Subword Embeddings Reveal Language Change

Yang Xu1 and Jiasheng Zhang2 and David Reitter3
1Department of Computer Science, San Diego State University
2College of Information Sciences and Technology, The Pennsylvania State University
3Google AI
y xu4@sdstate.edu, jzp5181@ist.psu.edu, reitter@google.com

Abstract
We propose an augmented word embedding model that better incorporates subword information with additional parameters that characterize the semantic weights of characters in composing words. Our model can reveal some interesting patterns of long-term change in Chinese.

Introduction
• Language change is reflected in how (words and phrases etc.) are composed.
  – Indo-European languages: becoming from synthetic to analytical.
  – Chinese: from single-character words to multi-character words.
• Can we use vector representations of words to characterize these patterns?

Theoretical Background
• Chinese: the relative dominance of the monosyllabic words (i.e., single character as a word) in ancient Chinese has shifted to biophonic words in modern Chinese.
  – Examples: 看 (to win) → 看許 (to win; victory); 看 (to help) → 看許 (to help).
• Most Indo-European languages: shifting from synthetic (i.e., analytic) to analytic (multi-word):
  – Examples: des Haunes (the house’s) → von dem Haus (of the house); Edith chanta (Edith has sung) (Haspelmath and Michaelis, 2017).
• Motivation: Can modern NLP techniques provide deeper insights into these types of shift?

Methodological Background
Word embedding models
• Word2Vec (Mikolov et al., 2013a)
  – Learning word vectors by predicting the target word given context words (continuous bag of words, i.e., CBOW), or predicting the context word given target word (skipgrams).
• CBOW: the learning objective is to maximize the negative log-likelihood $L_{CBOW} = \sum w \log p(w_i | C_i)$, where $w_i$ is the target word, and $C_i$ represents the surrounding context words.
• The probability $p(w_i | C_i)$ is formulated by a softmax function:
  $$p(w_i | C_i) = \frac{\exp(q(w_i, v_i))}{\sum_j \exp(q(w_j, v_i))}$$
  where $v_i = \frac{1}{|C_i|} \sum w_j$ in $C_i$.
• In practice, negative sampling is used instead of softmax to reduce the amount of computation:
  $$L_{CBOW} = - \log (1 + \exp(-s \cdot w_i)) - \sum_{w \neq w_i} \log (1 + \exp(s \cdot w_i))$$
• Logit = $\sum_{w \neq w_i} \log(p_i(w_i))$, and can be estimated by negative sampling similarly.

Incorporating subword information
Principle 1: Semantic compositionality
• Internal subwords of a word contain information about the word’s semantic meanings.
  The meaning of the whole is the sum of the parts.
  – Chinese example: “走马”的 education) can be inferred from the meanings of its first character “走” (to travel) and second character “马” (horse).
• Chen et al. (2015) proposed character-enhanced word embedding (CWE) model for Chinese, by replacing the context word vector $v_i$ with a weighted average vector that incorporates the character vectors. See eq. (2).

Method Dynamic subword-enhanced embeddings (DSE)
We propose DSE, a variant model based on CWE and fastText, which characterizes the semantic weights carried by characters in Chinese words.
• Associate each word with a scalar parameter $b$, indicating the weight of the word itself in predicting the co-occurred words within the context window.

Conclusions and Future Work
• The increasing trend of $b^h$ may reflect the modernization of Chinese language as the concepts and terminology in science and technology (and western culture) had been introduced since the 19th century, and more so ever after 1980s.
• Chinese has been evolving towards multisyllabic from monosyllabic, which is not just reflected in the frequency change, but also in terms of semantic weight.
• Going forward, we would like to see the results of applying similar method to other languages (Indo-European languages especially).
• For more detailed investigation, see our paper: Treat the Word As a Whole or Look Inside? Subword Embeddings Model Language Change and Typology, In Proceedings of the 21st International Workshop on Computational Approaches to Historical Language Change (LREC-2020).

Acknowledgements
The authors acknowledge support from the National Science Foundation, grant BCS-1734304 to D. Reitter.