《智能信息处理》课程考试

基于知识图谱的自然语言处理研究分析

陈梓怡

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 19 日

基于知识图谱的自然语言处理研究分析

陈梓怡

(大连海事大学 信息科学技术学院，辽宁 大连 116026)

摘　要　本文将从多角度对中外自然语言处理的发展进行对比分析。对多篇来自 CNKI、Web of Science 与自然语言处理相关的重大国际会议文献, 采用词频统计法、共现分析法相结合的方法, 利用知识图谱呈现统计结果。统计结果表明, 中外对自然语言处理的研究表现出极大的相似性, 研究内容都集中在信息抽取、人工智能、信息检索、机器翻译、机器学习等领域。但检索主题词的 选取、数据清洗时的主观性给研究带来误差。并对国内自然语言处理的发展提出建议。

关键词　自然语言处理 知识图谱 信息检索 机器学习；

中图法分类号　9725　　DOI 号 24861479086539-753

Research and analysis of Natural Language Processing based on knowledge graph

Chen Ziyi

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026，China)

Abstract　This paper will make a comparative analysis of the development of natural language processing at home and abroad from various perspectives. This paper presents statistical results of several papers from CNKI and Web of Science related to natural language processing by using the method of word frequency statistics and co-occurrence analysis. The statistical results show that there are great similarities between Chinese and foreign researches on natural language processing, and the research contents are all concentrated in the fields of information extraction, artificial intelligence, information retrieval, machine translation and machine learning. However, the selection of keywords and subjectivity in data cleaning bring errors to the research. Some suggestions on the development of natural language processing in China are given.

Key words　Natural Language processing; knowledge graph ;information retrieval ;machine learning

0　绪论

知识图谱，是结构化的语义知识库，用于迅速描述物理世界中的概念及其相互关系，通过将数据粒度从 document 级别降到 data 级别，聚合大量知识，从而实现知识的快速响应和推理。 当下知识图谱已在工业领域得到了广泛应用，如搜索领域的 Google 搜索、百度搜索，社交领域的领英经济图谱，企业信息领域的天眼查企业图谱等。交叉研究包含有：自然语言处理与语义 web、数据挖掘、机器学习、知识表示与推理、认知计算、信息检索与抽取； 信息抽取（infromation extraction）是知识图谱构建的第 1 步，其中的关键问题是：如何从异构数据源中自动抽取信息得到候选指示单元？ 信息抽取是一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术。

1 相关概念

1.1 本体

本体是对客观的事物以一种形式化的、客观的并且系统化的方式进行描述。本体由哲学领域发起，对现实世界的客观事物进行本质化的描述。它在哲学中的定义为对世界上客观存在物的系统地描述”,是客观存在的一个系统的解释或说明,关心的是客观现实的抽象本质。后来随着在人工智能、计算机以及网络领域中的应用发展,其定义也被融入了许多新的内容。其中最著名、被引用最为广泛的定义是由 Gruber 提出的：“本体是概念化的明确的规范说明”。Studer 对本体诸多定义进行概括分析后认为,本体论的概念包括四个主要方面:（1）概念化(conceptualization):客观世界现象的抽象模型,其表示的含义独立于具体的环境状态;（2）明确(explicit):概念及它们之间联系都被精确定义;（3）形式化(formal):精确的数学描述,计算机可读;（4）共享(share):本体中反映的知识是其使用者共同认可的,是相关领域中公认的概念集,它所针对的是团体而不是个体。本体的目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇相互关系的明确定义。

1.2 自然语言处理

自然语言处理(Natural Language Processing, NLP)是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系，但又有重要的区别。自然语言处理并不是一般地研究自然语言，而在于研制能有效地实现自然语言通信的计算机系统，特别是其中的软件系统。因而它是计算机科学的一部分。

1.3 知识图谱

知识图谱，是结构化的语义知识库，用于迅速描述物理世界中的概念及其相互关系，通过将数据粒度从 document 级别降到 data 级别，聚合大量知识，从而实现知识的快速响应和推理。当下知识图谱已在工业领域得到了广泛应用，如搜索领域的 Google 搜索、百度搜索，社交领域的领英经济图谱，企业信息领域的天眼查企业图谱等。交叉研究包含有：自然语言处理与语义 web、数据挖掘、机器学习、知识表示与推理、认知计算、信息检索与抽取； 信息抽取（infromation extraction）是知识图谱构建的第 1 步，其中的关键问题是：如何从异构数据源中自动抽取信息得到候选指示单元？ 信息抽取是一种自动化地从半结构化和无结构数据中抽取实体、关系以及实体属性等结构化信息的技术。

2 图谱制作与分析

2.1 文献数据来源

中文文献从 CNKI 中获取，检索条件为“KY=自然语言处理+NLP+natural language processing”，来源选择 SCI 来源期刊，EI 来源期刊，北大核心，年份选择 2008 到 2020 年，共得到 787 条结果。英文文献从 webofscience 核心合集中获取，检索条件为主题为 Natural Language Processing 或是 NLP（将两个主题的检索结果进行 and 组配），年份选择 2015 年到 2020 年，文献类型为 article，review，得到 6557 条结果。

本文研究方法包括统计分析法、对比分析法、共现分析法。研究中使用的工具有统计软件 Excel、可视化软件 UCINET 及其自带插件 NetDraw。

2.2 关键词共现分析

概念格的约简能够有效地提高概念格的堆护效率，关键词字体大小对应着相应的记录数目越多，共现网络具有结构性，表现了关键词之间的联系。如图 1，从 web of science 核心数据库的英文文献得到的关键词共现性图谱的结构中可以观察到，有以 ontology，convolutional neural network，sentiment analysis,text mining 为中心的四块。如图 2，从 CNKI 的中文文献得到的关键词共现性图谱的结构中可以观察到，有注意力机制与长短时记忆网络，知识图谱，文本分类与深度学习，语义句法分析四个分块。

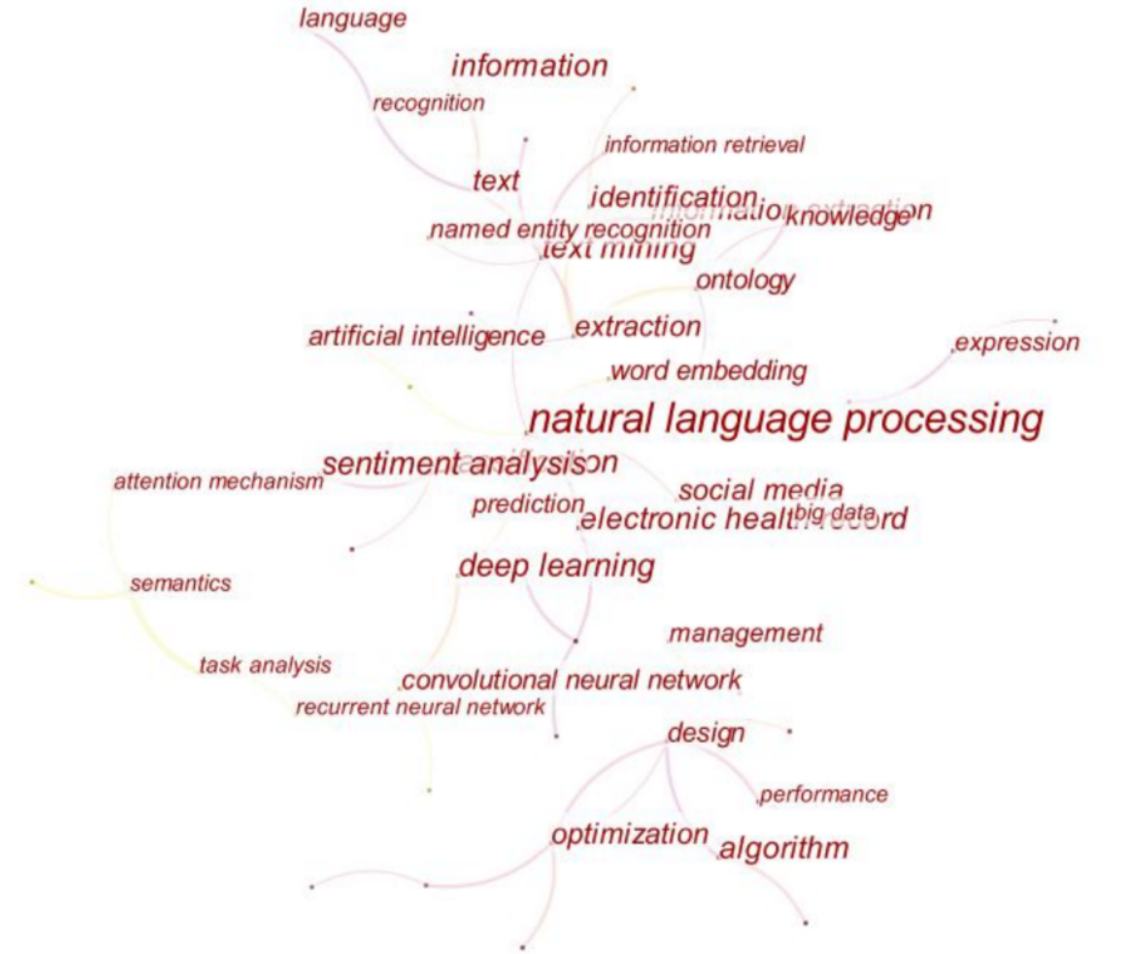

图 1 英文文献中关键词共现性图谱

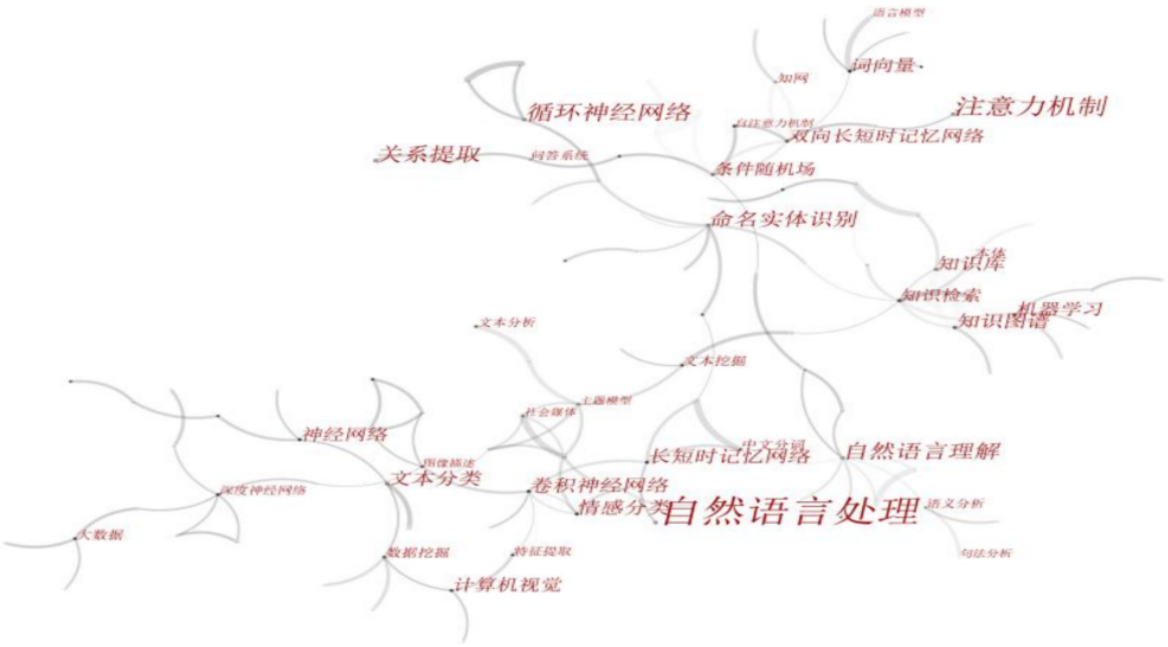

图 2 中文文献关键词共现性图谱

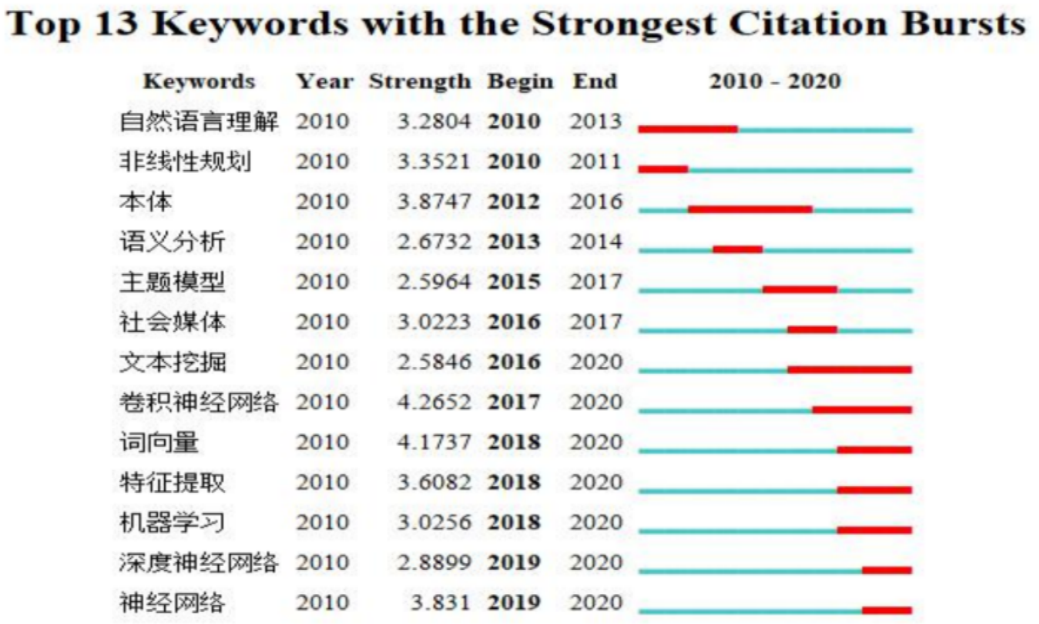

图 3 关键词突现

2.3 机构合作与领域合作分析

图 4 节点与地区字体大小对应着发文数量，中国大陆的发文数量仅次于美国，但中国大陆在机构合作关系图谱中处于边缘位置，与其他机构的合作力度不强。由图 5，可以看到医疗领域与图书情报领域关系密切，生物化学与生物信息联系紧密。

2.4 关键词突现分析

突现时间的最小单位设为一年，共有 13 个突现点。如图 3，关键词突现中，卷积神经网络、词向量、特征提取的关键词强度大，热度一直持续到 2020 年。

2.5 文献引用共现分析

图 6 中，能量圈越大的节点说明被引用的次数越多，可见，Mikolov Tomas(2013)，Pennington J(2014)，Mikolov T(2013)，Manning CD(2014)，Collobert R(2011)，LeCunY(2015)，Srivastava N(2014)，VaswaniA(2017)，Cho K(2014)，He KM(2016)能量圈较大，通过文献引用共现图的结构可以看到文献所引用文献的组合情况。

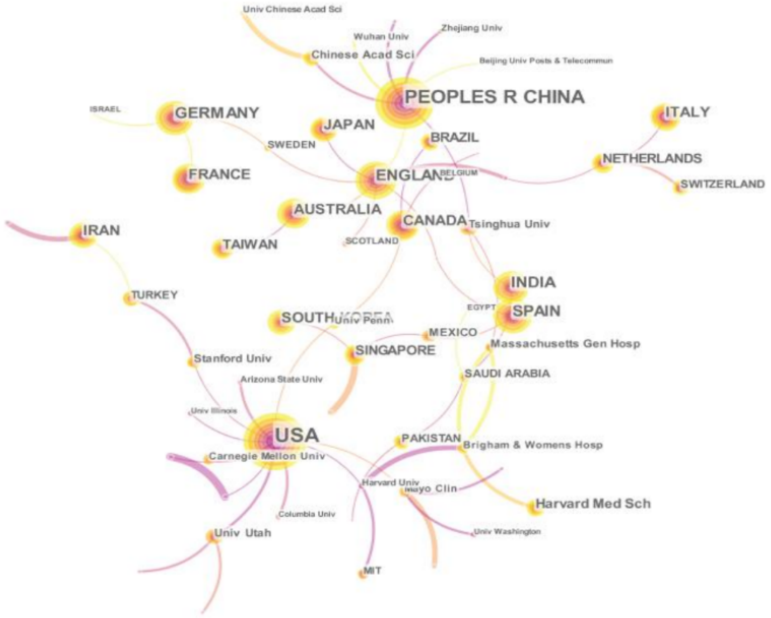

图 4 机构合作关系图

图 5 领域关系合作图

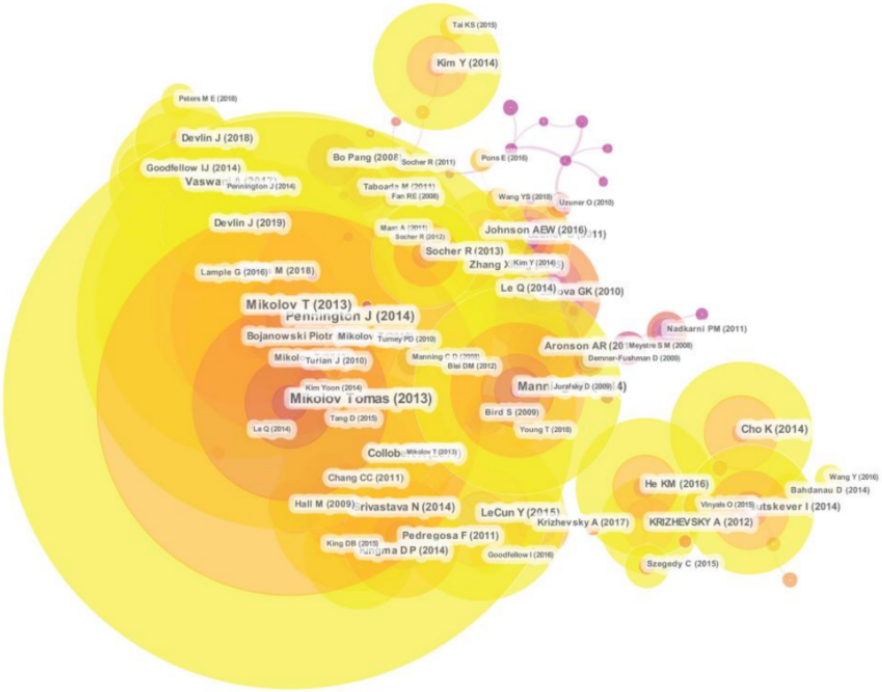

图 6 文献引用共现图

3 自然语言处理存在的问题与展望

本文利用词频统计法以及可视化的方法,对比分析了国内外自然语言处理领域内的年代分布、研究力量分布、研究内容分布。从整体上来看,国内与国外在自然语言处理研究领域具有很大的相似性,包括年代分布、研究力量分布、研究内容分布,同时也看出了国内在自然领域内的卓越成就,尤其是在国际顶级会议中的发文量(包括录用的长短论文)逐年增加,这说明,国内的研究成果得到了世界的赞同与认可。但是本文在研究过程中也存在以下不足:

(1) 在获取数据来源时,使用“自然语言处理”作为主题词搜索,实际上 NLP 是个很大的范围,用主题词“自然语言处理”很难涵盖“机器翻译”、“信息检索”等相关领域,如需做深入研究,还需要区分和界定这些关键词之间的联系,比如信息检索哪些方面用到了自然语言处理技术,它们之间交叉部分主要在哪里,这些问题对笔者以后深入研究提供了一些思维脉络;

(2) 选择的期刊数据库来源也有限,一些自然语言处理刊文量较大的国内期刊,比如《情报学报》,并没有收录到 CNKI;国外还有 EI、ISTP 等数据库也收录了部分的文献,因此数据来源难免有偏颇之处。

基于上文统计分析,笔者对国内自然语言处理技术的发展提出如下建议:

(1) 从年代分布来看,国内应该继续加强与国际合作交流,比如积极参加 NLP 相关的国际学术会议。虽然中国在自然语言处理技术方面取得不少成果,但跟国际相比还存在一定差距,从年代分析来看,中国要想在该领域内取得更进一步的发展就必须善于学习和借鉴国际一些先进技术或者成熟的理论,以免走重复路、走弯路,可以把更多的时间用到解决中国的实际问题中,比如热点关键词中的“中文信息处理”、“中文分词”、“维吾尔语”等充分体现中国需要运用自然语言处理技术解决众多自身的语言问题。既要积极地走出去,又要合理地引进来。

(2) 从研究力量分布来看,国内应加强与自然语言处理相关的不同学科之间的作者合作,不能仅仅停留在某个学科内,而应该进行较多的跨学科合作。自然语言处理和计算机科学联系最为紧密,并且和信息科学、语言学的研究密不可分,如果各学科各成一家,不能及时了解相互的研究进展,那么对整个自然语言处理的发展是很不利的。

(3) 抓住机遇,与时俱进。自然语言处理技术

是随着时代的发展不断发展的,随着云计算和大数据时代的到来,自然语言处理技术将会面对新的挑战,但挑战与机遇共存,科研者们应该善于运用工具,提出新的对策及应变方法。抓住机遇,乘势而上,自然语言处理技术一定会不断有新的突破,也一定会为社会、经济发展乃至整个民族振兴做出卓越的贡献

参考文献

[1]杨佳琦. 基于中文自然语言处理的糖尿病知识图谱构建[D].内蒙古科技大学,2020.DOI:10.27724/d.cnki.gnmgk.2020.000567.

[2]陈荟,邓晖,吴道婷.基于自然语言处理的教学设计学科知识图谱自动构建研究[J].中国教育信息化,2020(07):15-19.

[3]侯梦薇,卫荣,陆亮,兰欣,蔡宏伟.知识图谱研究综述及其在医疗领域的应用[J].计算机研究与发展,2018,55(12):2587-2599.

[4]王飞,陈立,易绵竹,谭新,张兴华.新技术驱动的自然语言处理进展[J].武汉大学学报(工学版),2018,51(08):669-678.DOI:10.14188/j.1671-8844.2018-08-002.

[5]刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.

[6]邱均平,方国平.基于知识图谱的中外自然语言处理研究的对比分析[J].现代图书情报技术,2014(12):51-61.