项目名称：基于集合概念的自然语言理解建模

摘要：

1. 立项背景和依据
2. 研究目的

本项目的研究目的是利用集合进行语义消歧（确定一个词在语境中的含义）

自然语言理解是现今一大热门研究方向，其目标是让计算机处理或“理解”自然语言，以完成有意义的任务，比如订机票、购物或同声传译等。

然而，自然语言含有大量歧义，这些歧义根据语境的不同而表现为特定的义项。比如汉语中的多义词，只有在特定的上下文中才能确定其含义，甚至存在故意利用无法确定的歧义营造幽默效果的用法。例如，给定一句话“苹果的创始人是乔布斯，它的logo是苹果”，在这句话中第一个苹果指苹果公司，而第二个苹果则指的是苹果这种水果。

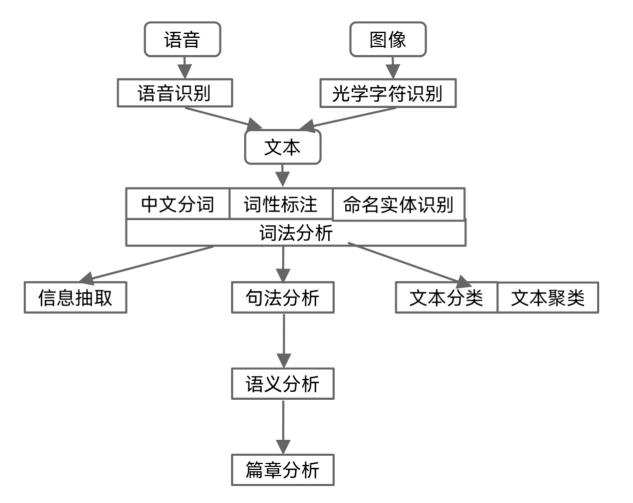
向量化作为一种将单词映射到数学空间的方法，同时保留其本身蕴含的信息。

目前的自然语言理解模型如BOW模型（词袋）、IT-IDF（词频-反向文档频率）、N-gram模型、神经网络语言模型、Word2vec、向量空间模型等，虽然解决了人工构建词典效率低、耗时费力的不足，其中部分模型更是应用了深度学习大大提高了效率，但依然在歧义句的理解上表现不佳。

我们意识到这些模型只是从统计学的角度得到词与词之间的关系，并不是真正理解了语义。故而

1. 自然语言处理研究现状

一个完整的NLP问题解决过程是一个层次化过程。



我们研究的词义消歧就是语义分析的内容。语义分析是要找出词的意思，并在词意的基础上拼接出一段完整话语的意思，从而确定语言所表达的真正含义或概念。而想要完成语义分析，词法分析和句法分析就是重中之重。词法分析是指找出词汇的各个组成部分，分析这些组成部分之间的关系，进而从中获得语言学的信息。句法分析是对句子和短语的结构进行分析，目的是找出词、短语等的相互关系及其在句中的作用。

中文词法分析已经比较成熟，基本达到了工业使用的水准。目前的分词工具有jieba，PKUSEG，THULAC，它们都支持词性标注。

jieba支持四种模式。精确模式，试图将句子最精确地切开，适合文本分析；全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。paddle模式，利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。

PKUSEG的优势在于多领域分词。不同于以往的通用中文分词工具，此工具包同时致力于为不同领域的数据提供个性化的预训练模型。根据待分词文本的领域特点，用户可以自由地选择不同的模型。pkuseg目前支持了新闻领域，网络领域，医药领域，旅游领域，以及混合领域的分词预训练模型。在使用中，如果用户明确待分词的领域，可加载对应的模型进行分词。如果用户无法确定具体领域，推荐使用在混合领域上训练的通用模型。

THULAC则具有分词能力强、分词准确率高、速度快的综合优势。利用集成的目前世界上规模最大的人工分词和词性标注中文语料库（约含5800万字）训练而成，模型标注能力强大。该工具包在标准数据集Chinese Treebank（CTB5）上分词的F1值可达97.3％，词性标注的F1值可达到92.9％，与该数据集上最好方法效果相当。速度较快。同时进行分词和词性标注速度为300KB/s，每秒可处理约15万字。只进行分词速度可达到1.3MB/s。

1. 词向量的发展（深度学习）

在自然语言处理领域，文本作为非结构化的字符数据，首先需要转化为可计算的数值数据，所以首先进行语法分析，既将文本分割为单独的单词，将单词作为文本的基本单位。而独热编码（既每个单词被表示为只有对应位置为1其余为0的向量）虽然具有简单性和健壮性的优点，然而单词表示之间没有相似性，互相之间没有关联，不包含任何语义语法信息，且在词汇表特别大的情况下，会使模型的计算量剧增造成维数灾难。所以针对这种情况有人提出了词向量--一种词的分布式表示方法。它是一个维度相对来说较低的稠密向量，也就是说它的大部分维度都为非0实数。自从词向量被提出并结合神经网络应用在自然语言处理子任务中，许多任务的准确率大大提升，例如命名实体识别、事件抽取和自动问答等。

词向量作为词的分布式表示方法自从1986年被Hinton提出后，经过多年的研究，产生了非常多词向量的生成模型。不同的模型由于其输入输出的不同，使得词向量具有不同的含义和影响。例如Skip-Gram模型中的词向量，词向量直接可以做简单的算数运算来类比词之间的相似性。

2006年，Geoffrey Hinton提出了深度神经网络反向传播（BP）算法。深度学习网络利用一个多层的神经网络，将原始数据从输入层开始，经过逐层非线性变化得到输出。从词向量到Word2Vec从RNN到LSTM，运用深度神经网络的强大拟合能力，结合从互联网获得的海量数据，自然语言处理技术日趋成熟，并取得了巨大的成功。

深度学习的概念源于人工神经网络的研究，其采用的模型主要是神经网络模型。受到人脑神经系统的启发，研究人员构造了一种模仿人脑神经系统的数学模型，成为人工神经网络（Artificial Neural Network,ANN)，简称神经网络。

感知器（Perceptron）是一种模仿生物神经元构建的数学模型，他具有四个重要的组成部分：输入、权重和偏置量、求和函数以及激活函数。

**参考文献**

1. 自然语言处理基础教程[专著]=A first course in natural language processing/王刚，郭蕴，王晨编著-北京：机械工业出版社
2. [深入解词向量-词向量的可视化\_amao1998的博客-CSDN博客\_词向量可视化](https://blog.csdn.net/amao1998/article/details/81742824)
3. [本课程详细讲解了word2vec的训练过程和方法, 不涉及数学公式, 简单易懂, 并且全程手写代码加强理解。\_哔哩哔哩\_bilibili](https://www.bilibili.com/video/BV11U4y1E7sK?spm_id_from=333.1007.top_right_bar_window_default_collection.content.click)
4. 项目研究内容
5. 语义消歧

无论在语法层次、句法层次、语义层次还是在语用层次，自然语言中都存在着大量歧义现象，如何消解这些歧义是自然语言处理面临的核心问题。

1. （语法消歧）下雨天留客天留人不留

如何划分词的边界是中文信息处理中的一个难题。例1有多种划分方式，如下所示：

1. 下雨天，留客天，留人不留？
2. 下雨天，留客天，留人不？留！
3. 下雨天留客，天留人不留。
4. （句法消歧）做手术的是张老师。

这句话有两种理解方式：一种是张老师生病了，需要做手术；另一种是张老师主刀医师，由他来做这场手术。

1. （语用消歧）小张欺负小李，因此我批评了他。

这句话实际上并不存在歧义，“他”作为人称代词虽然即可以指代小张又可以指代小李，但在“小张欺负小李”的这个语境下，小张犯了错误，“我”批评的是小张。

而针对多义词处理的语义消歧则是我们组的研究内容。

研究目标

第一，使计算机理解语义

三、项目特色及创新点

摆脱了传统的基于概率式的判别模型，不再是基于大量语料得到的词与词之间的关系，而是建立语义与语义之间的关系，从而大大提高自然语言模型在处理歧义句方面的准确性。

将词语的不同语义做成一个集合。

虽然是理解语义，但并不是构建词典，同样通过把文字表示成计算机能够运算的数字或向量来提高效率。

四、申请理由

1.团队条件

成员来自计算机科学与技术学院科学与技术专业，在校内接受过专业训练，因此团队具有较强的算法能力。团队成员均有获得学院二等以上奖学金的经历。大部分同学微积分，线性代数等数学基础专业课成绩优秀，具有较好的数学基础。

1. 前期准备工作

成员阅读了大量相关论文和书籍。

1. 项目实施方案
2. 语义消歧研究思路和方法

第一阶段，词法分析，进行中文分词和词性标注等词法分析任务。在Python中使用jieba中的paddle模式。

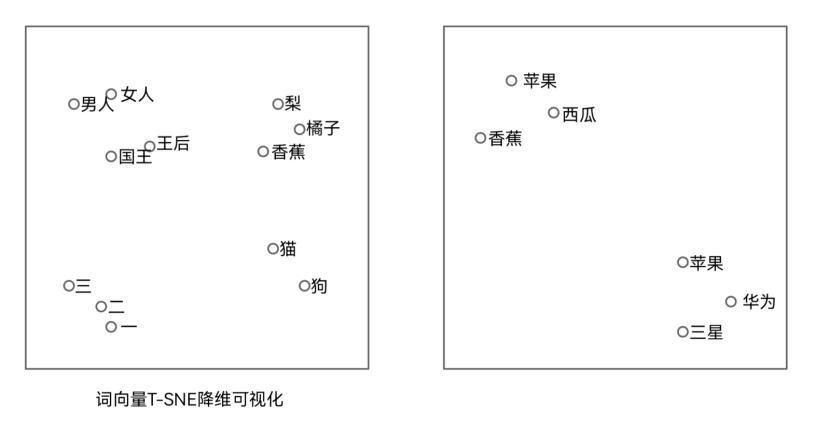
第二阶段，向量转化，借助Gensim库和scikit-learn来完成向量转化。构建语料库。

2.技术路线

词嵌入与Word2Vec

Word2Vec算法是机器学习在文本分析中最重要的应用，是一个非常有用的工具。Word2vec是用来把语料库中的单词构建成向量的工具，其优势是可以通过够建好的向量表达单词的语义信息。

Mikolov的论文中有一个著名的例子，用V代表单词的向量，使用向量执行运算：V（King）-V（Man）+V（Woman）的结果与V（Queen）向量非常接近。



Word2Vec通过理解上下文来工作。通过设定一个滑动窗口大小，并基于这个窗口大小，尝试根据周围的单词来识别输出单词的条件概率。

Word2vec的训练方式有两种，分别是Continuous Bag of Words（CBOW）模型和Skip Gram模型。CBOW方法是用周围词预测中心词，从而利用中心词的预测结果不断地调整周围的向量。训练完成后，每个词都会作为中心词，对周围词的词向量进行调整，从而获得整个文本里所有词的词向量。CBOW对周围的词的调整是统一的：求出的梯度的值会同样地作用到每个周围词的词向量当中去。所以CBOW的预测行为的次数跟整个文本的词数几乎是想等的（每次预测行为才会进行一次反向传播）复杂度是O（V）。而Skip-Gram是用中心词来预测周围的词。在Skip-Gram中，会利用周围的词的预测结果来不断地调整中心词的词向量。最终所有的文本遍历完毕之后，也就得到了文本所有词的词向量。

首先，利用jieba分词工具，对所分析句子进行基础分词。

对所有单词建立关键词对应索引的word\_2\_index字典，与索引对应关键词的index\_2\_word字典，每一个key与其value唯一对应（{‘key’：value}）。同时建立关键词对应1行n列矩阵的word\_2\_onehot字典。词表即我们所构建的全部词的集合，n为词表大小，也就是我们集合中词的个数，矩阵中与第value个位置的数值为1，其余位置为零。

然后以词为单位设定一个滑动窗口，按照设定好的规则为所得例句词语进行初始分析学习，记录所求真实值。规则即为滑动窗口长度，输入输出设置等。

准备一个行数为词典大小、列数为简化后大小（w1）和一个行数为简化后大小、列数为词典大小（w2）的矩阵，根据已获得信息进行简单初始化。然后对w1与w2进行训练。

下面以关键语句“今天天气真好” 为例，假设我们要分析关键词“天气”，词表大小为5000，简化大小为50。首先根据key‘天气’，在word\_2\_onehot表中找到对应矩阵，使其与w1进行矩阵乘法，即可获得与关键词唯一对应的简化后的1行50列矩阵，这个矩阵即为我们所求的词向量。

使词向量与w2进行矩阵乘法，可得到一个1行5000矩阵，对该矩阵进行softmax操作，使所得矩阵各个项的数值均小于1（softmax操作后矩阵中各个项之和等于1）。找到矩阵中各个项数值最大也就是最接近1的数，取到其的列数x。此时word\_2\_onehot表中第n列数值为x的词应为我们所要求的输出词。由于w1与w2刚建立时数据集不完善存在较大误差，于是设定loss值，用于记录所得值与真实值之间的误差。根据loss值重新对w1与w2进行修改调节，使其结果向真实值靠拢。

为了简化操作，我们可以利用word\_2\_index字典与index\_2\_word字典，基于word\_2\_onehot字典中有且只有唯一一位数为零的特性，我们可以利用word\_2\_index字典，格局关键字直接取到索引x，此时w1中第x行即为我们所要求的词向量。在寻找目标矩阵的过程中，获取经过softmax操做后数值最大数的行号后，可以直接根据此值利用index\_2\_word字典直接对应出所求关键词。

同时，我们可以设置一定阈值，对词向量间相似度超过阈值的词语归纳进统一集合中，并为每一个词设定集合标记，每一个词语可被不同集合收纳。在基础语句分析之后，根据词语标记找到其近义词集合，利用集合中的其他词对原句中的词语进行替换，然后判断替换后符合语法语义构成的句子，并按照一定的设定好的公式进行计算，最后选取最合适的集合，对所得结果进行二次校正。

可用热力图对词向量进行可视化。热力图，又名相关系数图。根据热力图中不同方块颜色对应的相关系数的大小，可以判断出变量之间相关性的大小。

Gensim库

Python的一个开源库，现发展成为世界上最大的NLP/信息检索Python库之一，兼具内存高效性和可扩展性。

Gensim的可扩展性体现为，它采用了Python内置的生成器和迭代器进行流式数据处理，所以数据集事实上并未完全加载到内存中。大部分信息检索算法都涉及矩阵分解或矩阵相乘运算。在Gensim中，这部分功能由numpy实现，而numpy底层是由FORTRAN/C实现的，并针对数学运算操作进行深度优化。由于所有繁重的工作都交由底层BLAS库完成，因此Gensim兼顾了C语言的执行效率以及Python的易用性。

六、项目进度安排

七、项目研究所需资源

八、项目经费预算与用途

九、项目完成预期成果