

施伟

+49(0)681 302 70027 ◊ w.shi@coli.uni-saarland.de ◊ www.weishi.de

Room 3.11 C7.2, Campus, Universität des Saarlandes

66123 Saarbrücken, Germany

研究兴趣

篇章关系, 文本分类, 关系抽取, 自然语言生成, 以及运用机器学习方法解决各种自然语言处理任务.

教育背景

萨尔大学, 萨尔布吕肯, 德国 2016.07 - 2020.10(expected)

博士研究生, 语言科学和技术系

导师: Prof. Dr. Vera Demberg

中国科学院大学, 北京, 中国 2013.09 - 2016.06

硕士, 计算机科学与技术专业, 中国科学院自动化研究所 (CASIA).

硕士课题: 基于深度语义表示的短文本情感分析研究

导师: 郝红卫研究员, 徐波研究员

武汉大学, 湖北, 中国 2008.09 - 2012.06

学士, 自动化专业, 动力与机械学院.

学士课题: 带时滞的多智能体分布式一致性研究

导师: 丁李副教授

研究经历

初级研究员 2016.07 - 现在

萨尔大学 @ SFB-1102

· SFB-1102 "Information Density and Linguistic Encoding"

Project B2: Cognitive Modelling of Information Density for Discourse Relations

- 系统分析了隐性篇章关系分类任务中常用的数据拆分设定, 验证了常用的数据分割设置会有导致错误结论的风险, 提出了交叉验证并验证了其对这个任务的必要性.
- 隐性篇章关系的人工标注十分困难, 而且机器学习方法对于显性关系的抽取具有较高的准确率. 由此我们反向翻译平行语料中的非英语 (法语, 德语, 捷克语) 文本, 利用其中的显性标签以获取准确率较高的自动标注隐性篇章关系英文数据, 在 Penn Discourse Treebank(PDTB) 上的实验结果也证实了所提出的跨语言数据增强 (Multi-lingual data augmentation) 的有效性.
- 模拟隐性篇章关系的人工标注流程中标注者先添加连词再标注关系的方法, 我们提出了带记忆力模块的 Seq2seq 分类模型, 在试图生成显性文本的同时预测隐性篇章关系, 在 PDTB 的不同的设置上均取得了 SOTA 的表现.
- Next Sentence Prediction(NSP) 是预训练模型 Bert 中的重要组成部分, 而获取准确的下文对于篇章关系来说也至关重要, 因为常见的 encoder 并不能成功掌握接下来的事件观点和主张. 我们通过对比 Bert 以及剔除 NSP 子任务后 Bert 的表现证实 NSP 在同领域以及跨领域的隐性篇章关系中的重要性.
- (Ongoing) BioDRB 是标注于生物医学领域的篇章关系数据集. 但由于领域的鸿沟, 现有的方法无法在 BioDRB 上获得较好的表现. 我们尝试使用信息检索和构建知识图谱的方式来获取实体三元组 (SPO triples), 通过显性篇章关系的标签以及无监督学习和带实体信息的预训练模型 (e.g. ERNIE) 来帮助领域的适应 (Domain Adaptation).

研究助理 2014.07 - 2016.06

中国科学院自动化研究所 (CASIA)

· 基于海量数据检索的短文本智能分析系统 (十二·五规划重大专项).

- 负责专题分析模块开发: 情感分析和关系抽取.

- 研究文本分析, 包含文本分类, 聚类和检索. 提出了带 Attention 的 LSTM 自编码联合学习模型, 在 SST 数据集五分类效果为 49.7%.
- 基于短文本内容的人物关系抽取技术研究 (横向课题, Y5D1441DE1)
- 互联网新闻定向网络爬虫
 - 特定网页爬虫: 解析网页 html, 获取链接 url, 爬取网页.
 - 全网新闻爬虫: 基于开源 Java 爬虫 WebMagic 框架, 识别种子网页, 采用 Redis 存储 url 信息, 分布式爬取种子网页链接地址, 解析标题, 时间, 内容存入数据库.

发表论文

- **Wei Shi** and Vera Demberg. *Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019, Poster), Hong Kong, China
- **Wei Shi** and Vera Demberg. *Learning to Explicitate Connectives with Seq2Seq Network for Implicit Discourse Relation Classification*. In Proceedings of The 13th International Conference on Computational Semantics (IWCS-2019, Oral), Gothenburg, Sweden.
- **Wei Shi**, Frances Yung and Vera Demberg. *Acquiring Annotated Data with Cross-lingual Explicitation for Implicit Discourse Relation Classification*. In Proceedings of the 7th Workshop on Discourse Relation Parsing and Treebank (DISRPT@NAACL-2019, Oral), Minneapolis, USA.
- Fangzhou Zhai, Vera Demberg, Pavel Shkadzko, **Wei Shi** and Asad Sayeed. *A Hybrid Model for Globally Coherent Story Generation*. In Proceedings of the 2nd Workshop of Storytelling (StoryNLP@ACL-2019, Oral), Florence, Italy.
- **Wei Shi**, Frances Yung, Raphael Rubino, and Vera Demberg. *Using Explicit Discourse Relation Connectives in Translation for Implicit Discourse Relation Classification*. In Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP-2017, Oral), Taipei, Taiwan.
- **Wei Shi** and Vera Demberg. *On the Need of Cross Validation for Discourse Relation Classification*. In Proceedings of The 15th Conference of the European Chapter of the Association of Computational Linguistics (EACL-2017, Poster), Valencia, Spain.
- Peng Zhou, **Wei Shi**, Jun Tian, Zhengyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. *Attention-based Bidirectional Long Short-term Memory Networks for Relation Classification*. In Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016, Poster), Berlin, Germany.

会议展示

- Katharina Spalek, Beate Bergmann, **Wei Shi** and Vera Demberg. *'Only' Increases Expectations for Causal Coherence Relations*. Talk presented at Discourse Expectations: Theoretical, Experimental, and Computational Perspectives (DETEC), September 27-28, 2019, Berlin, Germany,

学术活动

PC member of IEEE TCSS, CoNLL-2018, 2019, NAACL-HLT-2019, ACL-2019, 2020 etc.