天津大学

本科生毕业设计（论文）开题报告

题目：面向微博文本的中文分词

学　　院 智能与计算学部

专　　业 软件工程

年　　级 2016

姓　　名 郭培溟

学　　号 3016218085

指导教师 王赞

（不少于 2000 字，内容包括：课题的来源及意义，国内外发展状况，本课题的研究目标、研究内容、研究方法、研究手段和进度安排，实验方案的可行性分析和已具备的实验条件以及主要参考文献等。）

一、课题来源及意义

西方语言的书写习惯是词与词之间用空格隔开，而汉语书写连续，词和词之间无分界标志，不像英语等语言有天然的分割符号，但是词又是语言中能自由运用的基本单位，因此中文分词是中文信息处理的一个基础任务和研究方向。计算机处理中文文本时，首要任务就是对文本进行切分，即将“天津大学智能与计算学部”切分成“天津大学/智能/与/计算/学部”的过程。分词清晰之后才能对文本信息进行处理。只有做好了中文分词，我们才能进行中文语言的处理、搜索、翻译、检索等等。

众所周知，深度学习技术严重依赖具体的数据集、语料库。而目前常见的数据集（ctb、msr、pku）多为新闻领域。因此训练出来的分词器在新闻题材的文本上有较好的分词效果，但是在微博文本上的分词效果略显疲软，缺乏跨领域应用的泛化能力。而且模型大多较大，运算速度慢，实用价值低。本课题拟探究分词模型在微博文本的跨领域应用问题。在运算速度和分词效果上提高深度学习模型在微博文本的分词性能，更好的为舆情监控、微博评论生成等下游应用提供技术支持。

二、国内外发展状况

从算法分类的角度看，现有的分词算法分为三类：基于字符串匹配的分词方法、基于统计的传统机器学习分词方法以及基于深度学习的神经网络分词方法。

基于字符串匹配的分词方法，是中文分词任务较早期的解决方案，顾名思义，就是将待切分的句子与语料库进行字符串的匹配，若匹配成功，则进行子串的切分。依照匹配方向不同可分为前向匹配法、后向匹配法，依照优先匹配的长度又可分为最长匹配法和最短匹配法。基于匹配的分词方法简单易行，但是分词效果相比其他两种方式欠佳。

Xue(2003)等人首次将分词任务形式化为序列标注任务[2]，该工作用四种标签(tag)——B(begin)、M(middle)、E(end)和S(single)对句子中的每一个字符标注切分信息（如表1）。四种标签分别表示所标注的字符是词的开头、中间、末尾或者单字词。

序列标注学习任务是自然语言处理中最基础的结构化学习任务，在序列标注的模型中，两个串的各个节点单元需要严格一一对应，非常方便使用各种成熟的机器学习工具来建模和实现。Peng et al.(2004)和Tseng et al.(2005)

则自然地将标准的序列学习工具条件随机场(CRF)引入中文分词任务，再配合基于字或者词的特征模板和传统机器学习分类器，构成了统计时代的分词算法框架。

表 1　4 标签序列标注示例

天	津	大	学	/	智	能	/	与	/	计	算	/	学	部
B	M	M	E	/	B	E	/	S	/	B	E	/	B	E

B：开始，M：中间，E：末尾，S：单字词

随着word2vec[4]等方法的提出，将词嵌入(word embedding)这种分布式词向量表示方法引入到自然语言处理之后，深度学习开始席卷整个领域。深度学习模型不仅以缓解特征工程代价的优势而著称，而且效果上也有一定的提升，以F分数为评价结果，其分词效果也要普遍好于其他另外两种方式。采用蕴含有效的句法和语义信息的词向量、预训练语言模型（如bert等），具有强大学习能力的网络结构（LSTM、transformer等）处理中文分词任务是近年来的研究热点。

Zheng et al. (2013) 提出神经网络中文分词方法，首次验证了深度学习方法在中文分词任务上的可行性[5]。随后为了更加完整的建模分词上下文，Chen et al. (2015a) 提出了一种带有自适应门结构的递归神经网络抽取 n-gram 特征。同年 Chen et al. (2015b) 提出使用 LSTM 来捕捉句子中的长距离依赖，以此来解决滑动窗口的局限性。Zhang et al. (2016) 提出了一种基于转移的模型用于分词，并将传统的特征末班和神经网络自动提取的特征结合起来，从而获得分词精度的进一步提升[8]。Yang et al. (2017) 提出了使用多种预训练手段，挖掘文本的分词知识信息对分词效果进行提升。

此外，Zhang et al. (2018) 在深度学习模型中融入了字典信息提高了分词模型的跨领域应用能力。而 bert[11]模型的提出，刷新了各大自然语言处理任务榜单，其也在几大分词数据集上完爆之前的算法结果。

三、研究目标、内容和方法

研究目标：

本课题拟利用现有的深度学习技术，并结合科研前沿热点，加入字典信息和预训练语言模型 bert，利用知识蒸馏等手段压缩深度模型，设计并实现一款面向微博文本的速度快、效果好的中文分词器。

研究内容：

目前基于深度学习的中文分词技术主要针对具体数据集在某一领域（如新

闻）进行研究，缺乏跨领域应用的泛化能力，且模型较大，运算速度慢。本课题拟探究分词模型在微博文本的跨领域应用，利用bert等预训练语言模型、使用transformer[12]抽取句子长距离特征提高分词的准确度。并基于计算机科学中cache的思想，预保存常见的5-gram词向量特征，以及知识蒸馏[12]等技术压缩深度模型，提高分词速度。使用字典信息等来提高模型跨领域能力。力求在微博文本上的分词效果优于、快于“结巴分词”等常见的分词工具。

研究方法：

如下图1，首先结合bert,使用常见的分词数据集CTB、MSR、PKU训练出具有较好分词效果但是较复杂的分词器bertSegmentor，将其作用在大规模无标签数据集上(如giga-Data)，得到大规模有标注的数据。然后利用知识蒸馏等技术，并且加入字典信息训练得到参数量较少的分词模型Segmentor。由于模型中使用了5-gram词嵌入，基于计算机科学中cache的思想，可以将在大规模语料中频繁出现的5-gram词嵌入提前存储下来，使用过程中直接加载到模型中，以此来进一步提升模型速度。

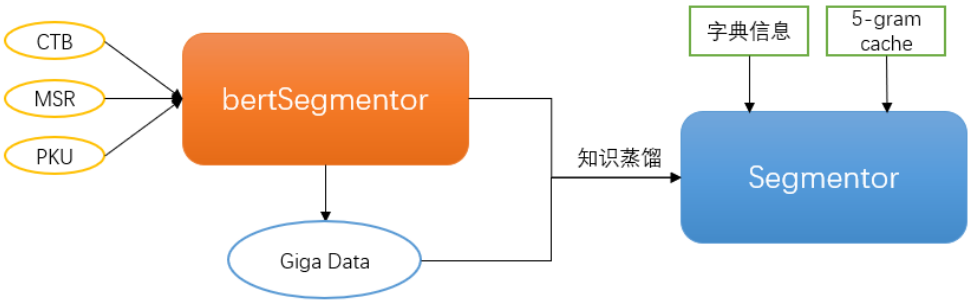

图 1　研究方法概念图

四、进度安排

时间	工作阶段	工作安排
2019年11月29日-2019年12月15日	启动与选题	论文选题并登记系统
2019年12月16日-2020年1月12日	开题	阅读文献，撰写开题报告
2020年2月17日-2020年3月8日	实验过程	1.设计实现bertSegmentor 2.设计并实验验证参数量较少的Segmentor模型结构
2020年3月9日-2020年3月29日	实验过程	结合字典信息，实验加入字典的方法，提高模型跨领域能力

2020年3月30日-2020年4月19日	实验过程	利用模型蒸馏等技术，过训练得到速度快效果好的分词模型
2020年4月20日-2020年4月26日	实验整理	整理实验结果、撰写论文大纲
2020年4月27日-2020年5月17日	论文撰写	毕业论文撰写及修改
2020年5月18日-2020年5月31日	论文提交	线上提交论文
2020年6月1日-2020年6月14日	论文答辩	线下答辩、线上论文二次提交
2020年6月15日-2020年6月28日	评估、评优	学校、学院论文抽样评估及论文评优

五、实验可行性和实验条件

实验可行性：

从训练模型角度看，bert的提出刷爆了众多自然语言处理任务榜单，直接使用bert和一层bi-lstm就可以在ctb60数据集上达到97.4的F分数，因此训练出高质量的分词模型并非难事。

从加入字典信息跨领域分词的角度看，目前已有相关方法的前沿研究。Zhang et al. (2018) 在深度学习模型中融入了字典模板提高了分词模型的跨领域应用能力。

从模型压缩的角度看，知识蒸馏[13,14,15]等方式早已经得到了实践的检验，在保证模型效果不大打折扣的情况下，训练出参数量较小的深度模型，以此来提高深度学习模型的运算速度。

从实验实现上看，pytorch等深度学习框架的开源和广泛使用，极大的降低了训练深度学习模型的工程代码门槛，简洁方便的使用教程和api使得我们可以快速搭建算法模型。

实验条件：

目前已具备了训练所需的数据集。

高性能台式机电脑1台(Intel i7 9700/16G/2T/Nvidia RTX 2080Ti-11G显存)。

六、参考资料

[1] Xue N, Shen L. Chinese word segmentation as LMR tagging[C]. Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17. Association for Computational Linguistics, 2003: 176-179.

[2] Peng F, Feng F, McCallum A. Chinese segmentation and new word

detection using conditional random fields[C]. Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004: 562.
[3] Tseng H, Chang P, Andrew G, et al. A conditional random field word segmenter for sighan bakeoff 2005[C]. Proceedings of the fourth SIGHAN workshop on Chinese language Processing. 2005.
[4] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
[5] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 647-657.
[6] Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for Chinese word segmentation[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 1744-1753.
[7] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for chinese word segmentation[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1197-1206.
[8] Zhang M, Zhang Y, Fu G. Transition-based neural word segmentation[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 421-431.
[9] Zhang Q, Liu X, Fu J. Neural networks incorporating dictionaries for chinese word segmentation[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[10] Yang J, Zhang Y, Dong F. Neural word segmentation with rich pretraining[J]. arXiv preprint arXiv:1704.08960, 2017.
[11] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems. 2017: 5998-6008.
[13] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural

network[J]. arXiv preprint arXiv:1503.02531, 2015.
[14] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.
[15] Fukuda T, Suzuki M, Kurata G, et al. Efficient Knowledge Distillation from an Ensemble of Teachers[C]. Interspeech. 2017: 3697-3701.

选题是否合适： 是□ 否□
课题能否实现： 能□ 不能□
指导教师（签字）
年 月 日

选题是否合适： 是□ 否□
课题能否实现： 能□ 不能□
审题小组组长（签字）
年 月 日