**读书报告**

王贵涛

自然语言处理的第六章主要介绍了向量语义与嵌入。

**词汇语义学（Lexical Semantics）**

词汇语义学讲的是如何表达一个单词的意思。我们需要建立一个模型来解释单词的意思，词义模型应该允许我们得出有用的推论，帮助我们解决与意义相关的任务，如回答问题、总结、检测释义或抄袭以及对话。

同一个单词也可能有完全不同的意思，这就需要我们解决这个单词在语境中到底表达什么意思。不同的单词也可能表达同一个意思，我们称为同义词（synonym），他们有同样的语义（propositional meaning）。

有些单词虽然不是同义词，但是相似词，判断两个词是不是相似词，有多相似对我们进行语义判断很有用。两个单词除了相似之外，还可以通过其他方式联系起来，如词语关联（word relatedness），表示两个物体是相关的。

词与词之间的常见亲缘关系是语义场（semantic field），指一组包含特定语义范畴的并相互之间具有结构关系的词。语义框架（semantic frame）指一组表示特定类型事件的透视图或参与者的单词。

单词还具有内涵（connotations），例如读者的感情、情绪、观点或评价。词汇的情感意义有三个重要维度，效价（valence）、唤醒（arousal）和支配（dominance）。

**向量语义学（Vector Semantics）**

一个词的意义在于它在语言中的使用，不应该使用某种逻辑语言来定义每个词，通过语言使用中的环境或分布来定义一个词。

向量语义是将一个词表示为多维语义空间中的一个点，表示单词的操作称为嵌入（embeddings）。

**单词和向量（Words and Vectors）**

词汇-文本矩阵是每个文本中每个词汇出现的次数，表示为一个向量空间。如果两个文档具有相似的单词，那它们的列向量也相似。

信息检索（Information retrieval）的任务是在文档集合中找到最符合查询的文档。

单词向量是行向量，指单词在每个文档中出现的次数。

最常用的单词-单词矩阵记录的是两个单词一同出现的次数。

**用余弦测量相似度（Cosine for measuring similarity）**

定义两个单词之间的相似度，最常见的方法是取向量之间夹角的余弦，余弦的范围是0-1，数字越大说明两个单词越相似。

**TF-IDF：矢量中的加权项（TF-IDF: Weighing terms in the vector）**

单词-单词矩阵的问题是虽然频繁出现在上下文的单词很重要，但有些单词往往无处不在，反而就不重要了。tf-idf算法平衡了这种冲突。

首先计算单词频率，加1后取对数。再计算单词出现的文档数量，用文档总数除后取对数。最后的权重由以上计算的两个结果相乘得到。

**点态互信息（Pointwise Mutual Information）**

我们假设两个单词的出现是独立的，可以用两个单词出现的频率相乘做分母，两个单词一起出现的频率做分子，再取对数，称为点态互信息（PMI）。这种做法需要一个庞大的语料库。如果遇到负值，可以取零，称为PPMI。

**Word2vec模型（****Word2vec）**

Word2vec模型的特点是向量短而密集。稠密向量的表现往往比稀疏向量要好，可以更好地避免过拟合。Word2vec指判断一个单词是否会出现在另一个单词附近。

我们采用skip-gram方法，将上下文当做正例，随机抽取词汇当做负例，使用逻辑回归训练分类器，使用回归权值进行嵌入。

skip-gram假设所以上下文单词都是独立的，因此可以将概率相乘。skip-gram需要的反例往往比正例要多，为每个正例配置k个反例。我们需要最大化目标词的相似性，减少负例的相似性，即最大化相似单词的点积，最小化负例的点积。

之后可以使用随机梯度下降来训练模型。窗口的大小等因素也会影响性能，窗口越大，需要的训练也就越多。

**嵌入的语义属性（Semantic properties of embeddings）**

向量语义模型有很多参数，如窗口的大小。更短的上下文窗口会导致更语法化的表示，往往是具有相同词性的语义相近的词。使用长窗口时，得到的往往是与主题相关但不相似的词。

如果两个词相靠近，说明两个词是一级联想。如果两个词有共同的邻居，说明他们是二级联想。

嵌入的另一个作用是捕获关系，向量的偏移可以使单词之间形成某种类推关系。嵌入还可以用来研究单词意思随时间的变化。

单词嵌入也会导致偏见的产生，基于人们的刻板印象，因此查询时会错误的降低一部分文档的权重。嵌入还会展示出人类推理的内联想，比如人们在观念中会将什么样的词联想在一起。通过目前的方法可以改善偏见，但并不能消除它。

历史嵌入可以用来衡量过去的偏见，对古老的种族刻板印象进行调查。