

人工智能与机器人先进技术丛书

智能摘要与 深度学习

Summarization and Deep Learning

高扬 著

 北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

人工智能与机器人先进技术丛书

智能摘要与 深度学习

Summarization and Deep Learning

高扬 著

 北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

版权专有 侵权必究

图书在版编目 (CIP) 数据

智能摘要与深度学习/高扬著. —北京: 北京理工大学出版社,
2019. 4

ISBN 978 - 7 - 5682 - 6902 - 5

I. ①智… II. ①高… III. ①人工智能 - 算法 - 研究 IV. ①TP18

中国版本图书馆 CIP 数据核字 (2019) 第 060902 号

出版发行 / 北京理工大学出版社有限责任公司

社 址 / 北京市海淀区中关村南大街 5 号

邮 编 / 100081

电 话 / (010) 68914775 (总编室)

(010) 82562903 (教材售后服务热线)

(010) 68948351 (其他图书服务热线)

网 址 / <http://www.bitpress.com.cn>

经 销 / 全国各地新华书店

印 刷 / 保定市中华美凯印刷有限公司

开 本 / 710 毫米 × 1000 毫米 1/16

印 张 / 9.75

字 数 / 142 千字

版 次 / 2019 年 4 月第 1 版 2019 年 4 月第 1 次印刷

定 价 / 48.00 元

责任编辑 / 张海丽

文案编辑 / 张海丽

责任校对 / 杜 枝

责任印制 / 李志强

图书出现印装质量问题, 请拨打售后服务热线, 本社负责调换

序

我常遇到学生这么说：“我觉得自己本科没学到什么东西，想在研究生阶段让自己在专业领域学习更多知识，对就业更有帮助”，或者“老师，我想读博士，做研究”，等等。总之，大多学生都很有上进心。

但是，有多少学生知道，上学这段时间究竟要学什么？研究什么？培养什么能力？具备哪些品质？仅比别人多上两年学，就真的可以在职场更有信心了吗？这个话题太大，我也担心自己打开话匣子就扯远了。那么我就从一个小例子引出我想说的观点，也希望真的对你有帮助。

有 A、B 两个同学：A 同学雄心满满，一上来觉得自己要研究一个大课题，一定要与众不同；B 同学相对保守，觉得我先找一个别人做过的上手看看，然后再决定自己做什么。一段时间之后，A 同学确实找到一个前沿课题，但却抱怨“参考资料太少，任务太新，以至于太多不完善，没法下手”。B 同学也如他所愿，找了一个容易上手的课题，但也在抱怨“大家都做这个，能做的都被别人做了，我也没什么可做了”。奇怪，怎么都不行了呢？

问题不在初心，而在于思维习惯。当出现问题时，A、B 同学都习惯于迁怪环境（除自己以外的原因统称环境），以至于怀疑最初的选择。有这种思维习惯的人，解决方案往往也从外界着手。比如，认为学术没意义，还得去工业界才能锻炼人；或者先换个课题吧，等等。可想而知，换了可能还会重蹈覆辙。

我的建议是，做自己人生的顶层设计，首先要对自己做一个定位：“我要做一个什么样的人？世界因为有我会有什么不同？”不要觉得这是大话或者套话，每个人都不同，即使你只能影响一个人，那也是存在的意义。比如，一个宿舍里，有一个人坚持每天上自习，慢慢地，同宿舍的其他人也会开始学习。其

次，因为这样的人生定位，我需要有怎样的认知，即什么最重要，需要坚持什么，放弃什么。之后，以信念、认知指导所需要培养的能力。我应该培养怎样的能力，掌握什么样的方法来符合我确定的认知和信念。确定了自己要培养的能力，就会明白，我该如何行动才能拥有这样的能力，具体的行动包括制定目标、规划年度计划甚至每周计划等。最后，为了达到这些制订的计划，我该如何利用环境或者资源帮助实现目标。以这样的思维方式，也就无暇顾及抱怨环境和别人的错，或者怀疑选择，因为这样对完成你的目标没有帮助，对完善你需要具备的能力无济于事，对你的信念没有价值，离你所要的人生越来越远。

例如，如果一个人的定位是“我要做一个有用的人”，那么，基于这样的定位，他可能需要有“利他”的认知，放弃“小我”的信念。这样的认知就会帮助他学习如何包容外界的缺陷或者他人缺点，并且关注提升自己核心业务的能力。进而，也就做出了1年、5年乃至更长时间的计划。那么，具体到我们的学习阶段，遇到了困难（要知道，不管你做什么选择，都会有困难），就知道自己如何思考，如何行动，如何利用环境来达到自己的目标。比如，当你发现自己的课题参考文献太少，可利用的资源太少，你可能会想，我的目标是2年内做出这个课题的创新工作，那我就需要用3个月的时间搜集现有的所有资料，并做出总结归纳，发现哪些东西不够完善。之后，我会用6个月的时间，通过搭建新系统、提出合理假设、找同行讨论等方法做好前期准备。最后用一年的时间做出哪怕一点点的创新工作。因为整个过程，你关注的不再是怎么做才完美，你关注的是离自己的目标接近了多少。同样，当遇到B同学的问题时，你就会想，环境在该领域的创新空间较小，我现阶段的能力不足以让我发现本领域的核心问题。要么直接提升自己的能力，让自己对问题理解的层次更深；要么可以从其他领域寻找灵感，迁移创新，或者将本领域了解到的知识服务于其他领域。条条大路通罗马，关键在于你所站的角度。

我常跟学生说：“学习不在于你学到了什么，而是你学到了如何学习。”我也督促自己能做到“授之以鱼，不如授之以渔”，与君共勉！希望我说了这么多，对你回答我最开始提问的问题有了思考，也有了答案。

前言

亲爱的同学，之所以在开篇说了那么多“鸡汤”，是希望同学们不管做什么方向的研究，具体做什么不是最重要的，重要的是在你前进的方向中学会如何学习、如何探究、如何掌握对你之后人生路更有用的方法。当然，在学习和研究的道路上，学会时间管理、自我精力管理，学习团队合作、与导师沟通，学会有效反馈的方法等，都会对你的学习、工作，甚至日常大有裨益。

阅读本书之前，读者需要具备少量的高等数学、线性代数、概率统计和编程的基本能力。补充读物包括机器学习、深度学习 (Deep Learning)、自然语言处理 (Natural Language Processing)，可以更好地帮助了解本书的内容。因此，本书更适合有相关背景的研究生和高年级本科生，尤其是具备少量自然语言处理方面知识的人士。

本书共 7 章，总体分为 3 个部分：第一部分包括第 1 章和第 2 章，介绍智能摘要产生的背景、需求、应用、分类以及文本摘要的基础知识；第二部分包括第 3~5 章，介绍信息抽取的基础知识，并且与深度学习方法结合讲述，还包括抽取式摘要经典和常用方法和思路；第三部分包括第 6 章和第 7 章，介绍文本生成的基本知识以及自动生成式摘要的常用方法及思路。

深度学习和文本摘要技术发展极其迅速，尤其自动摘要系统是自然语言处理领域较为复杂的分支之一。笔者通过有限的逻辑线索，将相关内容和技術串讲下来，难免有疏漏甚至错谬之处。若对本书有任何意见和建议，请与我们联系 (email:dpsummarization@163.com)。

非常感谢对本书出版提出建议的老师和同学们。尤其感谢王文博对第 7 章的贡献，李飞对 4.2 节的贡献，以及王洋对第 5 章的贡献。还要感谢魏林静、任

慕成、郭一迪、周宇翔等在前期投入的工作。

高扬

2019年3月

目录

此书献给我的母亲!

第一章 绪论

第二章 基础理论

第三章 实验方法

第四章 实验结果

第五章 讨论

第六章 结论

参考文献

附录

It was the best of times, it was the worst of times.

— *Charles Dickens, A Tale of Two Cities*

目录

第一部分 文本摘要技术

第 1 章 文本摘要概述	3
1.1 摘要技术的需求	3
1.2 自动文本摘要的应用	5
1.3 文本摘要分类	6
第 2 章 摘要评价方法	9
2.1 评价数据	9
2.1.1 抽取式摘要数据集	9
2.1.2 生成式摘要数据集	10
2.2 评价指标	13
2.2.1 自动评测	13
2.2.2 人工评测	16
2.2.3 半自动评测	16
2.3 总结	17

第二部分 信息抽取

第 3 章 文本表示	23
3.1 词级表示	23
3.1.1 Word2Vec	23

3.1.2	神经网络	33
3.2	句级表示	38
3.2.1	Paragraph Vector 模型	38
3.2.2	Skip-Thoughts 文档-句子表示	40
3.2.3	卷积神经网络	41
3.3	文档级表示	45
3.3.1	混合模型	45
3.3.2	BERT	46
第4章	命名实体识别	55
4.1	命名实体识别简介	56
4.1.1	命名实体概念	56
4.1.2	命名实体的用途	56
4.2	命名实体识别方法	57
4.2.1	基于规则的方法	58
4.2.2	支持向量机	58
4.2.3	最大熵模型	59
4.2.4	隐马尔可夫模型	61
4.2.5	条件随机场	63
4.2.6	Bi-LSTM+CRF 模型	64
第5章	抽取式摘要	67
5.1	无监督方法	68
5.1.1	启发规则方法	68
5.1.2	向量空间模型	70
5.1.3	基于图的模型	70
5.1.4	组合优化方法	73
5.2	有监督方法	76
5.2.1	分类模型	76
5.2.2	句子打分与排序	78

5.3	强化学习	80
5.3.1	基本原理	80
5.3.2	强化学习优化的抽取式摘要	83
5.4	总结	85

第三部分 文本生成

第6章	神经网络文本生成模型	91
6.1	语言模型	91
6.2	“编码-解码”网络框架	92
6.2.1	模型框架	92
6.2.2	注意力机制	95
6.3	“序列到序列”生成模型	96
6.4	网络训练	100
6.4.1	Sampled Softmax 算法	100
6.4.2	Beam Search 算法	101
6.4.3	随机梯度下降算法	103
6.5	面临的问题	104
第7章	生成式摘要	107
7.1	未登录词问题解决方案	107
7.1.1	复制机制	108
7.1.2	指针机制	108
7.2	生成重复词问题解决方案	113
7.2.1	注意力改进方法	113
7.2.2	覆盖机制	114
7.2.3	分散机制	115
7.3	生成错误关系问题	116
7.4	长文本摘要生成问题	118
7.4.1	注意力改进方法	118

7.4.2	卷积 seq2seq 模型	120
7.5	强化学习优化的生成式摘要	122
7.6	抽取器 + 生成器	124
7.6.1	抽取器 + 指针生成网络	125
7.6.2	强化抽取器 + 生成器	125
7.7	总结	125
附录 A	专用名词缩写	127
	索引	129
	后记	131
	参考文献	133

第一部分

文本摘要技术

第1章 文本摘要概述

如今是信息爆炸的时代，在互联网和物联网的世界里，每个人或者每个物体都可看作一个节点，理论上它可以跟任意一个节点互联互通，且信息能瞬间传达。我们比以往任何时期都能更容易地获取海量的信息，但也正因如此，在有限的时间内，我们根本无法处理所有的信息。信息无处不在，唾手可得，却也面临信息严重过载。那些重要或不重要的信息一起扑面而来，我们该如何处理？正如开篇引用的文字：“这是最好的时代，也是最坏的时代。”

1.1 摘要技术的需求

- 爆炸式的信息：2018年11月，IDC发布最新版《数据时代2025》白皮书^①，预测全球数据量将从2018年的33ZB^②增至2025年的175ZB（图1.1）。这就说明了目前我们所处的这个科技时代，信息数据的增长是爆炸性的。

- 互联网用户普及：2018年“互联网女皇”趋势报告指出，2017年全球互联网用户达36亿，虽然增速放缓，但是2017年互联网普及率高达49%，2018年则超过50%^③（图1.2）。

- 碎片化阅读：在移动互联网的时代，内容更加丰富多彩，获取信息渠道多样，从传统的文本、语音，变成了视频、短视频、直播等内容。在充斥了大量移动互联网应用的环境下，用户固定在某一个时间段完整阅读的可能性越来越

① <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-data-age-china-whitepaper.pdf>

② 1ZB相当于1万亿GB。

③ <http://hybg.cebnet.com.cn/upload/internet-trends-2018/internet-trends-2018-en.pdf>

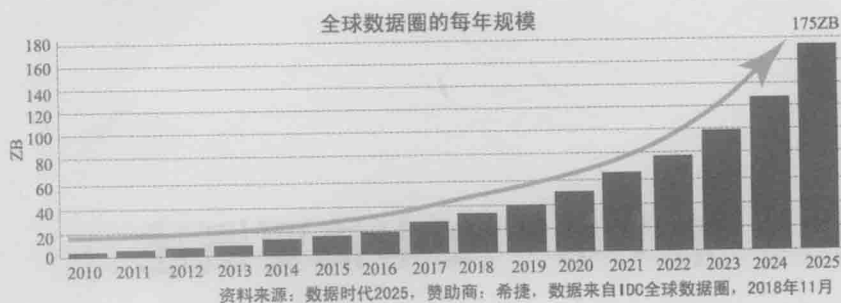


图 1.1 全球数据圈每年的规模

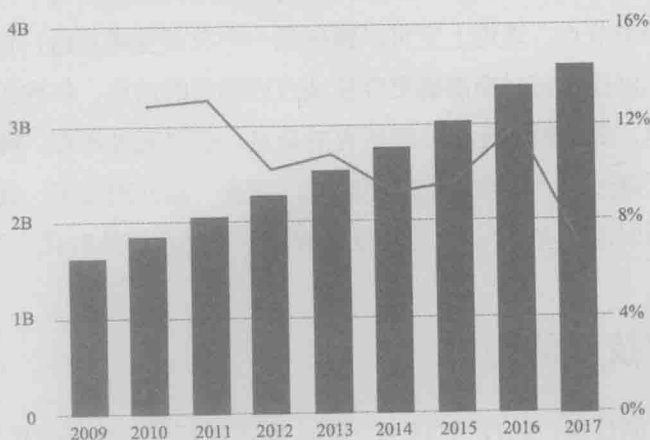


图 1.2 全球互联网用户及其年比增长

越小，如图1.3^①所示。因此，用户将越来越倾向于碎片时间阅读，更愿意关注经过整合之后重要和有趣的内容。

综上所述，在信息爆炸的形势和庞大用户群体碎片化阅读的趋势下，从海量信息中提取重要的内容，过滤那些冗余的重复信息，整合碎片化的信息，继而形成连贯的内容，已成为一个迫切的需求。鉴于此，自动文本摘要则提供了一个高效的解决方案。自动文本摘要旨在通过机器，自动输出简洁、流畅、保留关键信息的摘要。

^① <https://zhuanlan.zhihu.com/p/37067159>