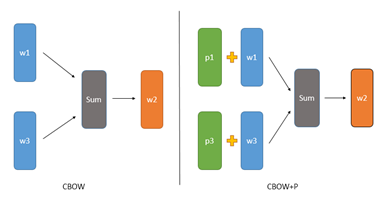
词向量是文本表示的重要组成。当前绝大多数词向量生成方法都是围绕着文字本身展开，利用上下文、权重、依存关系以及词的形态等信息优化词向量的训练。然而从语言学角度看，口语才是直接表义的，文字由于对口语的记录才有了语义。由于口语的直接表义，本文提出了pronunciation-enhanced word embedding(PWE)思想，把词的发音融入到词向量的训练，充分发挥语音和文字在表义上的作用。PWE可以和现有的许多模型相结合，也适用于不同的语言。

从文字与发音的关系来说，汉语、英语和西班牙语是三种不同类型的语言。汉语中，文字与发音没有任何关联。在英语中，由于英语的复杂发音规则，我们只能从词的拼写中大概猜出词的发音。对于西班牙语来说，几乎每个字母的发音都是固定的，我们可以根据词直接得到发音。因此，本文以汉语、英语和西班牙语为例，在word2ec的基础上融入PWE思想。此外，本文也将以中文为例，在CWE（character enhanced word embedding）模型的基础上融入PWE思想。CWE模型考虑到一些语言的词是由更加基本的单位组成，如中文的词是由字组成，因此其将这些基本单位也融入到词向量的建模中。



Word2vec包含CBOW和Skip-gram，基于CBOW的PWE模型如上图所示，其利用词的发音信息对词向量进行修正。在模型中，我们首先对拼音或者音标向量化，得到音向量。修正的词向量就是将原先的词向量和音向量相加，使得修正的词向量既包含文字信息，也包含发音信息。

以word2vec为例，CWE模型也是在输入层对词表示进行修正，将词所包含的字向量也加入到原先的词向量中。因此，基于CWE的PWE模型是要将词的字信息和音信息一同加入到词向量中，使修正词向量既包含字信息，也包含音信息。如下图所示

我们以汉语、英语和西班牙语为例在词相似性和文本分类实验上对PWE模型进行了评估。汉语中，词相似性实验的数据集是wordsim-240和wordsim-296，文本分类的数据集是网易、搜狗和复旦新闻分类数据；英语中，我们用MTurk-771、MEN、WS-353-SIM和WS-353-REL为词相似性实验数据集，用20Newsgroups新闻分类语料库作为分类数据集；西班牙语中，我们用WS-353和RG-65作为词相似性实验数据，用TASS 2017作为文本分类数据，TASS 2017是源于推特的情感分类语料库。实验结果表明，不论是词相似性实验还是文本分类实验，基于PWE思想的模型普遍优于基准模型，说明加入音信息后，词向量确有改进。。