Deep Learning for NLP 2020
Exercise 06 Solution

June 1, 2020

1 Pingo

- Which of the following statements about sense-disambiguated word embeddings are correct?
  - ☑ They differentiate better between two similarly written words with different meanings.
  - □ They differentiate better between two differently written words with the same meaning.
  - ☑ Their use is limited because they require more effort to set up.
- Which of the following statements apply on bilingual embeddings?
  - ☑ They are helpful for training neural networks for languages where few resources exist.
  - □ In order to train bilingual embeddings, one always needs parallel texts.
  - □ BiVCD relies on correlation analysis.
- Which of the following statements on syntactic embeddings are correct?
  - ☑ Contexts in word2vec do not include information about word order.
  - □ The structured skip-gram model is based on word contexts of variable length.
  - ☑ Dependency-based embeddings are an attempt at solving the problem of long dependencies.

2 Using Word Embeddings

2.1 Contextualized Embeddings

Multilingual BERT is trained by running standard BERT on the concatenation of multilingual text from the largest 100 languages in Wikipedia. Does this yield multilingual contextualized embeddings? Which problems do you see? Explain in up to 5 sentences.

Answer: The word embeddings obtained from multilingual BERT for mutual word translations are misaligned in a shared vector space, as the cross-lingual signals, e.g., parallel data and cross-lingual dictionary, are not involved during the training. In addition, languages with differing typological profiles are not equally hard for learning static and contextualized embeddings. For instance, German has higher inflectional morphology than Chinese. This hinders the success of zero-shot cross-lingual transfer between languages, where annotated training data is provided in English but none is provided in the language to which systems must transfer.

2.2 Handling out-of-vocabulary Words

When using pre-trained word embeddings, one can run into the issue that certain words/tokens are missing from the embedding vocabulary (so called out-of-vocabulary words), i.e. the embedding vector is unknown for these words.

Think of at least three possible solutions and state their advantages and disadvantages.

Solution	Advantage	Disadvantage
ignore and skip OOV words	easy to implement	loss of information (criticality depends on the task; in text classification it can go either way)
if available, use the provided `UNKNOWN` token	properly learned vector which should be distant from non-OOV vectors	loss of information, cannot distinguish different OOV words
use one random vector per OOV word	easy to implement	random vectors might be close to "real" vectors, needs knowledge about embedding vector space to be done right
use character or sub-word embeddings	distinguishable embeddings for different OOV words	implementation effort
look up the token on the internet, then average vectors of known words in search results	might work (depending on the token)	slow (not suitable for live systems)