Project Plan for Visual Embedding of Chinese

Shen Zhuoran, cmsflash@hku.hk
Yao Qingning, yaoqingning@hku.hk

September 25, 2018

Contents

1 Introduction

Contemporary natural language processing relies heavily on deep neural networks. This approach requires words or other linguistic units (e.g. characters, sentences, paragraphs, etc.) to be converted to vectors before being fed into neural networks. The process of converting linguistic units to their vector

representations is language embedding. Traditionally, language embedding models are mostly developed on alphabetical languages like English. In those languages, individual characters do not have meaning and the visual appearance of a linguistic unit is generally unrelated to its semantics. Therefore, traditional approaches for language embedding generally treat linguistic units being processed as black boxes and rely on their distributional properties to infer their meanings. In contrast, the Chinese writing system is morpheme-based and logographic. Hence, an individual Chinese character is meaningful and the meaning of it is related to its appearance. Based on these reasons, this paper will propose to discard the unrealistic black-box assumption for Chinese embedding by providing appearance information of characters to the embedding model.

There are several competing methods to incorporate visual information into the embedding process. Cao et al. (2018) proposed to use stroke n-grams to inform the embedder of the visual structure of a character. They achieved state-of-the-art performance on various tasks, including word similarity, named entity recognition, and text classification. Dai and Cai (2017) suggested another approach that uses convolutional neural networks, a highly successful family of deep-learning models in computer vision, to directly generate the vector representation of a character from a rendered image of it. However, despite the promising application of computer-vision techniques to natural language processing, they did not adopt modern model architectures in computer vision or test the approach on the most well-accepted dataset. Hence, this paper will propose to adapt the latest architectures from computer vision to language embedding and test the models on the largest and most well-accepted Wikipedia Chinese dataset.

2 Background

2.1 Language Embedding

Language embedding is one of the most fundamental tasks in natural language processing. It refers to the process transforming words or other linguistic units to vectors, which enables computers to process the meanings of the vocabularies instead of merely the textual representations. Language embedding is essentially the process of dimensionality reduction, which transforms the words from a high dimensional space (one dimension per word, very much like a one-hot vector) to a lower-dimensional space (e.g. 300D) by extracting common features and meanings out of words.

Language embeddings are not only used directly in tasks like word similarity and machine translation, it is also known to boost performance of tasks such as syntactic parsing. In addition, given the recent advancement in neural networks and deep learning, the continuous vector representation of words or documents are becoming especially useful.

2.2 Visual Embedding of Chinese

Cao et al. (2018) proposed to use stroke n-grams to inform the embedder of the visual structure of a character. The *Standard for Stroke Sequences of General Characters in Standard Chinese* establishes a set of rules that defines a unique sequence of strokes to compose each character in the *Table of General Standard Chinese Characters*. Cao et al. (2018) constructed a database mapping each Chinese character to its predefined sequence of strokes. Their method maps each character to a sequence of strokes per the database and extracts a stroke n-gram (An n-gram describes a sequence by the frequency of each possible subsequence of length n or shorter in the sequence. E.g., a 2-gram of "hello" could be 1*"h", 1*"e", 2*"l", 1*"o", 1*"he", 1*"el", 1*"ll", and 1*"lo".) from the sequence. Then, a feedforward

neural network takes the n-gram and produces a vector reprsentation of the character. Using the embeddings generated through this method in state-of-the-art models in downstream tasks lead to new states-of-the-art on various tasks, including word similarity, named entity recognition, and text classification.

Dai and Cai (2017) suggested another approach that uses convolutional neural networks, a highly successful family of deep-learning models in computer vision, to directly generate the vector representation of a character from a rendered image of it. However, despite the promising application of computer-vision techniques to natural language processing, they did not adopt the latest model architectures for computer vision or test the approach on the most well-accepted dataset.

3 Objective

Our objective is to improve Chinese embedding by adapting the latest advancements in computer vision, produce a visual embedder of Chinese with competitive performance, and apply and evaluate the model in major downstream tasks.

4 Methodology

We will have a thorough literature review of the current state of language embedding. Given our objective and the special characteristics of Chinese, we will investigate both language-independent theories and techniques, and methods specifically for Chinese. Special focus will be directed toward works on visual embedding of Chinese. Literature review will be a continuous process as we progress, though it will be of a heavier focus in early stages of the project.

After reviewing an appropriate amount of existing works, we will prepare baseline implementations of state-of-the-art models, to serve as baselines for comparison and for studying the implementation details like possible optimizations for language embedding tasks. We will try reaching out to the relevant researchers and request their code. If we cannot reach them or they are not willing to share code, we will re-implement the models in Python with `PyTorch` and try to reproduce their results. We have prepared a `PyTorch` template `beauty-net` for implementing the models.

Then, we will design, implement, revise, and optimize our approaches to push for the best performance. Our models will be empirically evaluated against the state-of-the-art models by using the generated embeddings in downstream tasks, like language modeling and text classification. Then we will collect the results and summarize our findings in future documentations.

5 Schedule and Milestones

September:

- Tasks: Background research and project planning
- Milestones: Finish project plan and set up project website

October/November:

- Tasks: Data preparation, further research, and reproduction of current method
- Milestones: Finish data collection, implement baselines

December/January:

- Tasks: Begin implementation and preliminary experiments
- Milestones: Working implementation, initial experiment results, interim report

February through April:

- Tasks: Revision and optimization of the methods, finalization of the project
- Milestones: Finalized implementation, experiment results and final report

References

Cao, Shaosheng, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. "Cw2vec: Learning Chinese Word Embeddings with Stroke N-Gram Information."

Dai, Falcon Z, and Zheng Cai. 2017. "Glyph-Aware Embedding of Chinese Characters." *arXiv Preprint arXiv:1709.00028*.