**Enriching Word Vectors with Subword Information**

**Piotr Bojanowski∗ Edouard Grave∗ Armand Joulin Tomas Mikolov**

Facebook AI Research {bojanowski,egrave,ajoulin,tmikolov}@fb.com

**Abstract**

Continuous word representations, trained on large unlabeled corpora are useful for many natural language processing tasks. Popular models that learn such representations ignore the morphology of words, by assigning a distinct vector to each word. This is a limitation, especially for languages with large vocabularies and many rare words. In this paper, we pro- pose a new approach based on the skip-gram model, where each word is represented as a bag of character n-grams. A vector representation is associated to each character n-gram; words being represented as the sum of these representations. Our method is fast, allow- ing to train models on large corpora quickly and allows us to compute word representations for words that did not appear in the training data. We evaluate our word representations on nine different languages, both on word similarity and analogy tasks. By comparing to recently proposed morphological word repre- sentations, we show that our vectors achieve state-of-the-art performance on these tasks.

**摘要**

在大型未标记语料库上训练的连续单词表示可用于许多自然语言处理任务。学习这种表示的常用模型通过为每个单词分配不同的向量即可忽略单词的形态。这是一个限制，特别是对于具有大词汇量和许多罕见词汇的语言。在本文中，我们提出了一种基于skip-gram模型的新方法。其中每个单词被表示为一系列n-gram字符，向量表示与每个n-gram字符相关联，这些表示的总和即为词向量。这种方法可以快速出训练大型语料库上的模型，并允许我们计算未出现在训练数据中的单词的单词表示。我们在九种不同的语言上评估我们的词向量，包括单词相似度和类比任务。通过与最近提出的形态词表示进行比较，我们表明我们的向量在这些任务上实现了最先进的性能。