文章编号：1003-0077（2017）00-0000-00

自然语言显式命题自动识别和解析方法

刘璐[1] 彭诗雅[1] 玉郴[1] 于东[1]
（1.北京语言大学 信息科学学院，北京 100083）

摘要：自然语言中包含很多显式命题，正确理解这些命题是理解文本信息的关键。正确识别显式命题并解析其中的关键成分有助于理清语言中的逻辑关系、辅助自然语言理解。该文基于百度百科数据构建了自然语言显式命题标注数据集，并提出两个研究任务：自然语言显式命题自动识别和命题关键成分解析。其中，显式命题自动识别任务判断一个自然语言句子是否是命题；命题解析任务从已获取的命题中解析出支撑该命题成立的重要成分。针对任务一，构建基于 BERT 的二分类模型；针对任务二，构建基于 BERT-BiLSTM-CRF 的序列标注模型。实验结果表明，模型在任务一的正确率达到 74.90%，超过基线模型 15.25%；在任务二的 F 值达到 90.74%，超过基线模型 17.69%。该文为下一步研究提供了可靠的标注数据集和基线方法。

关键词：显式命题；显式命题自动识别；命题关键成分解析

中图分类号：TP391 **文献标识码：**A

Automatic Recognition and Analysis of Explicit Propositions in Natural Language

Lu Liu[1], Shiya Peng[1], Chen Yu[1], Dong Yu[1]

(1. College of Information Science, Beijing Language and Culture University, Beijing, 100083, China)

Abstract：There are a large number of explicit propositions in natural language. These propositions contain the vital information about the text. Recognizing the proposition and analyzing the essential ingredients in propositions can help to clarify the logic behind the text, help computers to understand human language. Based on the Baidupedia, we built an explicit proposition corpus. And two tasks were proposed: the automatic explicit proposition recognition and the essential explicit proposition ingredients analysis. The first task determines whether a sentence is a proposition or not. The second task selects the essential ingredients in the propositions acquired before. For the task one, we constructed a classification model based on BERT. For the task two, we constructed a sequence labeling model based on BERT-BiLSTM-CRF. The experimental results show that the model got an accuracy as 74.95% over the baseline model 15.30% in task one, and model have got a F-value as 90.74% over the baseline model 17.69% in task two.

Key words: explicit propositions; automatic explicit proposition recognition; essential explicit proposition ingredients analysis

0 引言

自然语言中存在大量命题，这些命题中大都包含文本中的关键信息。正确理解这些命题可以辅助

收稿日期：201*-*-*；**定稿日期**：201*-*-*

基金项目：教育部人文社会科学研究青年基金项目(19YJCZH230)；国家社科基金重点项目(16AYY007)；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果(19YCX114)

表 1 标注范例（“1”代表是命题，“0”代表不是命题）

例句	是否命题	关键成分
S1：特点对于春日野穹异乎寻常的热爱，可能是妹控，也可能不是	0	
S2：所有子囊果是子囊壳的核菌纲真菌都归于球壳目	1	“子囊果是子囊壳的核菌纲真菌”和“归于球壳目”

自然语言理解，促进一些相关任务的发展，如文本摘要、阅读理解、文本推断等。命题和命题形式是形式逻辑的研究对象。在形式逻辑中，判断是对于思维对象有所肯定或否定的一种思维形式[1]，命题是判断的语言表达[2]。命题是表达了对事物情况有所肯定或否定的陈述句，它蕴含了或真或假的思想。如果一个句子是命题，它的陈述就一定能区分真假，否则，它就不是命题[2]。自然语言中，大部分命题都是由逻辑联结词引导的，本文将这类命题称为显式命题。《形式逻辑》[2]一书对命题类型做了细致的划分，我们关注其中的四个类型的命题在自然语言文本中的表现，这四种类型分别是性质命题、联言命题、选言命题和假言命题。

自然语言中有逻辑联结词的句子不一定是显式命题，显式命题也往往包含阻碍理解的冗余信息。本文提出两个任务：自然语言显式命题自动识别和命题关键成分解析。自然语言显式命题自动识别任务要求判断一个子句是否为命题，如表 1 中，S1 不是命题 S2 是命题；命题关键成分解析要求解析出支撑命题成立的关键成分，如表 1 中，S2 中的“子囊果是子囊壳的核菌纲真菌”和“归于球壳目”就是该命题的关键成分。我们基于百度百科[1]语料为两个任务建立了大规模人工标注语料库。

显式命题自动识别任务可以转化为单句的二分类问题。我们将传统的统计学习方法支持向量机（Support Vector Machine，SVM）[3]作为该任务的基线模型，构建双向长短期记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）[4]作为对照。此外，我们利用变压器的双向编码器表示（Bidirectional Encoder Representation from Transformers，BERT）[5]在本任务上微调作为分类模型之一。BERT 作为一种新型的语言模型，充分利用大量无监督数据将语言学知识隐含地引入特定任务中，在 11 个自然语言处理任务中达到目前最好的效果。BERT 在文本分类上表现优越[5]，适合作为显示命题自动识别任务的模型之一。

命题关键成分解析任务可以转化为序列标注问题。本文主要采用传统统计学习模型条件随机场（Conditional Random Field，CRF）[6]作为该任务的基线模型。我们构建目前流行的序列标注模型 BiLSTM-CRF（BiCRF）[7]，进一步探索 BiLSTM 对于特征的选择能力。另外，我们构建 BERT-BiLSTM-CRF(BBiCRF)用于该任务，充分利用 BERT 在处理百科语料上的优势。

实验结果表明，BERT 模型在显式命题自动识别任务上的正确率高达到 74.95%，超过基线模型 15.30%；BBiCRF 模型在命题关键成分解析任务上的 F 值达到 90.74%，超过基线模型 17.69%。

本文结构如下：第 1 节介绍国内外自然语言中命题的研究现状；第 2 节介绍本次涉及到的命题并逐一分析；第 3 节介绍数据标注工作，并分析标注结果；第 4 节和第 5 节在两项任务上分别设计模型进行实验，比较不同模型在数据集上的表现能力，同时对数据集的特点做出解释。最后一节对论文工作进行总结和展望。

1 相关工作

自然语言中的命题研究一直以来都受到各界学者的关注。国内关于形式逻辑的研究主要集中在逻辑学和语言学领域[8-10]。李先焜[11]定义了语言逻辑的概念，阐述了语言逻辑的地位，梳理了研究语言逻辑的方法，解释了语言逻辑的研究意义。周礼全[12]区分了正统逻辑和自然语言逻辑。龚启荣[13]强调了当代形式逻辑研究对人工智能发展的重要性，他认为当代形式逻辑是人工智能最合适的工具。高逢亮等人[8]从语义学、语法学、语用学和修辞学等不同领域展开研究，深入探讨了逻辑学和语言学之间的关系，肯定了逻辑学对于语言理解的重要性。

现有命题研究大多集中在哲学和语言学领域。在哲学领域，周文华[14]分析了现有的一些命题定义的特点，并给出了新的命题定义。他将命题定义为具有某种属性的对象。杨宏郝[15]阐述了

[1] https://baike.baidu.com/

表 2 性质命题、联言命题、选言命题和假言命题的例句及关键成分

命题类型	例句	关键成分
性质命题	T3：所有的白猫都是哺乳动物。	“白猫” “哺乳动物”
联言命题	T4：味噌，既可以做成汤品，又能与肉类烹煮成菜。	“味噌” “可以做成汤品” “能与肉类烹煮成菜”
选言命题	T5：成语释义海角天涯释义形容极远的地方，或彼此相隔极远，或者事物的尽头。	“海角天涯” “彼此相隔极远” “事物的尽头”
假言命题	T6：如果左耳先听到声音，那么听者就觉得这个声音是从左边来的，反之亦然。	“左耳先听到声音” “听者就觉得这个声音是从左边来的”

命题和判断之间的关系，他提出判断无逻辑结构可言，只有作为语句的命题和作为语句模式的命题形式才有逻辑结构。在语言学领域，沈园[16]讨论了逻辑判断基本类型划分，说明逻辑判断是与语法、语用等各方面有着密切联系而又相对独立的一个范畴。李勤[9]从逻辑中判断的分类出发，对句子语义中的命题进行了分类。韩铁稳[17]认为逻辑常项是区分逻辑思维形式的标志，对逻辑常项的语言形式进行了汇总。黄士平[18]认为逻辑常项是逻辑形式中不变的部分，将表述逻辑常项的语言形式直接出现称为显性形式，将未直接出现称为隐形形式，并分析了逻辑常项的隐性形式。

以上研究是从语言学角度和哲学角度对形式逻辑以及命题的一些研究。在计算语言学领域，与命题直接相关的研究较少，部分研究关注篇章句间关系，这种关系也和语言逻辑相关。

张牧宇等[19]总结了中文篇章及语义分析的特点，提出面向中文篇章句间关系的层次化语义关系体系,对句间关系类型进行详细描述,并在新闻语料上进行了标注。随后，张牧宇等[20]对中文篇章句间关系识别任务进行初步探索，他们根据文本单元间是否存在篇章连接词将关系分为显式篇章句间关系和隐式篇章句间关系。

基于逻辑学和语言学对命题类型划分的研究成果，本文提出自然语言显式命题自动识别任务和命题关键成分解析任务,并基于百度百科语料为两个任务建立了大规模人工标注语料库。

2 命题及其关键成分研究

2.1 命题研究

在自然语言中，命题是表达了对事物情况有所肯定或否定的陈述句,它蕴含了真或假的思想。如果一个句子是命题，它的陈述就一定能区分真假，否则，它就不是命题[2]。例如：

T1：所有声音都称之为音频。

T2：正因为如此，使得它不但能够与世界的顶尖时尚风潮同步，而且更能体现亚洲女性的娇柔与美感。

T1 是命题，T2 不是命题。T2 中代词“它”指代不明确，难以辨别真或假，因此该句不是命题。

根据一个命题本身是否包含有其他命题可把命题分为两类：一类是本身不包含其他命题简单命题，包括性质命题，关系命题；另一类是包含了其他命题的复合命题，包括联言命题，选言命题，假言命题，和负命题等。本研究针对其中的性质命题，联言命题，选言命题和假言命题展开（见表 2）。

2.1.1 性质命题

性质命题是一种简单命题，是断定事物具有某种性质的命题，由逻辑常项和逻辑变项组成。

逻辑变项：分为主项和谓项。主项是表示命题对象的概念，谓项是用来表示命题对象所具有或不具有的某种性质的概念。

逻辑常项：分为量项和联项。量项是表示命题对象数量的概念，联项是用来联系主项与谓项的概念。

表 2 的例句 T3 中，主项是“白猫”，谓项是“哺乳动物”。量项是“所有”，联项是“是”。

2.1.2 联言命题

联言命题是复合命题的一种，是断定事物的若干情况同时存在的命题。联言命题一般由联言

表 3 各类命题在文本中的逻辑常项

命题类型	文本中的逻辑常项
性质命题	一切……是，一切……都，所有……是，所有……都，一切……不是，所有……不是，有些……是，有的……是，某些……是，有些……不是，有的……不是，某些……不是
联言命题	既……又……，不但……而且，一方面……另一方面。虽然……但是
选言命题	或……或……，可能……也可能……，要么……要么，不是……就是
假言命题	如果……那么……,只要……就……,若……必……,只有……才……,不……不……,没有……没有……

肢和联言联结词组成。

联言肢：联言命题所包含的简单命题称为联言肢。

联言联结词：表达联言命题的逻辑联结词称为联言联结词。

表 2 的例句 T4 中，联言肢为“味噌可以做成汤品”和“味噌能与肉类烹煮成菜”。联言联结词为“既……又”。

2.1.3 选言命题

选言命题也是复合命题的一种，是断定事物若干种可能情况的命题。选言命题由选言肢和选言联结词组成。

选言肢：选言命题所包含的简单命题称为选言肢。

选言联结词：表达选言命题的逻辑联结词称为选言联结词。

表 2 的例句 T5 中，选言肢为“海角天涯形容彼此相隔极远”和“海角天涯形容事物的尽头”。选言联结词为“……或……或……”。

2.1.4 假言命题

假言命题也是复合命题的一种，是断定事物情况之间条件关系的命题。假言命题由假言肢和选言联结词组成。

假言肢：假言命题所包含的简单命题称为假言肢。

假言联结词：表达假言命题的逻辑联结词称为假言联结词。

表 2 的例句 T6 中，假言肢为“左耳先听到声音”和“听者就觉得这个声音是从左边来的”。假言联结词为“如果…那么…”。

2.1.5 四类命题在自然语言中的逻辑常项

逻辑常项是逻辑形式中不变的部分,即在同类型的逻辑形式中都存在的部分[18]。常项在思维逻辑形式中起决定作用,它是区分思维理辑形式的标志[17]。在上述四种命题中，性质命题的逻辑常项包括量项和联项，而联言命题、选言命题和假言命题中的联结词就是逻辑常项。

根据《形式逻辑》[2]一书中总结的文本中的常见的逻辑常项，我们汇总并补充了四类命题在文本中常见的逻辑常项，见表 3。

2.2 关键成分研究

自然语言中的显式命题也往往包含冗余信息。抽取命题的关键成分有利于理解自然语言。命题的关键成分是指构成该命题成立的最基本要素。

段士平[21]将语块定义为以整体形势存储在大脑记忆库中，并可以作为预制板块，供人们提取使用的多词单位。通俗的讲，语块的概念淡化了原有的词汇与语法之间的界限，不仅包括多词的搭配、句子框架、还可以扩大到句子甚至语篇。参考段士平[21]对“语块”定义，我们将命题的关键成分定义为构成命题最基本的几个语块。

性质命题的关键成分是指命题中构成主项和谓项的最小语块。例如，“白猫”和“哺乳动物”是性质命题“所有的白猫都是哺乳动物”的关键成分（见表 2，T3）。

联言命题的关键成分是指构成各联言肢中主项和谓项的最小语块。例如，“味噌”、“可以做成汤品”和“能与肉类烹煮成菜”是联言命题“味噌,既可以做成汤品,又能与肉类烹煮成菜”的关键成分（见表 2，T4）。

选言命题的关键成分是指构成各选言肢中主项和谓项的最小语块。例如，“海角天涯”、“彼此相隔极远”和“事物的尽头”是选言命题“成语释义海角天涯释义形容极远的地方，或彼此相隔极远，或者事物的尽头”的关键成分（见表 2，T5）。

假言命题的关键成分是指构成各假言肢中主项和谓项的最小语块。例如，“左耳先听到声音”和“听者就觉得这个声音是从左边来的”是假言

表 4 三种语料来源对比

语料来源	例句	说明
微博	加油啊们们哈哈在这么一个少女心爆棚的时刻谈理想真是不太合适啦不过当然还是有几两大事要办啦一年中超忙的但也会棒的啦~	语料格式相对随意，且包含有语气词“哈哈”，“啦”和“的啦”。内容没有针对特定的事物展开。
文学	许多总是劝时代把文学节目推掉，做一个白天的轻松点的节目，要不每晚十点才下班，没有正常的夜生活。	“许多”和“时代”是文学作品中虚拟创作的人物。文学作品中有很多虚拟的人和事物，且少有关于某些事物性质的描述。
百度百科	所谓“资本主义”是指资本主导社会经济和政治的意义。	百度百科结构完整，内容都是针对某一特定词条展开的。如本例中，有明确的描述事物“资本主义”，且有清楚“资本主义”的性质“资本主导社会经济和政治的意义”。

命题“如果左耳先听到声音，那么听者就觉得这个声音是从左边来的，反之亦然”的关键成分（见表 2，T6）。

3 自然语言中的逻辑命题挖掘和标注

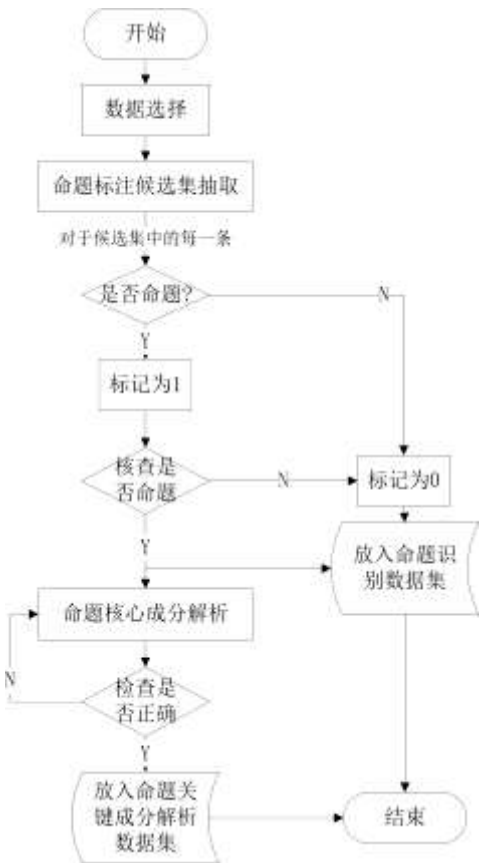

图 1 标注流程图

本次逻辑命题挖掘和标注任务主要分为数据选择和预处理、数据标注、数据标注结果分析几个步骤，图 1 是本次数据标注的整体流程。

3.1 数据选择和预处理

显式命题可能出现在任何类型的自然语言文本中。合适的语料来源能挖掘更多、更有效的语料。因此语料来源的选择至关重要，直接关系到标注的难度和成本。为此，我们收集、对比并分析了文学语料、微博语料[2]和百度百科语料（见表 4）。

微博是一种信息发布及社交网络平台[22]。用户可以通过发表微博来记录生活、分享心情、表达观点等。微博中抽取的句子的句式比较随意，且通常含有较多语气词（如表 4 中，“哈哈”，“啦”和“的啦”等）。文学语料虽然结构完整、文本规范，但其内容往往很难控制。大部分文学作品或多或少涉及到各类虚拟人物或事物，这导致难以判断文本的真假性。如表 4 中，“许多”和“时代”是文学作品中创作的虚拟人物，因此与他们相关事物的真假较难判断。除此之外，文学作品大多叙事，较少出现对某些事物有所肯定或否定的表达。百科语料是内容开放、自由的网络百科全书。与微博语料相比，其结构更为完整、规范，语句更通顺；与文学语料相比则更具真实性和普适性。如表 4 中，百度百科例句有明确的描述事物“资本主义”及性质“资本主导社会经济和政治的意义”，并能辨别真假。因此，我们百度百科作为数据来源。

本次研究选择 2,111,764 条百科词条对应的

[2] https://weibo.com

简短描述并去除了不完整的句子。我们把四种命题类型的常见逻辑常项作为触发词，从语料中抽取各类命题的候选集。筛选得到 24,337 条候选显式命题。我们招募 5 名具有语言学背景的研究生和本科生对 24,337 条显式命题集进行标注。具体流程在下一节中详述。

3.2 标注流程

标注工作分为两个步骤：第一步是对已有命题候选集中每一条数据进行“0-1”标注；第二步是对第一步中被标注为“1”的数据进行关键成分抽取。本次标注的总体流程如图 1 所示。

5 名语言学专业的硕士生和本科生参与了标注。整体的标注分为培训、试标注、正式标注三个环节。

在线下培训环节，介绍了标注任务、界定了命题的概念及不同类型命题的定义及划分。

在试标注环节，5 名标注人员对同样的 100 条候选的命题句进行“0-1”标注以确保标注人员完全理解了标注规范。一致性检验结果显示，5 名标注者之间的一致性较好（Fleiss’ kappa=0.69[23]），说明培训后的标注员理解了标注规范，可以进行正式标注。

正式标注环节可分为两个阶段：第一阶段是判断候选命题是否为命题，标注者要求对候选命题进行“0-1”判断，是命题的候选句标为“1”，不是命题的候选句标为“0”。为了保证标注结果的准确性，第二阶段有两项标注任务：对第一阶段中标注为“1”的数据进行的二次核查标注。若核查之后该句子依旧被认定为是符合规范的命题则标为“1”并标注命题关键成分；若核查之后认定该命题不符合规范，则标为“0”。我们从核查为 1 的数据中随机选择 5565 条数据进行第二阶段标注。第二阶段进行命题关键成分抽取，要求标注员按照标注规范标记命题关键成分。

全部的标注流程结束之后，将 24,337 条标注为“0”和“1”的句子作为显式命题自动识别任务的数据集，将标注了关键成分的 5,565 条显式命题作为关键成分解析任务的数据集。

3.3 标注结果及数据分析

我们从百度百科中抽取了 24,337 个候选命题进行标注，最终构建了包含 24,337 条有效数据的自然语言显式命题自动识别任务数据集和包含了 5,565 条数据的命题关键成分解析数据集。其中，自然语言显式命题数据集包含 14,625 条非命题和 9,712 条命题。

为之后实验需求，我们将所构建的数据集按照命题类型所占比例切割成训练集、验证集、测试集三个部分。具体数据情况如表 5 所示。

表 5 显示命题自动识别和关键成分解析数据集统计

任务类型	命题类型	训练集	验证集	测试集
显式命题自动识别	性质命题	6190	609	609
	联言命题	5763	567	567
	选言命题	1506	148	148
	假言命题	6878	676	676
	总计	20337	2000	2000
关键成分解析	性质命题	1553	170	170
	联言命题	1769	194	194
	选言命题	51	5	5
	假言命题	1378	131	131
	总计	4565	500	500

4 显式命题自动识别研究

显式命题自动识别是自动识别出某一句话是否为命题，可以抽象为“0-1”二分类问题。本文采用 SVM[3] 作为该任务的基线模型，还构建 BiLSTM[4]和在本任务上微调的 BERT[5]进作对照实验。评价指标为分类正确率。

4.1 显式命题自动识别模型

4.1.1 Support Vector Machine（SVM）

SVM 由 Cortes 和 Vapnik[3]于 1995 年首先提出。它在解决小样本、非线性及高维模式识别中表现很好，是现有机器学习中应用最广泛的一种分类算法。

我们基于 scikit-learn[3]实现 SVM 模型。SVM 输入特征为句子全部字向量的加和。本研究中使用的字向量由大规模百度百科语料预训练而来[24]。

4.1.2 Bi-directional Long Short-Term Memory（BiLSTM）

BiLSTM[4]是的基本思想是每一个训练序列都有向前和向后的两个 LSTM（Long Short-Term Memory）。这个结构将输入序列中每一个点前向和后向的信息拼接起来作为完整的上下文信息提供给输出层。我们将平均池化和最大池化的 BiLSTM 输出拼接起来作为句子的最终表示，将其

[3] https://scikit-learn.org/stable/modules/svm.html

提供给分类器

实验代码的实现基于 tensorflow 框架[4]实现，采用字级建模，使用预训练的字向量作为初始表示，字向量在训练过程中不断更新。

4.1.3 Bidirectional Encoder Representation from Transformers（BERT）

BERT[5]是一种新的语言表示模型。它的模型结构由多层的双向 Transformer 编码器[25]构成。BERT 支持对特定的任务进行微调，只需要添加一个额外的输出层，不需要对模型结构进行大量的修改。BERT 作为一种新型的语言模型，充分利用大量无监督数据将语言学知识隐含的引入特定任务中，在多项自然语言处理任务中达到了最优结果[5]。因此，BERT 模型能很好的移到显式命题自动识别任务上。

对于显式命题自动识别任务，我们使用了 Google AI 开源的 BERT 代码[5]，根据我们的任务调整部分参数，微调过程载入 Google AI 在维基百科数据上预训练的中文模型。

4.2 结果分析

首先，我们对比了 SVM、BiLSTM 和 BERT 三个模型在显式命题自动识别任务 2000 条测试集上的表现能力；其次，我们以准确率最高的 BERT 模型为例，详细地分析了该模型在不同长度和不同类型命题上的表现。

4.2.1 总体分析

表 6 三个模型在显式命题自动识别测试集的正确率

模型	正确率(%)
SVM	59.65
BiLSTM	66.80
BERT	**74.95**

表 6 展示了三个模型在显式命题自动识别任务上的结果。SVM、BiLSTM 和 BERT 在显式命题自动识别任务中的测试集上正确率分别为 59.65%、66.80%和 74.95%。其中，BERT 表现最好，比 SVM 高 15.30%，比 BiLSTM 高 8.15%。这可能是因为，BERT 是在基于大规模维基百科数据上预训练得到的模型进行微调进而进行分类，百度百科数据和维基百科数据都是半结构化的百科数据文本，因此他们在结构上、内容上都具有一定的相似。因而，BERT 在基于百度百科数据的显式命题自动识别任务上表现较好。

4.2.2 BERT 在不同类型命题的表现

图 2 展示了 BERT 在本次研究涉及到的四类命题上的识别正确率。从图中我们可以明显的看出，BERT 模型在联言命题的和假言命题的识别准确率分别达到了 80.60%和 78.38%，说明 BERT 对联言命题的和假言命题的识别能力比较高。相反的，BERT 模型对于性质命题和选言命题的识别能力较差，准确率仅有 72.74%和 71.30%。

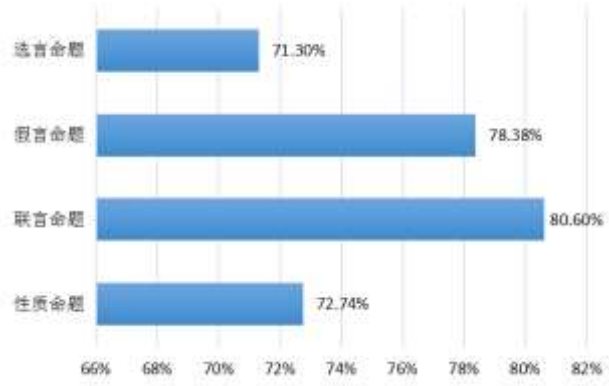

图 2 BERT 在不同类型命题上的正确率

4.2.3 BERT 在不同句长命题的表现

图 3 展示了 BERT 模型在不同长度命题上的显式命题识别正确率。从图中可以看出，BERT 模型的在不同长度命题上的表现并无明显规律，这证明本次所建立的数据难度是比较均匀的，不会受到句长因素的影响。

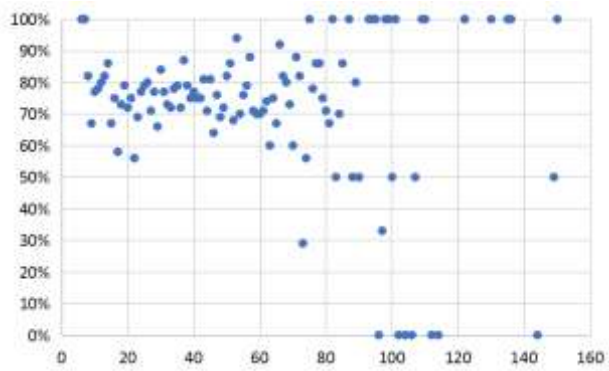

图 3 BERT 在不同命题长度区间上的正确率

5 命题关键成分解析研究

命题关键成分解析是从已有命题中提取出关

[4] https://tensorflow.google.cn/
[5] https://github.com/google-research/bert

键成分，可以抽象为序列标注问题。本文采用四位标记“BMES”对标注数据进行整理。每个关键成分的开头被标记为“B”，中间部分被标记为“M”，结尾部分被标记为“E”，关键成分为单字，则标为“B”，其余非关键成分被标记为“S”，如表 7。本文构建 CRF[6]作为基线模型并构建序列标注模型 BiLSTM-CRF[7]与之对比。最后，由于 BERT 模型在处理百科语料上的优势，我们构建 BERT-BiLSTM-CRF(BBiCRF)进行实验。评价指标为准确率、召回率和 F 值。

表 7 命题关键成分解析标记实例

例句	某些甲虫是中间寄主
标记	S S B E S B M M E

5.1 命题关键成分解析模型

5.1.1 Conditional Random Field（CRF）

CRF[6]是给定一组输入序列，预测另一组输出序列的条件概率分布的模型。CRF 在预测时考虑相邻数据的标记信息。

我们使用开源的工具包 CRF++[6]实现本实验。

5.1.2 BiLSTM-CRF（BiCRF）

基于神经网络的方法在序列标注任务中非常流行。Lample[26]等人基于提出了基于词和字符嵌入的 BiLSTM-CRF 命名实体识别模型。在序列标注任务中，基于深度学习的 BiLSTM 被用来提取特征，CRF 层为最终的预测标签添加一些约束。受到该模型的启发，我们使用基于字嵌入的 BiLSTM-CRF 作为命题关键成分解析的基准模型之一。

本文使用开源工具包 NCRF[7]实现了基于预训练字向量[24]的 BiLSTM-CRF 实验。

5.1.3 BERT-BiLSTM-CRF（BBiCRF）

最近 BERT 模型刷新了自然语言处理多项任务的最高记录。基于维基百科预训练的 BERT 模型在处理百科数据上具有优势，因此我们构建 BBiCRF 进行实验。BBiCRF 的主要思想是使用 BERT 模型在大规模语料库上预训练上下文相关的字向量表示替换 BiCRF 的字向量，最终利用 CRF 层的输出预测输入序列的标记。

我们基于在命名实体识别任务的开源工具包 BBiCRF[7]进行修改并实现命题关键成分抽取模型。

[6] https://taku910.github.io/crfpp/

[7] https://github.com/macanv/BERT-BiLSTM-CRF-NER

5.2 结果分析

首先，我们分析 CRF、BiCRF、BBiCRF 三个模型在命题关键成分解析任务 500 条测试集上的表现能力；其次，我们以 BERT 模型为例，详细地分析了该模型在不同命题类型数据上的表现能力；再次，我们选取部分实际数据案例，分析其在三个模型上的序列标注结果，详述每个模型识别的异同。最后，为验证关键成分的有效性，我们用 5565 条命题的关键成分替换显式命题识别数据集中的相应的命题，并在 BERT 上进行实验。

5.2.1 总体分析

表 8 展示了三个模型在命题关键成分解析任务上的实验结果。CRF、BiCRF、BBiCRF 在该任务测试集上 F 值分别为：73.05%、83.45%和 90.74%。其中，BBiCRF 表现最好，比 CRF 高 17.69%，比 BiCRF 高 7.29%，这得益于 BERT 在大规模维基百科上地预训练。而 BiLSTM 处理了长距离的依赖问题，加强了局部窗口的联系，这使得 BiCRF 模型得到的标注结果的准确率、召回率、F 值分别比仅用 CRF 高 3.98%、17.44%和 10.4%。

表 8 模型在命题关键成分解析测试集的表现（%）

模型	正确率	准确率	召回率	F 值
CRF	73.34	74.46	71.70	73.05
BiCRF	82.43	78.44	89.14	83.45
BBiCRF	**92.04**	**90.99**	**90. 50**	**90.74**

5.2.2 BBiCRF 在不同类型命题的表现

为进一步分析模型在命题关键成分解析任务上的表现，我们在图 4 中展示了 BBiCRF 在解析不同类型命题的关键成分时的实验结果。

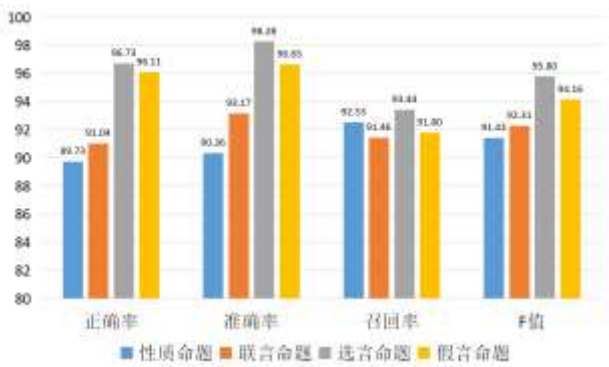

图 4 BBiCRF 在不同类型命题上的表现（%）

从图 4 中看出，BBiCRF 对选言命题的标注能力较高，标注准确率有 98.28%，F 值也达到 95.8%，这可能是由于所构建的数据集中选言命题的数量

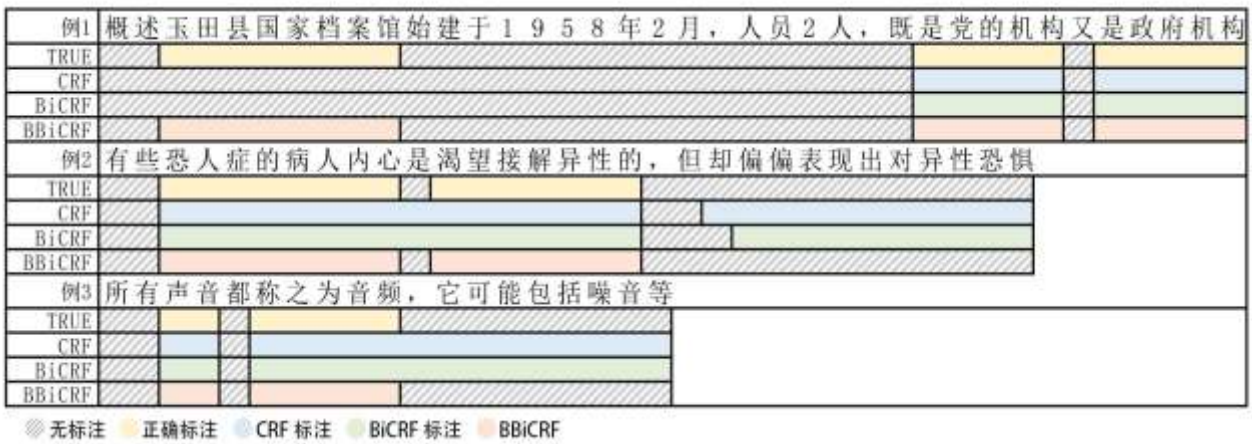

图 5 实例分析

较少。如何在样本量差距较大的情况下还能平均的学习到每个类型命题的特点并进行关键成分解析将是未来研究的重难点之一。我们也将在未来的数据集完善过程中增大选言命题的规模。除此之外，BBiCRF 对其他类型命题的标注结果比较平均，这说明数据集内每种类型命题在关键成分解析任务上难度比较平均。

5.2.3 实例分析

我们选取部分实际数据案例，分析它们在 CRF、BiCRF、BBiCRF 三个模型上序列标注的结果。图 5 展示了我们从测试集中选择的 3 条案例以及三个模型的解析结果。

对于例句 1，CRF 和 BiCRF 都没有将第一个关键成分“玉田县国家档案馆”标注出来，而 BBiCRF 能正确地将关键片段标注出来。与其他两个模型相比，BiCRF 能准确地标注出命题的几个关键成分，漏标关键成分的情况比较少。

对于例句 2，CRF 错误地将“却偏偏表现出对异性恐惧”标注为一个关键成分片段，BiCRF 则将“偏偏表现出对异性恐惧”标为一个关键片段。CRF 和 BiCRF 有时会将非关键片段标为关键片段，BBiCRF 能做出正确标注。

对于例句 3，CRF 和 BiCRF 在进行第二个关键片段标注时都错误地定位了片段的边界，将“称之为音频”标记为了“称之为音频，它可能包括噪音等”。以上两个模型能很好的识别到命题关键成分所在的位置，但是对于其边界判断还存在问题，未能再结束位置停止标注。相比前两个模型，BBiCRF 能很好地定位关键成分边界。

综合以上，CRF 和 BiCRF 在进行命题关键成分解析任务时会出现关键成分漏标、错误标注关键成分边界的情况，而 CRF 更是会出现将非关键成分标注为关键成分的情况。BBiCRF 则在这些问题有很好的表现。

5.2.4 BERT 在关键成分数据集的表现

为探究关键成分是否能代表命题中的关键信息，我们在显式命题自动识别任务上利用解析的关键成分进行了实验。

我们将显式命题自动识别数据集中 5565 命题替换为仅其关键成分，例如，用“招式 讲究大开大阖”替换原命题“所有的招式都讲究大开大阖，势大力沉，大气磅礴，没有任何用来诱惑取巧的花样”。

表 9 BERT 在命题自动识别测试集的表现

数据	数据量	正确率（%）
原始测试集	2000	74.95
替换后测试集	2000	**82.70**
原始被替换的数据	444	81.98
替换后的数据	444	**98.20**

我们用 BERT 在替换后的数据集上重新进行实验。实验结果见表 9。从表中可以看出，将部分命题替换为关键成分后，BERT 在替换后测试集上的正确率达到了 82.70%，比在原始数据测试集提升了 7.75%。在替换为关键成分的命题数据上，识别正确率更是达到了 98.20%，比在原始被替换的命题数据上提高了 16.22%。结果证明了我们提取的关键成分有效的提升了命题自动识别的正确率，命题中的关键成分可以代表命题的关键信息。

6 结语

本文提出自然语言显式命题自动识别任务和命题关键成分解析任务。显式命题自动识别是判断一个自然语言句子是否是命题。命题解析是从已获

取的命题中解析出支撑该命题成立的成分。为此，我们为两个任务建立了大规模人工标注数据集，分别是：包含 24,337 条有效数据的自然语言显式命题自动识别数据集和包含 5,565 条数据的命题关键成分解析数据集。通过构建基于统计学习和基于深度学习的模型，我们对自然语言显式命题自动识别和命题关键成分解析两个任务上进行了初探。实验结果表明，本文提出的深度学习模型能有效识别显式命题并解析显式命题的关键成分，为下一步研究提供了可靠的基线方法。本文数据已经公布在 https://github.com/blcunlp/Explicit-Propositions。

自然语言中的命题可以分为显式命题和隐式命题，本文主要研究的是自然语言中的显式命题，在未来的工作中我们将逐渐探索自然语言中的隐式命题。本次研究根据性质命题、联言命题、选言命题和假言命题四种类型的命题及其触发词在百度百科数据源上筛选并构建了数据集，未来我们将通过人工标注的方法尝试在更多的数据源上进一步扩大数据集规模，涵盖更多的命题类型，如关系命题和负命题等，为后续模型研究提供良好的数据基础。

本文的工作为显式命题自动识别和关键成分解析的研究提供了可以研究和改进的方向，包括：（1）扩展命题类型，目前涉及到四种显式命题，未来可在其他类型命题上进一步扩展；（2）扩充从自然语言文本中挖掘候选命题时用的触发词表，针对每一类命题，继续探索可用的触发词，以此扩大语料规模；（3）从更多的数据源中挖掘命题，扩大命题数据集的规模；（4）尝试隐式命题的挖掘；（5）进一步强化两个任务的模型，提高模型的学习能力。

参考文献

[1] 金岳霖等，形式逻辑[M]. 人民出版社，1979: 68.

[2] 华东师范大学哲学系逻辑学教研室. 形式逻辑(第 5 版)[M]. 华东师范大学出版社， 2015: 38-119.

[3] Cortes C , Vapnik V . Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[4] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[5] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[6] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. 2001: 282-289.

[7] Yang J, Zhang Y. NCRF++: An Open-source Neural Sequence Labeling Toolkit[J]. ACL 2018, 2018: 74.

[8] 高逢亮. 浅论形式逻辑对语言研究的作用[J]. 现代语文 (语言研究版), 2017 (12): 4-6.

[9] 李勤. 论句子语义中的命题[J]. 燕山大學學報 (哲學社會科學版), 2006, 7(1): 1-6.

[10] 马佩. 论直言判断的种类[C]// 逻辑学文集. 1978: 116-132.

[11] 李先焜. 语言、逻辑和语言逻辑[J]. 哲学研究, 1986(8):41-48.

[12] 周礼全. 形式逻辑和自然语言[J]. 哲学研究, 1993(12):29-35.

[13] 龚启荣. 当代形式逻辑及其在人工智能中的应用理论研究[M]. 电子工业出版社, 2011.

[14] 周文华. 命题的一个新定义与命题的同一性问题[J]. 云南大学学报(社会科学版), 2016, 15(4):49-55.

[15] 杨宏郝. 判断与命题辨析[J]. 学术论坛, 2000(1):20-23.

[16] 沈园. 逻辑判断基本类型及其在语言中的反映[J]. 当代语言学, 2000, 2(3):125-137.

[17] 韩铁稳. 浅谈逻辑常项的语言表达形式[J]. 思维与智慧：上半月, 1989, 3:10-11.

[18] 黄士平. 逻辑常项隐性形式初探[J]. 江汉大学学报：社会科学版, 1991(4):93-97.

[19] 张牧宇，秦兵，刘挺. 中文篇章级句间语义关系体系及标注[J]. 中文信息学报, 2014, 28(2): 28-36.

[20] 张牧宇，宋原，秦兵，等. 中文篇章级句间语义关系识别[J]. 中文信息学报, 2013, 27(6): 51-58.

[21] 段士平. 国内二语语块教学研究述评[J]. 中国外语, 2008, 5(4):63-67.

[22] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报, 2012, 26(1):73-84.

[23] Fleiss J L. Measuring nominal scale agreement among many raters[J]. Psychological bulletin, 1971, 76(5): 378.

[24] Li S, Zhao Z, Hu R, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 138-143.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[26] Lample G, Ballesteros M, Subramanian S, et al. Neural Architectures for Named Entity Recognition[C]//Proceedings of NAACL-HLT. 2016: 260-270.

刘璐（1994—），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:luliu.nlp@gmail.com

彭诗雅（1995—），第二作者，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:pengshiya_blcu@163.com

玉郴（1997—），第三作者，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:yuchen7312@gmail.com

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