人工智能中的语义分析技术及其应用

文/神州泰岳

中国国民经济和社会发展第十三个五年规划纲要指出，实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。

一、人工智能语义分析技术

语义分析（Semantic Analysis）是人工智能（Artificial Intelligence）的一个分支，是自然语言处理技术的几个核心任务，涉及语言学、计算语言学、机器学习，以及认知语言等多个学科，语义分析任务有助于促进其他自然语言处理任务的快速发展。人工智能中的语义分析技术，特别是深度学习（Deep Learning）技术近年来发展迅猛，已经在围棋对弈、自动驾驶、图像识别、语音识别等多个领域取得了突破性进展。

语义分析指运用各种方法，学习与理解一段文本所表示的语义内容，任何对语言的理解都可以归为语义分析的范畴。一段文本通常由词、句子和段落来构成，根据理解对象的语言单位不同，语义分析又可进一步分解为词汇级语义分析、句子级语义分析以及篇章级语义分析。一般来说，词汇级语义分析关注的是如何获取或区别单词的语义，句子级语义分析则试图分析整个句子所表达的语义，而篇章语义分析旨在研究自然语言文本的内在结构并理解文本单元（可以是句子从句或段落）间的语义关系。简单地讲，语义分析的目标就是通过建立有效的模型和系统，实现在各个语言单位（包括词汇、句子和篇章等）的自动语义分析，从而实现理解整个文本表达的真实语义。

二、语义分析技术

（一）基础技术（按照词语分析、句子分析、篇章分析来写）

分别从词汇级、句子级和篇章级三个层次描述语义分析相关技术。

1.词语级语义分析

词汇层面上的语义分析主要体现在如何理解某个词汇的含义，主要包含两个方面：词义消歧和词义表示

（1）词义消歧

词汇的歧义性是自然语言的固有特征。词义消歧根据一个多义词在文本中出现的上下文环境来确定其词义，作为各项自然语言处理的基础步骤和必经阶段被提出来。词义消歧包含两个必要的步骤：（a）在词典中描述词语的意义；（b）在语料中进行词义自动消歧。例如“苹果”在词典中描述有两个不同的意义：一种常见的水果；美国一家科技公司。对于下面两个句子：

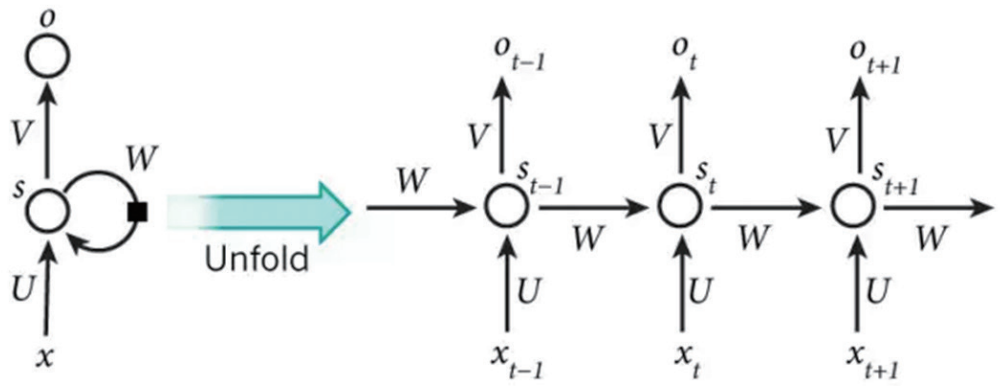
她的脸红得像苹果。

最近几个月苹果营收出现下滑。

词义消歧的任务是自动将第一个苹果归为“水果”，而将第二个苹果归为“公司”。从上面的例子中我们发现，词义消歧主要面临如下两个关键问题：（a）词典的构建；（b）上下文的建

北京神州泰岳软件股份有限公司

图：循环神经网络结构



模。

(2) 词义表示和学习

对于词义表示，早期的做法将某个词义表示为，从该词义在同义词网络中出现的位置到该网络根节点之间的路径信息。词义表示的另一个思路是将其数字化。最直观,也是到目前为止最常用的词表示方法是one-hot表示方法，这种方法把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为0，只有一个维度的值为1，这个维度就代表了当前的词。不难想象，这种表示方法存在一个重要的问题：任意两个词之间都是孤立的。造成的结果是：光从两个向量中看不出两个词是否有关系，即使这两个词是同义词，例如“计算机”和“电脑”、“上海”和“上海市”。

随着机器学习算法的发展，目前更流行的词义表示方式是词嵌入(Word Embedding,又称词向量)。其基本想法是：通过训练将某种语言中的每一个词映射成一个固定维数的向量，将所有这些向量放在一起形成一个词向量空间，而每一向量则可视为该空间中的一个点，在这个空间上引入“距离”，则可以根据词之间的距离来判断它们之间的(词法、语义上的)相似性。

2.句子级语义分析

句子级的语义分析试图根据句子的句法结构和句中词的词义等信息，推导出能够反映这个句子意义的某种形式化表示。根据句子级语义分析的深浅，又可以进一步划分为浅层语义分析和深层语义分析。

(1) 浅层语义分析

语义角色标注(Semantic Role Labeling，简称 SRL)是一种浅层的语义分析。给定一个句子，SRL的任务是找出句子中谓词的相应语义角色成分，包括核心语义角色(如施事者、受事者等)和附属语义角色(如地点、时间、方式、原因等)。

目前SRL的实现通常都是基于句法分析结果，即对于某个给定的句子，首先得到其句法分析结果，然后基于该句法分析结果，再实现SRL。

(2) 深层语义分析

深层的语义分析(有时直接称为语义分析,Semantic Parsing)不再以谓词为中心，而是将整个句子转化为某种形式化表示，例如：谓词逻辑表达式(包括lambda 演算表达式)、基于依存的组合式语义表达式(dependency-based compositional semantic

representation）等。以下给出了GeoQuery数据集中的一个中英文句子对，以及对应的一阶谓词逻辑语义表达式：

中文：列出在科罗拉多州所有的河流

英文：Name all the rivers in Colorado

语义表达式：answer(river(loc_2(stateid('colorado'))))

虽然各种形式化表示方法采用的理论依据和表示方法不一样，但其组成通常包括关系谓词（如上例中的loc_2、river等）、实体（如colorado）等。语义分析通常需要知识库的支持，在该知识库中，预先定义了一序列的实体、属性以及实体之间的关系。

3.篇章级语义分析

篇章是指由一系列连续的子句、句子或语段构成的语言整体单位，在一个篇章中，子句、句子或语段间具有一定的层次结构和语义关系，篇章结构分析旨在分析出其中的层次结构和语义关系。具体来说，给定一段文本，其任务是自动识别出该文本中的所有篇章结构，其中每个篇章结构由连接词，两个相应的论元，以及篇章关系类别构成。篇章结构可进一步分为显式和隐式，显式篇章关系指连接词存在于文本中，而隐式篇章关系指连接词不存在于文本中，但可以根据上下文语境推导出合适的连接词。对于显式篇章关系类别，连接词为判断篇章关系类别提供了重要依据，关系识别准确率较高；但对于隐式篇章关系，由于连接词未知，关系类别判定较为困难，也是篇章分析中的一个重要研究内容和难点。

（二）深度学习技术（深度学习在NLP中的研究内容）

在深度学习技术中，循环神经网络（Recurrent Neural Networks，RNNs）被证明在自然语言处理中是最有效的，下面将介绍循环神经网络。

RNNs的目的是使用序列来处理数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。RNNs之所以称为循环神经网路，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。理论上，RNNs能够对任何长度的序列数据进行处理。但是在实践中，为了降低复杂性往往假设当前的状态只与前面的几个状态相关。

RNNs已经在实践中被证明对NLP是非常成功的。如词向量表达、语句合法性检查、词性标注等。在RNNs中，目前使用最广泛、最成功的模型便是LSTMs(Long Short-Term Memory，长短时记忆模型)模型，该模型通常比vanilla RNNs能够更好地对长短时记忆模型依赖进行表达，该模型相对于一般的RNNs，只是在隐藏层做了手脚。

RNNs可以应用于语言模型与文本生成、文本分类、机器翻译等自然语言处理任务中。

三、面向业务建模的语义分析（介绍DINFO-OEC平台和技术）

（一）DINFO-OEC平台介绍

DINFO-OEC非结构化大数据分析挖掘平台，是中科鼎富（北京）科技发展有限公司研发的大数据产品，具有非结构化文本大数据的分析、

挖掘的超凡能力，是企业实施大数据战略的强大利器。

大数据中80%都是非结构化大数据，非结构化大数据因其中的业务对象、对象之间的关系等都蕴含在文本内容中，而文本内容来源繁多、表达方式灵活多样、存在着大量的歧义性，因此无法使用传统的BI工具等进行分析，无法直接服务于业务，实现业务价值。非结构化大数据是大数据处理的难点和热点。DINFO-OEC平台支持三位一体的多维度业务建模能力，结合自然语言处理、深度学习等统计文本挖掘算法，基于平台立体式的业务模型的智能语义感知技术，提供对非结构化大数据智能理解与自动化处理能力，实现文本知识的多维度的业务标签标记功能，将无序的非结构化信息转换为满足业务需求的结构化数据。DINFO-OEC平台支持与主流Hadoop、Spark等大数据平台实现对接，利用hadoop平台提供的分布式存储和Map/Reduce分布式计算能力，实现复杂、批量的大数据分析挖掘。利用Spark、kafka等提供的实时分布式计算能力，提供海量数据的实时分析计算能力，融合主流的搜索引擎技术，支持基于海量历史数据的交互式搜索功能。DINFO-OEC平台支持与常用的商业智能系统进行融合，实现结构化数据和非结构化数据的融合分析挖掘，最大化的挖掘大数据的业务价值，提供大数据分析挖掘支持下的业务创新。

（二）业务建模

业务建模技术，采用神州泰岳独创的“本体O-要素E-概念C”三位一体的专利技术（发明专利号201410155830.1）进行建模，将业务和语言分为两个不同层次建模。业务建模以本体论为核心，对业务知识进行规划，对业务规则进行建模配置，形成形式化的业务规则。业务建模技术支持业务与非结构化数据的语言表达分离，区分业务层次和语言层次进行分部建模。业务层次支持业务本体构建，支持业务要素发现与配置；语言层次支持语言概念的构建与维护，支持常用词汇库和同义库等建设。DINFO-OEC业务建模价值在于客户只需关注自身业务的描述，文本表示的多样性和歧义性等由系统来负责解决。

（三）平台特点（参考白皮书）

1.超凡的面向业务的非结构化数据建模能力

INDO-OEC业务建模，能把纷繁复杂的业务规则和灵活多样的语言表达习惯进行统一建模，从本体、要素和概念三个维度构建分析挖掘模型，有效地将“业务”描述与自然语言的表达进行分离，使得业务人员可以专注于自己擅长的业务需求及业务规则的建模，而无需考虑自然语言的歧义性、表达的多样性和复杂性等。

2.强大的非结构化分析挖掘能力

产品支持智慧语义感知算法，提供强大的自然语言理解相关分析算法，包括内容分类、聚类、主题分析、语义分析、实体识别、启发式搜索引擎、推荐引擎、摘要引擎等。

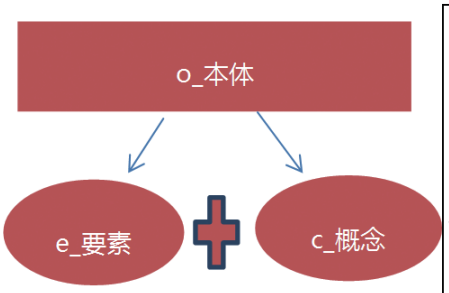
产品支持多种分析挖掘算法，包括C计算（提供概念的抽取、概念表达式挖掘、概念表达式匹配算法），S计算（提供常用的统计挖掘算法，包括但是不限于KNN、SVM、决策树等算法）和R计算（提供概念关联发现算法）。

3.丰富的多语种分析挖掘支持能力

系统内置了多语种分析挖掘算法。利用一套算法流程，实现多语种支持，语种扩展性好。新增加语种，不用修改算法。

多语种复用的能力。平台支持多语种业务规则保持一致的能力。业务规则（对应系统的本体树）的维护，只需维护中文简体版，无须维护其他语种的本体树，大大减少本体树维护工

图："本体O-要素E-概念C"三位一体的业务建模技术



1. 本体树 Ontology Tree
   - 业务分类，以及为每个类定制的挖掘策略
   - 挖掘策略采用"概念表达式"方式表示
2. 要素树 Element Tree
   - 业务相关的对象和属性
3. 概念树 Conception Tree
   - 表示时间、地点、值、人的情绪、态度等常用概念
   - 基础的语义资源，与业务无关

作量。

跨语种建模能力。平台支持用中文简体版，书写其他语种的本体树规则。修改、维护本体树类别，无须掌握其他语种。

4.卓越的大数据计算与存储平台集成能力

支持主流的Hadoop平台，支持Map/Reduce批量计算以及Spark实时计算，支持HDFS、Hbase、kafka等存取。支持的Hadoop平台包括Apache Hadoop、IBM BigInsights、华为FusionInsights、EMC Pivotal HD。

支持SOA集群架构，支持与Oracle、Mysql、DB2等主力数据库产品集成。

四、语义分析应用

（一）金融行业应用

人工智能的飞速发展，使得机器能够在很大程度上模拟人的功能，实现批量人性化和个性化地服务客户，这将给身处服务价值链高端的金融行业带来深刻影响，人工智能将成为决定银行沟通客户、发现客户金融需求的重要因素。它将对金融产品、服务渠道、服务方式、风险管理、授信融资、投资决策等带来新一轮的变革。人工智能技术在前端可以用于服务客户，在中台支持授信、各类金融交易和金融分析中的决策，在后台用于风险防控和监督，它将大幅改变金融现有格局，金融服务（银行、保险、理财、借贷、投资等方面）将更加地个性化与智能化。证券研报大数据云服务，是鼎富科技针对证券业、基金业研究人员、分析师推出的一款大数据云服务产品。系统提供SaaS服务，提供公告、研报的全网采集，以及事件结构化分析，提供研报一站式智能搜索，以及基于时间轴、基于信息锚点的大数据分析挖掘。系统能帮助分析师从大数据视角进行深度研究分析，提高工作效率。

（二）政府行业应用

舆情分析为政府、公安、社会等提供可自定义热点问题的舆情分析系统，信息出现的源头到产生的影响全程跟进分析，形成舆情影响波及范围、公众反响、不良舆论等内容的分析报告。

舆情分析能够大幅度缩短组织对互联网、论坛等电子信息渠道的公众舆论趋势的响应时间，通过关联分析能够帮助组织预测未来可能出现的状况并提前实施相关措施。

智慧传播云服务，是鼎富科技与腾讯网合

作推出的互联网信息监测预警平台，面向政府机构、企事业单位提供互联网信息监测、预警服务。舆情云项目的研发目的是为企业、政府、组织开发一款基于云服务的互联网舆情监测系统。该系统数据采集模块具有可配置、自动去重、垃圾过滤核心功能。系统分析挖掘功能采用智慧语义识别技术，保证了语义分析的准确性。系统可以按照客户需求进行舆情监测定制、统计报表定制和预警定制。

（三）客服行业应用

客服作为劳动密集型行业，对于一些大公司来说，成本依然很高。智能机器人客服的出现可以在很大程度上解决简单、重复性工作，帮助企业节省人工和坐席成本，提升运营效率。

小富机器人4.0是神州泰岳旗下一款智能客服机器人，它将开启全媒体时代的智能客服中心。小富机器人4.0有以下几个亮点：

亮点一：首创业务场景机器人

让机器人服从业务，而非业务屈从于机器人。客服、营销、外呼等业务，场景不同，业务逻辑也不同。小富4.0预设多种场景模式的业务框架，对应的知识类型和交互方式也有区分设计，可提供更专业、更具针对性的智能化服务。

亮点二：整体性业务建模，更具延展性

基于对业务的整体理解，而非Q&A的堆积。基于对具体场景的深刻业务理解，进行整体建模，具有完整的业务逻辑，机器人的思维延展性和可复用性大大增强，应答效率更高。

亮点三：差异化的知识类型表达体系

智能引导多轮会话，而非预设问题的反复跳转。小富4.0的业务知识体系化，并具有记忆能力，可基于业务逻辑自创造问答逻辑，智能地开展多轮引导式问答，让交互更自然、更具亲和力。

亮点四：智能碎片化知识加工

直接告知答案，而非仅告知答案所在的范围。小富4.0提供丰富的知识加工模式，可智能化地将结构化和非结构化的知识，碎片化为结构化的文档。应答客户提问时，可直接回馈用户的问题，而非给出一个答案所在区间。

此外，与小富机器人4.0同时展示的还有泰岳统一业务知识库系统，可提供知识自动加工和强大的知识图谱关联能力；泰岳客服大数据分析挖掘解决方案，可支持多层级业务类别自动分类和语义处理，为客户提供更智能、更高效的人工智能新体验。

五、语义分析及大数据发展趋势

人工智能技术及大数据已经成为新经济发展的动力，美国、欧洲、日本、中国等多个国家和地区均将大数据及人工智能作为国家战略。中国国民经济和社会发展第十三个五年规划纲要指出，实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。同时，2016年，国家发改委、科技部、工信部、中央网信办联合发布了《“互联网＋”人工智能三年行动实施方案》，首次单独为人工智能发展提出具体的策略方案，提出了人工智能发展的九大工程。2016美国白宫发布了《为人工智能的未来做好准备》(Preparing for the Future of Artificial Intelligence)和《国家人工智能研究与发展战略计划》(National Artificial Intelligence Research and Development Strategic Plan)两份重要报告。探讨了人工智能的发展现状、应用领域以及潜在的公共政策问题，提出了美国优先发展的人工智能七大战略方向及两方面建议，对我国人工智能产业发展具有重要的借鉴意义。

责任编辑：郭嘉凯
guojk@softic.com.cn