**哈尔滨工业大学**

**硕士学位论文开题报告**

**题 目：基于Attention机制的跨语言文本蕴含识别**

**院 （系） 计算机科学与技术**

**学 科 计算机科学与技术**

**导 师 郑德权**

**研 究 生 王贺伟**

**学 号 17S103167**

**开题报告日期 2018.9.6**

**研究生院制**

目录

[1. 课题来源及研究的背景和意义 3](#_Toc518126459)

[1.1 课题的来源 3](#_Toc518126460)

[1.2 课题研究的背景和意义 3](#_Toc518126461)

[2. 国内外在该方向的研究现状及分析 3](#_Toc518126462)

[2.1 国外研究的现状 3](#_Toc518126463)

[2.2 国内研究现状 4](#_Toc518126464)

[2.3 国内外文献综述的简析 4](#_Toc518126465)

[2.4 重要数据集和评测 5](#_Toc518126466)

[3. 主要研究内容 5](#_Toc518126469)

[3.1 基于LSTM句子表示学习的文本蕴含识别 5](#_Toc518126470)

[3.2 基于注意力机制的句子表示文本蕴含识别 6](#_Toc518126471)

[3.3 基于注意力机制的词对匹配文本蕴含识别 6](#_Toc518126472)

[3.4 跨语言文本蕴含识别 6](#_Toc518126473)

[4. 已完成的研究工作 7](#_Toc518126474)

[5. 研究方案及进度安排，预期达到的目标和取得的研究成果 8](#_Toc518126475)

[5.1 研究方案 8](#_Toc518126476)

[5.2 预期达到的目标和取得的研究成果 10](#_Toc518126477)

[5.3 进度安排 10](#_Toc518126478)

[6. 为完成课题已具备和所需要的条件和经费 11](#_Toc518126479)

[7. 预计研究过程中可能遇到的困难和问题，以及解决的措施 11](#_Toc518126480)

[8. 参考文献 12](#_Toc518126481)

# 课题来源及研究的背景和意义

## 课题的来源

本课题来源于之前在线Mooc 自动评分项目。

## 课题研究的背景和意义

随着自然语言处理的发展，对文本表层信息的理解已经无法满足人们的需求，如何对文本进行语义层面的理解，或者如何获得文本之间的推理关系对自然语言处理落地产品起着重要的作用，文本之间的相互关系又称为文本蕴含，在自然语言处理的很多实际应用中，经常需要进行文本的相似或者蕴含匹配，比如在问答系统中，当用户询问了“谁是网易公司的CEO”时，系统需要对问题和文本“创立网易公司的是谁。”进行匹配，来给出答案“丁磊”。同样，问题与候选答案，候选答案和相关文档都存在蕴含推理关系。

文本蕴含的概念有Dagan和Glickman在2004年提出，文本蕴含是研究判断前提和假说关系的问题，关系分为蕴含，对立，中立，假设前提是p,假说是h,蕴含指的是由p可以推理h 正确，对立指的是由p可以推理h不正确，中立是指由p无法判断h 的真假。文本蕴含的概念应该与复述区分开来，复述是两个陈述彼此之间能相互推导，而文本蕴含是单向的有p 可以推断h 的真假；同时文本蕴含与文本相似应该区别开来，虽然传统方法常常将文本相似作为判断文本蕴含重要的特征，但是很相似的句子也有可能是对立关系，因此不能和文本蕴含等价。

无论是问答系统、信息抽取，学生评价系统、机器翻译评价，自动文本摘要运用广泛，这些应用都面临着一个主要的难题：自然语言在表达形式上的多样性，即相同含义可以用不同的陈述形式来表达，同时需要一些广泛的知识才能推断文本之间的关系。

在关系抽取领域，使用文本蕴含技术扩展了抽取的模板，极大的丰富了目标关系的表达形式，同时这个也可以应用在意图识别中，跟数据库中文本进行文本蕴含的匹配能提高系统的泛化能力；在学生评分任务中，学生回答和标准答案之间的蕴含关系可以指示学生回答的完成情况, Nielsen等人据此利用文本蕴含技术建立了一套学生作业评分系统。

同时文本蕴含其实有三个研究研究问题，第一个就是本文的研究课题也就是文本蕴含识别，第二个就是文本对中蕴含关系的抽取，第三个就是文本蕴含的生成，文本蕴含本身就是要实现一个基于文本的推理引擎用于语义相关的实际任务。

本文主要研究两个方面，一个方面是中文文本蕴含，英文的文本蕴含已经到比较高的准确率，中文文本蕴含有数据集少和词汇资源缺乏的问题，另一方面是研究文本蕴含在意图识别的应用，对于问答系统中长文本的潜在意图进行判断。

# 国内外在该方向的研究现状及分析

## 国外研究的现状

近几年得益于斯坦福大学发布了更大的人工标记数据集SNLI，越来越多的基于深度学习的方法应用到文本蕴含的识别中，一系列的基于深度学习的模型远远超过之前的传统模型。首先Bowman等人在发布57k 的SNLI 数据集的同时给出了两个基准方法，一种是基于多特征分类，抽取训练集中的词对间Blue值，句子长度差，各词性重叠词个数，假设句一元和二元特征，句子间词性一元和二元特征等，构造分类器分类，一种是基于LSTM 句子表示分类的方法，构建两个独立的LSTM 网络分别获得固定长度的句子表示，将其拼接输入到多层神经网络分类器中，在SNLI上的效果分别达到78%,77%。Rocktäschel等首次将注意力机制引入文本蕴含识别，注意力机制首先在机器翻译中使用，目的是获得句子间词和词的对齐关系，Rocktäschel等认为句子中的部分信息就足够用于蕴含的识别，采用word-by-word attention 机制，类似于机器翻译中的使用，准确率达到83.5%。Wang and Jiang 等人在Rocktäschel 的基础上继续显示的加强前提和假设中词到词的匹配，Parikh et al 等人把问题重新定义，认为不需要对句子表示进行编码，单纯的利用attention机制找寻句子间构成的词对，把整句的文本蕴含问题分解为词到词的匹配，相比使用LSTM 减少了模型复杂度，Munkhdalai and Yu等在Wang and Jiang 等人的基础上使用更复杂的LSTM 模型，同时结合了之前的词匹配和序列编码模型，进一步提升了准确率。最新的Qian Chen等人进一步加强了序列编码表示模型，提出加强的LSTM模型，同时发现把句法分析信息加入到蕴含识别中能进一步提高准确率，能达到88% 的准确率。

## 国内研究现状

在国内，进行文本蕴含研究的学者有：北京大学中文系袁毓林教授[21]，中科院自动化所的宗成庆、张玉洁等，鲁东大学的罗琳、刘金凤等[22]，清华大学朱小燕老师的课题组也多次参与了英文文本蕴含的国际评测，取得了不错的成绩。北京大学教育技术系的贾继友教授对智能教学系统中的文本蕴含应用进行了研究，北京大学的徐幸、王厚峰对中文文本蕴含的推理模型进行了研究，赵红燕等人则提出用多特征进行文本蕴含识别[23]。

## 国内外文献综述的简析

在已有的各种文本蕴含识别和推理方法中，都不可避免地需要依赖各种知识，其中主要是所谓的文本蕴含知识，这些知识一般按照规则的形式来表示，也即“蕴含规则”，包括词汇短语蕴含规则、模板蕴含规则（描述的是带有变量的文本模板之间的蕴含关系）。而这些蕴含规则的获取也主要依赖于现有的各种知识以及这些知识在现实语料中的应用。这些知识包括：各种词典：如WordNet、FrameNet、VertNet等，利用这些词典，不仅可以计算词和词之间的相似度，还可以获取词的各种语义信息等。

然而基于知识库计算词语相似度的方法有着较大的局限性，对于知识库中没有的词以及词语歧义问题无法进行解决。此外传统基于词语相似度的文本蕴含识别方法都没有考虑词语的位置信息，这样导致距离远的两个词和距离近的两个词在一个句子中的所占比重一样。

尽管经过这几年的积累，英文的文本蕴念已经有了相当规模的知识库，例如：“DIRT”、“WikiRules”、“FRED”、“Downward entailing operators”等，而且这些知识库也为英文的文本蕴含识别技术研究建立了坚实的基础。但由于各种知识的不足，以及知识的不充分利用，还有已标注的文本蕴含对训练语料的规模限制，都成为文本蕴含识别的性能不高的重要原因。

此外，深度学习在文本蕴含中的应用才刚刚起步，亟待研究者去尝试深度学习在文本蕴含识别中的效果。

## 重要数据集和评测

文本蕴含的研究少不了相关评测和标准数据集的出现，标准数据集能很好

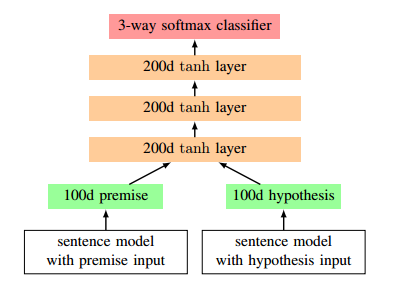
## 反映模型的能力，其中比较著名的就是RTE评测，主要处理来自问答系统，关系抽取，信息检索和文摘这四个与文本蕴含相关的领域，来自SemEval 的数据集SICK 提供4500 条数据作为训练数据，比较重要的就是Bowman等人在发布57k 的SNLI 数据集，人工标注了以下三种标签：entailment, contradictory，neural ，主要是来自对照片的文字描述，这个数据集中内容相比无限制的RTE 较为简单，不需要太多步骤的推理得到，引起了深度学习在文本蕴含识别问题的研究高潮，最近斯坦福又发布了新的Muti-NI 数据集，和SNLI 格式相同，加入一些口语和书面的文本，体裁上和SNLI不同。

## 最新的相关评测有两个，一个是CCL2018 中文蕴含识别，提供20000条中文蕴含文本对，一个是CIKM 跨语言文本匹配，提供20000条源语言为英文的文本对，以及1400 目标语言为西班牙语的文本对，对5000条目标语言为西班牙语的文本对识别。

# 主要研究内容

本文主要研究使用深度学习技术识别文本蕴含，其中基于LSTM 的表示模型和注意力机制的词对匹配模型是研究的主要方法，最新的方法中引入了加强的LSTM架构特别为了识别文本蕴含，同时又尝试将wordnet 知识和句法知识引入神经网，在关注单语的文本蕴含识别技术的同时，又关注跨语言文本蕴含识别，尝试提升在中文，西班牙文等缺乏大规模标记数据集情况下的识别能力。

## 基于LSTM句子表示学习的文本蕴含识别

Bowman等人在发布SNLI 的同时提出了一个baseline方法，分别通过LSTM 或者CNN 获得两个句子的表示，之后对其进行拼接，再通过多层神经网，最后通过softmax 进行预测，这是典型的孪生网的结构，Rocktäschel等又在这的基础上提出把两个表示连接在一起，先让前提句p通过一个LSTM，取之后时刻中记忆层的数据作为另一个LSTM的初始值，这个LSTM 的输入是假设句H。这一类方法主要是通过学习整个句子的语义表示来进行多分类。

## 基于注意力机制的句子表示文本蕴含识别

Rocktäschel等人认为虽然之前句子表示学习能把一个边长的句子表示成一个向量，但是固定长度的向量对句子的信息表达能力有限，Rocktäschel等人因此首次在蕴含识别任务上引入注意力机制，动机在于想找到加假设句H 中的 第k 个词与前提句P中最相关的词构成对齐, 这样能更加强的糅合两个句子向量，最终获得假设句H 相对于P的向量表示，然后进行多分类

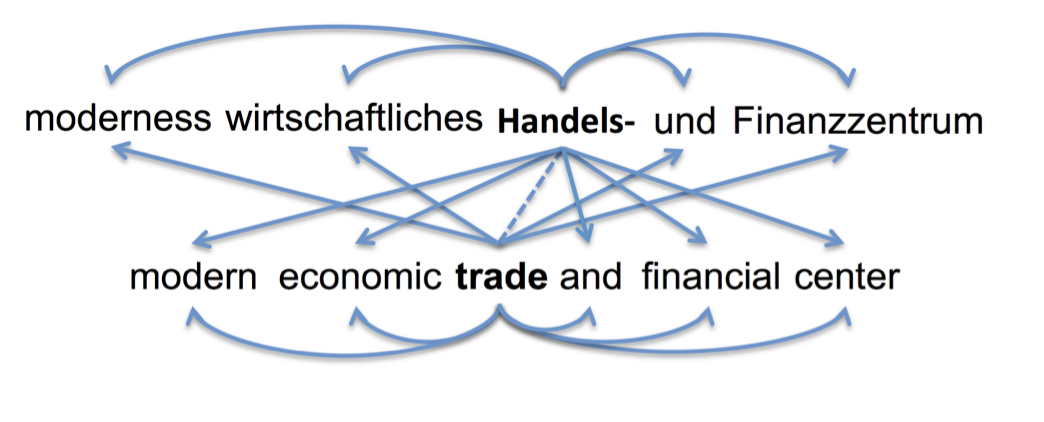
## 基于注意力机制的词对匹配文本蕴含识别

Parikh et al.(2016) 指出NLI 问题之前的解法是获得句子向量去做匹配，其实很多情况下不需要获得整个句子的语义，只需要比较特定词和局部词对的关系即可，跟之前attention 机制的直觉相同，这个模型不需要关心词序，纯粹的应用attention 机制，找到前提句中的词和假设句中词的软对齐关系，把这些词对向量分别进行线性操作，最后merge 在一起，进行分类，这个模型用较少的参数达到了较好的结果。

## 跨语言文本蕴含识别

之前的评测中涉及中文，西班牙语的文本蕴含识别，相比英语，中文和西班牙语的人工标记语聊很有限，这样就想到采用跨语言的研究方法，提高在中文和西班牙语上的识别准确，但是区别是他需要外部的词典，但是我们只有平行语料，可以对比两种方法的差异，以下训练双语词向量的常用方法。

1. bilingual mapping：先独立的得到双语词向量，然后根据双语词向量学习单语之间词向量的转换
2. monolingual adaptation：这一种是通过资源丰富的源语言预训练词向量，来影响目标语言词向量生成的过程，无监督的从平行语料中找到对齐关系
3. bilingual training：前面两种方式都是在修改源语言预训练表示，但是这种方式是同时学习词在双语的词表示，例如"good"虽然是英语但是在双语训练的时候和中文“很好”有着聚集效应



# 已完成的研究工作

目前完成了国内外文本蕴含识别的方法的调研工作，实现并改进A Decomposable Attention Model for Natural Language Inference，并在CIKM 数据集上进行试验测试，取得了较好结果。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 语言 | 词汇数量 | 未登录词数量 | 训练集合准确率 | 开发集准确率 | 参数 |
| English | 2806 | 3 | 0.907 | 0.895 | 100轮 |
| spanish | 5432 | 418 | 0.89 | 0.860 | 100轮 |

结果分析如下：

1. 现在直接使用训练集中的目标语言训练结果不好，原因可能是翻译源语言中带来的噪声，因此需要使用跨语言方法

2. 通过分析数据，不匹配：匹配 比例为 3：1，有类别不平衡问题

3. 2万的数据量还是较少需要引入更多的外部知识信息

# 研究方案及进度安排，预期达到的目标和取得的研究成果

## 研究方案

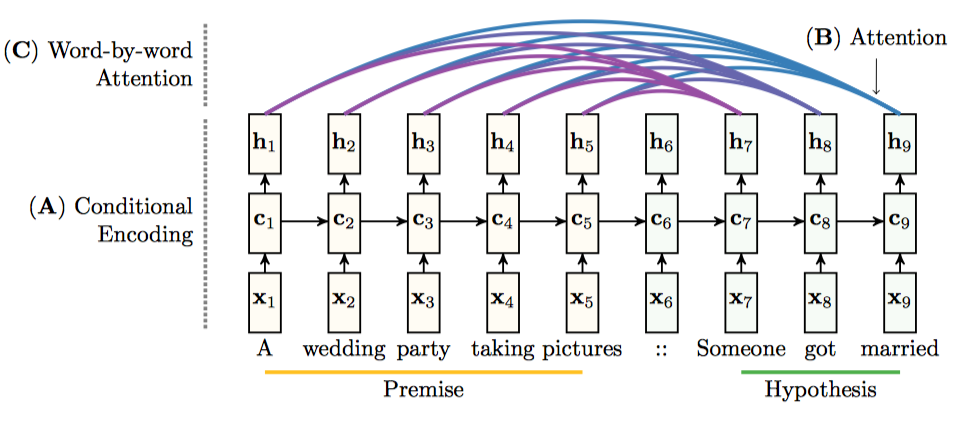
5.1.1基于注意力机制句子表示的文本蕴含识别

Rocktäschel等人在之前LSTM 句子表示模型的基础上，引入了word-by-word 注意力机制，目的在于编码假设句的时候局部的融合前提句的信息，融合的方式是如下图中的B，因为认为隐层状态向量表示之前句子的信息能力有限，这里引入Attention机制，B中对最后输出向量加入上一个句子的输出(h1,h2 …. h\_5) 的attention表示，这样更能引入premise 的信息，C进一步在第二个句子每一时刻的生成过程中引入之前句子的信息，动机来自attention 在机器翻译中的应用，寻找假设句中的词和前提句中所有词的对齐关系，对于假设句的第k个词，有对应输入对齐向量，其中ak 表示 第k个词的对齐向量，aij 表示 对应前提句P 中的第j个字的权重, hs 代表前提句的第j个隐层向量。



下面是怎么求解这个权值aij ，这个权值是对齐权值ekj 的归一化值，而这个ekj 是结合了前提句中第j个词的隐层输出 hjs ，第二个LSTM隐层输出hkt ，以及上一个时刻的对齐向量 hk-1a 共同表示。





5.1.2基于注意力机制的词对匹配文本蕴含识别

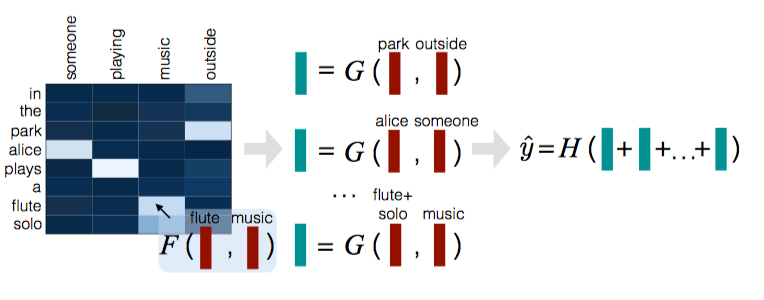
Parikh et al.(2016) 指出NLI 问题之前的解法是获得句子向量去做匹配，甚至抛弃了LSTM，其实很多情况下不需要获得整个句子的语义，只需要比较特定词和局部词对的关系即可,这个模型把整个文本蕴含的问题拆分为一个词的匹配问题，单纯的学习attention 对齐，分为如下三步：

1. Input representation: 在最基础的模型中，直接使用词向量序列表示这个句子，Parikh et al.(2016) 也提出可以使用 inter-sentence Cheng et al. (2016) 句子表示，在下图为蓝色标记
2. Attend: 这一步通过 F 函数获得词对的Attention权重，可以从下图看到，(park,outside) park 是promise 中的词，outside 是 h 中的词，其实这里park 对应的应该是 h 中所有词乘以其的权重 ,例如这里 (ai,bi) 构成了词对，这里ai是p中的每个词，这里G(a,b) 是直接把两个词的向量拼接在一起，仍然的到跟之前个数相同的向量



1. Aggregate： 这一步是把同一个句子的向量G 相加，这样得到两个句子向量v1,v2, 把这两个向量从到分类器 H 中进行分类





## 预期达到的目标和取得的研究成果

通过一系列的实验，对比文本蕴含识别各方法的优缺点，提高文本蕴含识别的准确率，实现跨语言的文本蕴含，使用迁移学习和双语词向量的方式提升中文蕴含识别能力，在新数据集下Mutil-NLI 给出新的结果，同时在中文数据集下取得跟英文数据集下相似的结果。

## 进度安排

|  |  |
| --- | --- |
| 时间 | 进度 |
| 2018.05~2018.07 | 调研国内外文本蕴含识别方法，实现成熟基础的基于attention的模型 |
| 2018.08~2018.10 | 尝试跨语言的文本蕴含，使用迁移学习和双语词向量的方式提升中文蕴含识别能力。 |
| 2018.09~2018.12 | 通过使用最新的Transformer模型提升单语言模型效果 |
| 2019.01~2019.03 | 尝试将WordNet，FrameNet 等外部资源引入神经网 |
| 2019.04~2019.05 | 整理之前工作，尝试在新数据集下Mutil-NLI 给出新的结果，同时在中文数据集下取得跟英文数据集下相似的结果 |
| 2019.06~2019.07 | 论文撰写，准备答辩 |

# 为完成课题已具备和所需要的条件和经费

已具备条件：

1. 斯坦福文本蕴含语料SNLI
2. CIKM 预料
3. CCL 2018 文本蕴含语料
4. Pytorch及GPU Titan X

# 预计研究过程中可能遇到的困难和问题，以及解决的措施

在现有的复杂模型上的改进难度比较大，同时目标语言文本蕴含训练数据不足，导致深度学习方法产生较严重的过拟合。可以限制神经网络的复杂度，设计精简的网络结构来避免过拟合问题。同时在怎么向神经网络中引入WordNet，FrameNet 知识上现有可参考模型较少。

# 参考文献

[1] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference[J]. Computer Science, 2015.

[2] Wang S, Jiang J. Learning Natural Language Inference with LSTM[C]// Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016:1442-1451.

[3] Parikh A P, Täckström O, Das D, et al. A Decomposable Attention Model for Natural Language Inference[J]. 2016:2249-2255..

[4] Tim Rocktaschel, Grefenstette E, Hermann K M, et al. Reasoning about Entailment with Neural Attention[J]. 2015.

[5] Liu P, Qiu X, Chen J, et al. Deep Fusion LSTMs for Text Semantic Matching[C]// Meeting of the Association for Computational Linguistics. 2016:1034-1043.

[6] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for Natural Language Inference[J]. 2016:1657-1668.

[7] Chen Q, Zhu X, Ling Z H, et al. Natural Language Inference with External Knowledge[J]. 2017.

[8] 郭茂盛, 张宇, 刘挺. 文本蕴含关系识别与知识获取研究进展及展望[J]. 计算机学报, 2017, 40(4):889-910.

[9] Cheng J, Dong L, Lapata M. Long Short-Term Memory-Networks for Machine Reading[J]. 2016.

[10] Shi C, Liu S, Ren S, et al. Knowledge-Based Semantic Embedding for Machine Translation[C]// Meeting of the Association for Computational Linguistics. 2016:2245-2254. [11] Shen T, Zhou T, Long G, et al. DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding[J]. 2017.

[12] Bowman S R, Gauthier J, Rastogi A, et al. A Fast Unified Model for Parsing and Sentence Understanding[J]. 2016.

[13] Bordes A, Usunier N, Garcia-Duran A, et al. Translating Embeddings for Modeling Multi-relational Data[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:2787-2795.

[14] Chen Q, Zhu X, Ling Z H, et al. Recurrent Neural Network-Based Sentence Encoder with Gated Attention for Natural Language Inference[J]. 2017:36-40.

[15] Lin Z, Feng M, Santos C N D, et al. A Structured Self-attentive Sentence Embedding[J]. 2017.

[16] Pennington J, Socher R, Manning C. Glove: Global Vectors for Word Representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014:1532-1543.

[17] Shi C, Liu S, Ren S, et al. Knowledge-Based Semantic Embedding for Machine Translation[C]// Meeting of the Association for Computational Linguistics. 2016:2245-2254.

[18] Mehdad Y, Negri M, Cabrio E, et al. Using lexical resources in a distance-based approach to rte[C]//Proceedings of the TAC 2009 Workshop on TE. November. 2009, 17.

[19] Burchardt A, Frank A. Approaching textual entailment with LFG and FrameNet frames[C]//Proc. of the Second PASCAL RTE Challenge Workshop. 2006.

[20] Wang Z, Hamza W, Florian R. Bilateral Multi-Perspective Matching for Natural Language Sentences[J]. 2017.