ELCo: Bridging Emoji Mashup and Lexical Composition

Yang Zi Yun

Project Supervisor: A/Prof. Min-Yen Kan

PhD Advisor: Yisong Miao

Agenda for today's talk (30 min + 10 min)

"Bridge at Night" Emoji

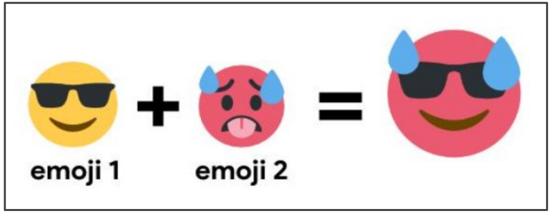
- Problem statement (2 min)
  - Emoji mashup
  - Lexical composition
- Related work (2 min)
- Dataset (4 min)
  - ZWJ dataset
  - ELCo-AN dataset
- Benchmark (2 min)
- Ranking Problem (6 min)
- Evaluation (Research Questions) (12 min)
- Summary (2 min)
- QnA & Discussions (10 min)

Problem Statement

ELCo: Bridging Emoji Mashup and Lexical Composition

Emoji mashup (emoji + emoji)

A Twitter Celebrity!



https://twitter.com/dsandler/status/884116974367907841
https://twitter.com/emojimashupbot

Lexical Composition (word + word)

The composition of **words**.

- shift the meanings of the constituent words.
- introduce implicit information.

- Verb-Particle Constructions
  - carry vs carry on
- Light Verb Constructions
  - make in “make a decision”
- Noun Compound Literality
  - flea in “flea market”
- Noun Compound Relations
  - “olive oil” is made of olives
  - “baby oil” is made for babies
- Adjective-Noun (AN) Attributes
  - TEMPERATURE is conveyed in “hot water”, but not in “hot argument”.

Shwartz, V., & Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition.

ELCo: Bridging Emoji Mashup and Lexical Composition

Representing concepts using emojis *by making more sense.*

We have a few preliminary studies in literature.

Our research gap.

Input: A lexical composition (phrase) *w*, an Emoji vocabulary *V*.

Output: A sequence of emojis E, $(e_1, e_2, .. e_n)$, $e_i \subseteq V$ which is able to uncover the implicit semantics in *w*.

Related Work

Natural Language (NL) Interface for Emoji

Emoji embeddings

- Skip-gram method: 700 emojis
- Emoji2vec: 1661 emojis
- EmojiNet: 2389 emojis

We have **3633 emojis** in Unicode Emoji 14.0.

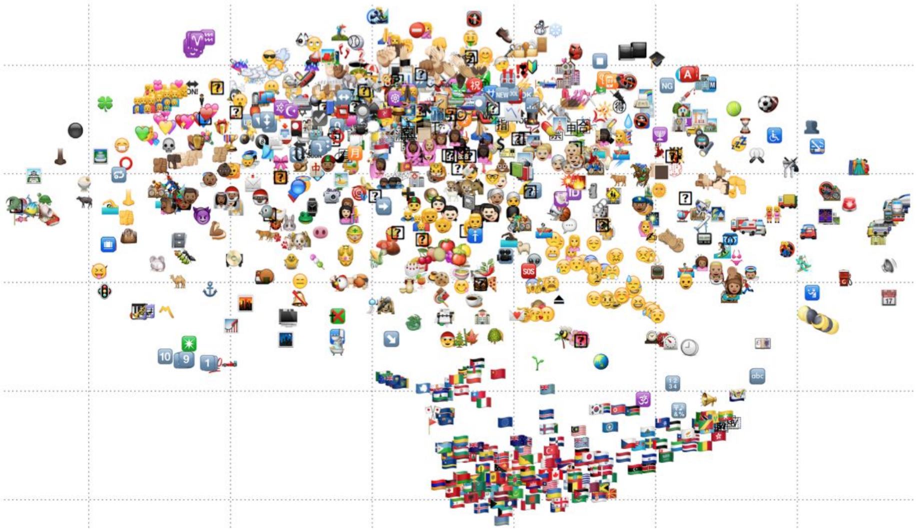


Figure 3: Emoji vector embeddings, projected down into a 2-dimensional space using the t-SNE technique. Note the clusters of similar emojis like flags (bottom), family emoji (top left), zodiac symbols (top left), animals (left), smileys (middle), etc.

Eisner, Ben, et al. "emoji2vec: Learning emoji representations from their description." *arXiv preprint arXiv:1609.08359* (2016).

Emojinating as a baseline

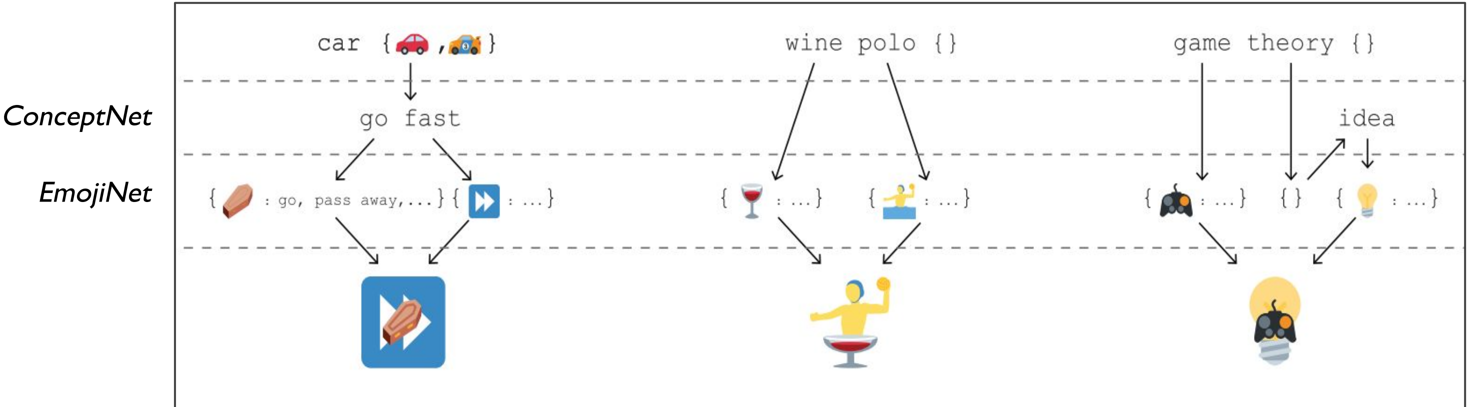


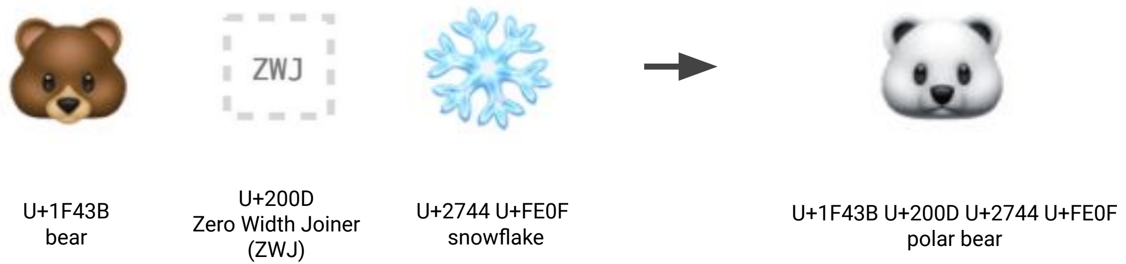
Figure 4: Generation of visual representations (T2) for three concepts: *car*, *wine polo* and *game theory*

- Current **state-of-the-art** method for representing concepts using emojis.
- Emojinating is basically dictionary lookup, it does not require any training.
- We replicated it as **baseline**.

Cunha, J. M., Martins, P., & Machado, P. (2018). How shell and horn make a Unicorn: Experimenting with visual blending in emoji.

Dataset

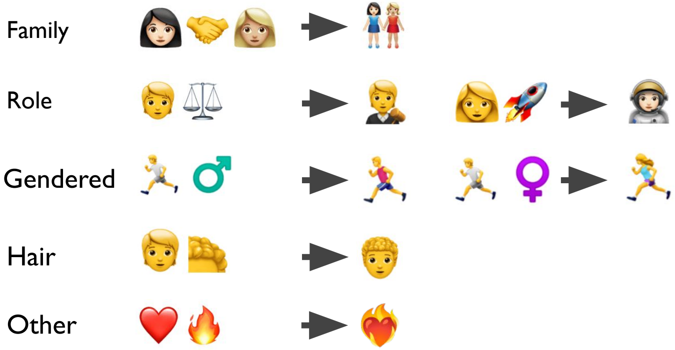
Emoji ZWJ Sequences



- **Zero Width Joiner (ZWJ)** is like an invisible glue character.
- Emoji ZWJ Sequence: a sequence of emojis joined together by ZWJ.

Dataset: ZWJ dataset

<u>ZWJ emoji categories</u>



We select **33 ZWJ emojis** out of 1353 emojis that gives meaningful emoji composition.

Table 3.1: Statistics of Emoji ZWJ Sequences Version:14.0

Category	# of ZWJ emojis	# of unique ZWJ emojis
Family	332	32
Role	360	**20**
Gendered	572	0
Hair	72	0
Others	13	**13**

33 ZWJ emojis

IRB-approved data collection

1. Choose the correct attribute.

cool drink

EMOTIONALITY

GENEROSITY

IMPORTANCE

TEMPERATURE

2. Key in an emoji sequence.

cool drink

3. Rate the baseline output.

	1	2	3	4	5
cool drink 🆒🍹	○	○	○	○	○

Our dataset: ELCo-AN dataset

Table 3.2: Samples of ELCo-AN dataset annotation

AN concept	Attribute	Human annotations	Emojinating Output	Average ratings
short flight	DURATION	⏳✈️ and 🤏✈️	🚋✈️	2.5
short story	DURATION	⏳📖 and 🤏🗣️	🚊📔	2.4
short money	QUANTITY	🤏💵 and 🤷‍♀️💸💸💸	🚋💰	2.3
short supply	QUANTITY	🤏📦 and 🛖❌	🚡🖌️	1.8
short hair	LENGTH	🤏💇‍♀️ and 💇‍♀️➡️👩‍🦰	🚡👱	1.6

- Dataset
  - Consists of **210 Adjective-Noun(AN) concepts**, which covers 45 adjectives and 77 attributes.
  - 1663 total annotations, an average of **7.92 annotations** per concept.
  - Average length of emoji sequences: 2.59.
  - Average ratings of baseline model: **2.32** out of 5.
- Annotators: 40 NUS students and they were paid fairly.
  - Each annotate 41.5 concepts on average.
  - We allow a large spectrum of true responses.

Benchmark

Benchmarking on baseline model

- We aim at estimating the difficulty of our task.
- We do it by performing the **emoji generation task** on both ZWJ and ELCo-AN datasets using our **baseline model: Emojinating**.

Table 4.1: Sample outputs of EMOJINATING model on ZWJ dataset

ZWJ concept	Golden ZWJ annotation	Output of EMOJINATING	Path in ConceptNet
judge	🧑⚖️	🤾‍♀️💅	judge →pass sentence
teacher	🧑🏫	🏫🎓	teacher →school students
office worker	🧑💼	🏢👷	-

Too literal and not making sense.

Benchmark results

Table 4.2: Benchmark result of Emojinating model on 33 ZWJ concepts and 210 ELCo-AN concepts (with 1663 responses).

	ZWJ	ELCO-AN
Both emojis matched	0/33 (0 %)	11/1663 (0.66 %)
One emoji matched	10/33 (30 %)	315/1663 (18.94 %)
No emoji matched	23/33 (70 %)	1337/1663 (80.40 %)

similar distribution

- Emoji generation task is challenging for SOTA.

Ranking Problem

Ranking Problem

Generation Problem

OR

Ranking Problem

❓ Too challenging for SOTA.

✔ Simplistic settings.

- We re-formalize our problem under a simplistic ranking setting.
- To evaluate the intrinsic property of emoji & lexical composition.

Ranking Problem

Sampling

Using the concept *"big group"*

Ground-truth ranking:

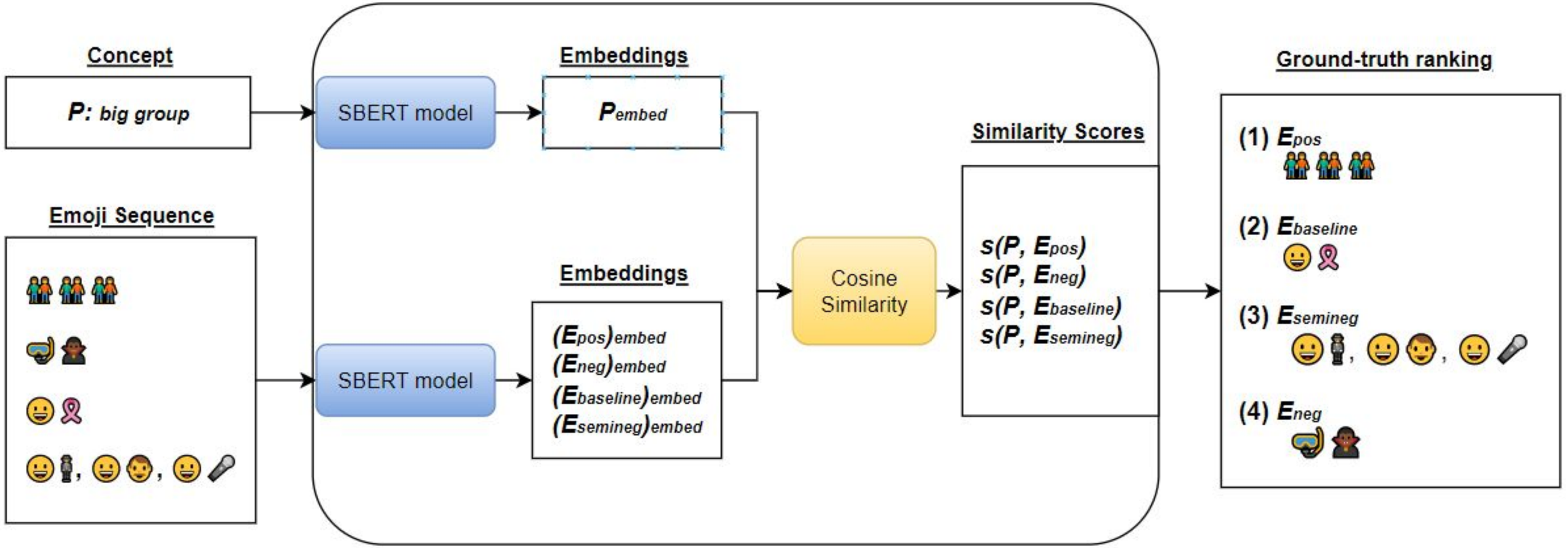
1. 🧑‍🤝‍🧑🧑‍🤝‍🧑🧑‍🤝‍🧑
   (Positive sample)
2. 😀🎗️
   (Baseline sample)
3. 😀🕴️, 😀👨, 😀🎤
   (Hard-negative sample)
4. 🤿🧛🏾‍♂️
   (Easy/random negative sample)

1. **Positive sample (ground truth)**, E_{pos}
2. **Baseline sample (plausible)**, $E_{baseline}$
3. **Negative sample**, $E_{easy-negative}$
   - Semi-hard negative sample, $E_{semineg}$
   - Easy negative sample, E_{neg}

Ranking Problem

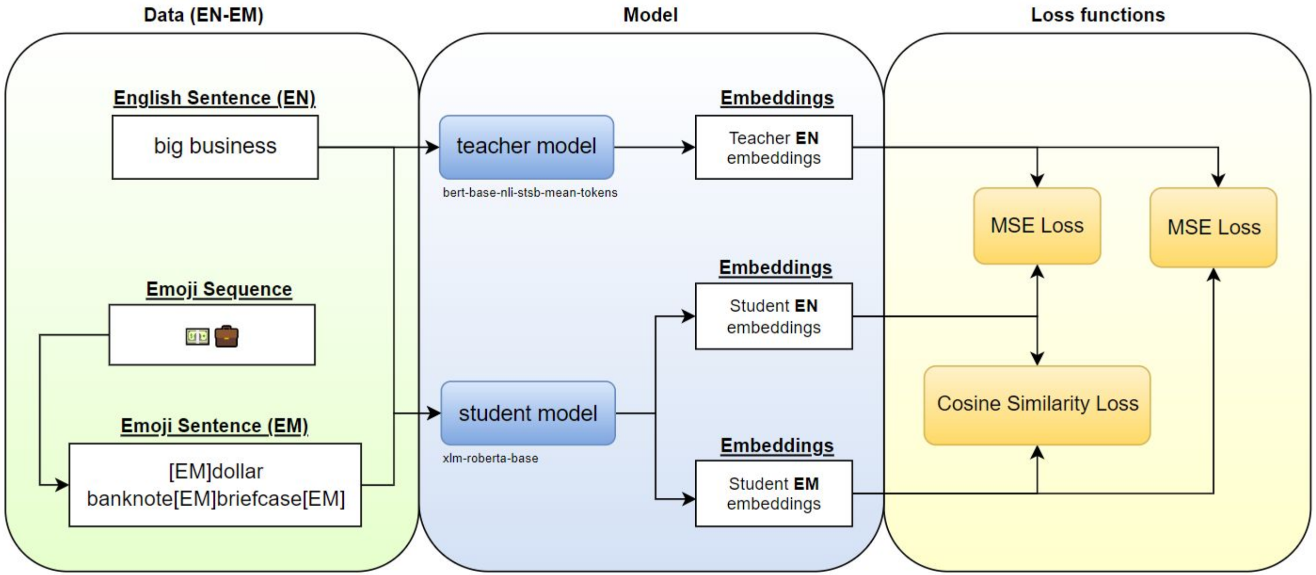
Ranking Problem

Unsupervised approach



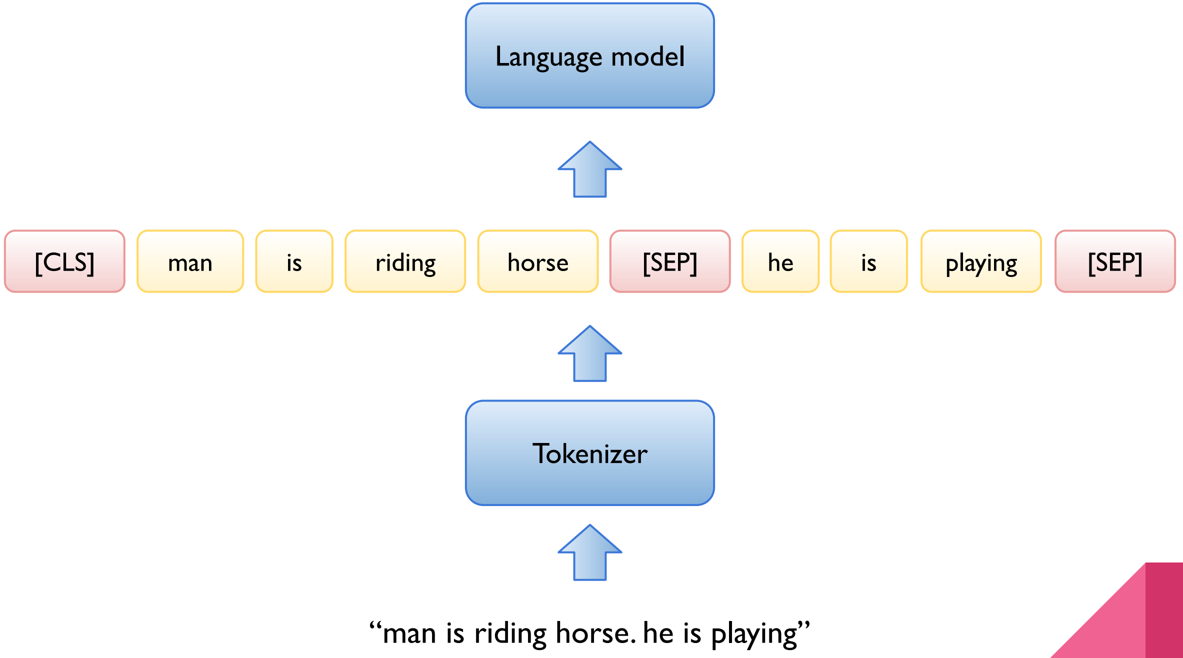
ELCoM Model

Supervised training

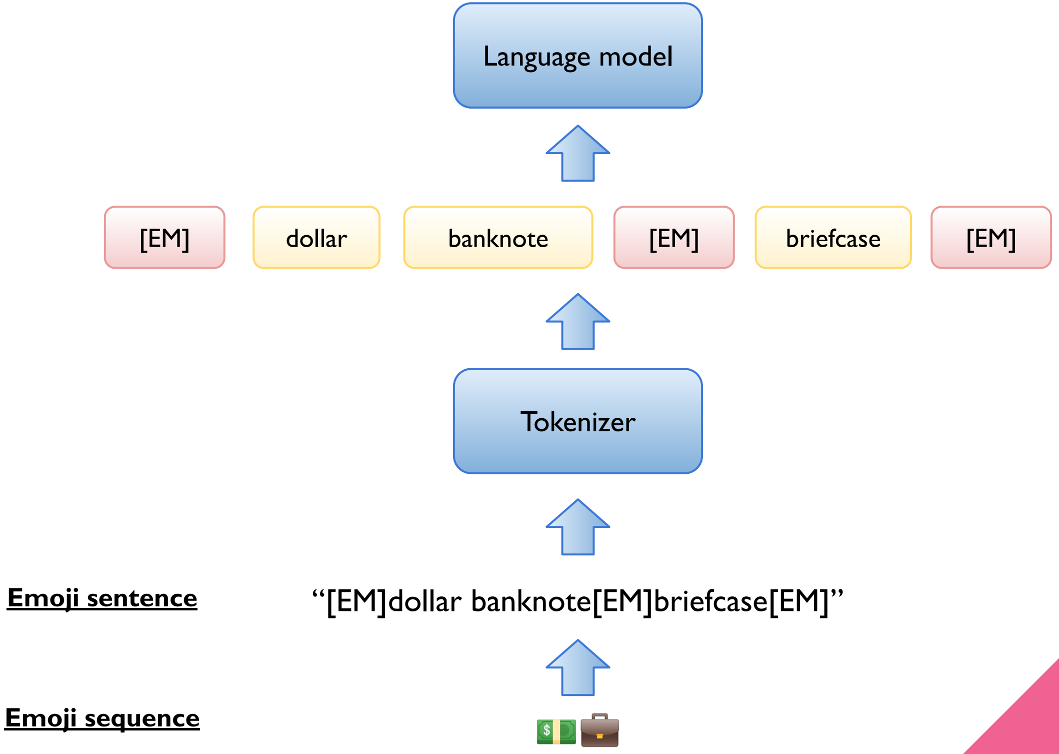


- Multilingual training: teacher–student model
- Multiple training objectives: MSE loss & Cos-sim Loss
- Emoji connector: special token [EM]

Background of special tokens



Special token [EM]



Evaluation

Research Questions

RQ1: What is the performance of vanilla Pretrained Language Model(PLM) in predicting emoji compositions?

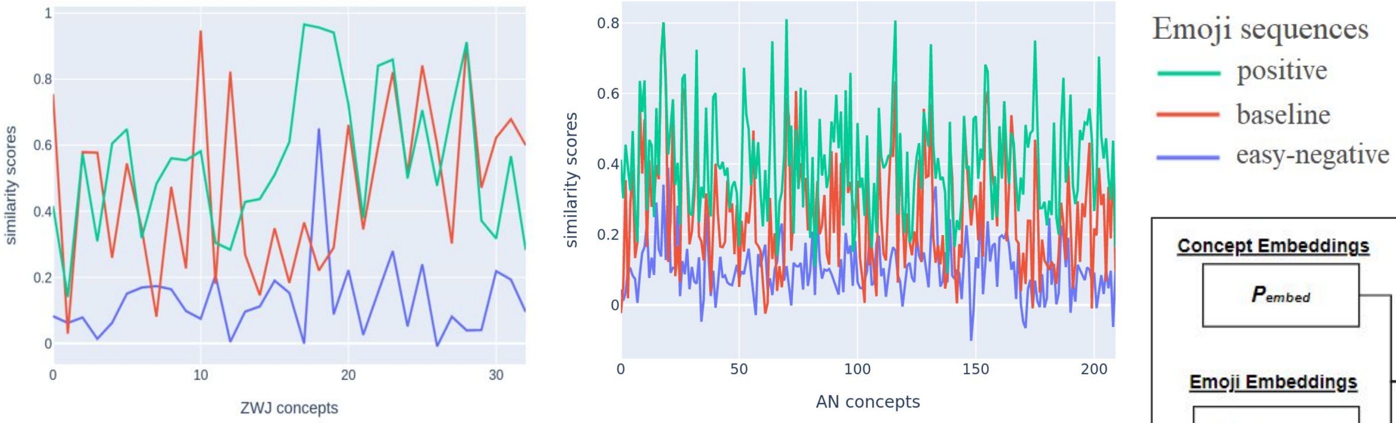
RQ2: Can PLM be optimized on our ELCo-AN dataset?

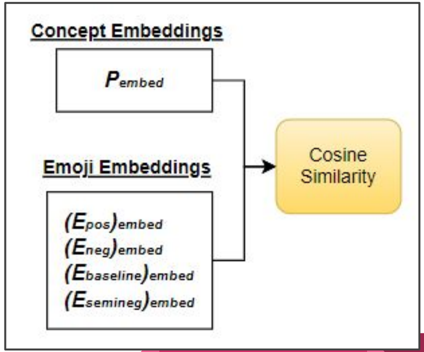
RQ3: What has been learned by the model from training on our ELCo-AN dataset?

RQ1

What is the performance of vanilla Pretrained Language Model(PLM) in predicting emoji compositions?

RQ1: What is the general performance of vanilla PLM in predicting emoji compositions?

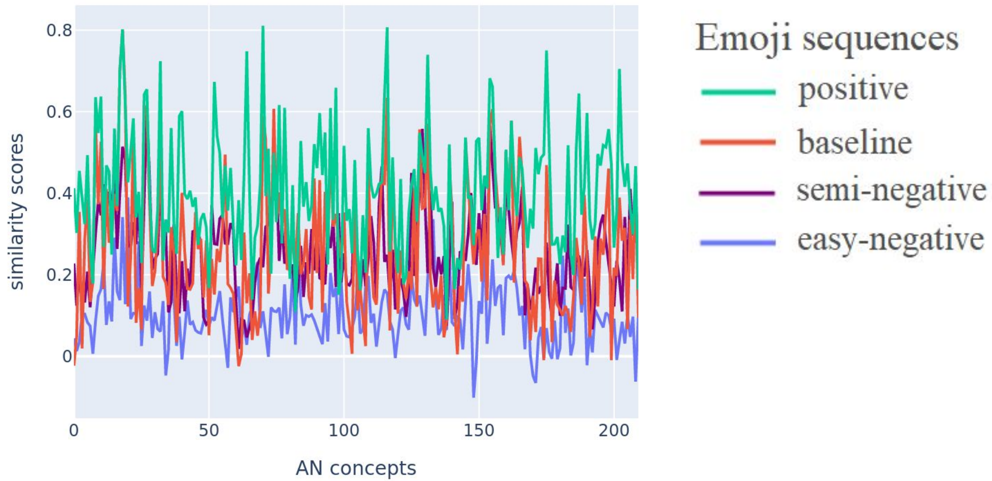




$E_{positive} > E_{baseline} > E_{easy-negative}$

- SBERT (and PLM) can understand lexical composition and emoji composition to a certain extent.

RQ1: What is the general performance of vanilla PLM in predicting emoji compositions?



To even stress test the model, we include $E_{semi-negative}$.

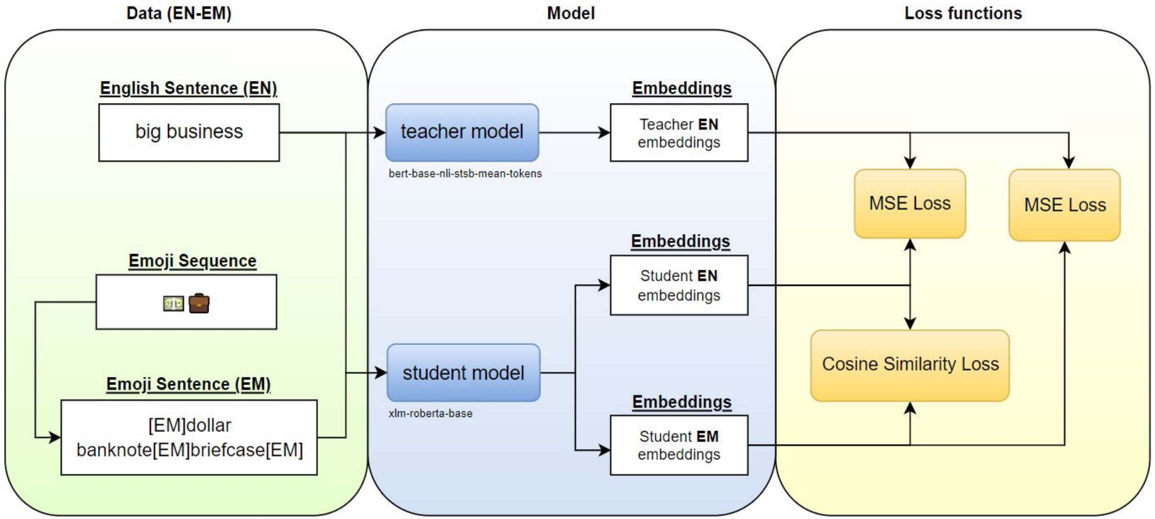
$E_{positive} > E_{baseline} \approx E_{semi-negative} > E_{easy-negative}$

- $E_{semi-negative}$ is challenging to the model.

RQ2

Can PLM be optimized on our ELCo-AN dataset?

RQ2: Can PLM be optimized on our ELCo-AN dataset?



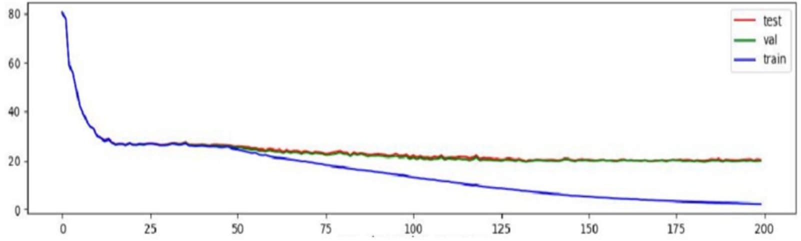
ELCoM model

Evaluation metrics:

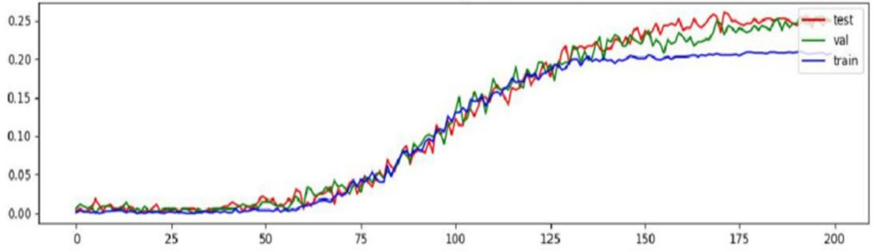
- MSE Loss
- Cosine similarity score
- Translation score
- Information Retrieval score

RQ2: Can PLM be optimized on our ELCo-AN dataset?

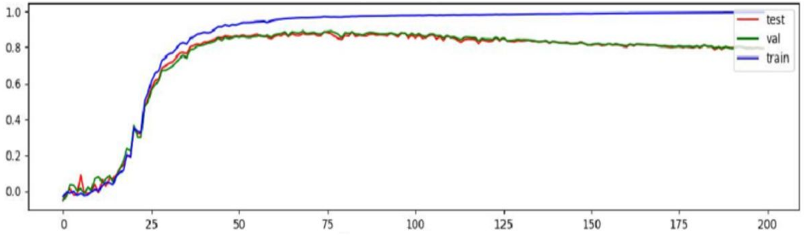
MSE Loss



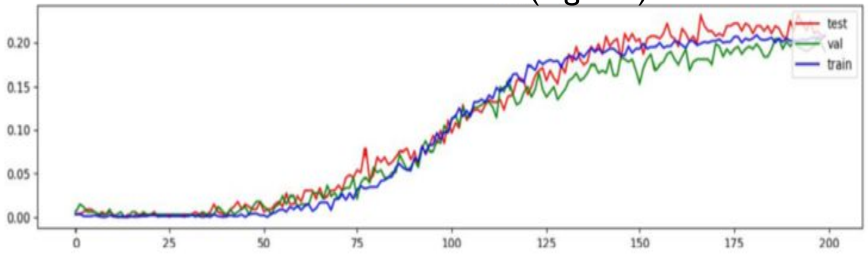
Translation score (src2trg)



Cosine similarity score

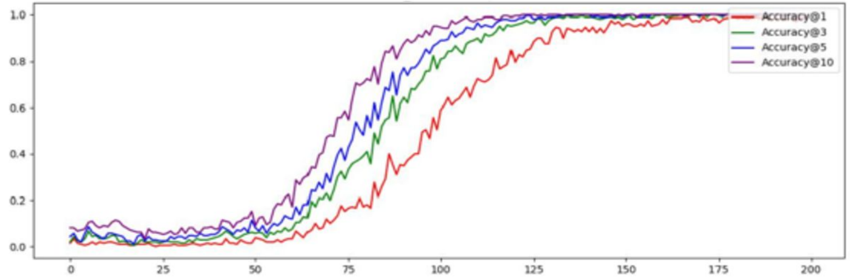


Translation score (trg2src)

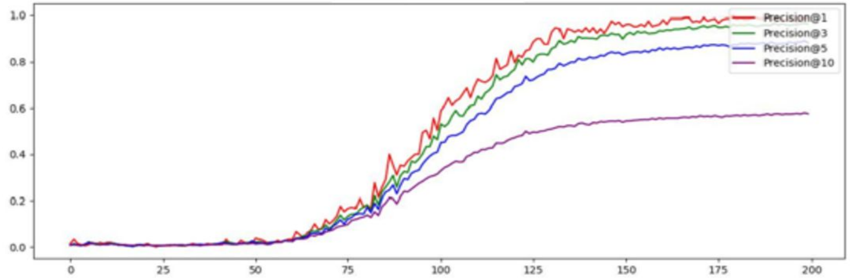


RQ2: Can PLM be optimized on our ELCo-AN dataset?

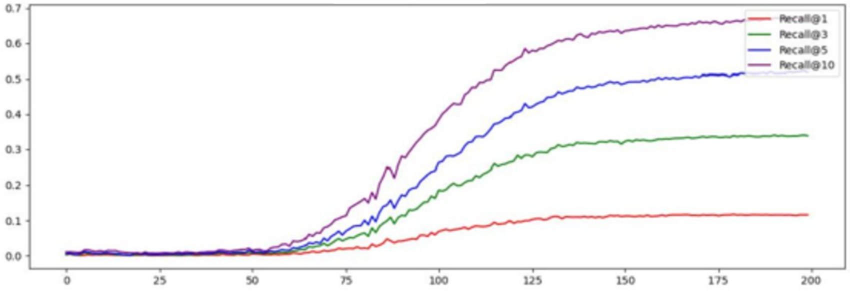
Accuracy



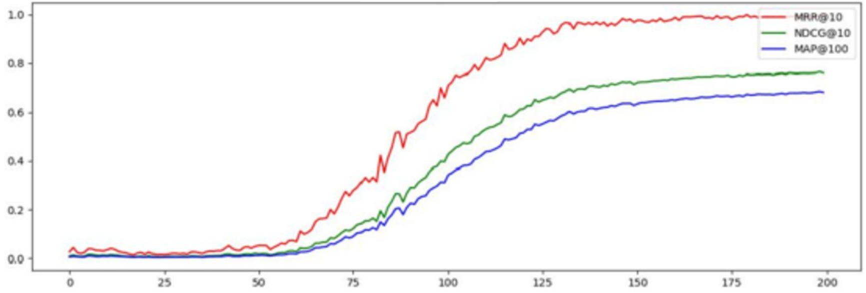
Precision



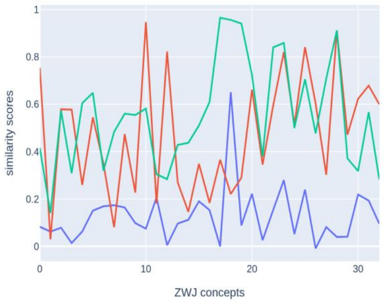
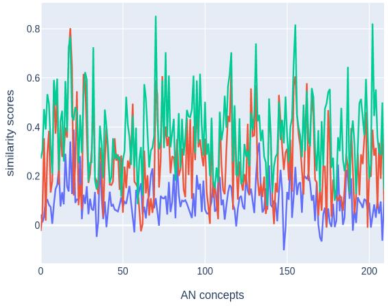
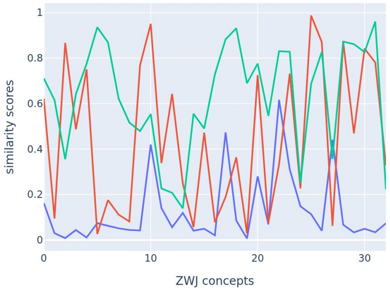
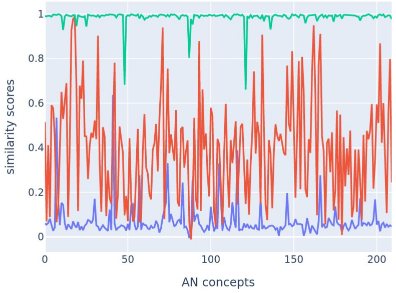
Recall



MRR, NDCG, MAP



RQ2: Can PLM be optimized on our ELCo-AN dataset?

	ZWJ dataset	ELCo-AN dataset
Vanilla SBERT model		
ELCoM model		

Emoji sequences
- randneg
- baseline
- pos

RQ3

What has been learned by model from the training on our ELCo-AN dataset?

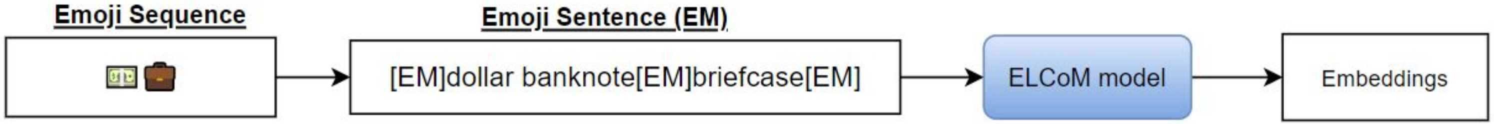
- RQ3.1 What has been learned by the special token [EM]?
- RQ3.2 What is the impact of emoji ordering on our ELCoM model?

RQ3.1

What has been learned by the special token [EM]?

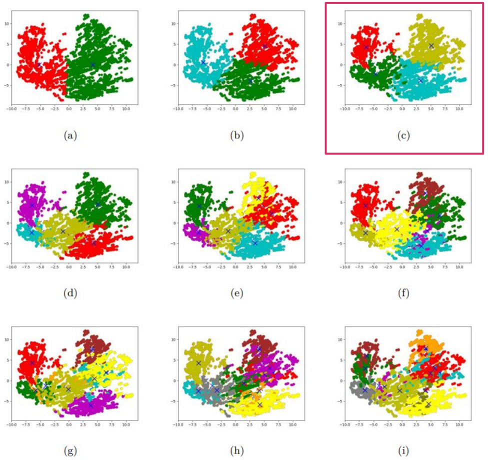
RQ3.1 What has been learned by the special token [EM]?

Extract and clustering [EM] token



Based on the idea of **contextual word embeddings:**

- we extract [EM] token embeddings
- perform KMeans Clustering
  (k = number of clusters)



k = 4

Figure 7: KMeans clustering visualisation results of [EM] embeddings for number of cluster k = 2 to 10.

RQ3.1 What has been learned by the special token [EM]?

- The total number of [EM] tokens extracted from 1663 parallel data (EN-EM) is **5944**.

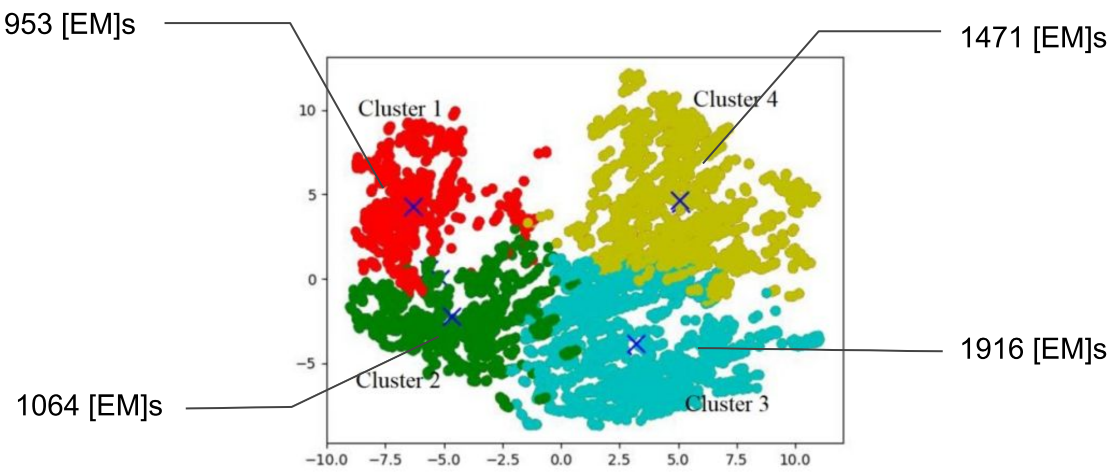
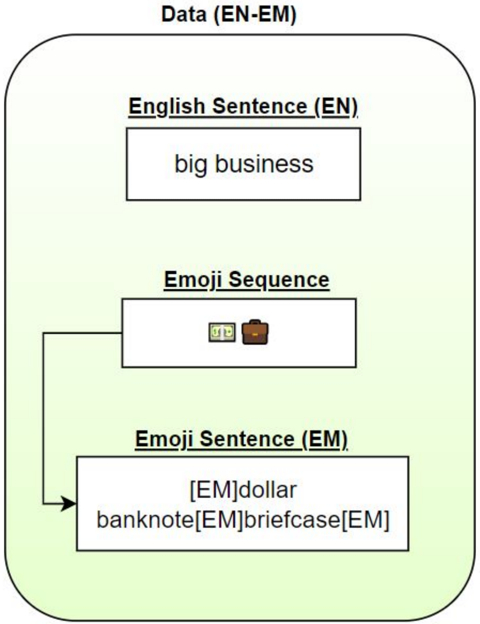


Figure 5.7: KMeans clustering of [EM] embeddings for number of cluster k = 4.

RQ3.1 What has been learned by the special token [EM]?

What does each cluster represent?



Each [EM] token can be associated to:

1. [EM] **position** in the Emoji sentence

   $[EM]_0$dollar banknote$[EM]_1$briefcase$[EM]_2$

2. The original **English sentence (EN)**

   big business

3. The **Emoji sentence (EM)** that the [EM] token belongs to

   [EM]dollar banknote[EM]briefcase[EM]

RQ3.1 What has been learned by the special token [EM]?

1. [EM] **position** in the Emoji sentence

- Hypothesis: The position of [EM] token might be a feature that affects the embeddings learned.
- For each cluster, we count the occurrence of each [EM] position.

Table 5.8: Evaluation of position: Count of [EM] token in each cluster according to their position in emoji sequence.

[EM] position	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Total
$[EM]_0$	283	(29.7%)	445	(27.7%)	517	(27.0%)	419	(28.5%)	1664
$[EM]_1$	283	(29.7%)	444	(27.7%)	517	(27.0%)	418	(28.4%)	1662
$[EM]_2$	265	(27.8%)	418	(26.1%)	480	(25.1%)	383	(26.0%)	1546
$[EM]_{>2}$	122	(12.8%)	297	(18.5%)	402	(21.0%)	251	(17.1%)	1072
Total	953		1604		1916		1471		5944

- Observation: For each cluster, the distribution of $[EM]_0$, $[EM]_1$, [EM], and $[EM]_{>2}$ are mostly even.
- Conjecture: The position of [EM] token might not be related to its learning of embeddings.

RQ3.1 What has been learned by the special token [EM]?

2. The original **English sentence (EN)**

- Adjectives

Table 5.9: Top 5 adjectives by frequency in each clusters and their count.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Adj@Top1	short, 282	dirty, 222	high, 273	clear, 177
Adj@Top2	low, 197	wrong, 210	hot, 225	warm, 159
Adj@Top3	dull, 122	dark, 145	big, 212	fresh, 112
Adj@Top4	dry, 103	far, 104	common, 151	cool, 106
Adj@Top5	thin, 86	foreign, 88	deep, 145	right, 105

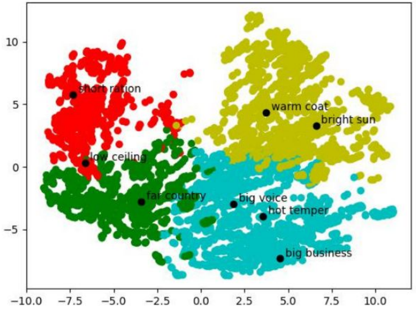


Figure 5.8: Sample points from each cluster annotated with their English sentence.

3. The **Emoji sentence (EM)** that the [EM] token belongs to

- Emoji

Table 5.10: Top 5 Emoji word by frequency in each clusters and their count.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Emoji@Top1	pinching hand 🤏, 189	cross mark ❌, 258	men holding hands 👬, 168	speaking head 🗣️, 168
Emoji@Top2	down-right arrow ↘️, 63	red question mark ❓, 138	fire 🔥, 161	thumbs up 👍, 149
Emoji@Top3	expressionless face 😑, 54	microbe 🦠, 63	man standing 🧍‍♂️, 125	brain 🧠, 99
Emoji@Top4	down arrow ⬇️, 52	new moon 🌑, 59	dashing away 💨, 114	sparkles ✨, 91
Emoji@Top5	sun ☀️, 46	house 🏠, 58	package 📦, 114	light bulb 💡, 87

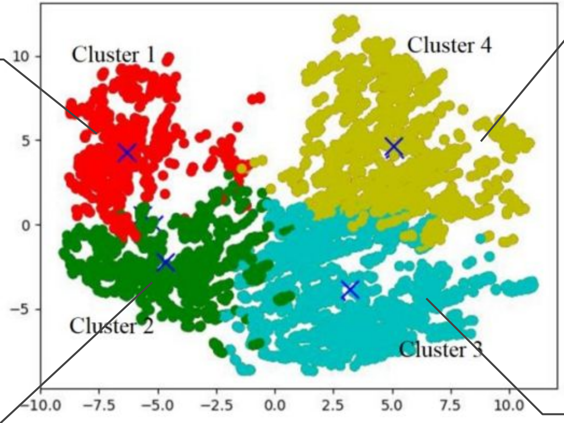
RQ3.1 What has been learned by the special token [EM]?

What does each cluster represent?

small in dimension

short, low

🤏 ↘ ⬇️



positive sentiments

clear, warm, fresh

🗣️, 👍, ✨

negative sentiments

dirty, wrong, dark

❌, ❓, 🌑

big in dimension

high, hot, big

👬, 🧍

RQ3.1 What has been learned by the special token [EM]?

4. Distance from cluster centroids.

- Emoji sequence
  - We list out the Top 15 samples that are closest to each cluster centroids.
  - Observation:

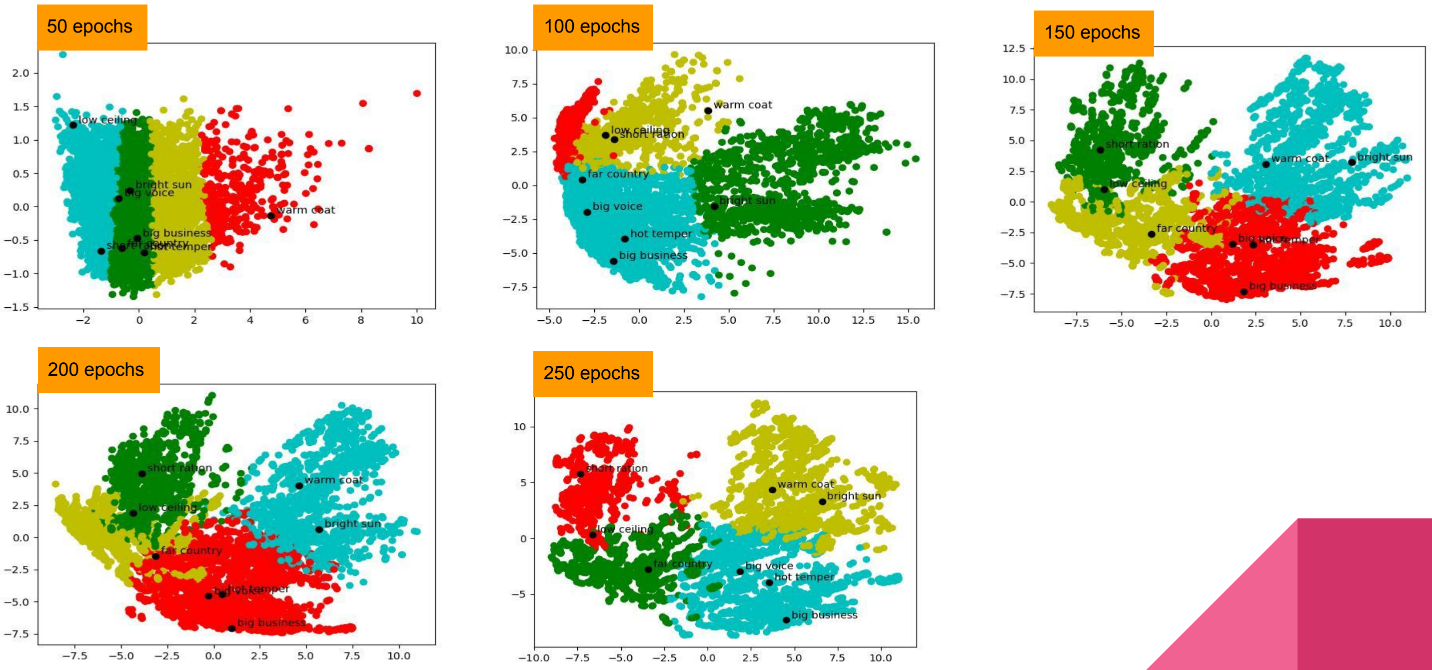
    Cluster 3 has a consistent pattern, in which most Emoji sequences consist of **repeating emojis**.
  - For example,

    “dark purpose”: 😈😈😈😈 and “deep concentration”: 😵😵😵😵

This is a good evidence that the model learned some latent features of emoji compositionality.

RQ3.1 What has been learned by the special token [EM]?

How is the learning process of [EM] embeddings?



RQ3.1 What has been learned by the special token [EM]?

- We managed to identify some evidence of patterns being learned by the [EM] token embeddings,
  - which is highly correlated with the English sentence and Emoji involved,
  - but not correlated with the position of [EM] tokens.

- We believe that more patterns of emoji compositionality could be discovered given a larger or more diverse dataset.

RQ3.2

What is the impact of emoji ordering on the model?

RQ3.2 What is the impact of emoji ordering on our ELCoM model?

How important is emoji order in general, or specifically for Adjective-Noun compounds?

Linguistics research on emoji syntax:

- Positions of emojis 🎁🎁🎈 are interchangeable. (McCulloch & Gawne, 2018)
- 😂😂👌 is more common than the order 👌😂😂 (Steinmetz, 2014)
- ❤️✈️✈️✈️✈️🧑 means "love send-fast I" (Herring & Ge, 2020)
  - Object-Verb-Subject (OVS) ordering is the most frequent ordering.

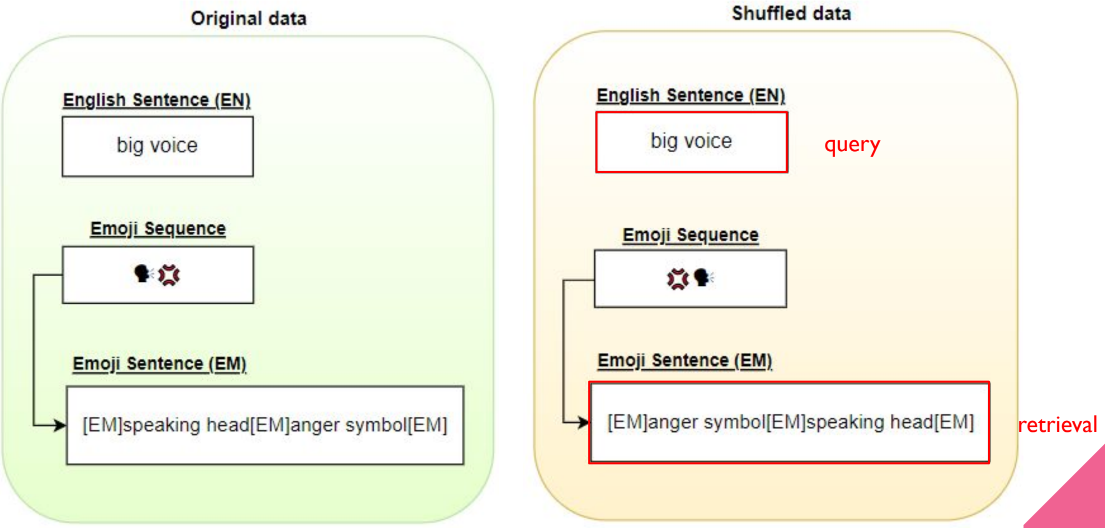
Emoji do not have a fixed syntax in the same way language does. For Adjective-Noun compound, we assume that humans can interpret the emoji sequences even after changing the order of emojis.

https://www.semanticscholar.org/paper/Emoji-Grammar-as-Beat-Gestures-McCulloch-Gawne/b2e0175bca066f3c293c56d7a5c6ff4f324e45f5
https://time.com/2993508/emoji-rules-tweets/
https://www.researchgate.net/publication/342283337_Do_Emoji_Sequences_Have_a_Preferred_Word_Order
https://makingnoiseandhearingthings.com/2016/12/07/do-emojis-have-their-own-syntax/

RQ3.2 What is the impact of emoji ordering on our ELCoM model?

Experiment: Perturb our ELCo-AN dataset.

- We **shuffle** the emoji ordering.
- We evaluate the model on **information retrieval** tasks.



RQ3.2 What is the impact of emoji ordering on our ELCoM model?

Hypothesis:

- The performance drop is NOT significant, because position feature is NOT important as learned before.

General expectation:

- The model will decrease in performance given shuffled emoji sequences.
- This is because our ELCoM model is trained on the original sequence.
- Language model such as transformer is sensitive to word positioning.

Table 5.12: Comparison table of the ElCoM model performance on information retrieval setting for original AN dataset vs shuffled AN dataset.

	Original AN dataset			Shuffled AN dataset		
k	Accuracy@k	Precision@k	Recall@k	Accuracy@k	Precision@k	Recall@k
1	99.52%	99.52%	13.18%	93.33%	93.33%	12.24%
3	99.52%	95.71%	37.60%	99.05%	86.83%	33.93%
5	100.00%	86.29%	55.85%	99.52%	77.62%	50.09%
10	100.00%	54.86%	70.04%	99.52%	49.38%	63.12%

Observation:

- The model performance decreases for the shuffled dataset.

RQ3.2 What is the impact of emoji ordering on our ELCoM model?

Error Type 1: Error due to dataset properties

Query (AN concepts)	Ground truth retrieval (Emoji sentences, EM)	Retrieved Emoji sentences @ Top3	Supposed query of Retrieved Emoji sentence @ Top1
big voice	'[EM]speaking head[EM]loudspeaker[EM]' '[EM]loudspeaker[EM]speaking head[EM]' '[EM]speaking head[EM]studio microphone[EM]' '[EM]microphone[EM]grinning face[EM]' '[EM]megaphone[EM]loudspeaker[EM]' '[EM]anger symbol[EM]speaking head[EM]' 💢🗣️	'[EM]speaking head[EM]anger symbol[EM]' 🗣️💢 '[EM]speaking head[EM]loudspeaker[EM]' '[EM]man standing[EM]flexed biceps[EM]'	hot argument 🗣️💢
ineffectual therapy	'[EM]cross mark[EM]spiral notepad[EM]skull[EM]' '[EM]person getting massage[EM]thumbs down[EM]' '[EM]ear[EM]person[EM]cross mark[EM]cross mark[EM]' '[EM]woman health worker medium-dark skin tone[EM]mouse face[EM]' '[EM]woman gesturing NO[EM]pill[EM]' '[EM]deaf woman[EM]ear[EM]thumbs down[EM]unamused face[EM]' '[EM]hospital[EM]thumbs down[EM]'	'[EM]pill[EM]woman gesturing NO[EM]' '[EM]deaf woman[EM]ear[EM]thumbs down[EM]unamused face[EM]' '[EM]ear[EM]person[EM]cross mark[EM]cross mark[EM]'	wrong medicine

Wrong retrieval

Cases where ELCoM model fails at Accuracy@1 in shuffled AN dataset.

RQ3.2 What is the impact of emoji ordering on our ELCoM model?

Error Type 1: Error due to dataset properties

Correct retrieval

Query (AN concepts)	(Emoji sentences, EM)	Retrieved Emoji sentences @ Top3	Supposed query of Retrieved Emoji sentence @ Top1
big voice	'[EM]speaking head[EM]loudspeaker[EM]' '[EM]loudspeaker[EM]speaking head[EM]' '[EM]speaking head[EM]studio microphone[EM]' '[EM]microphone[EM]grinning face[EM]' '[EM]megaphone[EM]loudspeaker[EM]' '[EM]anger symbol[EM]speaking head[EM]' 💢🗣	'[EM]speaking head[EM]anger symbol[EM]' 🗣💢 '[EM]speaking head[EM]loudspeaker[EM]' '[EM]man standing[EM]flexed biceps[EM]'	hot argument 🗣💢
ineffectual therapy	'[EM]cross mark[EM]spiral notepad[EM]skull[EM]' '[EM]person getting massage[EM]thumbs down[EM]' '[EM]ear[EM]person[EM]cross mark[EM]cross mark[EM]' '[EM]woman health worker medium-dark skin tone[EM]mouse face[EM]' '[EM]woman gesturing NO[EM]pill[EM]' '[EM]deaf woman[EM]ear[EM]thumbs down[EM]unamused face[EM]' '[EM]hospital[EM]thumbs down[EM]'	'[EM]pill[EM]woman gesturing NO[EM]' '[EM]deaf woman[EM]ear[EM]thumbs down[EM]unamused face[EM]' '[EM]ear[EM]person[EM]cross mark[EM]cross mark[EM]'	wrong medicine

Cases where ELCoM model fails at Accuracy@1 in shuffled AN dataset.

RQ3.2 What is the impact of emoji ordering on our ELCoM model?

Error Type 2: Error by the model

Query (AN concepts)	Ground truth retrieval (Emoji sentence)	Retrieved Emoji sentences @ Top3	Supposed query of Retrieved Emoji sentence @ Top1
dark glass	'[EM]glass of milk[EM]sunglasses[EM]' '[EM]wine glass[EM]new moon[EM]' '[EM]glass of milk[EM]black circle[EM]'	'[EM]angry face with horns[EM]sunglasses[EM]' '[EM]angry face with horns[EM]spade suit[EM]' '[EM]glass of milk[EM]black circle[EM]'	dark purpose
deep sigh	'[EM]double exclamation mark[EM]thinking face[EM]' '[EM]wind face[EM]face with steam from nose[EM]' '[EM]sad but relieved face[EM]face exhaling[EM]confused face[EM]' '[EM]face exhaling[EM]hole[EM]'	'[EM]disappointed face[EM]face with steam from nose[EM]' '[EM]wind face[EM]face with steam from nose[EM]' '[EM]sad but relieved face[EM]face exhaling[EM]confused face[EM]'	dry critique

😞😤 Wrong retrieval

Cases where ELCoM model fails at Accuracy@1 in shuffled AN dataset.

Conclusion:

- The model is robust in handling shuffled emoji sequences.
- This is a desired property if the emojis are not in a particular ordering.

Summary

Summary of key contributions

(1) To address the lack of understanding of implicit emoji semantics, we propose a novel research problem in **bridging lexical composition and emoji mashup** (ELCo). 🔍

(2) We overcome the emoji data scarcity problem by collating **ELCo-AN dataset**, which is a 210 Adjective-Noun compound to Emoji sequences parallel dataset. 📊

(3) We test the capability of **vanilla PLM/SBERT models** in ranking emoji samples given a concept. We show that PLM is good at distinguishing positive and random negative emoji pairs, but relatively weak when being tested on semi-hard negative samples. 🔬

(4) We propose **ELCoM model** to learn our task. We fine-tune the model on ELCo-AN dataset, showing that our task is under the current model's capacity. 🔬

(5) We perform **model interpretation and behavioral analysis**. We find out which latent features being learned or not by our model. 🔬

Future Work

"Bridge at Night" Emoji

- **Dataset**: Expand our ELCo dataset by including more variants of lexical compositions.
- **Model**: Further experiment with our ELCoM model on more datasets and improve its interpretability.
- Finding better **emoji representations** (textual and visual) and **augmentation** approaches.
- **Move towards a generation task**
  a. Explore **PLM prompting** methods to utilize knowledge in PLM for our task.
  b. Using **commonsense knowledge graph** dynamically to uncover meaningful information.

Any Questions?

Thank you! Let's talk! 🙇‍♀️

Spare Slides.

Following slides are **not** going to be presented orally, but they do provide complementary information to the main slides.

RQ3.1 What has been learned by the special token [EM]?

[EM]dollar banknote[EM]briefcase[EM]

↓ *ELCoM model*

extract [EM] token embeddings

↓

KMeans Clustering

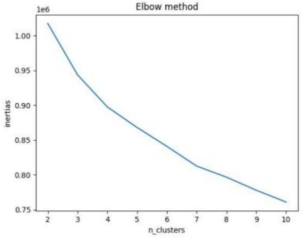


Figure 5.6: Elbow method to determine the optimal number of cluster

→

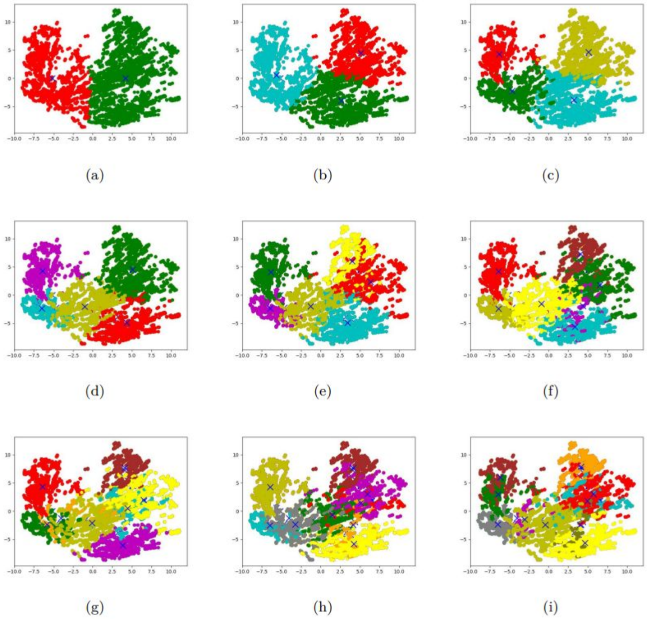


Figure 7: KMeans clustering visualisation results of [EM] embeddings for number of cluster k = 2 to 10.

RQ3.1 What has been learned by the special token [EM]

3. The **Emoji sentence (EM)** that the [EM] token belongs to

- Emoji (word-level)

Table 5.10: Top 5 Emoji word by frequency in each clusters and their count.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Emoji@Top1	pinching hand 🤏, 189	cross mark ❌, 258	men holding hands 👬, 168	speaking head 🗣️, 168
Emoji@Top2	down-right arrow ↘️, 63	red question mark ❓, 138	fire 🔥, 161	thumbs up 👍, 149
Emoji@Top3	expressionless face 😑, 54	microbe 🦠, 63	man standing 🧍, 125	brain 🧠, 99
Emoji@Top4	down arrow ⬇️, 52	new moon 🌑, 59	dashing away 💨, 114	sparkles ✨, 91
Emoji@Top5	sun ☀️, 46	house 🏠, 58	package 📦, 114	light bulb 💡, 87

- Emoji sequence (sentence-level)

[EM]$_0$dollar banknote[EM]$_1$briefcase[EM]$_2$

Observation: [EM]s at different positions but from the same Emoji sequence tends to be clustered to the same group, with up to 99% of the EN-EM samples achieving this criterion.

RQ3.1 What has been learned by the special token [EM]?

4. Distance from cluster centroids.

- Adjectives

Table 5.11: Top 5 adjectives by nearest distance from cluster centroids for each cluster.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Adj@Top1	short	dry	big	hot
Adj@Top2	thin	thin	deep	high
Adj@Top3	cool	full	dark	clear
Adj@Top4	inadequate	low	high	cool
Adj@Top5	dull	dirty	immediate	warm

Table 5.9: Top 5 adjectives by frequency in each clusters and their count.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Adj@Top1	short, 282	dirty, 222	high, 273	clear, 177
Adj@Top2	low, 197	wrong, 210	hot, 225	warm, 159
Adj@Top3	dull, 122	dark, 145	big, 212	fresh, 112
Adj@Top4	dry, 103	far, 104	common, 151	cool, 106
Adj@Top5	thin, 86	foreign, 88	deep, 145	right, 105

- Emoji sequence
  - We list out the Top 15 samples of which [EM] token belongs to that are closest to the cluster centroids.
  - Observation: Cluster 3 has a consistent pattern, in which most Emoji sequences consist of repeating emojis. For example, "dark purpose": 😈😈😈😈 and "deep concentration": 😵😵😵😵
  - This is a good evidence that the clustering represents some latent features of emoji compositionality.

Sample Parallel data

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
big business	[EM]dollar banknote[EM]briefcase[EM]	💵💼

Parallel data

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
big voice	[EM]speaking head[EM]anger symbol[EM]	🗣💢

Shuffled

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
big voice	[EM]anger symbol[EM]speaking head[EM]	💢🗣

query

retrieval

RQ3.2 What is the impact of emoji ordering on the model?

Parallel data

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
big voice	[EM]speaking head[EM]anger symbol[EM]	🗣️💢

Shuffled

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
big voice	[EM]anger symbol[EM]speaking head[EM]	💢🗣️

query

expected correct retrieval

Parallel data

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
hot argument	[EM]anger symbol[EM]speaking head[EM]	💢🗣️

Shuffled

English sentence (EN)	Emoji sentence (EM)	Emoji sequence
hot argument	[EM]speaking head[EM]anger symbol[EM]	🗣️💢

retrieval

RQ2: Can PLM be optimized on our ELCo-AN dataset?

Loss & Evaluation metrics

Loss	Evaluation Metrics	Data	Number of samples
MSE loss	MSE loss	Parallel data	997 train, 333 validate and 333 test samples
Cosine similarity loss	Cosine similarity score	All samples, including positive, baseline & negative samples.	2193 train, 732 validate and 732 test samples
-	Translation Score	Parallel data	997 train, 333 validate and 333 test samples
-	Information Retrieval Score	Parallel data (Queries, corpus, related corpus)	210 queries (EN sentences) and 1757 corpus (EM sentences)

- We train ELCoM model for around 200 epochs.
- AdamW optimizer with learning rate of 2e-5, eps of 1e-6.
- Warmup Linear scheduler with 5000 warm-up steps.

Benchmark baseline on ZWJ dataset

Number of ZWJ emojis tested: 33

- Both emojis match: 0/33
- One emoji matches: 10/33
- None matches: 23/33

Table 3.1: Statistics of Emoji ZWJ Sequences Version:14.0

Category	# of ZWJ emojis	# of unique ZWJ emojis
Family	332	32
Role	360	**20**
Gendered	572	0
Hair	72	0
Others	13	**13**

	name	zwj	emojinating	conceptnet
0	health worker	👦⚕️	👩‍⚕️ 👷	['health', 'worker'] -> [':woman_health_worker:', ':construc
1	judge	👦⚖️	🤾‍♀️ 💅	['pass', 'sentence'] -> [':woman_playing_handball:', ':nail
2	pilot	👦✈️	🚜 👨‍✈️	['land', 'plane'] -> [':tractor:', ':man_pilot:']
3	farmer	👦🌾	🐖 🚜	['farm', 'land'] -> [':pig:', ':tractor:']
4	cook	👦🍳	🌡️ 🥖	['measure', 'flour'] -> [':thermometer:', ':baguette_brea
5	person feeding baby	👦🍼	🙎‍♀️ 🍼	['person', 'feeding'] -> [':person_pouting:', ':baby_bottl
6	student	👦🎓	:question_mark: 👩‍🏫	['question', 'teacher'] -> [':question_mark:', ':woman_teac
7	singer	👦🎤	🤒 🤒	['sore', 'throat'] -> [':face_with_thermometer:', ':face
8	artist	👦🎨	🎨 🤳	['paint', 'portrait'] -> [':artist_palette:', ':selfie_med
	teacher	👦🏫	🏫 🎓	['school', 'students'] -> [':school:', ':graduation_cap:
	factory worker	👦🏭	🏭 👷	['factory', 'worker'] -> [':factory:', ':construction_wo
	technologist	👦💻		
	office worker	👦💼	🏢 👷	['office', 'worker'] -> [':office_building:', ':construc
	mechanic	👦🔧	🈂️ [illegible]	['service', 'car'] -> [':Japanese_service_charge_button:
	scientist	👦🔬	:question_mark: 👳	['question', 'theories'] -> [':question_mark:', ':person_we
15	astronaut	👦🚀	🚀 🛶	['space', 'travel'] -> [':rocket:', ':canoe:']
16	firefighter	👦🚒	🚬 🤵🏿 ♂️	['smoke', 'jacket'] -> [':cigarette:', ':man_in_tuxedo_dark