一起走进自然语言处理的世界

原创：Evan AI遇见机器学习 3月11日

自然语言处理简介

自然语言处理（Natural Language Processing，简称NLP）就是用计算机来处理、理解以及运用人类语言(如中文、英文等)，它属于人工智能的一个分支，是计算机科学与语言学的交叉学科，又常被称为计算语言学。由于自然语言是人类区别于其他动物的根本标志。

没有语言，人类的思维也就无从谈起，所以自然语言处理体现了人工智能的最高任务与境界，也就是说，只有当计算机具备了处理自然语言的能力时，机器才算实现了真正的智能。

语言有五个特性：
1. 规律性和例外：比如我们的语法，比如中文中的构词来源，是有规律的，但也有很多其他的例外，比如特殊搭配。

2. 组合性：不同的句子或词或字，组合的方式不同，就具有不同的意义。

3. 递归性：比如汉语所谓的字本位现象，这个特性造成了语言非常复杂，

4. 比喻性：语言的本质就是产生新的语言进行表示，其实都是在做比喻。所以，比喻性是语言非常重要的特性。语言的理解跟世界知识是密切相关的，如果你撇开了知识这些东西谈语言，其实都是无从谈起的。

5. 交互性：好理解，就表示的语言之间的交互。

这些特性使得我们把语言放在计算机上，变得非常具有挑战性，但是同时研究好这些特性，就会让我们的机器更加的能理解语言。

一、自然语言方向简介

自然语言处理（简称NLP），是研究计算机处理人类语言的一门技术，包括：
1. 句法语义分析：对于给定的句子，进行分词、词性标记、命名实体识别和链接、句法分析、语义角色识别和多义词消歧。

2. 信息抽取：从给定文本中抽取重要的信息，比如，时间、地点、人物、事件、原因、结果、数字、日期、货币、专有名词等等。通俗说来，就是要了解谁在什么时候、什么原因、对谁、做了什么事、有什么结果。涉及到实体识别、时间抽取、因果关系抽取等关键技术。

3. 文本挖掘（或者文本数据挖掘）：包括文本聚类、分类、信息抽取、摘要、情感分析以及对挖掘的信息和知识的可视化、交互式的表达界面。目前主流的技术都是基于统计机器学习的。
信息检索：对大规模的文档进行索引。可简单对文档中的词汇，赋之以不同的权重来建立索引，也可利用1，2，3的技术来建立更加深层的索引。在查询的时候，对输入的查询表达式比如一个检索词或者一个句子进行分析，然后在索引里面查找匹配的候选文档，再根据一个排序机制把候选文档排序，最后输出排序得分最高的文档。

4. 机器翻译：把输入的源语言文本通过自动翻译获得另外一种语言的文本。根据输入媒介不同，可以细分为文本翻译、语音翻译、手语翻译、图形翻译等。机器翻译从最早的基于规则的方法到二十年前的基于统计的方法，再到今天的基于神经网络（编码-解码）的方法，逐渐形成了一套比较严谨的方法体系。

5. 问答系统： 对一个自然语言表达的问题，由问答系统给出一个精准的答案。需要对自然语言查询语句进行某种程度的语义分析，包括实体链接、关系识别，形成逻辑表达式，然后到知识库中查找可能的候选答案并通过一个排序机制找出最佳的答案。当然，现在还有VQA这种和图像结合的问答，也很有趣的。

6. 对话系统：系统通过一系列的对话，跟用户进行聊天、回答、完成某一项任务。涉及到用户意图理解、通用聊天引擎、问答引擎、对话管理等技术。此外，为了体现上下文相关，要具备多轮对话能力。同时，为了体现个性化，要开发用户画像以及基于用户画像的个性化回复。

以上讲解主要来自微软亚洲研究院周明博士的总结。

二、自然语言处理主要技术

自然语言处理大概有五类技术，分别是：

1. 分类：文字的序列，我们要打印标签，这是我们常做的最基本的自然语言处理。

2. 匹配：两个文字序列都匹配，看它们匹配的程度，最后输出一个非负的实数值，判断这两个文字序列它们的匹配程度。

3. 翻译：把一个文字序列，转换成另外一个文字序列。

4. 结构预测：你给我一个文字序列，让它形成内部结构的一个信息。

5. 序列决策过程：在一个复杂的动态变化环境里面，我们怎么样不断去决策。比如描述序列决策过程的马尔可夫随机过程，这是一个有效的、非常常用的数学工具。

我们看自然语言处理的大部分问题，基本上做得比较成功、实用的都是基于这样的技术做出来的。比如：分类，有文本分类、情感分析；匹配，有搜索、问答、单轮对话、基于检索的单轮对话；翻译，有机器翻译、语音识别、手写体识别、基于生成方法的单轮对话；结构预测，有专名识别、词性标注、语意分析；序列决策过程，有多轮对话。

三、自然语言处理资料推荐

1. 关于书籍：《数学之美》--吴军，科普且生动形象，入门必备；《统计学习方法》--李航，这个讲述基础机器学习算法，这是值得看的；《统计自然语言处理》--宗成庆，经典好书，可以详细看。

2. 关于综述：自然语言处理综述（https://arxiv.org/abs/1708.05148），这个综述主要是深度学习在NLP的应用和发展，值得一看的；关于自然语言生成的综述（https://arxiv.org/abs/1703.09902v1），讲述自然语言生层的各种方式和应用。

3. 关于教程：Stanford nlp公开课-cs224n，这是一个关于自然语言的公开课，对于入门来说还是非常推荐的。

关于其他资料：A Survey on Dialogue Systems（https://arxiv.org/abs/1711.01731），这是一个关于对话系统的综述，对这个方向感兴趣的小伙伴可以多看下。

多看论文，做实验，多看论文，做实验，写论文.....

四、总结及未来方向

及未来方

关于用深度学习处理自然语言处理：近些年来深度学习的火爆，很多学者把深度学习应用于自然语言处理。但是除了翻译我知道比较大的进展，其他很多方向解决的效果还很不理想，这个需要我们去努力，需要看清深度学习的局限，也需要挖掘深度学习的无限可能。

自然语言处理的终极目标是机器能够做自然语言处理，理解人类的语言。具体来说有两个方面：像人一样能够去说话；像人一样能去阅读。那很多时候，我们在做学问就应该拿这些目标作为导向，做实验，先要思考怎么才能先达到像人类那样处理语言。

看好的方向：我认为联结主义和符号主义的结合的方式，是实现机器理解和处理人类语言的更好方式，一下子说可能比较抽象，大家可以看下两个链接：

专访深度好奇创始人吕正东：通向理解之路（http://www.sohu.com/a/153274496_465975），讲述联结主义和符号主义结合的三种方式-- 表示层面，操作层面和知识层面，论文推荐看下。大家也可以看看吕正东老师和李航老师的最近的文章，很多论文也就是表达的这个观点。

AI科学突破的前夜，教授们应当看到什么？（https://zhuanlan.zhihu.com/p/32852022）这篇中讲述的观点更加的明确，AI和突破，自然语言处理的突破，应该来自神经网络和符号处理的结合！

欢迎关注我们，看通俗干货！