Treat the Word As a Whole or Look Inside?
– Subword Embeddings Model Language Change and Typology

57th ACL 2019 Florence ANNUAL MEETING July 28th - August 2nd of the Association for Computational Linguistics

Presented in LChange 2019
The 1st International Workshop on Computational Approaches to Historical Language Change (co-located with ACL 2019), Florence, Italy

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 use a variant of word embedding model that incorporates *subword* information to characterize the degree of compositionality in lexical semantics. Our models reveal patterns of long-term change in multiple languages: Indo-European languages put more weight on subword units in newer words, while conversely Chinese puts less weights on the subwords, but more weight on the word as a whole.

Background

- Subword units play roles in determining word meanings differently across languages.
- Indo-European languages: a word consists of *morphemes*, e.g., root, affixes.
- Chinese: a word consists of *characters* (字).

Two empirical observations

- Chinese words: *monosyllabic* → *bisyllabic*.
  - Examples: 胜 (to win) → 胜利 (to win; victory); 助 (to help) → 帮助 (to help).
- Indo-European languages: *synthetic* (single-word) → *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)
- Word embedding technique can provide new evidence.

Word2vec and variants

- Word2vec (Mikolov et al., 2013a) learn dense word vectors by:
  - Predicting target word given context words (CBOW)
  - Predicting context word given target word (Skipgram)

Variants: incorporating subword information

Character-enhanced word embedding (CWE) (Chen et al., 2015)

- The meanings of characters contribute to meaning of the word.
  - Example: "教育" (education) = "教" (to teach) + "育" (to raise).
- Method: replace context word vector $\boldsymbol{v}_k$ with a weighted average character vectors. See eq. (1).

fastText (Bojanowski et al., 2017)

- Utilize n-grams to deal with sparsity.
  - Example: $\vec{v}_{\texttt{love}} = \vec{v}_{\texttt{love}} + \vec{v}_{\texttt{<lo}} + \vec{v}_{\texttt{lov}} + \vec{v}_{\texttt{ove}} + \vec{v}_{\texttt{ve>}}$
- Method: represent the word as the sum of its n-gram vectors. See eq. (2).

CWE includes character embeddings

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

fastText includes n-gram embeddings

$$\boldsymbol{x}_i = \boldsymbol{v}_i + \sum_{t=1}^{N_i}\boldsymbol{c}_t \quad (2)$$

Method

Dynamic subword-enhanced embeddings (DSE)

- A variant model based on CWE and fastText characterizing the semantic weights carried by subword units in the word.
- Associate each word w with a parameter h^w
  - **Meaning of $\boldsymbol{h^w}$**: How informative a word itself is in predicting its neighbor words.
  - **Meaning of $\boldsymbol{1-h^w}$**: How informative the subword units in w are in predicting its neighbors.

Average Embedding in DSE

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

Model architectures

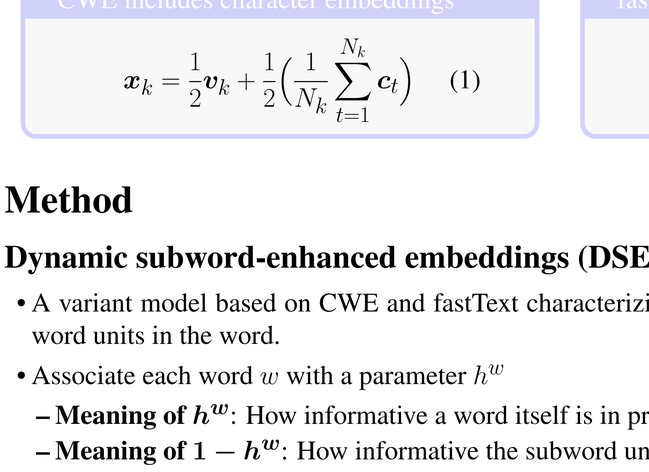


Figure 1: DSE-CBOW

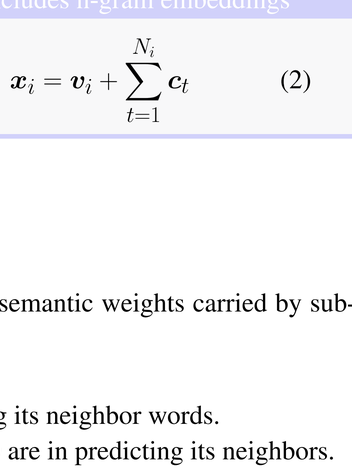


Figure 2: DSE-SG

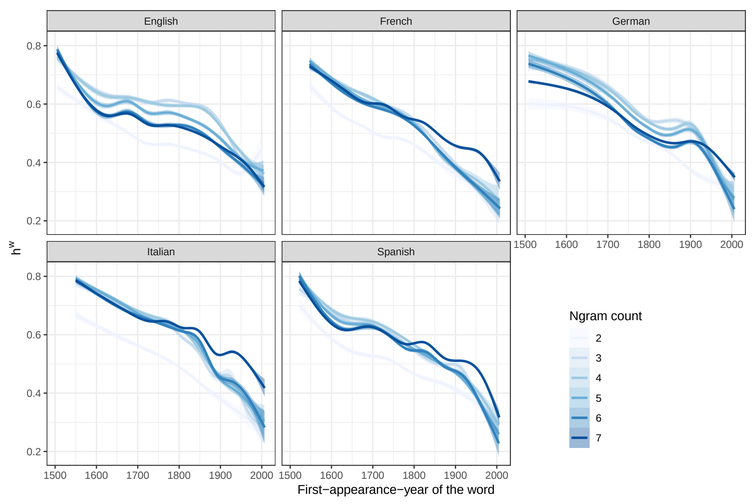
Hypothesis

- h^w of a word should correlate to its relative "age".
- Particularly, newer Chinese words should have larger h^w than those older words.

Measure the age of a word

- *First-appearance-year*: the earliest year that a word appears according to 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: $h^w \sim$ first-appearance-year



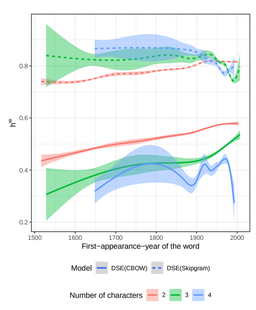


Figure 3: Up: h^w *increases* with first-appearance-year in Chinese words. Left: h^w *decreases* with first-appearance-year in five Indo-European languages.

- **Opposite $h^w \sim$ year relationships**:
  - Subword units in Chinese (i.e., characters) carry more semantic weight in older words than in newer words.
  - In five Indo-European languages, subword units (i.e., n-grams) carry more semantic weight in newer words than in older words.

Ruling out word frequency effect

- Word frequencies might be a confounding effect that explains the $h^w \sim$ year correlation.
- Fit linear model with additional parameter: $h^w \sim$ year + freq. β_{year} remains significant.
- Two-step verification: $m' = h^w \sim$ freq; $m = residuals(m') \sim$ year. β_{year} of the second model remains significant.
- Therefore, the observed $h^w \sim$ first-appearance-year is statistically reliable.

Result: Evaluation of Embeddings

Language	Model	Similarity	Analogy
Chinese	DSE-CBOW	0.597	0.666
	CWE	0.605	0.668
	DSE-SG	0.583	**0.651**
	fastText	0.591	0.588
English	DSE-CBOW	0.659	0.302
	CWE	0.669	0.324
	DSE-SG	0.705	**0.356**
	fastText	0.702	0.338

- Two tasks: word similarity (WordSim-353) and word analogy (questions-words).
- Compared with CWE and fastText, with our own implementations (by disabling the h^w parameters).
- DSE-CBOW compared with CWE; DSE-SG compared with fastText.
- In general, DSE performs decently well.

Result: Case Study

Older words	h^w	Newer 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

Older words	h^w	Newer words	h^w
acid, 1517	0.73	**acid**osis, 1907	0.07
		oxo**acid**s, 1953	0.07
compare, 1524	0.86	**compar**ison, 1659	0.61
		comparatives, 1810	0.14
human, 1504	0.87	trans**human**ism, 1955	0.50
locking, 1600	0.77	un**lock**able, 1854	0.11

Conclusions and Future Work

- Chinese language: characters play less semantic roles in newer words than older ones.
- Indo-European languages: newer word place more semantic weight on subword units.
- **Chinese words are treated more as a whole semantic unit "synthetically", while words in Indo-European languages require more "analytically" attention into the subword level.**
- Future work: Investigate other Eastern-Asian languages; Use roots and affixes instead of n-grams.

Acknowledgements

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