




DAY, A. D., and **MANKOS, C. F.** Computer Information Systems Department (Computer Science). *Sentence Compression Using Emoji Summarization*.

Automatic text summarization is a category of algorithms that aim to produce a small set of representative information from a larger input document. There are two general categories of automatic summarization: extractive and abstractive. Extractive summarization uses sentences and phrases that are already present within the document in order to produce an outline. While this method can produce representative summaries, it does have the drawback of limiting the vocabulary that can be used in the summary. On the other hand, abstractive summarization tries to understand the information within the document and summarize it by creating new phrases and sentences. In our previous research [1], we have implemented two extractive summarization methods: Term Frequency-Inverse Document Frequency and TextRank and presented the results of running the algorithms on three different corpora: Moby-Dick by Herman Melville, a selection of Reuters news articles, and a selection of posts on Reddit. In this research project, we aim to explore more abstractive methods for summarization.

In the domain of Natural Language Processing (NLP), there has been a recent shift to abstractive summarization. The most probable cause of this shift is the development of word embeddings [2]. Word embeddings allow machines to produce representative, fixed-length vectors from single tokens, words. This was a massive breakthrough mainly because many machine learning algorithms take a fixed-length vector as their input. Most previous attempts of word embeddings were lossy or produced vectors of an unwieldy length. Another benefit of this vector representation is that it allows a direct numerical comparison between the words. This work has been expanded from words to both emoji [3] and sentences [4] in our research.

Our project aims to tackle abstractive summarization in a new and novel way by compressing a sentence into a series of emojis. Emojis are a pictographic language that is commonly used on the internet and within text messages. In the latest emoji standard, there are 2,823 characters. The motivation for using emojis was two-fold. First, emojis are very information-dense. Thus, this allows us to compress large chunks of the sentence into just a single emoji. Second, emojis have no formal grammar. This cuts out a large issue with most neural machine translation in, so we can focus on the adequacy rather than the fluidity of the produced translation. There are two possible uses for our algorithm. First, producing a summary of a document using emoji could prove to be a quick way to let someone decide if this topic is interesting to them. Second, it could help with the understanding of emoji on social media platforms by going backwards from a series of emojis to a sentence.

The algorithm devised develops a sentence translation in three discrete steps. In the first step, the sentence is split into a combination of n-grams. N-grams are chunks of the input sentence comprised of words that appear next to each other. For example, one n-gram from the sentence “Rock music approaches at a high velocity” could be “Rock music”. In the second step, these n-grams are transformed into their vector representation. In the third step, the emoji with the closest vectorized description is chosen to represent that n-gram. By combining these emojis together, a predicted translation is produced. The results are shown below:

Table 1: Input sentence and corresponding output emoji from our algorithm	
Input Sentence	Output Emoji
Rock music approaches at high velocity.	
Christmas music rings from the clock tower	
It isn't perfect but it is a start	

We have devised an n-gram grouping algorithm that predicts an emoji sequence based on an input sentence. The results of this algorithm have not been formally verified due to the lack of scoring techniques. However, empirical evaluations of the results have been positive.

Faculty Sponsor: Dr Soo Kim

References

- [1] DAY, A., AND KIM, S. A comparison of automatic extractive text summarization techniques. In 34th Annual Spring Conference of the Pennsylvania Computer and Information Science Educators (PACISE) (Apr. 2019), PACISE, pp. 98, 102.
- [2] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (2013), pp. 3111–3119.
- [3] EISNER, B., ROCKTASCHEL, T., AUGENSTEIN, I., BOSNJAK, M., AND RIEDEL, S. emoji2vec: Learning emoji representations from their description. In Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media (Austin, TX, USA, Nov. 2016), Association for Computational Linguistics, pp. 48–54.

[4] PAGLIARDINI, M., GUPTA, P., AND JAGGI, M. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics (2018).