文章编号：1003-0077（2017）00-0000-00

中文句法异构蕴含语块标注和边界识别研究

金天华[1] 姜珊[1] 赵美倩[1] 刘璐[1] 于东[1,2]
（1.北京语言大学 信息科学学院，北京 100083；2.北京语言大学 语言资源高精尖创新中心，北京 100083）

摘要：文本蕴含是自然语言处理的难点，其形式类型复杂、知识难以概括。早期多利用词汇蕴含和逻辑推理知识识别蕴含，但仅对特定类型的蕴含有效。近年来，利用大规模数据训练深度学习模型的方法在句级蕴含关系识别任务上取得优异性能，但模型不可解释，尤其是无法标定引起蕴含的具体语言片段。本文研究文本蕴含成因形式，归纳为词汇、句法异构、常识三类，并以句法异构蕴含为研究对象。针对上述两个问题，提出句法异构蕴含语块的概念，定义其边界识别任务。本文制定句法异构蕴含语块标注规范，建立标注数据集。在此基础上，分别建立基于规则和基于深度学习的模型，探索句法异构蕴含语块的自动识别方法。实验结果表明，本文提出的深度学习模型能有效发现蕴含片段，为下一步研究提供了可靠的基线方法。

关键词：文本蕴含；句法异构；语块标注

中图分类号：TP391 **文献标识码：**A

Chinese Chunked-based Heterogeneous Entailment Parser and Boundary Location

Jin Tianhua[1], Jiang Shan[1], Zhao Meiqian[1], Liu Lu[1], Yu Dong[1,2]
(1. College of Information Science, Beijing Language and Culture University, Beijing 100083, China; 2. Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China)

Abstract：Recognize textual entailment(RTE) is a difficult task for natural language processing. In early period, RTE was based on lexical knowledge and logical reasoning knowledge, which only worked on specific type. Recently, deep learning models become a mainstream. But this approach can not locate concrete linguistic fragments in sentences. This paper divides textual entailment into three main categories: lexical entailment, chunked-based heterogeneous entailment and common-sense entailment. So we propose the concept of chunked-based heterogeneous to define the task of recognizing chunk boundary. We establish a chunks annotation standard and labeled dataset. Then we build a rule-based model and a deep learning model respectively to explore the automatic recognition methods. The experimental results show that the deep learning model proposed in this paper can discover the entailment fragments effectively and provide a reliable baseline for the following research.

Key words: textual entailment; syntactic heterogeneous; chunks-labeling

0 引言

文本蕴含定义为一对文本之间的有向推理关系[1]，其中蕴含前件记作 P（Premise），蕴含后件记作 H（Hypothesis）。文本蕴含识别（Recognizing textual entailment, RTE）是基于语义理解，对两个句子之间的蕴含和矛盾关系做出判断的任务。文本蕴含作为语义理解的基础

收稿日期：201*-*-*；**定稿日期**：201*-*-*

基金项目：北京语言大学语言资源高精尖创新中心（TYR17001J）；国家社科基金重点项目（16AYY007）；中央高校基本科研业务费专项资金资助项目（北京语言大学梧桐创新项目：17PT05）

任务，可以建立起不同文本之间的语义推理关系网，促进关系识别、事件抽取、自动文摘等任务的发展，同时在问答、文本挖掘、阅读理解、信息检索等应用领域发挥关键作用。

文本蕴含识别早期的研究工作[2][3][4]多从词汇蕴含角度出发，探索近义词、上下位词、整体和部分等词汇关系在文本蕴含识别中的应用。然而单纯词汇蕴含并不能完全涵盖文本蕴含的所有范畴。目前对文本蕴含成因的定量研究仍处于初步阶段。另一方面，近年来，随着 SICK[5]、SNLI[6]、MultiNLI[7]等数据集的提出，用机器学习方法建立 end-to-end 模型判断整句的句法蕴含关系成为研究热点[8][9][10]。此类模型可以有效判断整句级别的蕴含关系，但无法确定引起蕴含的关键片段位置，其结果缺乏可解释性。因而大大削弱了其应用价值。

针对第一个问题，本文将蕴含成因归纳为词汇蕴含、句法异构蕴含、常识和社会经验三种类型。我们翻译并校对了 SNLI 数据集中的 3766 条蕴含句对数据，由人工对其蕴含成因类型进行标注，其中词汇蕴含仅占 31.5%，说明词汇蕴含只是蕴含的一种类型。常识和社会经验占比为 29.1%，由于常识的概念模糊，包含的信息粒度大，因而不在本文讨论范围内。标注结果中，句法异构导致的蕴含占比最多，达到占 39.4%，故本文以此为研究对象。

所谓句法异构蕴含，是指通过语言的位移、添加、删除、替换等手段[11]对 P 的形式进行有选择地筛选和强调得到 H，P 和 H 的句法变化使得它们在语义上具有蕴含关系，则 P 和 H 是句法异构蕴含。如下文 T1、T2 的两组例句就是句法异构蕴含。

值得一提的是，句法异构蕴含与复述有本质区别。句法异构蕴含不追求语义信息的完整性和一致性。分析发现，句法异构蕴含会保留或概括 P 中需要强调的、不可省略的部分，而删除不需要强调的部分。例如，T1 的 H 省略了 P 的地点状语“在蓝色卡车旁边”，突出强调了动词性谓语“拍摄”，这两句话具有句法异构蕴含关系。T2 的 H 省略了 P 的谓语“拍摄”和宾语“电影”，而 H 的谓语和宾语是由 P 的地点状语“在蓝色卡车旁边”充当。P 和 H 是句法异构的，他们之间也具有句法异构蕴含关系。

T1:P:一群人在蓝色卡车旁边拍摄电影。
H:一群人在拍摄电影。
T2:P:一群人在蓝色卡车旁边拍摄电影。
H: 一群人在蓝色卡车旁边。

本文研究导致蕴含现象的句法异构类型，通过观察大量蕴含句对，分析归纳得出句法异构类型包括结构变化和省略变化。结构变化又分为成分抽取、从句抽取、语序变化；省略变化分为省略修饰语和省略中心语。

针对第二个问题，本文需深入语料内部确定引起整句级别蕴含关系的关键片段，认为这些关键片段可以被称为句法异构蕴含语块。语块的概念最早由 Skechan 提出[12]，指兼具词汇和句法特征的半固定的语言结构。在本文中，句法异构蕴含语块是 P 和 H 中句法成分或句法结构不同，且具有蕴含关系的部分。蕴含语块可以是句中充当句法成分的词、短语，甚至是整个单句或者复句中的某个小句。例如“香甜的苹果—苹果”“漫长的夜晚—夜晚”都属于从“adj+的+n”到“n”的变化，那么“adj+的+n”和“n”就分别是 P 和 H 的句法异构蕴含语块。

显然，句法异构蕴含语块的确认依赖于蕴含成因的研究。从机器学习角度来说，句法异构蕴含语块的识别问题可以转化为边界识别问题。本文主要采用深度学习模型，处理整合 P 和 H 的蕴含信息用于识别蕴含边界下标。受 Wang[20]的启发，我们利用 match_LSTM 计算获得包含 P 和 H 蕴含信息的表示向量，作为 Ptr-Net 的输入，进而寻找蕴含边界。

本文首先介绍国内外蕴含类型研究，在此基础上针对句法异构蕴含现象进行分析总结，归纳得到句法异构蕴含类型。接着介绍我们在蕴含语块标注方面的工作，用标注结果归纳得到一套简单有效的规则系统，并将该规则系统与深度学习模型应用于语块边界自动识别，分析比较两者者在实验上的有效性，并对论文工作进行总结和展望。

1 相关工作

现有的文本蕴含数据集都是为了解决文本蕴含问题而开发，然而目前没有专门研究蕴含类型成因的数据集。早期文本蕴含评测 RTE-1 至 RTE-3[13][14][15]及 SciTail[16]将文本蕴含视为二分类任务，句子对之间只存在蕴含和中立两种关系。近年来的大规模数据集，如 SNLI，MultiNLI 等，把文本蕴含关系分为“蕴含”“矛盾”“中立”三种，以供学界研究文本蕴含的整体类型。截止本文写稿期间，我们尚未看到单独讨论蕴含成因类型的研究和讨论句子内部导致蕴含关系的语言片段的研究。

在英文研究领域，Ido Dagan 和 Oren Glickan[17]从宏观角度把英语蕴含关系分成五类：Axion rule（公理），Reflexivity（自反性），Monotone extension（单调性扩张），Restrictive extension（限制性扩张），Transitive Chaining（传递链）。这些概念较为抽象，不便理解，在具

体标注过程中难以实践。

在中文研究领域，RITE-3 任务针对中文语料提出了 19 类蕴含现象和 9 类矛盾现象[18]，包含了近义词、反义词、上下位词等词汇类别和从句、时态等句法类别。任函[19]提出了面向汉语文本推理的语言现象标注类别，包含了 20 个类别的语言现象体系，同样包含了同义词（近义词）、上下位词、反义词等词汇类别，该类别体系以词汇为主，句法特征的内容不多，仅有一个结构变化，较为笼统。

以上研究是从语言学角度对蕴含类型进行区分，没有考虑数据的实际情况，容易出现某些类别数据稀疏的情况。因此，本文将数据处理和蕴含类型相结合，利用现有数据集，深入语料寻找导致蕴含关系的语言片段，探究蕴含现象成因。

2 句法异构蕴含成因研究

我们根据汉语句法特点把句法异构蕴含的成因归纳成两类：一，结构变化：成分抽取、从句抽取、语序变化；二，省略变化：省略修饰语、省略中心语。这两个类别不一定独立存在的，可以同时存在。句法异构蕴含成因类型汇总如表 1 所示。

2.1 结构变化

汉语以语序和虚词作为主要语法手段[20]，语序变化可以同时改变句子的表层结构和深层结构，也就是既改变句子的形式，又改变句子的意义。除了语序变化外，成分抽取、小句抽取也属于结构变化。

语序变化：“语序”不仅是表示语法结构、语法意义的形式，也是言语表达或修辞的手段[21]。语序变化类句法异构蕴含就是指由语法结构内部成分的线性顺序发生变化导致的蕴含。例如：

T3：P：三个女人和一个小女孩在和小狗玩。

H：与小狗玩耍的女人们。

P 属于“施受谓”语序，施事是“三个女人和一个小女孩”，受事是“一只小狗”，“谓”指谓语“玩”。在 H 中受事“小狗”谓语“玩耍”被提前到施事“女人们”前面。同时，H 把一个陈述句变成了短语。

T4：P：一家人正走在一些很大的独立的几何雕塑下面。

H：人们在一些非常大的雕塑下行走。

P 属于“主动——施谓”语序，“动”指动词，“谓”指谓词，在动词后面有一个表示地点的状语，H 把句尾的地点状语提前到动词前面，两句话的语序发生了改变。

成分抽取：从 P 中把主谓宾结构的某一部分抽取出来，单独成句。被抽取出来的结构如果是一个定中结构，有可能变成一个简单的主谓句，也有可能变成一个存在句。例如：

T5：P：一个穿着黄色毛衣的年轻人看着那张上面摆着各种花的桌子。

H：这里有个人。

P 的主语“一个穿着黄色毛衣的年轻人”被抽取出来，省略修饰后单独成句，H 是一个表示人物存在的句子“这里有个人”。

T6：P：一个穿着黑色裤子没穿衬衫的男孩儿正在玩一个白色的气球。

H：男孩穿着黑色裤子。

P 的主语“一个穿着黑色裤子没穿衬衫的男孩儿”被抽取出来，省略部分修饰语后变成一个简单的主谓句 H，“男孩穿着黑色裤子”。

小句抽取：在有多个小句的复句中抽出某一个小句，单独成句，一般情况下，我们会选择保留包含完整信息的小句，而省略作为从属地位补充信息的小句。例如：

T7：P：男人和女人在海滩上漫步，身后是绚丽的晚霞。

H：一个男人和一个女人在海滩上散步。(NULL)

P 是由一个主谓小句和一个表示背景信息的小句构成的，H 省略了表示背景信息的小句。

T8：P：小男孩在哭，因为他被雪球击中了。

H：小男孩在哭。(NULL)

同理，P 由一个包含了完整信息的主谓小句和一个表示原因的小句构成，H 省略了表示原因的小句。

2.2 省略变化

语言具有递归性，相同或不同的语言结构层层嵌套，结构规则重复使用而不会造成结构上的混乱[11]。基于语言递归性，省略部分结构而得到蕴含现象也属于句法异构蕴含。省略变化主要有省略中心语、省略修饰语两类，这容易与上一节的小句抽取混淆。两者之间的区别主要在于他们作用于不同的语言单位。小句抽取是在复句中进行，而省略则是在某一简单句内部进行。

省略中心语：在偏正结构中，省略了核心谓词，而保留修饰语。被保留的修饰语可以是形容词性成分，地点状语、时间状语等等。例如：

T9：P：年长的白人女子在她的厨房做蛋糕。

H：一位老太太在厨房里。

P 是“主谓宾”结构，在主语“一位年长的白人女子”和谓语“做”之间有地点状语“在她的厨房”，H 省略谓语和谓语的宾语“蛋糕”，只保留主语和地点状语。

T10：P：一群人划独木舟穿过热带雨林。

H：一群人正在划独木舟。

P 中有 2 个谓词性短语“划独木舟”和“穿过热带雨林”，在这里“穿过热带雨林”可以看作是中心谓词，“划独木舟”是表示方式的方式状语，H 省略了中心谓语，保留主语和方式状语，并在方式状语前加上表示动作持续的“正在”，构成一个新的主谓句。

省略修饰语：在偏正结构中省略修饰性成分，保留中心语。与上面的省略中心语相对。被省略的修饰语可以是表示地点、时间、工具的状语，也可以是表示事物性状的形容词性成分。

T11：P：一个男人在晴天晾衣服。

H：男人晾晒衣服。

H 省略了时间状语“在晴天”。

T12：P：穿着黑色衬衫的吧台服务员用一台大机器做咖啡。

H：吧台侍者在做咖啡。（省略工具）

H 省略了人物修饰语“穿着黑色衬衫的”和表示工具的状语信息“用一台大机器”。

此外，句法异构蕴含的成因不一定独立存在。比如 T13 中，P 的主语“穿着红色连帽衫的男孩”被提取出来，单独成句为 H，这属于成分抽取引发的蕴含。同时，P 中的“红色连帽衫”和 H 中的“红色衣服”属于上下位词造成的蕴含。并且，P 和 H 中，“穿着红色连帽衫（红色衣服）的男孩”和“男孩穿着红色衣服（红色连帽衫）”属于由语序调换造成的蕴含。文本蕴含语料中类似的实例说明了蕴含成因是混合的，不是单一的。

T13：P：穿着红色连帽衫的男孩走在人行道上。

H：男孩穿着红色衣服。

表 1 句法异构蕴含成因类型

成因		例句	说明
结构变化	语序变化	P：三个女人和一个小女孩在和小狗玩。 H：与小狗玩耍的女人们。	由语法结构内部成分的线性顺序发生变化导致的蕴含。
	成分抽取	P：一个穿着黑色裤子没穿衬衫的男孩儿正在玩一个白色的气球。 H：男孩穿着黑色裤子。	从 P 中把主谓宾结构的某一部分抽取出来，单独成句。
	小句抽取	P：男人和女人在海滩上漫步，身后是绚丽的晚霞。 H：一个男人和一个女人在海滩上散步。（NULL）	在有多个小句的复句中抽出某一个小句，单独成句。
省略变化	省略中心语	P：年长的白人女子在她的厨房做蛋糕。 H：一位老太太在厨房里。	在偏正结构中，省略了核心谓词，而保留修饰语。
	省略修饰语	P：穿着黑色衬衫的吧台服务员用一台大机器做咖啡。 H：吧台侍者在做咖啡。	在偏正结构中省略修饰性成分，保留中心语。

3 句法异构蕴含的语块边界标注

我们从英文开源数据集 SNLI 选取了一部分数据，将其翻译成中文，筛选出其中结构清晰、表达合适的 4000 条蕴含数据进行了人工标注。经过校对后，获得有效标注结果为 3766 例。具体方法和流程在本章中详述。

3.1 数据选择

我们的数据来源于英文开源数据集 SNLI。一方面，目前尚未出现大规模中文文本蕴含数据集，在 2012 年发布的 RITE-2 的几个中文数据规模太小，并且不太容易获取，使用不方便，而英文领域有多个大规模开源数据集，例如 SNLI、MultiNLI，获取和使用都很方便。另一方面，文本蕴含本质上是一种语义关系，不同语言之间的蕴含成因会有共同之处，我们可以借助于英文数据来研究中文蕴含。

SNLI[6]是目前主流的文本蕴含数据集，其中的数据全部是依靠众包（Crowdsourcing）人工生成的真实文本，语言形式灵活多样，数据质量较

高，不会存在明显的语法错误。SNLI 的数据规模巨大，拥有 560152 条训练数据和 10000 条测试数据，每条数据包含一句 Premise 和一句 Hypothesis，以及一个关系标签，有充足的语料挑选余地。标注过程中需要考虑句子长度，句子过长，结构复杂，分析困难；句子过短，信息太少，不具有标注价值。SNLI 的 Premise 平均长度为 14.1 个单词，Hypothesis 的平均长度为 8.3，长度适中，便于人工标注。

基于以上，我们将 SNLI 的部分训练数据翻译成中文，挑选出长度在 5~35 个汉字之间，结构清晰，表达符合汉语用语习惯的句子进行人工标注和分析。

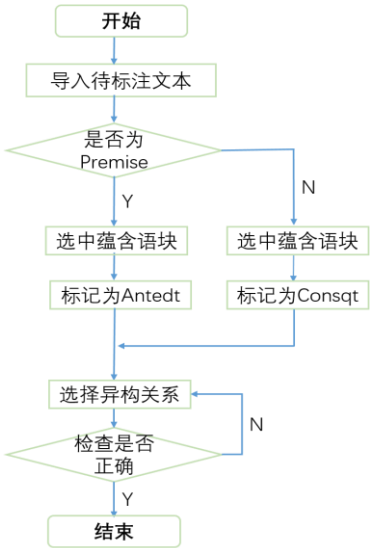


图 1 BRAT 标注过程

3.2 标注方法

本文标注工作实质上是在已知蕴含关系的基础上确定句法异构语块边界。标注员首先要看完原句 P 和蕴含句 H，对句子表达的内容有一个了解。根据 H 的内容回到 P 中寻找相关内容，分别标注出 P 和 H 的句法蕴含语块。

根据句法异构蕴含的类型划分标注语块的类型。省略类的蕴含语块往往是一个定中短语或状中短语；结构变化的蕴含语块类型多样，小句抽取的蕴含语块是复句中的小句，我们可以用逗号作为划分依据；成分抽取的蕴含语块是句中某个完整的句法成分，若句法成分前有修饰语，那么语块也要包括修饰语；语序变化的蕴含语块较为特殊，需要结合具体语料划分。

本文使用基于 Web 的文本标注工具 BRAT 进行蕴含语块标注，流程如图 1 所示。导入待标注文本，选择原句 P 和蕴含句 H 中的蕴含语块，分别标记为“Antedt”和“Consqt”。连接“Antedt”和“Consqt”，在弹出的对话框中为两个语块选择相应的句法异构关系。如果有标注错误，双击“Antedt”或“Consqt”或者关系类型，移动、添加、删除标注内容。标注结果由 BRAT 自动保存，如图 2 所示。完成整个文件中的数据标注后，得到一个后缀名为.ann 的文件。

为了提高标注语料的一致性，在第一次标注结束两周后，我们按照最终标准对数据进行了二次标注。最后，分析提取得到的句法异构蕴含语块，人工校对修改，得到最后的标注结果。这在一定程度上解决了多人标注引起的不一致问题，提高了蕴含语块标注的准确性。

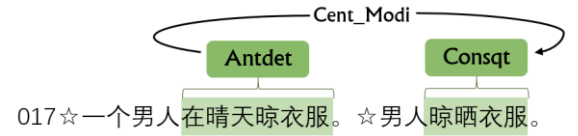


图 2 BRAT 标注示例

3.3 标注结果及数据分析

我们总共筛选出 4000 条蕴含数据，获得有效标注结果 3766 例。其中句法异构蕴含有 1483 例，占 39.4%；词汇蕴含 1188 例，占 31.5%；常识蕴含 1095 例，占 29.1%。

最后我们又针对句法异构蕴含进行语料扩充，总共标注句法蕴含 2000 例，结构变化类 463 例，占比 23.15%；省略类 1537 例，占比 76.85%。

可以看到，文本蕴含主要还是通过词汇关系和句法异构产生的，其中句法异构略多于词汇关系，而在句法异构蕴含中又是以省略类为主，结构变化导致的蕴含较少。

表 2 蕴含分类统计

蕴含成因类型分类			扩展句法异构蕴含		
蕴含分类	数量	百分比	句法异构蕴含分类	数量	百分比
词汇蕴含	1188	31.5%	结构变化类	463	23.2%
常识蕴含	1095	29.1%	省略类	1537	76.8%
句法异构蕴含	1483	39.4%	合计	2000	100%
合计	3766	100%			

4 句法异构蕴含边界识别研究

4.1 句法规则识别方法

通过解析句法异构蕴含语块对的词性和句法依存分析，我们总结出了一套句法异构蕴含的规则系统。在依存句法体系中，“HED”指的是核心关系，通常是小句的谓语，“SBV”指的是主语，“VOB”指的是宾语，“IOB”指的是间接宾语，“POB”指的是后置定语，“ATT”指的是定语，“ADV”指的是状语，“COO”表示两个重复的成分。本文的句法异构中可以有规则匹配的类型归纳如下：

1.“被”、“把”语块：

我们通过匹配句子中的标志字“被”和“把”，并判断“被”和“把”在语块中担任“ADV”成分，则认为此语块为“被”结构或“把”结构语块。

LTP 中的句法依存分析结果，“被”字语块一般被解析为如下结构：

(1) [ATT]* + FOB + 被 + [[ATT]* + POB] + HED

“把”字语块句法依存分析的主体结构为：

(2) [SVB] + 把 + [ATT]* + POB + HED + [[ATT]* + VOB]

“被”字语块蕴含的句法依存结构如图 3：

S1：大象正被一个男人骑着。

S2：人在骑大象。

“把”字语块蕴含的句法依存结构如图 4：

S3：走过街道，把它打扫干净。

S4：清扫街道。

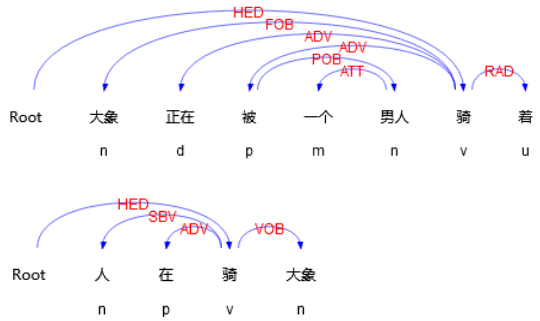


图 3 S1、S2 句法结构

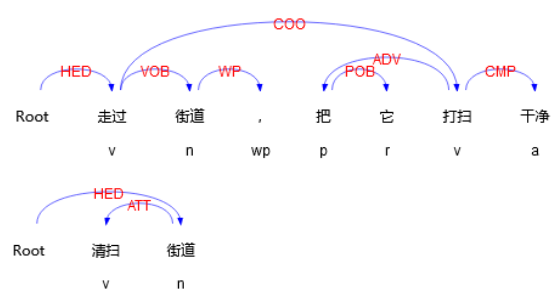


图 4 S3、S4 句法结构

2. 普通语块：

与“被”字语块和“把”字语块对应，一般语块的句法依存分析的主体结构为：

(3) SBV + HED + VOB

蕴含语块对中，HED 必须一致或有蕴含关系，并且 FOB 和 VOB，POB 和 SBV 一致或 H 句中的主体结构中某成分被省略。

H 省略 P 中并列的信息：即句法依存分析树的结构中，P 有多个 HED，H 缺少 P 中标记为 COO 部分的子树。语块对的主体结构如下：

(4) P: SBV + [HED^P]* + VOB

H: SBV + [HED^H]* + VOB

其中[HED^H]* ∈ [HED^P]*，示例如下：

S5：坐在滑板上在乡间滑行。

S6：坐在滑板上。

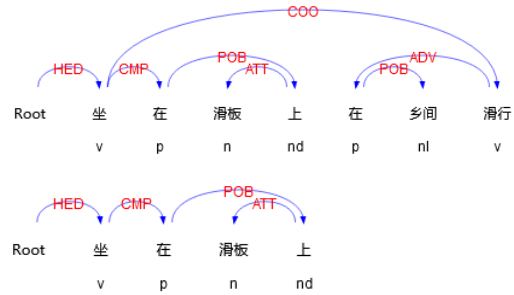


图 5 S5、S6 句法结构

H 省略 P 中修饰的信息：蕴含句对的 HED 相同，H 中缺少 P 中的一个或几个 ATT 成分，其他成分相同，语块对的句法结构表示为：

(5) P: [ATT_1^P]* + SBV + HED + [ATT_2^P]* + VOB

H: [ATT_1^H]* + SBV + HED + [ATT_2^H]* + VOB

其中，[ATT^H]* ∈ [ATT^P]*，P 和 H 可以省略某一句子成分，且 P 的信息包含 H 的信息。例句如下：

S7：一个亚裔小女孩儿。

S8：一个小女孩儿。

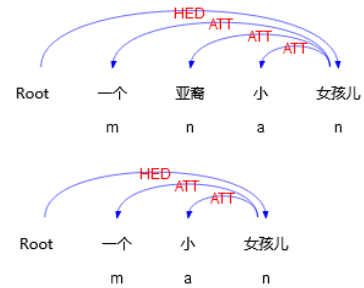


图 6 S7、S8 句法结构

H 只保留了 P 中的 HED，省略其它的句法成分。语块对的句法结构表示为：

(6) P: [[ATT]* + HED^P]*

H: HED^H

表 3 句法规则覆盖度

规则	被、把语块	HED 省略	ATT 省略	HED 提取	合计
覆盖度	4．07%	5.53%	16.99%	22.67%	48.37%
有效性	93.1%	88.71%	91.85%	98.63%	--

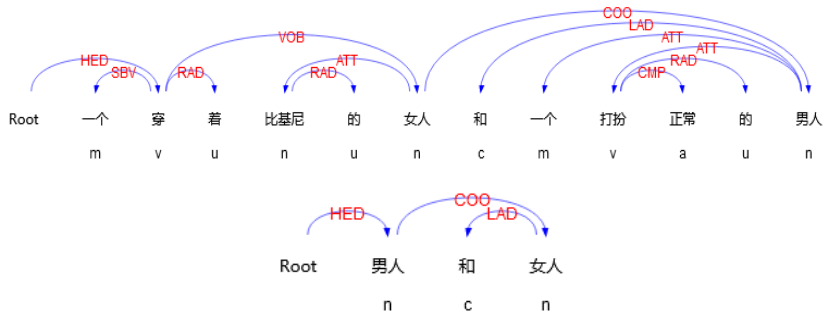


图 7 S9、S10 句法结构

其中，P 的结构为一组或多组修饰语加核心词，且[HED^H]∈[HED^P]，如果[HED^H]包含多个短语，则用“和”连接。例句示例如下：

S9：一个穿着比基尼的女人和一个打扮正常的男人。

S10：男人和女人。

按照上述 6 条规则自动抽取蕴含语料，每条规则抽取的数量与数据库中语块总数的比值为相应规则的覆盖度，每条规则抽取得到的语块数量与数据库中符合此规则的语块数量的比值为相应规则的有效性，为规则的具体评价。规则覆盖度评价如表 3 所示。句法异构的句对结构转化多样，句法成分位置灵活，以及同义词及上下位词的替换使得我们难以用规则概括所有的句法异构蕴含。本文总结规律性强、较为常见的蕴含语块对，确保了抽取数据的有效性，但由于规则限制比较严格，未能覆盖全部数据。本节规则识别的结果为进一步的深度模型实验提供了参考标准。

4.2 基于深度学习方法的实验

4.2.1 模型

本文采用深度学习模型处理整合 P 和 H 的蕴含信息，识别蕴含边界下标。基于 Wang[22]的模型，如图 8 所示，此模型主要分为两个模块：match_LSTM 和 Pointer Network(Ptr-Net)。

Wang[23] 针对文本蕴含任务提出了 match-LSTM 模型，用来判断 P 是否蕴含 H。与 Wang[23]工作不同的是，我们没有利用 match-LSTM 判断 P 和 H 的蕴含类型，而是计算获得包含 P 和 H 蕴含信息的表示向量，作为 Ptr-Net 的输入。Ptr-Net 由 Vinyals[24]提出，它采用 attention 机制作为指针，选择输入序列的位置下标作为输出。在此我们采用 Ptr-Net，在整合了 P 和 H 蕴含信息的向量中寻找蕴含边界。

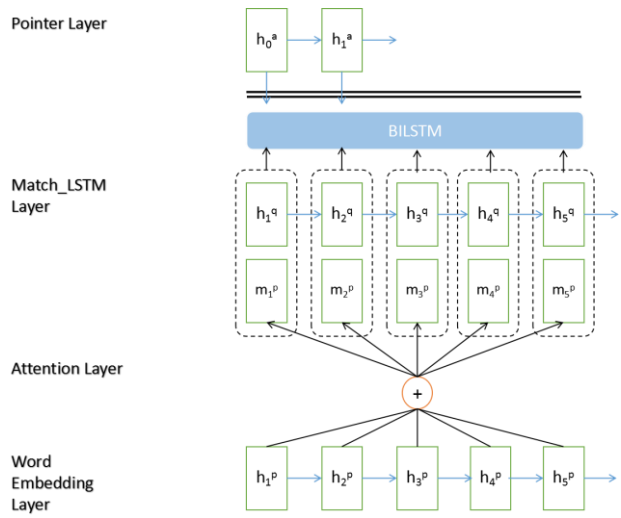


图 8 模型结构图

4.2.2 实验设计与分析

我们在 SNLI 数据库中选取 2000 条句法异构类型的蕴含对，采用前文的规则进行人工标注。其中，训练集包含 1700 条数据，测试集包含 300 条数据。实验代码基于 tensorflow 框架，采用边界正确率作为评价指标。我们分别统计了 P 和 H 蕴含片段的前后正确率及总体正确率，实验结果如表 4 实验结果表 4 所示。

表 4 实验结果

模型	P 整句	H 整句	P-H 整句
LSTM+Ptr	65.40%	68.83%	68.71%
match_LSTM+Ptr	72.61%	74.75%	74.42%

实验结果如所示，对于两个模型 P 和 H 两个蕴含边界识别总正确率分别为 68.71%、74.42%，

P 的蕴含边界正确率分别为 65.40%、72.61%，H 的蕴含边界正确率分别为 68.83%、74.75%。由实验结果知，模型对于 H 的蕴含片段识别能力略高于 P，Attention 机制显著的提高了模型的正确率。

本文首次提出句法异构蕴含边界识别问题，并且首次采用深度学习模型探索端到端识别蕴含边界的可能性。我们对比了 LSTM+Ptr-Net 和 match_LSTM+Ptr-Net 两个模型，前者使用 LSTM 为序列建模，后者在 LSTM 的基础上增加了 Attention 机制。

5 结语

本文通过标注蕴含句对，分析总结句法异构蕴含类型，归纳句法异构蕴含规则，并对该规则有效性进行验证，结果表明基于规则的方法可以为进一步的深度模型试验提供参考标准。本文用深度学习模型识别蕴含语块边界，在小规模中文语料上提供了可靠的基准线。本文的实验代码和数据已经公布在 Github 网站。网址如下：https://github.com/blcunlp/CCHEP

与整句级别的蕴含识别任务相比，本文在句法异构蕴含识别上的正确率还有待提高。我们计划进一步探讨句法异构蕴含规则，扩大规则覆盖范围，为深度学习模型提供更为可靠的外部知识。

本文的工作为日后蕴含成因分析与语块标注研究提供了可供改进的方向，其中包括：（1）提高语块标注的准确性，解决因错误标注带来的语块边界不清问题；（2）扩展蕴含成因类型，现有句法异构蕴含类型还能继续扩充，因常识和社会知识造成的蕴含也值得深入分析；（3）扩展句法异构蕴含规则，现有规则较为简单，对中文特殊句式的研究不够深入，未能覆盖到大部分句法异构蕴含现象。

参考文献

[1] 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.

[2] Bos J, Markert K. Combining Shallow and Deep NLP Methods for Recognizing Textual Entailment[J]. Proc of the Pascal Rte Challenge, 2005:65--68.

[3] Rodolfo Delmonte, Sara Tonelli, Marco Aldo Piccolino Boniforti, et al. VENSES – A Linguistically-Based System for Semantic Evaluation[J]. Lecture Notes in Computer Science, 2005, 3944:344-371.

[4] 刘茂福，李妍，顾进广．基于统计与词汇语义特征的中文文本蕴涵识别[J]．计算机工程与设计，2013,34(5):1777-1782.

[5] Marelli M, Menini S, Baroni M, et al. A SICK cure for the evaluation of compositional distributional semantic models[C]// Language Resources and Evaluation Conference. 2014:A-696.

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

[7] Williams A, Nangia N, Bowman S R. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference[J]. 2017.

[8] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[J]. 2018.

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

[10] Gong Y, Luo H, Zhang J. Natural Language Inference over Interaction Space[J]. 2017.

[11] 叶蜚声 徐通锵．语言学纲要[M]．北京大学出版社，2006:110-111.

[12] Skehan P. A Cognitive Approach to Language Learning. Oxford Applied Linguistics.[M]. 上海外语教育出版社，2000.

[13] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge[C]// International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. Springer-Verlag, 2005:177-190.

[14] Bar-Haim R, Dagan I, Dolan B, et al. The Second PASCAL Recognising Textual Entailment Challenge[J]. Proceedings of the Pascal Challenges Workshop on Recognising Textual, 2006, 3944:177-190.

[15] Magnini B, Magnini B, Dagan I, et al. The third PASCAL recognizing textual entailment challenge[C]// Acl-Pascal Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007:1-9.

[16] T Khot, A Sabharwal, P Clark. SciTaiL: A Textual Entailment Dataset from Science Question Answering. 2018.

[17] Dagan I, Glickman O. Probabilistic textual entailment: Generic applied modeling of language variability[J]. 2004.

[18] Matsuyoshi S, Miyao Y, Shibata T, et al. Overview of the NTCIR-11RecognizingInference in Text and Validation (RITE-VAL) Task[C]// Proceedings of the11thNTCIR Conference . 2014:223-232.

[19] 任函．面向汉语文本推理的语言现象标注规范研究[J]．河南科技学院学报，2017(7):75-78.

[20] 陆俭明．现代汉语语法研究教程[M]．北京大学出版社，2013:8-12.

[21] 范晓．关于汉语的语序问题（一）[J]．汉语学

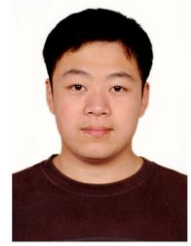
习,2001(05):1-12.
[22] Wang S, Jiang J. Machine Comprehension Using Match-LSTM and Answer Pointer[J]. 2016.
[23] Wang, Shuohang, Jiang, Jing. Learning Natural Language Inference with LSTM[J]. 2015:1442-1451.
[24] Vinyals O, Fortunato M, Jaitly N. Pointer Networks[J]. Computer Science, 2015.

金天华（1995—），第一作者，硕士研究生，主要研究领域为计算语言学，自然语言处理。

E-mail:lyhjth@126.com

姜珊（1995—），第二作者，硕士研究生，主要研究领域为自然语言处理。

E-mail:ccjiangshan@yeah.net



于东（1982—），通信作者，博士，副教授，主要研究领域为自然语言处理。

E-mail:yudong_bluc@126.com