第十六章 NLP

1 什么是自然语言处理?

自然语言处理是人工智能和语言学领域的分支学科。此领域探讨如何处理及运用自然语言；自然语言处理包括多方面和步骤，基本有认知、理解、生成等部分。

其中基本的处理步骤有文本挖掘、分词标注、词性标注、实体命名识别以及各种分支下的处理等。**而常说的词向量则是将词语相对化地化成数值概率形式(你可以看作低维度的词语化作高维度的文本内容空间)，从而利用各种数学方式进行处理的过程。**

2 什么是文本挖掘?

文本挖掘有时也被称为文字探勘、文本数据挖掘等，大致相当于文字分析，一般指文本处理过程中产生高质量的信息，而后提供给后续的进程来利用这些高质量的信息的步骤及行为。高质量的信息通常通过分类和预测来产生，如模式识别。文本挖掘通常涉及输入文本的处理过程（通常进行分析，同时加上一些衍生语言特征以及消除杂音，随后插入到数据库中），产生结构化数据，并最终评价和解释输出。

'高品质'的文本挖掘通常是指某种组合的相关性，新颖性和趣味性。典型的文本挖掘方法包括文本分类，文本聚类，概念/实体挖掘，生产精确分类，观点分析，文档摘要和实体关系模型（即，学习已命名实体之间的关系）。文本分析包括了信息检索、词典分析来研究词语的频数分布、模式识别、标签\注释、信息抽取，数据挖掘技术包括链接和关联分析、可视化和预测分析。本质上，首要的任务是，通过自然语言处理（NLP）和分析方法，将文本转化为数据进行分析。

常见的后续处理有使用词袋模型进行特征抽取，并且使用类似于 Word2Vec 之类的词嵌入方式进行特征分析抽取。

3 常见的中文分词有哪些?

中科院计算所NLPIR http://ictclas.nlpir.org/nlpir/

ansj分词器 https://github.com/NLPchina/ansj_seg

哈工大的LTP https://github.com/HIT-SCIR/ltp

清华大学THULAC https://github.com/thunlp/THULAC

斯坦福分词器 https://nlp.stanford.edu/software/segmenter.shtml

Hanlp分词器 https://github.com/hankcs/HanLP

结巴分词 https://github.com/yanyiwu/cppjieba

KCWS分词器(字嵌入+Bi-LSTM+CRF) https://github.com/koth/kcws

ZPar https://github.com/frcchang/zpar/releases

IKAnalyzer https://github.com/wks/ik-analyzer

哈工大的分词器：主页上给过调用接口，每秒请求的次数有限制。

清华大学THULAC：目前已经有 Java 、 Python 和 C++ 版本，并且代码开源。

斯坦福分词器：作为众多斯坦福自然语言处理中的一个包，目前最新版本3.9.2， Java 实现的CRF算法。可以直接使用训练好的模型，也提供训练模型接口。

Hanlp分词：求解的是最短路径。优点：开源、有人维护、可以解答。原始模型用的训练语料是人民日报的语料，当然如果你有足够的语料也可以自己训练。

结巴分词工具：基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图 (DAG)；采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

字嵌入+Bi-LSTM+CRF分词器：本质上是序列标注

ZPar分词器：新加坡科技设计大学开发的中文分词器，包括分词、词性标注和Parser，支持多语言。但是已经许久没有更新过，所以无法处理。

4 为什么需要中文分词?

中文分词是中文文本处理的一个基础步骤，也是中文人机自然语言交互的基础模块。不同于英文的是，中文句子中没有词的界限，因此在进行中文自然语言处理时，通常需要先进行分词，分词效果将直接影响词性、句法树等模块的效果。当然分词只是一个工具，场景不同，要求也不同。

在人机自然语言交互中，成熟的中文分词算法能够达到更好的自然语言处理效果，帮助计算机理解复杂的中文语言。竹间智能在构建中文自然语言对话系统时，结合语言学不断优化，训练出了一套具有较好分词效果的算法模型，为机器更好地理解中文自然语言奠定了基础。

更加通俗地来讲，单字很多时候没办法表达语义信息，但是词可以表达。分词相当于预处理的过程。能够使得后面涉及到语义有关的分析的时候，更加准确。

5 词性标注是什么意思？为什么需要它，句法标注呢?

词性标注（part-of-speech tagging）,又称为词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或者其他词性的过程。

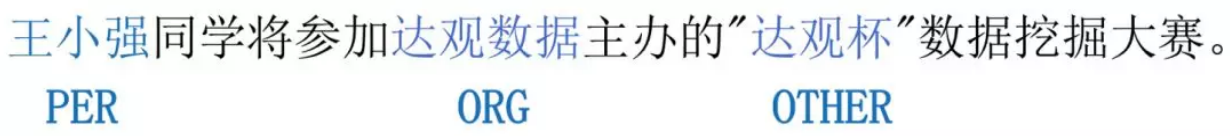
词性标注是很多NLP任务的预处理步骤，如句法分析，经过词性标注后的文本会带来很大的便利性，但也不是不可或缺的步骤。

例如，在处理一些类似于文本分类的任务的时候，词性标注就显得并没有那么需要了

句法标注，一般被称为依存句法分析。一般是对语料文本进行句法分析和标注，形成树库（tree bank）语料。对后续的语料分析起到很好应用。

6 命名实体识别是什么？有哪些主流的算法?

命名实体识别**(NER)**又称作专名识别，是自然语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。**NER**系统就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，比如产品名称、型号、价格等。因此实体这个概念可以很广，只要是业务需要的特殊文本片段都可以称为实体。



学术上NER所涉及的命名实体一般包括3大类（实体类，时间类，数字类）和7小类（人名、地名、组织机构名、时间、日期、货币、百分比）。

实际应用中，NER模型通常只要识别出人名、地名、组织机构名、日期时间即可，一些系统还会给出专有名词结果（比如缩写、会议名、产品名等）。货币、百分比等数字类实体可通过正则搞定。另外，在一些应用场景下会给出特定领域内的实体，如书名、歌曲名、期刊名等。

NER是NLP中一项基础性关键任务。**从自然语言处理的流程来看，NER可以看作词法分析中未登录词识别的一种，是未登录词中数量最多、识别难度最大、对分词效果影响最大问题。**同时NER也是关系抽取、事件抽取、知识图谱、机器翻译、问答系统等诸多NLP任务的基础。

常见的算法有：？

7 语言模型有什么？它有哪些领域?

简单地说，语言模型就是用来计算一个句子的概率的模型，也就是判断一句话是否是人话的概率?

语言模型是假设一门语言所有可能的句子服从一个概率分布，每个句子出现的概率加起来是1，那么语言模型的任务就是预测每个句子在语言中出现的概率，对于语言中常见的句子，一个好的语言模型应该得到相对高的概率，对不合语法的句子，计算出的概率则趋近于零。如果把句子 equation.svg 看成单词的序列 equation_1.svg ，那么语言模型可以表示为一个计算 equation_2.svg 的模型。语言模型仅仅对句子出现的概率进行建模，并不尝试去理解句子的内容含义。

之前特别火热的 BERT 模型也是一种语言模型。包括但不仅限于 ELMo、GPT等，都是同样的语言模型。

那么如何计算一个句子的概率呢？给定句子（词语序列）

$$S = W_1, W_2, \dots, W_k$$

它的概率可以表示为：

$$P(S) = P(W_1, W_2, \dots, W_k) = p(W_1)P(W_2|W_1) \dots P(W_k|W_1, W_2, \dots, W_{k-1})$$

语言模型的应用——主要用于语音识别和资料压缩的领域当中，以及用于信息搜索等。

8 概率图模型的概述?

通过计算条件概率来计算模型参数，因此也叫概率模型。而如果我们在概率模型的基础上，使用基于图的方法来表示概率分布，那么我们称这种模型为概率图模型。

在概率图模型的表达中，结点表示变量，结点之间直接相连的边表示相应变量之间的概率关系。如果结点之间没有边，我们就认为这两个变量是没有概率关系的，即这两个变量的出现都是独立的。

根据图模型的边是否有向，概率图模型通常被划分为有向概率图模型和无向概率图模型。

9 马尔科夫过程的定义是什么？它更常用于哪些方面?

是一类随机过程。 它的原始模型**马尔可夫**链，由俄国数学家A.A.**马尔可夫**于1907年提出。 **马尔可夫过程是**研究离散事件动态系统状态空间的重要方法，它的数学基础**是随机过程**理论。

常用于?

10 文本分类的核心是什么？它有哪些对应的方法?

文本分类（Text Classification）在NLP领域里是一个很普通而应用很广的课题，指计算机将一篇文章归于预先给定的某一类或某几类的过程。主要的应用领域为网页分类、微博情感分析、用户评论挖掘、信息检索、Web文档自动分类、数字图书馆、自动文摘、分类新闻组、文本过滤、单词语义辨析以及文档的组织和管理等。

目前，文本分类已经有了相当多的研究成果，比如应用很广泛的基于规则特征的SVM分类器，以及加上朴素贝叶斯方法的SVM分类器，当然还有最大熵分类器、基于条件随机场来构建依赖树的分类方法。在传统的文本分类词袋模型中，在将文本转换成文本向量的过程中，往往会造成文本向量维度过大的问题，当然也有其他的压缩了维度的一些分类方法。还有一些是基于人工的提取规则，甚至是hard coding方式。这样不利于算法的推广。近些年随着深度神经网络（Deep Neural Network, DNN）的兴起，人们开始尝试用DNN解决文本分类的问题。

常用的方法有：Fasttext、TextCNN、TextRNN、RCNN、Char CNN、Char RNN、HAN、Dynamic Memory Network、Entity Network.

11 文本分类的分类器结构有什么?

使用朴素贝叶斯算法、向量空间距离测度分类算法、K最近邻分类算法、支持向量机、神经网络算法模型、决策树分类算法模型、Bagging 算法、Boosting 算法模型、基于TFIDF的Rocchio算法等。其中效果最好的是bilstm。

12 贝叶斯和SVM，哪个对于文本分类有更好的效果?

朴素贝叶斯(native bayes)文本分类模型是一种简单而高效的文本分类模型,但是它的属性独立性假设使其无法表示现实世界属性之间的依赖关系,影响其分类性能。SVM 只适用于有限文本，因为文本过多以后，在实际的应用中会导致内存炸裂。但是它的优点是能够解决在神经网络方法中无法避免的局部极值问题。而且它能够同时适用于稠密特征矢量与稀疏特征矢量两种情况。

13 信息检索是什么？它与知识图谱有什么相似的地方?

信息检索也称情报检索，就是利用计算机系统从海量文档中找到符合用户需要的相关文档。面向两种或两种以上语言的信息检索叫做跨语言信息检索（cross-language/trans-lingual information retrieval）。

两者是可以存在可以融合的地方，知识图谱是结合了实体识别关系识别以及知识计算、知识推理等基础技术的一种高级技术。而信息检索更加面向于信息本身。

14 自动文摘的意义和常用方法是？为什么需要自动文摘?

网络上巨大的信息量使得信息检索的难度加大，而信息摘要对于信息的发布者，使用者以及搜索引擎都有着重要的作用。如果能提供给用户简短的文本摘要则可以帮助用户快速地找到所需要的信息，提供给搜索引擎则可以提高检索速度。

常用方法有针对抽取式文摘的基于单一因素的摘要方法、基于启发式规则、基于图排序方法、基于整数线性规划(ILP)方法、基于神经网络方法以及基于次模函数的方法的。

还有就是生成式文摘，也就是基于形式化语义表示，基于短语选择与拼凑，还有就是基于深度学习的序列转换模型等。

主要是引用于信息检索。

15 问答系统的基本组成是怎样的？目前已经达到什么层次了?

传统的问答系统根据以下的流程进行工作：（1）问题解析（2）信息检索（3）答案抽取。

其中问题解析自然包括分词、词性标注、句法分析、命名实体识别、问题分类、问题拓展等。

而信息检索则是以问题解析的结果作为输入，并且从底层知识库中返回一系列相关的排序后的文档。

而答案抽取，则是利用了部分自动文摘的思想，从文档中抽取出自己想要最终的答案。

目前最为出色的阅读理解系统是微软亚洲研究院提供的 nlnet，目前在 SQuAD 2.0 排名第一。而如果是SQuAD 1.0版本，则是由Google 的 BERT (其实是一种预训练语言模型)引领风骚。已经远远超过人类的表现。

常见的模型为，输入层，处理层，验证层，输出层。一般已经有双模型。一个模型处理题目与答案内容，一个模型处理来自文章的信息。然后耦合为一个模型。

15.1 关键词抽取是什么?

如字面的意思，关键词抽取的意思就是抽取当前文档的关键词，或者是海量文档中的所有的关键词。而目前比较流行的是无监督关键词抽取。因为有监督的方法一般都会带来高昂的人工成本。但是如果为了提高效率的话，也会考虑采用半监督的关键词抽取模型。

常用的方法是有基于统计的 TF-IDF 抽取、基于词图模型的方法，还有就是基于主题模型的关键词抽取。

16 机器翻译常见模型结构是怎样的？为什么它能够起效?

最常见的模型是 Seq2seq 模型。

模型主要由编码器及解码器构成。而编码器和解码器一般都是由RNN类网络构成，常用 LSTM。主要是因为使用了 RNN 能够自适应输入和输出。当然你选择任何类型的网络都没关系。这只是一个范式。

你也可以考虑使用对偶学习的范式，这是一个小样本数据量都能够发挥很大作用的模型范式。

17 词向量是怎么回事？为什么它能够起效?

NLP 里面，最细粒度的是词语，词语组成句子，句子再组成段落、篇章、文档。所以处理 NLP 的问题，首先就要拿词语开刀。

举个简单例子，判断一个词的词性，是动词还是名词。用机器学习的思路，我们有一系列样本(x,y)，这里 x 是词语，y 是它们的词性，我们要构建 f(x)->y 的映射，但这里的数学模型 f（比如神经网络、SVM）只接受数值型输入，而 NLP 里的词语，是人类的抽象总结，是符号形式的（比如中文、英文、拉丁文等等），所以需要把他们转换成数值形式，或者说——嵌入到一个数学空间里，这种嵌入方式，就叫词嵌入（word embedding)，而 Word2vec，就是词嵌入（ word embedding) 的一种

18 BERT是什么？和之前的ELMo有什么区别?

使用某种模型预训练一个语言模型看起来是一种比较靠谱的方法。从之前AI2的 ELMo，到 OpenAI的 fine-tune transformer，再到Google的这个BERT，全都是对预训练的语言模型的应用。

BERT这个模型与其它两个不同的是，它在训练双向语言模型时以减小的概率把少量的词替成了Mask或者另一个随机的词。

而这个操作是为了提高模型自身的泛化能力。

而 ElMo 也是类似的内容，但是在层数上，以及Transformer上有更加优越的地方。实际上，ELMo 所需要消耗的资源远小于BERT所需要的资源。

19 想要提高文本的泛化能力，有什么办法?

你可以考虑使用BERT、ELMo、GPT等可以提高泛化能力的范式

20 目前 NLP 的学习范式有哪些？它们都是怎么运作的?

目前常见的学习方式还是集中在解码器与编码器这种方式上的各种变形。

BERT 也可以称作新的范式。以及前文提到的Dual learning 也是一种范式。还有隔壁的 CV 的 GAN 也可以说是一种范式。

21 如何去除停用词？为什么要去除停用词？哪些停用词表比较好用?

使用停用词表即可去除停用词。使数据清洗更为干净，停用词会对模型的结果产生影响。常见的停用词表有哈工大维护的，百度维护的。这些可以到 GItHub 上进行查询。

22 文本的特征工程怎么做?

1. 你可以考虑使用词袋模型(Bag of Word)和TF-IDF 模型、还有N元 词袋模型(Bag of N-Gram Model)等。
2. 还是用上 Word Embedding 也就是上述的 word2vec。
3. 你可以考虑以下这种特殊的特征方法：

   1、26个字母大小写数目的统计，52维的特征；

   2、文本中出现的城市名称，这也能构成一类特征；

   3、统计文本数据中出现次数最高的30个词语，给出每个句子中这些词语出现次数，这就又多了个30维的特征；

   4、统计常用的那些标点符号的出现次数；

23 注意力机制为什么会起效?

概括地说，在神经网络实现预测任务时，引入注意力机制能使训练重点集中在输入数据的相关部分，而不是无关部分。

注意力是指人的心理活动指向和集中于某种事物的能力。比如说，你将很长的一句话人工从一种语言翻译到另一种语言，在任何时候，你最关注的都是当时正在翻译的词或短语，与它在句子中的位置无关。在神经网络中引入注意力机制，就让它也学会了人类这种做法。

注意力机制最经常被用于序列转换（Seq-to-Seq）模型中。如果不引入注意力机制，模型只能以单个隐藏状态单元，如下图中的S，去捕获整个输入序列的本质信息。这种方法在实际应用中效果很差，而且输入序列越长，这个问题就越糟糕。

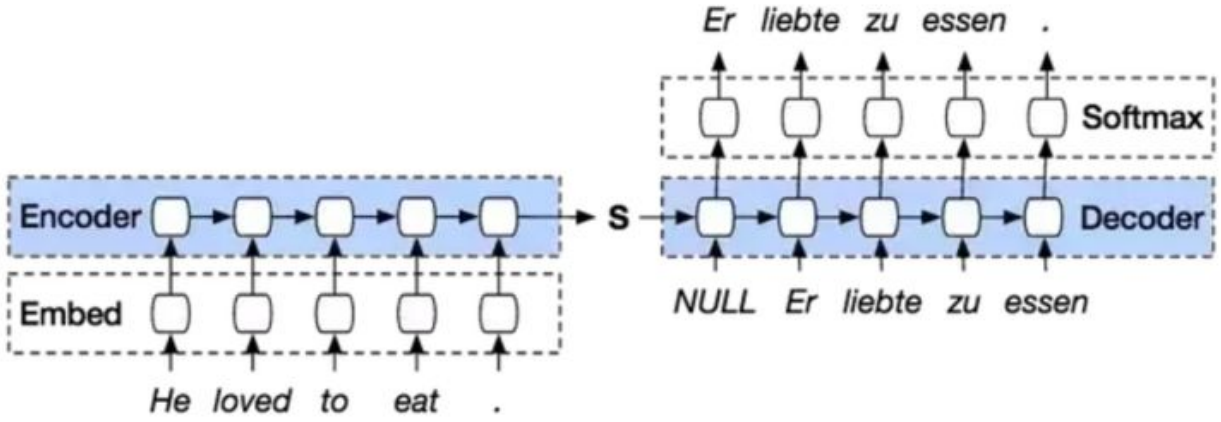


图1：仅用单个S单元连接的序列转换模型

注意力机制在解码器（Decoder）运行的每个阶段中，通过回顾输入序列，来增强该模型效果。解码器的输出不仅取决于解码器最终的状态单元，还取决于所有输入状态的加权组合。

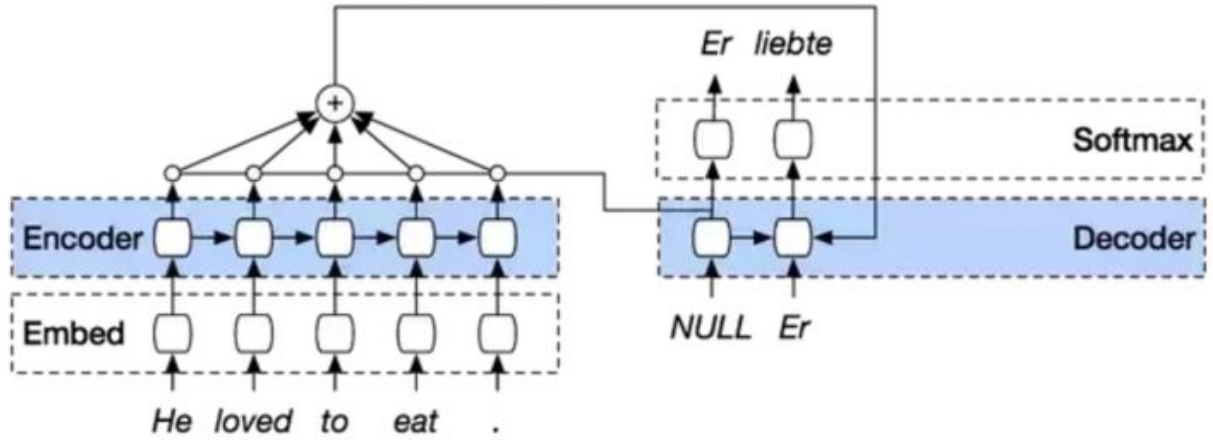


图2：引入注意力机制的序列转换模型

注意力机制的引入增加了网络结构的复杂性，其作为标准训练模型时的一部分，通过反向传播进行学习。这在网络中添加模块就能实现，不需要定义函数等操作。

24 自注意力机制与外部注意力机制的区别?

自注意力机制可以仅通过自身的信息来更新学习参数。并且进行对齐。

你可以观察两者公式的区别：

自注意力机制：

$$\mathbf{a} = softmax(\mathbf{w_{s2}} tanh(W_{s1} H^T))$$

外部注意力机制：

Global：

$$a_i = softmax(v_a^T tanh(W_a[\mathbf{h_i}; \overline{\mathbf{h}}_\mathbf{t}]))$$

Local：

$$a_i = softmax(\mathbf{h_i^T} W_a \overline{\mathbf{h}}_\mathbf{t})$$

这两者都是在 NLP 使用率比较高的 soft-attention。而 Hard-attention 更常用于图像领域。

而自注意力机制则是不需要引入 $\overline{\mathbf{h}}_\mathbf{t}$ 也就是外部信息的。

25 目前常见的应用方向有哪些?