第三周周记

周一	
完成内容	阅读论文《面向短文本的命名实体识别》
内容描述	1. 了解了短文本区别于一般长文本的几个难点 2. 了解了短文本规范的方法 3. 了解了短文本的命名实体方法
未解决问题	无

周二	
完成内容	阅读了论文《面向信息抽取的中文命名实体识别研究》
内容描述	1. 中文命名实体识别概述 2. 命名实体识别的方法（基于规则的方法、基于统计的方法、规则与统计相结合的方法） 3. 了解统计模型（HMM、MEM、CRF）
未解决问题	无

周三	
完成内容	阅读了论文《面向企业信息检索的中文分词系统的研究与实现》
内容描述	1. 了解了中文分词技术发展现状，课题的研究内容 2. 了解中文分词的关键技术（中文分词算法、分词词典机制） 3. 了解中文分词的难点（通用词表和切分规范、切分歧义、未登录词识别）
未解决问题	无

周四	
完成内容	阅读了论文《统计机器翻译中的中文分词策略研究》
内容描述	1. 了解了中文分词的串行式融合的基本思想 2. 了解了基于多种分词的词对齐融合的基本思想
未解决问题	无

周五	
完成内容	继续阅读论文《统计机器翻译中的中文分词策略研究》
内容描述	1. 了解了基于多种分词的判别式词对齐融合基本思想 2. 了解了基于多种分词的语言模型融合的基本思想 3. 了解了融合单语和双语知识的中文分词模型的基本思想
未解决问题	无

周末	
完成内容	了解中文分词的几种算法
内容描述	1. 最小匹配算法 2. 最大匹配算法 3. 逐字匹配算法 4. 联想-回溯法

未解决问题	无

工程汇总	
完成任务	如上所述
任务描述	如上所述
代码量	无
未解决问题	无

论文汇总	
论文列表	[1] 《面向短文本的命名实体识别》 [2] 《面向信息抽取的中文命名实体识别研究》 [3] 《面向企业信息检索的中文分词系统的研究与实现》 [4] 《统计机器翻译中的中文分词策略研究》
论文摘要	步入信息时代，人们对语言间翻译的需求与日俱增。传统的基于人工的翻译已经远远不能满足人们的需求，而机器翻译，持别是统计机器翻译，因其良好的自动学习能力和较好的翻译效果逐渐受到人们青睐。 ＂词＂是语言中能独立运用的最小语言单位。与英语等语言不同，在书写汉语句子时，词与词之间没有分隔标记。因此，对中文文本进行词的识别，即中文分词，就成为了构建汉语相关的机器翻译（如汉英机器翻译）的一个重要预处理步骤。长期以来，人们对中文中＂词＂的定义没有达成共识。已有的工作表明，不同的自然语言处理任务对分词有不同的需求，在单语意义下性能较好的分词工具，未必能在双语应用（如机器翻译）中得到较好的性能。因此，需要重新考虑机器翻译所采用的中文分词策略。 中文分词对统计机器翻译的影响是非常复杂的，主要体现在全局和局部两方面： １．从全局看，机器翻译流程的各个步骤性质不同，对中文分词的需求也可能是不同的。已有的工作忽略了中文分词对机器翻译的全局影响，在优化中文分词时假设各个步骤采用的中文分词是同一种分词，并在此假设下对中文分词进行整体优化。这一做法容易导致机器翻译系统的性能由于使用非最优的分词组合而受损。 ２．从局部看，机器翻译流程的每一个步骤中也有对分词粒度的选择问题：粗粒度的分词能捕捉较多的上下文信息，细粒度的分词能缓解模型的稀疏性，两者各有利弊。已有的大部分工作忽略了中文分词对机器翻译的局部影响，在各个步骤中仅使用一种分词粒度，这一做法可能导致该步骤的性能由于所选择的分词粒度的不合适而受损。 针对中文分词对统计机器翻译影响的复杂性和已有工作的不足，本文提出了在汉英（英汉）统计机器翻译中融合多种中文分词的框架，以充分利用多种分词中包含的＂多样性＂和＂互补性＂的知识，主要工作如下： １．针对中文分词对统计机器翻译对的全局影响，本文提出了一种串行式分词融合策略。在机器翻译流程的不同步骤中分别采用不同的中文分词，W得到有利于提高翻译性能的中文分词组合。串行式融合策略缓解了统计机器翻译系统的性能由于使用非最优的分词组合而受损的问题。 ２．针对中文分词对统计机器翻译对的局部影响，本文提出了一种并行式分词融合策略。在同一步骤内部W多种分词作为输入，利用蕴含在多种分

	词中多样和互补的知识，提高各步骤的性能：在词对齐阶段，将基于多种分词的词对齐结果W启发式的方式进行融合，并针对这一后发式算法在建模、搜索、训练方面的局限性，进一步在判别式模型的框架下提出了基于多种分词的判别式词对齐模型，形式化地定义了多分词环境下判别式词对齐模型的建模、搜索、训练等问题；为了提高英汉机器翻译中语言模型的调序能为，在解码时融合基于多种分词的语言模型。并行式融合策略缓解了统计机器翻译系统某步骤的性能由于所选择的分词粒度的不合适而受损。 3．针对仅已有工作提出的利用词对齐知识来学习面向机器翻译的分词的不足，本文提出了一种融合单语知识和双语知识的分词方法。该方法有效地利用了双语词对齐的结果和单语分词工具的分词结果，利用序列标注模型，学习出一个独立的面向机器翻译的分词模型。该方法缓解了已有方法仅利用词对齐知识学习分词模型的不足
未解决问题	无

下周任务	
工作	继续查找相关的论文和继续阅读相关文献。
论文	继续寻找与中文分词和命名实体识别相关的论文
其他	无
汇总	了解更多与我的课题相关的知识

日期:2018/01/15 - 2018/01/20