天津大学

# 本科生毕业设计（论文）任务书


**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

## 一、原始依据（不少于200字，包括设计或论文的工作基础、研究条件、应用环境、工作目的等。）

中文分词就是将一个中文字符序列的句子切分成词序列的过程。西方语言的书写习惯是词与词之间用空格隔开，而中文不实行按词连写，词之间没有天然的分割符号，但是词又是语言中能自由运用的基本单位，因此分词成了汉语自动分析的十分重要又必不可少的第一道工序。

近年来深度学习在很多自然语言处理任务中都达到了很好的效果，基于深度学习模型的中文分词器的准确率也一再提升。Bert、ELMO 等语言模型的提出也再一次刷新了中文分词的准确率。跨领域分词、加入字典信息等问题也是近一两年来自然语言处理的研究前沿和热点。此外，知识蒸馏、过训练等技术也为如何获得参数量更少、运算速度更快的深度模型开辟了道路。

本课题拟利用现有的深度学习技术，并结合科研前沿热点，设计并实现一个在微博文本上速度快、效果好的中文分词器。

## 二、参考文献

[1] 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.

[2] 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.

[3] Yang J, Zhang Y, Dong F. Neural word segmentation with rich pretraining[J]. arXiv preprint arXiv:1704.08960, 2017.

[4] Zhang Q, Liu X, Fu J. Neural networks incorporating dictionaries for chinese word segmentation[C]. Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[5] 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.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in neural information processing systems. 2017: 5998-6008.

[7] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[8] 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.
[9] Cai D, Zhao H. Neural word segmentation learning for Chinese[J]. arXiv preprint arXiv:1606.04300, 2016.
[10] Wu W, Meng Y, Han Q, et al. Glyce: Glyph-vectors for Chinese Character Representations[J]. arXiv preprint arXiv:1901.10125, 2019.

## 三、设计（研究）内容和要求（不少于200字，包括设计或研究内容、主要指标与技术参数，并根据课题性质对学生提出具体要求。）

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，并利用 bert 等模型提高分词的准确度，使用知识蒸馏等技术压缩深度模型，提高分词速度、利用字典信息提高跨领域分词能力。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。

要求学生能够利用现有的深度学习库，快速、准确的实现神经分词模型。阅读自然语言处理顶级论文并复现论文结果。严谨、认真、按时的完成毕业设计课题，设计并实现一款面向微博文本的又快又好的中文分词器。

指导教师（签字）

年　月　日

审题小组组长（签字）

年　月　日