Jen Jui Liu
Jared Rydalch

Symbolic Generation Proposal

We want to make a system that takes as input English words and generates a symbolic representation of that word in a coherent system. Hieroglyphics, early Nordic writing, and Chinese are some examples, as each concept or word had a distinct symbol associated with it. We want to use some form of lexical relation learning in English, so that more complicated words can be somewhat represented by a combination of simple concepts. This should allow the system to take in a word's definition and stitch together symbols for simple concepts into a new symbolic representation of the more complicated word. Alternatively, if no symbol exists for any of the word's definitions, then the system will generate a new symbol.

Therefore, the system will consist of 4 parts. First, a system for breaking down English words into component parts. This will likely involve vector space representations of words, although it is possible that such representations are outdated and there exists a better method.

Second, a system to generate simple symbols. We haven't planned exactly how to do this, but a GAN trained on Chinese handwriting could potentially work, but simpler or more complex systems may work just as well as this system is only designed to generate simple, unique building blocks for the language. If all else fails, we could provide simple definitions for a few starter words and cut this system out entirely.

Third, a system to store and manage previously generated symbols. This would contain not only the visual representations of the words that symbols have been generated for, but also some form of reduction so that queried words could be reduced via synonyms to the same concept. This will be vital in the fourth and final system.

Finally, the fourth system will be designed to query the library of symbols contained in the third system and stitch them together based on the representation generated by the first system. This might be as simple as collecting pieces and creating as concise a two-dimensional packing of them as possible, or more complicated measures with possibilities for variations of symbols.

The core of this system's creativity comes from the fourth system. This system needs to break down inputs into a genotype of sorts and generate a coherent phenotype. Ideally, these generated symbols can be re-used in future generations, gradually creating a coherent language of sorts. Exploring the relationship between words is a poet's bread and butter, and this system will do something similar.

Potential resources:
https://arxiv.org/abs/1509.01692 (lexical relation learning)
https://arxiv.org/pdf/1301.3781.pdf (word representations in vector space)
http://www.iapr-tc11.org/mediawiki/index.php?title=Harbin_Institute_of_Technology_Opening_Recognition_Corpus_for_Chinese_Characters_(HIT-OR3C) (handwritten Chinese characters dataset)
https://ctwdataset.github.io/ (real world examples of Chinese text)

https://machinelearningmastery.com/how-to-develop-a-generative-adversarial-network-for-an-mnist-handwritten-digits-from-scratch-in-keras/ (Example of how to use a GAN to create artificial handwriting. This might not be what we're looking for, but could be somewhat related.)