基于高斯分布和汉字组件特征的中文词表示学习

易洁，钟茂生，刘根，王明文

江西师范大学 计算机信息工程学院

论文摘要

现有的词表示学习方法都是将每个词映射到低维空间中的一个点，致使词语语义表示过于刚性，该文使用一种基于密度的分布式嵌入式表示，并给出一种学习高斯分布空间表示的方法，以更好地捕获关于表示及其关系的不确定性，比点积余弦相似度更自然地表达词语的不对称性。同时，针对中文汉字本身特点，将组成汉字的组件即子汉字的语义信息加入词表示训练。与现有方法对比，该文的模型性能在词语相似度或下游任务等方面有更好的效果，且能更好表达词语的不确定性。

系统模型

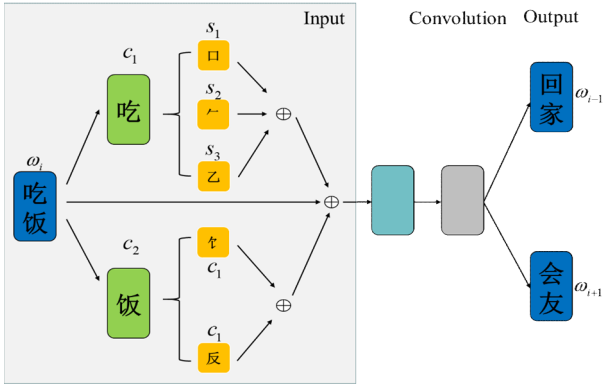

图1 整体模型示意图

论文简介

为了对语言进行建模，词语语义是第一步，只有在此基础上，我们才能进行后续的自然语言处理任务，如命名实体识别、文本分类、情感分析、问答等。现有的词嵌入模型，如word2vec，是将每个词表示成一个定长的向量，后续许多其他词表示方法也大都在此基础上展开。

虽然这些方法在很多下游任务上都被证明是很有效的，但是这些方法将词语表示为空间中的一个点，致使词语语义表示过于刚性，不能很好的捕获词语与相关的目标概念之间的不确定性。点向量之间一般是通过欧式距离、余弦距离或者点积进行相似度计算比较，它们都没有提供对象之间的非对称比较，如包含表示（inclusion）或者蕴含表示（entailment）。此外，现有方法大多将词语作为最小单位，忽略了字的形态信息。与其他语言不同，汉字是一种象形字，本身由偏旁部首等部件组成，这些偏旁部首组件包含着丰富的语义信息，构成字的部首和除部首之外的组件都可以增强字的语义信息。

本文将高斯分布和汉字部件信息两个方法结合，使用高斯分布来表示词语，并针对中文汉字的特点，将汉字的部件信息融合进去进行词表示模型训练。然后将得到的模型在词语相似度、文本分类、命名实体识别等任务上进行评估，实验结果表明本文模型学习得到的词表示模型具有较好的效果。

算法原理

将每个词语使用高斯分布进行表示，每个词都对应一个函数：

$$f(x) = N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

对于词语ω，本文使用组件特征向量和中心词向量来估计均值向量μ_ω：

$$\mu_\omega = \frac{1}{m+1}\left(\sum_{j=1}^{m} s_j + v_m\right)$$

使用期望似然核（expected likelihood kernel）作为能量函数E来计算两个单词ω_f和ω_g之间的相似度：

$$E(f,g) = \int f(x)g(x) = (f,g)_{L2}$$

模型损失函数L为：

$$L_m(v,c_p,c_n) = max\left(0, m - logE(v,c_p) + logE(v,c_n)\right)$$

其中c_p为正样本，c_n为负样本，m为边缘参数。

实验仿真

- 训练数据。 本文使用中文维基百科语料库，去除了文本中的英文、纯数字和一些非法字符，将繁体字转为简体字，去除停用词后再使用jieba工具进行分词，最后得到包含163,606,948个词语的语料库。本文直接使用JWE模型代码 中从HTTPCN爬取的汉字组件文件，其中包含了20,879个字符，13,253个组件特征和218个部首，其中7,744个汉字由多个组件构成，214个汉字的组件就是其偏旁部首。
- 词语相似度。 本文使用的是两个通用的评测数据集——wordsim-240和wordsim-297，数据集中每一对词语都包含人工评价的词语相似度分数，其中240数据集有1个词语没有在语料库中出现，297数据集有4个词语没有出现。使用模型对两个数据集中词语进行相似度打分，最后计算与人工评价分数的Spearman相关系数。
- 文本分类和命名实体识别。 本文使用的是网上已有的文本分类和命名实体识别项目，将skip-gram、高斯分布、JWE和本模型训练得到的词向量作为预训练模型输入，最后比较下游任务效果。其中文本分类任务使用的基线方法是CNN模型，命名实体识别任务使用的基线方法是BiLSTM+CRF模型。
- 给定一个词语，将skip-gram、高斯分布、JWE、本模型得到的相似度最高的前10个词语进行比较。

实验结果如下：

表一：词语相似度比较结果

	Wordsim-240	Wordsim-297	Avg
Skip-gram	0.523	0.564	0.544
高斯分布	0.550	0.543	0.547
JWE	0.521	**0.616**	0.569
本模型	**0.576**	0.570	**0.573**

表二：下游任务效果

	文本分类	命名实体识别
Skip-gram	96.84	87.70
高斯分布	96.82	88.72
JWE	96.20	87.95
本模型	**97.22**	**89.41**

表三：相似度最高的前10个词语

目标词	Skip-gram	高斯分布	JWE	本模型
苹果	小米	黑莓	好像变	苹果公司
	黑莓	坚果	苹果汁	小米
	坚果	葡萄	苹果公司	黑莓
	洋葱	小米	坚果	蓝莓
	红莓	手机	蛇果	水果
	树莓	苹果电脑	番茄	柠檬
	香蕉	苹果公司	苹果电脑	葡萄
	蓝莓	水果	冰淇淋	坚果
	水果	草莓	拒阅	华硕
	红富士	橘子	水果	红米
小米	苹果	红米	红米	红米
	坚果	苹果	小吱	玉米
	红米	黑莓	玉米	骁龙
	手机	玉米	魅族	坚果
	手机操作系统	手机	釉米	四核
	魅族	智能手机	华为	魅族
	黑莓	土豆	金山公司	手机
	土豆	骁龙	坚果	苹果
	智能手机	华为	雷军	双核
	银耳	坚果	中国移动通信	乔布斯

论文结论

本文提出使用高斯分布来表示词语，能够更好地捕获关于表示及其关系的不确定性，同时基于中文汉字特点加入组件特征，最后在词语相似度计算、文本分类和命名识别识别任务以及词语定性分析上均取得较好的效果。但同时也存在一些问题，如有些汉字的组件对汉字本身并没有语义促进作用，如“智”，其组件特征“矢”“日”“口”对其语义并没有帮助，反而产生噪音，所以今后可以考虑计算汉字组件特征的贡献度，或者加入五笔结构特征来进一步提高模型效果。另外，除了单维高斯分布，还会尝试使用高斯混合分布来进行词语多义性的表示。

西安电子科技大学 XIDIAN UNIVERSITY