Word Representation: Theoretical Explanation of an Empirical Formula

Leonel Escapita, Diana Licon, Madison Anderson,
Diego Pedraza, and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
El Paso, Texas 79968, USA
mranderson2@miners.utep.edu, laescapita@miners.utep.edu
dlicon2@miners.utep.edu, dapedraza@miners.utep.edu, vladik@utep.edu

1. Formulation of the Problem

- Computer representation of natural-language words should reflect how related these words are.
- This relation can be described, e.g., by the number X_{ij} of times when word i appears in the context of word j.
- It is desirable to find out how such characteristics depend on the properties of the words i and j.
- This way we will be able to predict, e.g., X_{ij} for pairs (i, j) for which we do not know X_{ij}.
- At present, these characteristics are determined by training a neural network.
- There is also a reasonably good approximate analytical formula $\ln(X_{ij}) \approx b_i + \widetilde{b}_j + w_i \cdot \widetilde{w}_j$.
- Here, b_i and $\widetilde{b}_j$ are numbers, w_i and $\widetilde{w}_j$ are vectors, and $a \cdot b$ is dot (scalar) product.

2. Formulation of the Problem (cont-d)

- The values b_i, $\widetilde{b}_j$, w_i and $\widetilde{w}_j$ can be found by using the Least Squares method:

$$J \stackrel{\text{def}}{=} \sum_{i,j} f(X_{ij}) \cdot (b_i + \widetilde{b}_j + w_i \cdot \widetilde{w}_j - \ln(X_{ij}))^2 \to \min .$$

- The efficiency of this method depends on the appropriate choice of the weight function $f(X)$.
- Empirical data shows that the most efficient is power law $f(X) = X^a$.
- How can we explain this empirical fact?

3. Our Explanation

- The values X_{ij} depend on the size of the corpus.
- If we consider twice smaller corpus, each value X_{ij} will decrease approximately by half.
- In general, if we consider a λ times larger corpus, we will get new values which are close to $\lambda \cdot X_{ij}$.
- The word representation should depend only on the words, not on corpus size.
- So, the resulting representation should not change if we replace X_{ij} with $\lambda \cdot X_{ij}$.
- Of course, if we replace X_{ij} with $\lambda \cdot X_{ij}$, the weights will change.
- However, this does not necessarily mean that the resulting representations will change.

4. Our Explanation (cont-d)

- Namely, for any $c > 0$, optimizing any function J is equivalent to optimizing the function $c \cdot J$.
- Example: the richest person on Earth is the richest whether we count his richness in dollars on in pesos.
- If we replace $f(X)$ with $c \cdot f(X)$, we will get $c \cdot J$ instead of J, so we will get the same representations w_i.
- Thus, we can have $f(\lambda \cdot X) = c \cdot f(X)$, and the resulting representation will be the same.
- So, the invariance with respect to corpus size means that for every $\lambda > 0$, there exists c depending on λ for which

$$f(\lambda \cdot X) = c(\lambda) \cdot f(X).$$

5. Our Explanation (cont-d)

- It is known that all measurable solutions to the functional equation $f(\lambda \cdot X) = c(\lambda) \cdot f(X)$ are power laws $f(X) = A \cdot X^a$.
- As we have mentioned, the use of such weights is equivalent to using the weights $f(X) = X^a$.
- This explains why the power law weights work the best.
- Power law weights are the only ones for which the resulting representation does not depend on the corpus size.

6. How to Prove the Result about the Functional Equation

- This result is easy to prove when the function $f(X)$ is differentiable.
- Suppose that $f(\lambda \cdot X) = c(\lambda) \cdot f(X)$.
- If we differentiate both sides with respect to λ, we get
$$X \cdot f'(\lambda \cdot X) = c'(\lambda) \cdot f(X).$$
- In particular, for $\lambda = 1$, we get $X \cdot f'(X) = a \cdot f(X)$, where $a \stackrel{\text{def}}{=} c'(1)$, so $X \cdot \dfrac{df}{dX} = a \cdot f$.
- We can separate the variables if we multiply both sides by $\dfrac{dX}{X \cdot f}$, then we get $\dfrac{df}{f} = a \cdot \dfrac{dX}{X}$.
- Integrating both sides of this equality, we get $\ln(f) = a \cdot \ln(X) + C$.
- By applying $\exp(x)$ to both sides, we get
$$f(X) = \exp(a \cdot \ln(X) + C) = A \cdot X^a, \text{ where } A \stackrel{\text{def}}{=} e^C.$$

7. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.