开题报告

## 一、毕业设计题目的来源

SemEval-2017 task12 : Clinical TemEval

## 二、毕业设计选题的目的和意义

电子病历是医疗机构生成的针对于医疗活动过程中文字、图表等数据的数字化信息，而且也是便于转储、管理和传输的医疗记录，其中的内容是由医务人员撰写的与患者开展医疗有关的过程实录，包括病程记录、出院小结等部分。电子病历中包含了大量丰富的医疗知识，通过分析即可得到诸如疾病的患病特征、 用药情况以及治疗方式等各项之间的潜在联系。这样的知识数据可以对医疗问题决策提供有建设性的帮助， 并且还可以为用户建立个性化的健康模型。

从2007年起，临时信息提取的研究就已经备受关注，近年来人们尝试将信息提取系统运用在临床医疗数据上，来帮助医生或临床医疗研究人员研究临床事件的时间线，例如发病症状、病情严重程度以及治疗规程等等。在2015年和2016年Clinical TemEval发布的任务中，参与人员已经很好地完成了临床报告中关键内容的提取：时间表达，事件表达以及时序关系等等。Clinical TemEval 2017 提出了一个新的方向：领域的适应性，即在一种疾病的临床报告文本上训练的模型是否能适用于另一种疾病。这两种疾病的患病群体的医学报告在表达上有很大的不同，这给研究带来了很大的挑战，也给未来的医疗文本提取的普适性研究带来了巨大的潜力。

自从Clinical TemEval 2017发布任务以来，共有11支队伍提交了他们的结果。分析报告表明，这些队伍都运用了共同的一些特征，例如n-grams、词表、词性标记、医学语言系统（UMLS）等等。在时间表达、事件表达和时序关系的提取中相较于基于规则提取的baselines都有所提升。但是，数据表明，跨领域提取的研究仍然是一个巨大的挑战。相对于针对同一领域数据训练和测试的准确率来说，跨领域提取的准确率下降了不小于20%的百分比。所以，在医学领域中，建立一个稳定而高效的医疗信息提取系统仍然需要大量的研究工作。未来，将会有更多不同种类的领域适应性技术运用在医学文本中，来弥补跨领域文本提取表现的不足。

自从Clinical TemEval 2017发布任务以来，共有11支队伍提交了他们的结果。分析报告表明，这些队伍都运用了共同的一些特征，例如n-grams、词表、词性标记、医学语言系统（UMLS）等等。在时间表达、事件表达和时序关系的提取中相较于基于规则提取的baselines都有所提升。但是，数据表明，跨领域提取的研究仍然是一个巨大的挑战。相对于针对同一领域数据训练和测试的准确率来说，跨领域提取的准确率下降了不小于20%的百分比。所以，在医学领域中，建立一个稳定而高效的医疗信息提取系统仍然需要大量的研究工作。未来，将会有更多不同种类的领域适应性技术运用在医学文本中，来弥补跨领域文本提取表现的不足。

## 四、毕业设计的主要内容、研究方法、技术路线及可行性分析

### 1、主要内容：

#### （1）提取事件表达（ES）。

例如，给一段临床医学报告文本：

*April 23, 2014: The patient did not have any postoperative bleeding so we will resume chemotherapy with a larger bolus on Friday even if there is slight nausea.*

需要提取的内容：

ES:

* bleeding
* resume
* chemotherapy
* bolus
* nausea

#### （2）提升领域的适应性

在结肠癌患者的病理报告上训练，在脑癌患者的病理报告上做预测。有监督和非监督两种方式。非监督型训练方式不在训练数据中加入任何脑癌患者的病理报告文本，而监督型训练方式则在训练数据中加入前10位脑癌患者的病理报告文本，比较和观察测试结果。（监督训练需要在确保申请到脑癌患者病理报告文本之后才能进行。）

### 2、研究方法和技术路线：

利用模型融合技术，融合不同的模型和特征。

针对模型，可选择实验的方法有：条件随机场、支持向量机、规则、决策树、神经网络等；针对特征，可选择的内容有：n-grams、词向量嵌入、词表、词聚类、词性标记、语义类别、UMLS等；针对领域适应性，可选择的方法有：给脑癌训练数据赋予更大的权重、在输入的词表中加入未知数据、使用预训练的与领域无关的词向量嵌入等。

### 3、可行性分析：

目前各实验方法均可以找到开放源码，在本研究中可在前人的基础上，结合多种实验方法，比较不同方案的优劣性，在实际实验过程中，将考虑更多的可能性，关注领域的最新进展，寻找最佳方案，参考不同方案实现方式，完成系统设计，并将最终测试结果与前人方案进行比较。

## 五、毕业论文的进度安排

2018年1月14日—2018年3月5日 确定选题，查阅相关文献，进行相关文献翻译，撰写开题报告；

2018年3月6日—2018年3月15日 收集临床医疗文本的数据及预处理；

2018年3月16日—2018年4月15日 相信确定论文核心研究内容，确定研究方法，初步进行系统设计；

2018年4月16日—2018年5月5日 完成系统设计并撰写毕业论文初稿；

2018年5月6日—2018年5月15日 修改、完善系统和毕业论文，提交指导老师审核；

2018年5月16日—2018年5月23日 完成毕业设计的定稿，整理毕业设计的全部资料；

2018年5月23日—2018年6月6日 提交所有材料，准备答辩。

## 六、毕业论文的主要参考文献

[1] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aur´elie N´ev´eol. 2017. LIMSI-COT at SemEval- 2017 Task 12: Neural architecture for temporal information extraction from clinical narratives. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 595–600. http://www.aclweb.org/ anthology/S17-2098.

[2] Andre Lamurias, Diana Sousa, Sofia Pereira, Luka Clarke, and Francisco M Couto. 2017. ULISBOA at SemEval-2017 Task 12: Extraction and classification of temporal expressions and events. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1016–1020. http://www.aclweb.org/ anthology/S17-2179.

[3] Po-Yu Huang, Hen-Hsen Huang, Yu-WunWang, Ching Huang, and Hsin-Hsi Chen. 2017. NTU-1 at SemEval-2017 Task 12: Detection and classification of temporal events in clinical data with domain adaptation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1007–1010. http://www. aclweb.org/anthology/S17-2177.

[4] Artuur Leeuwenberg and Marie-Francine Moens. 2017. KULeuven-LIIR at SemEval-2017 Task 12: Crossdomain temporal information extraction from clinical records. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval- 2017). Association for Computational Linguistics, Vancouver, Canada, pages 1027–1031. http:// www.aclweb.org/anthology/S17-2181.

[5] Yu Long, Zhijing Li, Xuan Wang, and Chen Li. 2017. XJNLP at SemEval-2017 Task 12: Clinical temporal information extraction with a hybrid model. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1011–1015. http://www.aclweb.org/ anthology/S17-2178.

[6] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 Task 12: A framework for cross-domain clinical temporal information extraction. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval- 2017). Association for Computational Linguistics, Vancouver, Canada, pages 1021–1026. http:// www.aclweb.org/anthology/S17-2180.

[7] Sarath P R, Manikandan R, and Yoshiki Niwa. 2017. Hitachi at SemEval-2017 Task 12: System for temporal information extraction from clinical notes. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1002–1006. http://www.aclweb.org/ anthology/S17-2176.

[8] William Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal Annotation in the Clinical Domain. Transactions of the Association for Computational Linguistics 2:143–154.

[9] Marcia Barros, Andr´e Lam´urias, Gonc¸alo Figueir´o, Marta Antunes, Joana Teixeira, Alexandre Pinheiro, and Francisco Couto. 2016. ULISBOA at SemEval-2016 Task 12: Extractions of temporal expressions, clinical events and relations using IBEnt. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1196>.

[10] Cyril Grouin and V´eronique Moriceau. 2016. LIMSI at SemEval-2016 Task 12: machine-learning and temporal information to identify clinical events and time expressions. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1190>.

[11] Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and YonghuiWu. 2016. UTHealth at SemEval-2016 Task 12: Temporal information extraction from clinical notes - uthealth’s system for the 2016 clinical tempeval challenge. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1201>.

[12] Peng Li and Heng Huang. 2016. UTA MLNLP at SemEval-2016 Task 12. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1197>.

[13] Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 Task 12: Extraction of temporal information from clinical documents using machine learning techniques. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1192>.

[14] Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. GUIR at SemEval-2016 Task 12: Temporal information extraction from clinical narratives. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics. <https://aclweb.org/anthology/S/S16/S16-1194>.

[15] Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, Wendy W Chapman. 2015. BluLab Temporal Information Extraction for the 2015 Clinical TempEval Challenge. <https://aclweb.org/anthology/S/S15/S15-2137>.

[16] Hegler Tissot, Genevieve Gorrell, Angus Roberts, Leon Derczynski, Marcos Didonet Del Fabro. 2015. UFPRSheffield: Contrasting Rule-based and Support Vector Machine Approaches to Time Expression Identification in Clinical TempEval. <https://aclweb.org/anthology/S/S15/S15-2141>.

七、其他要求

无