Enhancing word embeddings
Positionality, subword sizes, and hyphenation

Vítěz Novotný
witiko@mail.muni.cz

Faculty of Informatics, Masaryk University

October 13, 2020
Introduction

- SOTA DNN LMs are accurate but also slow, black-boxed, and monolithic.
- Word embeddings of SNN LMs [1–4] provide strong baselines for many tasks:
  1. semantic text similarity [5]
  2. text classification [6]
  3. information retrieval [7]
- Word embeddings produce systems that are fast, interpretable, and modular.
- MIR@MU research group develops and maintains Gensim [8]:
  - Essential Python NLP library: 2.6k article citations and 11.2k stars on GitHub
  - Contains hardware-accelerated implementation of Word2Vec and fastText SLL LMs. [9]
  - Perfect tool to prototype, implement, and evaluate enhanced word embeddings.
Positionality

- Word2Vec [1, 2] and fastText [4] CBOW models are trained to minimize the distance between the mean of context word embeddings and the masked word embedding:

\[
\text{The quick brown \textcolor{blue}{???} jumps over the lazy dog.}
\]

- However, the position of words in context is not taken into account.

- Mikolov et al. [10] achieved SOTA on English Word Analogy task using position-dependent weighting. However, no open-source implementation exists.

V. Novotný • Enhancing word embeddings • October 13, 2020
Subword sizes

- Unlike Word2Vec [1, 2], fastText [4] embeds not only words, but also subwords. This speeds up training and allows inference of embeddings for unknown words.

- Previous work reports optimal subword sizes only for English and German. Our experiments suggest:
  1. 5% improvement on Czech Word Analogy task with optimal subword sizes over defaults.
Hyphenation

- Hyphenation splits words into subwords based on morphology or phonology.

<table>
<thead>
<tr>
<th>NO HYPHENATION</th>
<th>DEFAULT HYPHENATION</th>
<th>CUSTOMIZED HYPHENATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>would have created surprise. Owing to the proximity of the Hay Market, the number of establishments of bad character, the preponderance of</td>
<td>would have created surprise. Owing to the proximity of the Hay Market, the number of establishments of bad character, the preponderance of</td>
<td>would have created surprise. Owing to the proximity of the Hay Market, the number of establishments of bad character, the preponderance of</td>
</tr>
<tr>
<td>the number of establishments of bad character, the preponderance of</td>
<td>ments of bad char&lt;e&gt;acter, the preponderance of</td>
<td>trading and working</td>
</tr>
</tbody>
</table>

- Hyphenating fastText should decrease model size and speed up training.
Sounds fun?

- Take a look at our bachelor’s and master thesis topics:
  - Positional weighting of fastText word embeddings (bachelor’s thesis, diploma thesis)
  - ... or come up with your own thesis topic!
- Join us at the PV212 seminar this Thursday at 10 AM (CET) over Zoom, where we will dive into the details of our word embedding experiments.
Bibliography


MUNI

FACULTY

OF INFORMATICS