Subword Embeddings Reveal Language Change

CogSci 2019
The 41st Annual Meeting of the Cognitive Science Society
Montreal, Canada

Yang Xu[1] and **Jiasheng Zhang**[2] and **David Reitter**[3]

[1]Department of Computer Science, San Diego State University

[2]College of Information Sciences and Technology, The Pennsylvania State University

[3]Google AI

yxu4@sdsu.edu, jpz5181@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 predominance of the monosyllabic words (i.e., single character as a word) in ancient Chinese has shifted to bisyllabic words in modern Chinese.
  - Examples: 胜 (to win) → 胜利 (to win; victory); 助 (to help) → 帮助 (to help).
- Most Indo-European languages: shifting from synthetic (single-word) to analytic (multi-word):
  - Examples: des Hauses (*the house's*) → von dem Haus (*of the house*); Edith chanta (*Edith sang*) → Edith a chanté (*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, a.k.a., CBOW), or predicting the context word given target word (skipgram).
- CBOW: the learning objective is to maximize the negative log-likelihood $L_{\text{CBOW}} = \sum_{i=1}^{T} \log p(w_i|C_i)$, 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:

$$p(w_i|C_i) = \frac{\exp(u_i^\intercal \cdot \boldsymbol{v}_c)}{\sum_{j \in V} \exp(u_j^\intercal \cdot \boldsymbol{v}_c)} \quad \text{where } \boldsymbol{v}_c = \frac{1}{|C_i|} \sum_{w_k \in C_i} \boldsymbol{v}_k \tag{1}$$

- In practice, negative sampling is used instead of softmax to reduce the amount of computation:

$$L_{\text{CBOW}} \approx -\log\left(1 + \exp(-u_i^\intercal \cdot \boldsymbol{v}_c)\right) - \sum_{n \in N_{i,c}} \log\left(1 + \exp(u_i^\intercal \cdot \boldsymbol{v}_n)\right)$$

- $L_{\text{skipgram}} = \sum_{i=1}^{T} \sum_{w_k \in C_i} \log p(w_k|w_i)$, and can be estimated by negative sampling similarly.

Incorporating subword information

Principle 1: Semantic compositionality

- Internal subword units of a word contain information about the word's semantic meanings. The meaning of the whole is the sum of the parts.
  - Chinese example: "教育" (education) can be inferred from the meanings of its first character "教" (to teach) and second character "育" (to raise).
- Chen et al. (2015) proposed character-enhanced word embedding (CWE) model for Chinese, by replacing the context word vector $\boldsymbol{v}_k$ with a weighted average vector that incorporates the character vectors. See eq. (2).

CWE includes character embeddings

$$\boldsymbol{x}_k = \frac{1}{2}\boldsymbol{v}_k + \frac{1}{2}\left(\frac{1}{N_k}\sum_{t=1}^{N_k} \boldsymbol{c}_t\right) \tag{2}$$

in which N_k is the number of characters in word w_k, and $\boldsymbol{c}_k$ is the vector of the tth character.

fastText includes n-gram embeddings

$$\boldsymbol{x}_i = \boldsymbol{v}_i + \sum_{t=1}^{N_i} \boldsymbol{c}_t \tag{3}$$

in which N_i is the number of n-grams in word w_i, and $\boldsymbol{c}_t$ is the tth n-gram.

Principle 2: Reducing sparsity

- In some morphologically rich languages, one word can have multiple forms that occur rarely.
- (Bojanowski et al., 2017) proposed fastText model: Learn representations for all n-grams and represent the word as the sum of its n-gram vectors. See eq. (3).
- English example: *love* = `<lo, lov, ove, ve>`. Then the vector of *love*, $\vec{v}_{\texttt{love}}$, is computed as $\vec{v}_{\texttt{love}} + \vec{v}_{\texttt{<lo}} + \vec{v}_{\texttt{lov}} + \vec{v}_{\texttt{ove}} + \vec{v}_{\texttt{ve>}}$

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 h^w, indicating the weight of the word itself in predicting the co-occurred words within the context window.
  - **Meaning of h^w**: How informative a word itself is in predicting its neighbor words.
  - **Meaning of $1 - h^w$**: How informative the subword units in a word (i.e., characters in the case of Chinese) are in predicting the word's neighbors.

Average Embedding in DSE

$$\begin{cases} \boldsymbol{x}'_k = \boldsymbol{h}_k^{\boldsymbol{w}} v_k + (1 - \boldsymbol{h}_k^{\boldsymbol{w}})\left(\frac{1}{N_k}\sum_{t=1}^{N_k} c_t\right), & \text{replacing the } x_k \text{ in CWE, eq. (2)} \\ \boldsymbol{x}'_i = \boldsymbol{h}_i^{\boldsymbol{w}} v_i + (1 - \boldsymbol{h}_i^{\boldsymbol{w}})\sum_{t=1}^{N_i} c_t, & \text{replace the } x_i \text{ in fastText, eq. (3)} \end{cases} \tag{4}$$

in which N_k, N_i, and c_t have the same meanings as in eqs. (2) and (3)

- Smaller h^w means that subword units plays larger role in composing the meaning of word.
- **Difference** from CWE and fastText: the semantic weights of subword units are *dynamically* modeled, i.e., learned from data, instead of being fixed.

Hypothesis

- The semantic weights of characters in a Chinese word should depend on how "new" the word is.
- h^w should be smaller in *older* words, but larger in *newer* words.

Measure the age of a word

- Use *first-appearance-year*: the earliest year that a word appears in the Google Books Ngram dataset (GBN).
- Examples: "爱人" (*love* + *person* = *lover*) first appears in the year of 1804 (AD), while "爱心" (*love* + *heart* = *love*) first appears in 1981. Thus, "爱人" is an older word than "爱心".

Result: h^w ∼ first-appearance-year

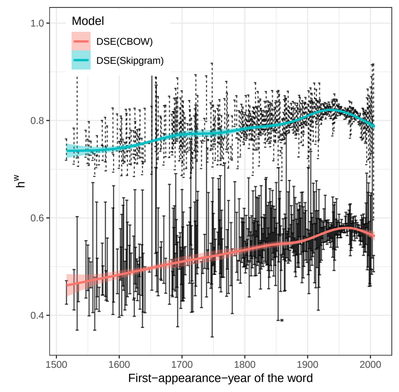

Figure 1: h^w increases with first-appearance-year for all words in GBN. Errorbars are 95% CIs.

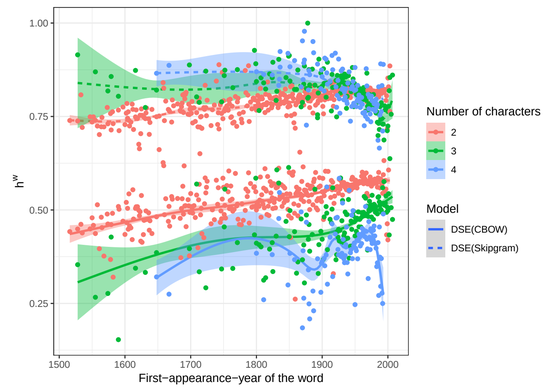

Figure 2: h^w increases with first-appearance-year for words that consist of 2, 3, and 4 characters respectively.

- **Hypothesis supported**: Subword units (i.e., characters) in Chinese carry more semantic weight in early (older) words than in modern (newer) words.

Result: Case Study

Table 1: Case study: older words are on the left with smaller h^w values; newer words are on the right, with larger h^w values. Words in a row share a same character.

Earlier words	h^w	Later words	h^w
安全(secure), 1581	0.75	安打(base hit), 1959	0.85
安定(settled), 1632	0.72	安检(security check), 1987	0.87
组成(consist of) 1568	0.67	课题组 (research group) 1988	0.86
覆盖 (cover), 1747	0.69	盖帽(block), 1972	0.91
把握(hold), 1591	0.69	拖把 (mop), 1985	0.86

- For example, the original meaning of "安" (safe) plays less role in modern words such as "安打", "安检" etc.
- When these words are used, the chunk of characters are more considered as a *whole* semantic unit, and the original meanings of the individual characters are less referred to.
- The magnitude of h^w is related to the part-of-speech tag of words: many new words (with larger h^w) are nouns of terminology.

Conclusions and Future Work

- The increasing trend of h^w 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 1900s
- 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 1st International Workshop on Computational Approaches to Historical Language Change*. (ACL 2019).

Acknowledgements

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