Logogram Language Generator

Selim Şeker **Yavuz Durmazkeser** **Güldeste Selen Dal** **Nurullah Cebeci**

Abstract

"In a written language, a logogram or logograph is a written character that represents a word or morpheme." Logograms appears in many ancient languages like Cuneiforms, Egyptian hieroglyphs and Chineese characters. And also appears in pop-culture like the alien language in the sci-fi movie Arrival. With the inspiration of these examples, we are proposing a system that generates logogram language.

1 Introduction

Main task that we aim to solve with this project is building a system that generates logographic symbols from actual words in different languages. At the end of this project we, hopefully, expect to contribute the art scene with generating realistic logogram languages to tell better stories.
To achieve such system, we separated the problem in to parts. First is the multilingual embeddings, and the second is image/symbol generation.

2 Related Work

2.1 Multilingual Embeddings

After Mikolov et al.(2013) introduces successful models to compute continuous vector representations of words, lots of methods developed to embed words and language in a vector space. Like, ELMo (Peters, Matthew E. et al. 2018), GloVe (Pennington et al. 2014) etc. But these methods are for the monolingual embedding. To map multiple languages in one vector space we need multilingual embeddings. MUSE: Multilingual Unsupervised and Supervised Embeddings (Lample et al. 2018) and UMWE: Unsupervised Multilingual Word Embeddings (Chen et al. 2018) are good examples for our problem.

2.2 Image Generation

In the last decade lots of works accomplish on the image generation problem with Variational Autoencoders (Kingma et al. 2014), (Makhzani et al. 2016) and GANs (Goodfellow et al. 2014).

3 The Approach

For the baseline method we will construct the model in Figure 1. To the embed the language, we will use the unsupervised multilingual embedding. For the image generation part, first we're going to do an unsupervised pre-training reconstruction process with Ominiglot dataset and variational autoencoder for provide a style information to model from different symbols and alphabets (Figure 2). After the pre-training process we will simply extract the encoding part from the VAE and feed the embedding vector through the decoder. Since these are the baseline methods, first we will do set of experiment and observe the results to make better progress then improve the baselines.

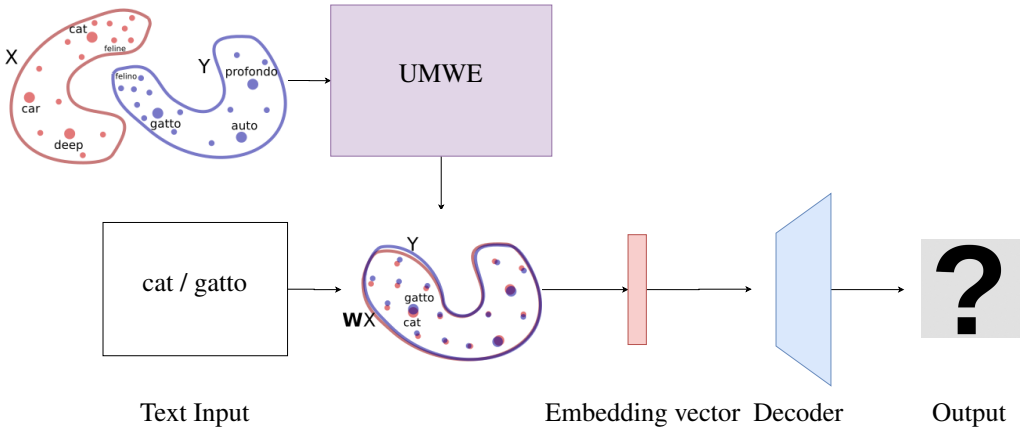

Figure 1: Logogram Language Generator Model

4 Experimental Evaluation

For the language data we will use different sets of languages to observe different outcomes. For example Romance languages (Spanish-French-Italian-Romanian) which is set of languages that have high lexical similarity and more distinct languages like Turkish-English-Finnish etc.
On the image generation part, as mentioned above, we will use Omninglot dataset to gain different concept and features from different alphabets.

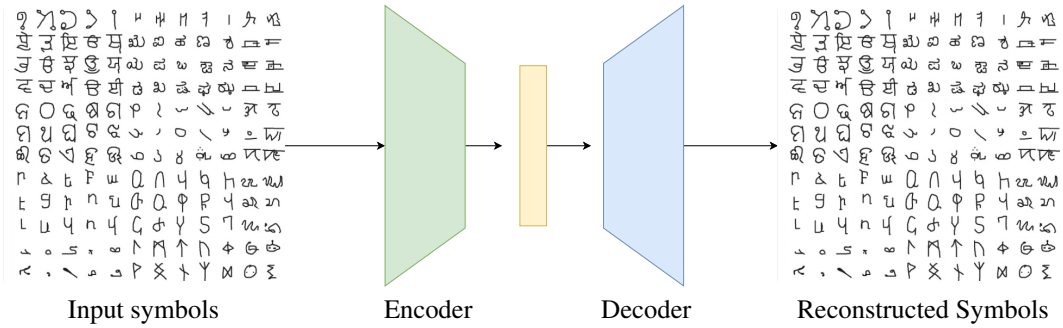

Figure 2: Variational Autoencoder for generating logogram symbols

5 Work Plan

Activity	Deadline
Complete the literature search	09/26/2020
Reproduce results of a baseline approach	10/17/2020
Make improvements and prepare presentation	11/07/2020
Showcase	11/14/2020

References

[1] https://en.wikipedia.org/wiki/Logogram

[2] Mikolov, T., Chen, K., Corrado, G.S., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. CoRR, abs/1301.3781.

[3] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep contextualized word representations. ArXiv, abs/1802.05365.

[4] Pennington, J., Socher, R. Manning, C. D. (2014). Glove: Global Vectors for Word Representation.. EMNLP (p./pp. 1532–1543), .

[5] Conneau, A., Lample, G., Ranzato, M., Denoyer, L., Jégou, H. (2018). Word Translation Without Parallel Data. ArXiv, abs/1710.04087.

[6] Chen, X., Cardie, C. (2018). Unsupervised Multilingual Word Embeddings. EMNLP.

[7] Kingma, D.P., Welling, M. (2014). Auto-Encoding Variational Bayes. CoRR, abs/1312.6114.

[8] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I.J. (2015). Adversarial Autoencoders. ArXiv, abs/1511.05644.

[9] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y. (2014). Generative Adversarial Networks. ArXiv, abs/1406.2661.

[10] https://github.com/brendenlake/omniglot