationship.
> Sent1: This is 👔 📈.,Sent2: This is big business.
> ...

$\text{BERT}_{\text{base}}$	$\text{RoBERTa}_{\text{base}}$	$\text{RoBERTa}_{\text{large}}$	$\text{BART}_{\text{large}}$	ChatGPT
55.0	55.8	62.9↑	62.9↑	60.0

Table 7: **(RQ4)** Performance comparison of ChatGPT and other models (both under zero-shot setting); ↑ denotes a model outperforming ChatGPT.

Table 7 shows that ChatGPT's zero-shot performance only outperforms smaller models, underperforming larger ones in EmoTE. Its errors include: **(1) Ignoring visual cues,** missing the fullness implied in 🚰 and 🥛 for *"full glass"*; **(2) Struggling with compositional patterns,** as seen when it fails to associate 🤓 👨‍💻 👨‍🏫 👨‍🎓 with *"full attention"*, a shortcoming our fine-tuned models overcome by decoding such semantic list compositions.

Nonetheless, in Figure 6, we observe that ChatGPT tends to yield balanced results across different strategies in EmoTE, demonstrating a particularly impressive performance in Metaphorical representation, which poses the greatest challenge for other models. This capability could potentially be ascribed to its training on a more expansive corpus.

7. Conclusion and Future Work

We make a novel contribution towards the understanding of emoji composition by presenting the ELCo dataset. We unveil that emoji composition leverages various strategies, spanning from simple representations for tangible concepts to more sophisticated metaphorical approaches for abstract ones. An empirical evaluation highlighted the difficulties that contemporary models face when deciphering emoji compositions, particularly struggling with metaphorical compositions. Even though ELCo fine-tuning enhances comprehension, cases involving commonsense knowledge and visual information remain unresolved.

This research paves the way for future exploration. While our study primarily focused on comprehension, extending to a generative setting where models comprehend and generate emoji compositions could be beneficial. It is also intriguing to contextualize comprehension within textual surroundings that provide social and environmental grounding.

Ethics Statement

Our data collection is approved by institutional review board (IRB). IRB reviewed on our experimental design and research procedures, to ensure that the research involves no more that minimal risks to the research participants. In particular, we ensure that none of the phrases involve sensitive topics or would not elicit strong negative responses. We also ensure that research participants' privacy and the confidentiality of their research data will be protected.

As emoji sequence composition becomes increasingly prevalent among users, it is essential to account for cultural differences in emoji interpretation. We acknowledge that we have not considered the varied nuances across cultures and the diverse emoji representations on different operating systems (OS). We believe that a more comprehensive exploration of these facets can deepen our grasp of emoji compositions.

Acknowledgement

The authors are grateful to their advisor, Prof. Min-Yen Kan, for his great support for this project. This includes his contributions to the initial ideation and project scoping, the technical insights he provided during our frequent meetings, his detailed feedback on our manuscript, and his unwavering support despite earlier rejections. Without him, this paper could not have been published.

The authors also extend their thanks to João Miguel Cunha and Qing Lyu for their technical consultation via emails about their related works. The authors also thank anonymous reviewers for their valuable comments.

8. References

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval 2018 task 2: Multilingual emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *The World Wide Web Conference*, pages 251–262.

Neil Cohn, Jan Engelen, and Joost Schilperoord. 2019. The grammar of emoji? constraints on communicative pictorial sequencing. *Cognitive research: Principles and implications*, 4:1–18.

João Miguel Cunha, Pedro Martins, and Penousal Machado. 2018. Emojinating: Representing concepts using emoji. In *Workshop Proceedings from The 26th International Conference on Case-Based Reasoning (ICCBR 2018), Stockholm, Sweden*, volume 185.

Justyna Częstochowska, Kristina Gligorić, Maxime Peyrard, Yann Mentha, Michał Bień, Andrea Grutter, Anita Auer, Aris Xanthos, and Robert West. 2022. On the context-free ambiguity of emoji. In *International Conference on Web and Social Media*.

Mark Davis and Peter Edberg. 2015. Unicode emoji. *Unicode Technical Standard*, 51.

Alex Day, Chris Mankos, Soo Kim, and Jody Strausser. 2020. Confet: An english sentence to emojis translation algorithm. In *Proceedings of the 35th Annual Spring Conference of the Pennsylvania Computer and Information Science Educators*. Pennsylvania Computer and Information Science Educators.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Alona Fyshe, Gustavo Sudre, Leila Wehbe, Nicole Rafidi, and Tom M Mitchell. 2019. The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping*, 40(15):4457.

Matthias Hartung. 2015. *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Zheng Hua, Daoxin Li, Li Zhang, Marianna Apidianaki, and Chris Callison-Burch. 2022. Is "my favorite new movie" my favorite movie? probing the understanding of recursive noun phrases. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5286–5302, Seattle, United States. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 152–161.

Noa Na'aman, Hannah Provenza, and Orion Montoya. 2017. Varying linguistic purposes of emoji in (Twitter) context. In *Proceedings of ACL 2017, Student Research Workshop*, pages 136–141, Vancouver, Canada. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Abu Awal Md Shoeb and Gerard de Melo. 2020. EmoTag1200: Understanding the association between emojis and emotions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8957–8967, Online. Association for Computational Linguistics.

Abu Awal Md Shoeb and Gerard de Melo. 2021. Assessing emoji use in modern text processing tools. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1379–1388, Online. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Arne Skjærholt. 2014. A chance-corrected measure of inter-annotator agreement for syntax. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–944.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI Conference on Artificial Intelligence*.

Yuancheng Tu and Dan Roth. 2012. Sorting out the most confusing english phrasal verbs. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 65–69.

Philipp Wicke and Marianna Bolognesi. 2020. Emoji-based semantic representations for abstract and concrete concepts. *Cognitive Processing*, 21.

Sanjaya Wijeratne, Lakshika Balasuriya, A. Sheth, and Derek Doran. 2017a. A semantics-based measure of emoji similarity. *Proceedings of the International Conference on Web Intelligence*.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017b. Emojinet: An open service and api for emoji sense discovery. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

9. Language Resource References

Barbieri, Francesco and Ronzano, Francesco and Saggion, Horacio. 2016. *What does this emoji mean? a vector space skip-gram model for twitter emojis*. ELRA (European Language Resources Association). Calzolari N, Choukri K, Declerck T, et al, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016); 2016 May 23-28; Portorož, Slovenia. Paris: European Language Resources Association (ELRA); 2016. p. 3967-72. PID http://hdl.handle.net/10230/33776.

A. Direct emoji embedding

In this paper, we have represented emojis with their English descriptions. The reason is that BERT model cannot tokenize emojis (due to its use of WordPiece[3]). To ensure a consistent approach across models, we opted to use emoji descriptions as a practical alternative, despite RoBERTa and BART having the capability to encode emojis. We have also tried using these methods, but as documented below, they return subpar results.

We consider using emoji descriptions as a suitable proxy, considering that word embeddings are already trained to integrate semantics during the pre-training phase. Employing emoji descriptions yields performance comparable to direct emoji representation, especially in simpler tasks like sentiment classification.[4]

If using direct emoji embedding for our EmoTE task, the accuracy scores are only 50.0 for RoBERTa-base, 51.0 for RoBERTa-Large, and 52.3 for BART-Large, all without fine-tuning. Even after fine-tuning, the accuracy scores are 56.6, 53.5, and 52.1, respectively. These figures are notably lower than those achieved with our existing methodology. This suggests that these models might not have adequate direct emoji representations for our emoji composition study, which demands a more refined semantic representation.

[3] https://github.com/huggingface/transformers/issues/12190

[4] https://towardsdatascience.com/emojis-aid-social-media-sentiment-analysis-stop-cleaning-them-out-bb32a1e5fc8e