**A SIMPLE BUT TOUGH-TO-BEAT BASELINE**

**FOR SENTENCE EMBEDDINGS**

**Sanjeev Arora, Yingyu Liang, Tengyu Ma**

Princeton University {arora,yingyul,tengyu}@cs.princeton.edu

**Abstract**

The success of neural network methods for computing word embeddings has motivated methods for generating semantic embeddings of longer pieces of text, such as sentences and paragraphs. Surprisingly, Wieting et al (ICLR’16) showed that such complicated methods are outperformed, especially in out-of-domain (transfer learning) settings, by simpler methods involving mild retraining of word embed- dings and basic linear regression. The method of Wieting et al. requires retraining with a substantial labeled dataset such as Paraphrase Database (Ganitkevitch et al., 2013).

The current paper goes further, showing that the following completely unsuper- vised sentence embedding is a formidable baseline: Use word embeddings com- puted using one of the popular methods on unlabeled corpus like Wikipedia, rep- resent the sentence by a weighted average of the word vectors, and then modify them a bit using PCA/SVD. This weighting improves performance by about 10% to 30% in textual similarity tasks, and beats sophisticated supervised methods in- cluding RNN’s and LSTM’s. It even improves Wieting et al.’s embeddings.

This simple method should be used as the baseline to beat in future, especially when labeled training data is scarce or nonexistent.

The paper also gives a theoretical explanation of the success of the above unsupervised method using a latent variable generative model for sentences, which is a simple extension of the model in Arora et al. (TACL’16) with new “smoothing” terms that allow for words occurring out of context, as well as high probabilities for words like *and*, *not* in all contexts.

**摘要**

通过神经网络方法来计算词嵌入的成功促进了用于生成较长文本（例如句子和段落）的语义嵌入的方法。令人惊讶的是，Wieting等人（ICLR'16）表明，这种复杂的方法表现非常优异，尤其是在领域外任务（迁移学习）中，通过一些简单的方法包括单词嵌入的微调再训练和基本的线性回归。 Wieting等人的方法需要使用标记数据集进行再训练，例如段落语句数据库（Ganitkevitch et al。，2013）。

目前的论文更进一步的表明以下的完全无监督的句子嵌入方法是一个强大的基线：即使用一种流行的方法，在未标记的语料库如维基百科上计算词向量，利用词向量的加权平均值代表句子，然后使用PCA / SVD进行微调。这种加权方法在文本相似性任务中将性能提高了大约10％到30％，并且击败了复杂的监督方法，包括RNN和LSTM。它甚至改进了Wieting等人的嵌入方法。这个简单的方法应该被用作未来的基线，特别是当标记的训练数据很少或不存在时。

本文还对使用潜在变量生成模型进行句子的上述无监督方法的成功进行了理论解释。Arora等人对模型进行了简单扩展，具有新的“平滑”方式，即允许出现在上下文之外的单词，以及所有上下文中的高频率单词如and，not。