文章编号：1003-0077（2017）00-0000-00

中文矛盾语块数据集构建和边界识别研究

李博涵[1] 姜姗[1] 刘畅[1] 于东[1]
（1.北京语言大学 信息科学学院，北京 100083）

摘要：正确理解文本矛盾是自然语言理解的一项基础性问题。目前的研究大多针对矛盾识别任务，深入文本内部探究矛盾产生原因的工作较少，且缺乏专门的中文矛盾数据集。该文在前人矛盾研究基础上，提出矛盾语块的概念，将其划分为 7 种类型，并根据标注规范构建了包含 16,224 条数据的中文矛盾语块（CCB）数据集。基于此数据集，利用序列标注及抽取式阅读理解类模型开展矛盾语块边界识别实验，以检验模型对矛盾内部语义信息的理解能力，结果显示阅读理解类模型在该任务上的性能优于序列标注模型。该文通过三个角度对影响语块边界识别的因素进行分析，为文本矛盾后续研究工作提供可靠的数据集和基线模型。

关键词：，自然语言理解；文本矛盾；矛盾语块

中图分类号：TP391 **文献标识码：**A

Research on Data Set Construction and Boundary Recognition of Chinese Contradictory Blocks

LI Bohan[1], JIANG Shan[1], LIU Chang[1] and DONG Yu[1]

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

Abstract :

Correctly understanding textual contradiction is a fundamental problem in natural language understanding. Most existing researches focus on contradiction detection tasks, but less work explores the causes of contradictions deep into the text. Moreover, fine-grained Chinese contradictory corpus is not available. Based on the contradiction types summarized by predecessors, we further clarify the concept of contradictory blocks, propose labeling guideline and build a Chinese Contradiction Block (CCB) dataset. Experiments on boundary recognition of contradictory blocks using sequence labeling models and extractive machine reading comprehension models on the dataset probe the ability of models understanding internal contradictory semantic information. We analyze the factors affecting the correct identification of block boundaries, and provide a baseline for follow-up research on this task.

Key words: Natural Language Understanding; Text Contradiction; Contradictory Block

0 引言

正确解析文本矛盾是自然语言理解中一项重要环节，其定义为：当拥有共同实体的两个句子极大概率不能同时成立时，这两个句子互相矛盾[1]。文本矛盾与多种自然语言处理任务有关：它是自然语言推断中三种文本关系类型之一；判断两个句子是否矛盾的任务被称为矛盾识别；而矛盾识别又可以应用到谣言检测[2]、新闻矛盾检测[3]，以及社交媒体上观点和情绪的矛盾检测[4]等相关领域中。

上述文本矛盾检测的任务通常围绕矛盾类型的特征展开，如用基于语义规则的方法[5]、逻辑推理的方法、应用编辑距离的对齐方法和反向模式匹配的方法来对矛盾关系进行识别[6][7]。随着大数据时代的到来，神经网络模型被运用在矛盾

收稿收稿日期：0-0-0；**定稿日期**：0-0-0

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（19YCX116）

识别任务上，如 Vijay 等人[8]使用 LSTM[9]在三个公开数据集上进行了矛盾检测的实验。然而，矛盾的现有工作大多关注文本矛盾的检测，并未探究引发矛盾现象的具体片段，使得矛盾检测任务缺乏可解释性。目前精细标注矛盾片段的数据集十分缺乏，因此无法基于该类数据集对矛盾做进一步探究。

针对以上问题，我们提出矛盾语块的概念，定义矛盾语块为导致句子矛盾的最小语言单元。同时为矛盾语块制定了标注规范，并根据该规范对中文自然语言推断数据集（Chinese Natural Language Inference，CNLI）[10]中的矛盾语料进行标注，构建了中文矛盾语块（Chinese Contradiction Block,CCB）数据集。我们在此数据集上进行了一系列矛盾语块边界识别的实验，以检验模型对文本矛盾的理解程度，验证模型能否精准定位导致文本矛盾的片段，分析影响模型识别矛盾语块边界的因素，也为该领域的进一步工作和相关实验提供了前提条件。

本文的内容分布如下：本章对本文工作进行简要介绍，第一章为相关工作，说明文本矛盾的工作的进展与相关数据集等。第二章为文本矛盾的界定、类型划分与标注规范。第三章重点描述了数据集的构建，包括数据集的选择以及量化分析等。第四章提出中文矛盾语块边界识别任务，介绍本文使用的序列标注与抽取式阅读理解两类模型。第五章在矛盾语块数据集上进行矛盾语块边界识别实验，对结果进行分析，并探究了影响实验效果的几类因素。第六章对本文工作进行总结。

1 相关工作

文本矛盾是一种复杂的语言现象，也是自然语言理解中的基本任务。2006 年 Harabagiu 等人[8]认为文本矛盾可由否定、反义、语义三个方面导致，并构建这三类矛盾的检测系统。De Marneffe 等人[1]在此基础上对文本矛盾进一步定义，认为文本矛盾的前提是有共同名词实体存在，且当给定两个句子大概率不能同时成立时，文本矛盾现象产生。该工作将文本矛盾分成 7 个类别，并基于其构造的矛盾数据集进行了矛盾识别的实验。Ritter 等人[11]强调外部知识对矛盾检测任务的影响，同时借用函数思想提出了针对外部知识的矛盾检测模型。

文本矛盾的数据分为两类：自然语言推断类数据集和专门的矛盾数据集。前者来源于自然语言推断任务，该任务数据集中的“矛盾”类别可抽取出来独立作为矛盾数据集使用。这类数据集包括 RTE 与 RITE 等自然语言推断评测数据集[13-15]、SNLI[16]、MultiNLI[17]、CNLI[8]数据集，和社交媒体领域的 PHEME 数据集[18]等等。第二种文本矛盾数据集是为矛盾检测任务设计的专用数据集，比如 Sem2012 评测数据集[19]和 SemEval2014 评测数据集[20]、从 RTE-3 数据集与维基百科等来源中抽取构建的综合数据集[1]，等等。

基于上述数据集进行矛盾识别是文本矛盾领域的常见任务。矛盾识别的经典方法是化用逻辑推理中矛盾的思想，用信息抽取的方法将句子中的信息表示为逻辑表达式的集合，基于这个集合推理判断语句是否矛盾[7][12]；也有学者利用依存句法树、编辑距离等方法评估句子相似度与词语的共现情况，如相似度极高的两句话中如果有一部分词序不同，则很有可能为矛盾关系[6]。刘茂福等人[21]将矛盾句对整合为事件图谱，利用事件图谱的方法在大数据中抽取矛盾信息。另外，文本矛盾中诸如命名实体识别、数字、反义词、否定等重要特征被广泛运用，作为分类的重要依据[22][8]。但有时准确界定文本矛盾所需的部分特征无法直接通过原文获取，需要外部知识的辅助，因此有工作通过矛盾领域的词向量[23]、“联想”搜索[24]、网络挖掘[25]等方法获取 WordNet[26]等外部知识库的信息，提升模型判断能力。

矛盾识别工作可扩展至多种应用。Badache 等人[2]运用矛盾检测系统，针对社交软件上的动态和回复，进行矛盾的强度评分。Lendvai P 等人[4]对社交软件中的谣言进行了分类，并加入人工特征辅助谣言检测的过程。Karimi H 等人[3]提供了针对新闻篇章的假新闻检测系统，而 Zadrozny[27]的矛盾检测模型可以运用在药物说明书的勘误工作里。

2 矛盾语块界定及标注规范

矛盾特征如反义词、否定词等在文本矛盾中较为常见。但在矛盾任务中，大多数工作聚焦于句子级别的矛盾识别，仅将特征作为辅助手段，而没有深入思考矛盾产生的深层原因。本文对矛盾句对内部片段归纳总结，提出矛盾语块的概念，探究其如何导致矛盾语义。

表 1 矛盾的不同类型

类型	例句	
反义	例 1	T1：一个男人在晚上走在一条繁忙的街道上，转过身**微笑**着。 T2：有一个人独自**生闷气**。
否定	例 2	T1：去年，克林顿**签署**了禁止联邦承认同性恋婚姻的立法，然后在竞选中吹嘘。 T2：禁止联邦承认同性婚姻的立法**从未被签署过**。
数字	例 3	T1：让受赠人 A 从 LSC 基金中获得 **3 万美元**，并以每箱 100 美元的成本卖出。 T2：受赠人 A 只能从 LSC 获得高达 **10000 美元**。
事件	例 4	T1：一条橙色的猫沿着牡丹花旁边的小路**奔跑**。 T2：一只猫在树林里**睡觉**。
句法	例 5	T1：**洛克希德公司宣布收购诺斯洛格**，完成国防工业合并。 T2：**诺斯洛格公司收购洛克希德公司**的价格非常低廉。
词义不兼容	例 6	T1：小女孩坐在桌子旁吃**香肠**。 T2：年轻女孩在吃**蛋糕**。
外部知识	例 7	T1：在罗德奥大道和代顿路的**拐角处**抓住它。 T2：罗德奥大道和代顿路的位置**完全平行**。

2.1 矛盾语块的界定与类型划分

本文参考 De Marneffe 等人[1]的定义，认为包含共同实体是两个句子矛盾的前提，且当它们所述内容极大概率不能同时成立时，矛盾成立。两个句子之间的共同实体包含了相同的语义信息，不为句间矛盾关系的判断提供帮助。排除掉此类共同事件后，两个句子剩余片段将包括不同的信息，这类片段信息导致句对矛盾，且句对矛盾类型受剩余片段形式的影响。因此定位此类片段是分析矛盾成因的关键。

为了进一步探究及分析上述片段，我们针对它提出矛盾语块的概念，定义其为*导致句对矛盾的最小语言单元*。矛盾语块是一类符合日常搭配习惯的文本片段，它能够完整表达句间矛盾的核心，且不包括对矛盾推断无用的冗余信息。由矛盾语块的不同类型可推出文本矛盾的不同类型，因此对语块的类型分析可借鉴文本矛盾类型的研究。本文参考 De Marneffe 等人[1]对文本矛盾 7 种类型的分类，对矛盾语块进行了相应的划分，如表 1 所示。

反义 矛盾语块的意义相反导致句子矛盾，如反义词。如例 1 中两个句子拥有共同实体“人”，T1 的“微笑”与 T2 的“生闷气”是反义词，描述了这个男人的相反情绪，导致该句对矛盾。

否定 矛盾语块带有明确的否定词，如“没有”、“不”、“从未”等等，否定词导致语义转折构成了整句的矛盾。如例 2 中 T2 的“从未”是 T1 中“签署”的否定，两者意思相左。

数字 矛盾语块中出现明确的数字，如数量、时间等，且数字的不一致导致两句话矛盾。如例 3“受赠人获得的收益”在两句话中数量不同，在 T1 与 T2 中分别为 3 万和 10000 美元，构成矛盾。

事件 由句中谓语语义不兼容造成的整句矛盾。例如例 4 中的猫在 T1 与 T2 中的动作分别是“奔跑”“睡觉”，两句的谓语显然无法同时发生，构成矛盾。

句法 两个句子句法结构不同，导致文本矛盾的产生。如例 5 中 T1 为洛克希德公司收购诺斯洛格公司，而 T2 却为诺斯洛格公司收购洛克希德公司，在谓语相同的情况下，主语与宾语位置调换，造成句对矛盾。

词义不兼容 句对中矛盾语块对共同实体的位置、状态、性质等描述存在差异，导致整句矛盾。词义不兼容表示词义不同，而非语义相反或否定，即该类型与前文的“反义”和“否定”类型不同。如例 6 中女孩所吃食物分别为“香肠”和“蛋糕”，两词既不是反义词，也没有否定词修饰，但它们词义互斥，导致了句对矛盾。

外部知识 部分矛盾句对较为复杂，难以从字面意思直接得出结论，需要借助外部知识来辅助判断。比如需要常识才能推导出例 7 中 T1 的“拐角处”表示两条街有交叉，与 T2 中它们“完全平行”相矛盾。

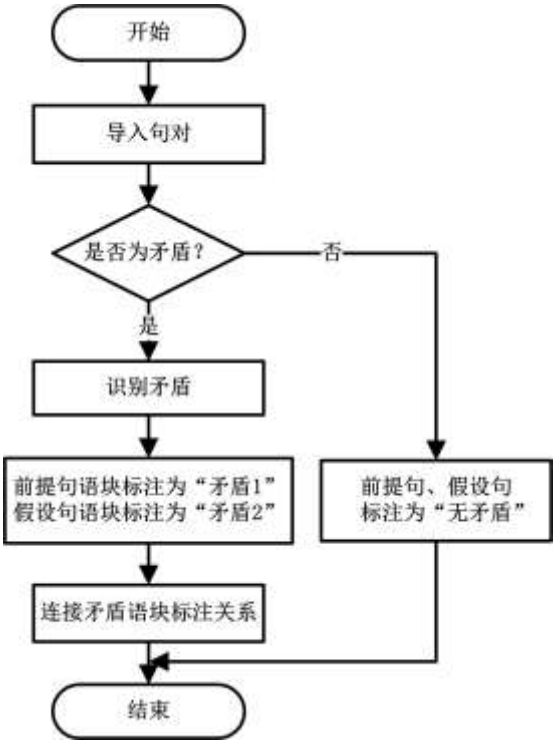

图 1 数据标注流程

2.2 矛盾语块标注规范

本文对标注工作制定如下标注规范：

标注内容 本文标注工作的内容是根据矛盾语块的定义与分类，在矛盾句对上对矛盾语块和语块类型进行标注。矛盾语块是导致整句间矛盾关系的最小语言片段，需要在两句话内分别进行标注，语块类型则需要参考语块的内容结构，在反义、否定、数字、事件、句法、词义不兼容、外部知识等 7 种类型中进行选择。

标注流程 标注流程如图 1 所示，首先导入待标注的文本，标注员需要通读并对比给定句对，确定句对间具有矛盾关系，并为矛盾类型进行大体的归类，之后划出句对中的矛盾语块并勾选语块类型；若判定句对不存在矛盾关系，则在两个句子上分别标注为“无矛盾”。

平台 采用轻量级在线文本标注工具 Brat（Brat Rapid Annotation Tool）[28]对中文矛盾语料进行结构化标注，网页效果如图 2 所示，标注结果以.ann 文件的格式由 Brat 自动保存在服务器端。

一致性 为保证标注一致性，标注中对标注员进行系统培训，并提供试标注环节便于其熟悉标注平台及标注规范，正式标注期间对标注内容定期进行简单抽样检测，正确率达到 90%以上。

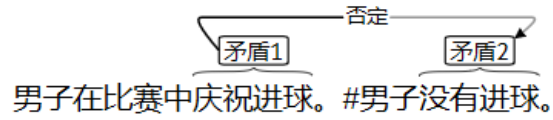

图 2 数据标注示例

3 数据集构建

我们从 CNLI 数据集中选取了部分矛盾类别的语料，对其进行人工标注，经过校对之后，获得 16,224 条有效的矛盾句对。

3.1 数据选择

CNLI 数据集为 SNLI 与 MultiNLI 两个数据集的部分语料经过翻译及人工校准而成，数据总量达 11 万，其中训练集、验证集和测试集的数量分别为 9 万、1 万和 1 万条，文本蕴含、矛盾与中立的数据比例大致为 1：1：1。每条数据都包括一个前提句和一个假设句，以及“蕴含”、“矛盾”或“中立”的关系标签。

其中，矛盾类型数据的数量为 35,696 条，其前提句平均长度为 20 字，假设句的平均长度为 14 字。该数据长度适中，不会由于句式结构过于复杂对标注和分析造成干扰，也不会因为长度太短而失去研究价值。我们规定前提句与假设句在数据集中分别作为矛盾句对中的 T1 与 T2，按照前文的标注规范对 CNLI 的矛盾类型数据进行了人工标注，构建中文矛盾语块（Chinese Contradiction Block, CCB）数据集。

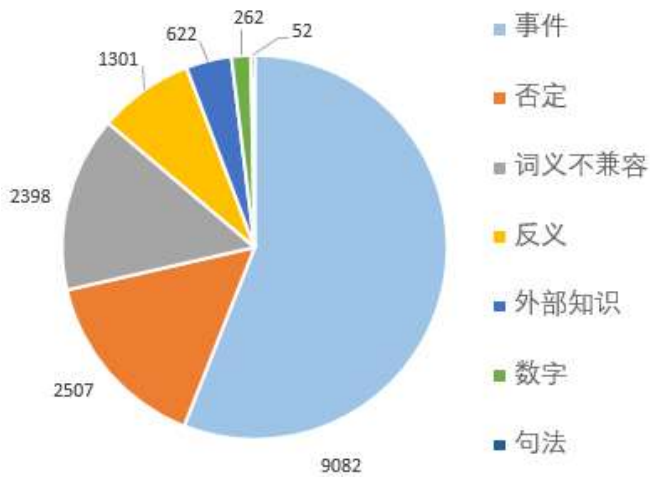

图 3 CCB 数据集中矛盾语块类型分布

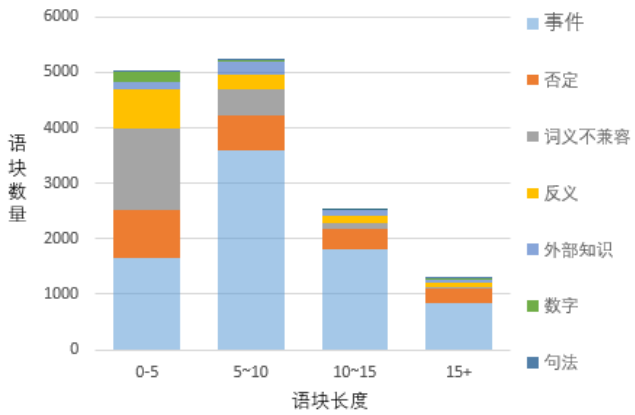

图 4 不同语块类型的长度及数量分布

3.2 数据集量化分析

CCB 数据集包含 16,224 条有效数据，其训练集包含 14,224 条句对，验证集和测试集各包含 1,000 条句对。数据集内各类别语块数量占比及长度如图 3、图 4 所示。通过分析可知，CNLI 数据集矛盾类别的语料中，半数以上句子内部的矛盾语块都为“事件”类型，占总量的 55.98%；“否定”与“词义不兼容”次之。从语块长度分布来看，长度范围在 0-10 的语块数量最多，其中非“事件”类型的语块长度大多在 0-5 之间，而事件类型的语块集中于 5-10。这是因为“事件”类型同时涉及到谓语和宾语，导致该类型语块较长。语块长度大部分小于 20，约占 86.34%；超过 20 个字的句子较少，约占 13.66%。

4 中文矛盾语块边界识别方法

本文将基于 CCB 数据集进行矛盾语块边界识别的实验。中文矛盾语块边界识别是预测给定矛盾句对中矛盾语块边界位置的任务，该任务需考虑语块边界信息的表示方法，即如何将语块边界的位置转化为模型能够表示的形式。本文采用序列标注和抽取式阅读理解两类模型，将边界信息分别转化为序列标签信息和语块始末索引，通过实验检验了不同类型的神经网络模型对矛盾语块边界识别的性能，并进一步分析了实验中影响语块边界识别效果的因素。

序列标注是将文本序列的每一个字符映射至分类标签集合中的任务。本文采用序列标注模型，将矛盾语块边界信息转化为序列标签信息，实验结果中带相同类型标签的连续子序列即为矛盾语块。

抽取式阅读理解任务为给定一段文本和一个问题，从文本中摘取一个片段作为问题的答案。本文将句对中的一句作为文本，另一句作为问题，利用问题的整句信息在文本中寻找对应的矛盾片段。之后将两句调换位置，寻找另一句的矛盾片段。对前提句与假设句中的矛盾片段预测是独立进行的，用两次分别预测的结果共同作为句对矛盾语块边界的索引。

4.1 序列标注模型

神经网络条件随机场模型 Jie Yang 等人在 2018 年提出用于神经序列标注的名为 NCRF++[29]的工具包，旨在通过条件随机场[30]推理层快速实现不同的神经序列标记模型。本实验中选择了其中的 BiLSTM 与条件随机场结合而成的神经网络条件随机场模型（Neural Conditional Random Field, NCRF)进行实验。NCRF 模型由两部分组成：序列层和推理层。序列层使用 BiLSTM 作为编码器来捕获从左到右和从右到左的序列信息，获得序列的特征表示。推理层接收这些特征表示，把序列映射到标签数量大小的向量上，通过条件随机场对每个单词的标签概率进行建模，即利用特征信息为序列中的每个词打上标签。

Bert-NER 模型 BERT 模型（Bidirectional Encoder Representations from Transformers）[31]是一个多层双向 Transformer[32]编码器，利用自注意力机制，来预训练文本的深层双向表示。预训练的 Bert 表示通过微调（fine-tuned）和改变输出层的方法在阅读理解、自然语言推断、情感分析等多个 NLP 任务上取得了最优成绩。本实验中我们选择应用于命名实体识别任务的 Bert-NER 模型作为序列标注模型进行矛盾语块边界识别实验。

4.2 抽取式阅读理解模型

MatchLSTM-PointNet 模型 Wang 等人[33]在 2017 年针对阅读理解任务提出了结合 MatchLSTM 和 Pointer Net 的模型。我们将原模型中加在前提句上的注意力机制同时应用于前提句与假设句形成 MatchLSTM-PrtNet 模型。该模型中的 MatchLSTM 作为编码器负责对序列信息建模，获得前提句和假设句双向注意力机制加权的向量表示，通过这种改进获得两个句子之间充分的交互和匹配信息；Pointer Net[34]作为解码器根据 MatchLSTM 生成的句子向量生成矛盾语块边界下标，直接定位矛盾语块范围。

表 2 NCRF 和 BERT-NER 实验结果

	准确率（Acc）	精确率(P)	召回率(R)	F1 值(F)
NCRF	76.84%	0.4975	0.4470	0.4709
Bert-NER	80.07%	0.5649	0.5331	0.5486

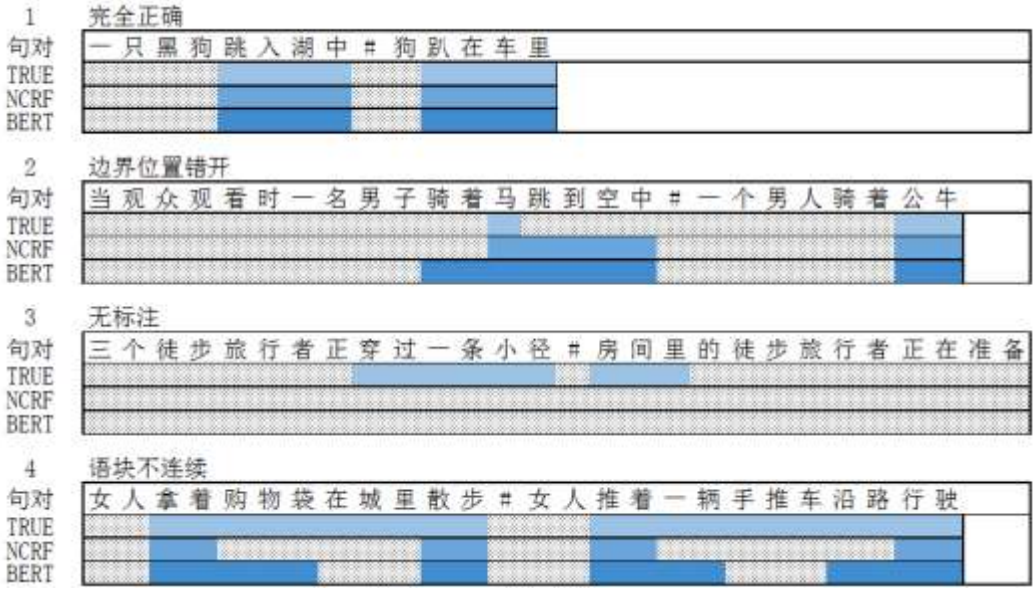

图 5 NCRF 和 BERT-NER 错误分析

Bert-MRC 模型 本文选择 Bert 经过微调后应用于 Squad 数据集[35]上的阅读理解模型，即 Bert-MRC 模型，进行矛盾语块边界识别实验。此模型将 Bert 作为编码器，取出其最后一层的信息送入推理层，得到句中每个字作为矛盾语块边界元素的概率，将概率最大的两个字分别作为始末边界，将它们在句中的索引作为输出。

5 实验结果与分析

5.1 实验设计

本文的矛盾语块边界识别是一项新的任务，与识别句对是否矛盾的工作[5]有所不同，因此尚无相关工作提供可与之比较的基线模型。我们将语块边界信息转化为序列标签和语块始末索引，使用较为通用的序列标注类模型和阅读理解类模型在 CCB 数据集上进行实验，验证模型对矛盾语块的理解能力，比较模型预测语块边界的性能。两类模型的不同输出导致其评价方式不同，直接进行类间比较没有意义。因此在类型内部比较的基础上，用完全匹配准确率统一 4 个模型的评价方式，直观对比两类模型在矛盾语块边界识别任务上的效果，并从虚词、矛盾语块类型、句子长度三个角度分析了语块识别准确率的影响因素。

5.2 序列标注模型

在两个序列标注模型中，将前提句和假设句用“#”拼接为一句作为输入。将矛盾语块边界信息转化为 BIO 格式[36]的序列标签信息，即用 B 表示语块的初始文字，用 I 表示语块内其他部分，用 O 表示句中非语块的片段。模型输出同样采用 BIO 格式标签表示预测语块的边界信息。

本文使用准确率、精确率、召回率与 F1 值四种指标对 NCRF 与 Bert-NER 两个序列标注模型的预测结果进行评价，如表 2 所示。NCRF 模型作为基线，取得了 76.84%的正确率；Bert-NER 模型比 NCRF 模型高 3.23%。且 Bert-NER 模型的精确率、召回率、F1 值均高于 NCRF 模型。产生差距的原因是 NCRF 模型作为一个简单的序列标注模型，仅通过 BiLSTM 对输入序列进行编码，没有对句子间的语义交互信息进行建模，因此不能很好的处理矛盾语块边界识别任务，实验性能一般。而 Bert-NER 模型包含了多层应用自注意力机制的 Transformer 模型，可以针对输入序列进行句间和句内的交互信息计算及充分的信息融合，同时

表 3 MatchLSTM-PrtNet 和 BERT-MRC 实验结果

模型	完全匹配（EM）	精确率（P）	召回率（R）	F1 值
MatchLSTM-PrtNet	13.11%	0.5643	0.46	0.511
Bert-MRC	40.20%	0.8387	0.7981	0.8179

预训练好的语言模型也携带了丰富的语义信息。另外，经过在矛盾数据集的微调训练，模型对矛盾关系的理解力进一步增强，因此实验性能突出，其预测内容更能准确判断矛盾语块边界。

我们在测试集中随机抽取了 100 条数据对 NCRF 模型与 Bert-Ner 模型的标注情况进行了具体分析。若将真实语块与预测语块的始末索引完全相同视为正确样例，则可以归纳出三种常见错误类型：边界错开、无标注与语块不连续，见图 5。

边界位置错开 模型预测的矛盾语块边界与真实矛盾语块边界没有完全重合，存在若干字的误差。

无标注 模型未对句子进行标注，导致文本的语块信息为空白。这种句子占比较少，且通常为长度较长、句法结构复杂的句子，这为识别语块边界造成了一定的困难。

标注语块不连续 模型预测的矛盾语块在中部断裂。这是因为模型在预测矛盾语块边界时，将最后一层序列信息中每个字符向量通过分类层独立地映射到相应的标签集合，导致语块内部的一些字符被识别成语块外的元素。经过统计发现，部分情况下将此类数据中的多个语块拼接可组成真实语块，这启发我们使用抽取式阅读理解模型进行矛盾语块边界识别实验。

图 6 将 NCRF 与 Bert-NER 模型的表现进行对比。被正确预测的句子通常长度较短，且句子简单清晰，前提句和假设句的文本相似度很高，只有部分位置的词和短语不一致。三种错误类型中，边界位置错开的情况最多，且该类型上 NCRF 的错误率明显高于 Bert-NER，相比于后者，NCRF 不能很好地精确定位矛盾语块的始末位置。无标注与语块不连续两种错误较为少见。经过统计，模型未进行标注的句子通常较长，长度大于 20。此类句子通常有更复杂的句法结构和更高的语义理解难度，而模型的理解能力有限，未能做出相应标注。在语块不连续类型中，NCRF 模型的错误数量小于 Bert-NER 模型，经过统计发现，这是因为 NCRF 更倾向于将不够确定的字段一并划入语块中，预测的语块长度较长。从另一个侧面导致其边界位置错开错误类型数量更多，但语块不连续问题较少。

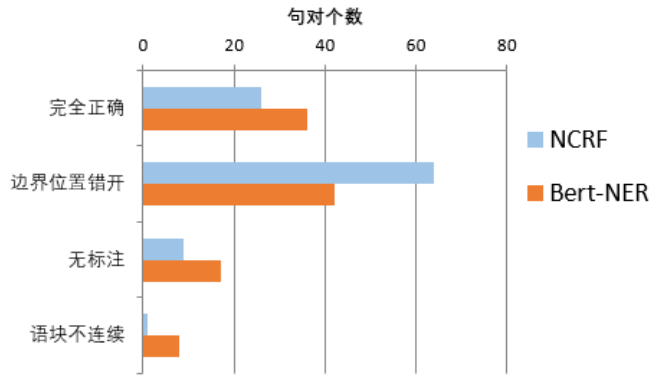

图 6 NCRF 和 BERT-NER 结果分析

5.3 抽取式阅读理解模型

本文使用完全匹配（Exact Match，EM）与 F1 值两种指标对两个阅读理解模型的预测结果进行评价，如表 3 所示。基线 MatchLSTM-PrtNet 模型的完全匹配正确率为 13.11%；Bert-MRC 模型的正确率为 40.2%，超过前者 27.09%。MatchLSTM-PrtNet 模型是基于外部注意力机制的匹配模型，可以获取句子间的交互信息：通过一个句子对另一个句子计算注意力权重，以此建模包含了两个句子之间相关程度的匹配向量，并利用该匹配向量生成矛盾语块边界下标。而 Bert-MRC 模型内大量的 Transformer 结构应用了自注意力机制，因此模型可以获取到丰富的矛盾语义信息，并将矛盾语义信息映射为两个句子上的矛盾语块。通过 F1 值也可看出，Bert-MRC 模型的综合性能优于 MatchLSTM-PrtNet 模型，对矛盾语块信息的提取能力更为突出。

与序列标注模型不同，抽取式阅读理解本身的输出格式规避了序列标注模型的部分错误类型：它的输出是矛盾语块的起始和结尾索引，因此输出结果是连续的，不存在语块不连续的错误；模型对每个字作为边界始末元素的概率进行处理，选出概率最大的字作为结果，因此不存在无标注的错误。对于抽取式阅读理解模型而言，导致预测错误的原因只有边界位置错开的问题。通过对比两模型的正确率可知，Bert-MRC 模型的边界位置错开的概率低于 MatchLSTM-PrtNet 模型。

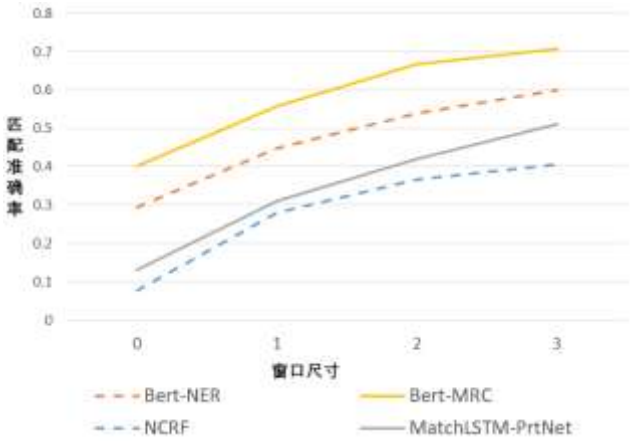

图 7 松弛实验结果，虚线代表序列标注类模型，实线代表阅读理解类模型

5.4 模型对比分析

由于序列标注模型与抽取式阅读理解模型的原理、输出与评价方式不尽相同，我们使用完全匹配的准确率作为统一的评价方式，直观衡量各模型的表现。该种评价方式要求模型对一个句对中的前提句与假设句的矛盾语块同时作出完全正确的判断，对模型的性能要求较高。

本文针对边界错开的错误类型，对 4 种模型的实验结果统一进行松弛处理，在真实边界结果两侧开尺寸为 1-3 的窗口，界定如果模型的预测边界在软边界中即为匹配成功。松弛操作后 4 种模型的准确率如图 7 所示，窗口大小为 0 时的表现是松弛前各模型的完全匹配准确率。整体来看，增加窗口对模型完全匹配的准确率有明显帮助，每新增加一个尺寸的窗口，模型准确率平均增加 17.32%、9.8%、5.8%，说明预测边界与真实边界索引误差为 1 的错误样例占较大比例。对比 4 类模型的表现可以看出，在增加窗口前表现更为优秀的 Bert-MRC 模型，增加窗口后性能依然稳定，准确率始终高于其他模型。

图中实线与虚线分别代表序列标注模型和阅读理解模型，可以看到相同类别内的两个模型准确率相差较大，以 Bert 为编码器的 Bert-NER、Bert-MRC 模型准确率明显高于以 BiLSTM 为编码器的 NCRF、MatchLSTM-PrtNet 两个模型。这证明了包含自注意力机制的预训练语言模型对矛盾语义理解的有效性。

从模型类别的角度分析，当模型使用相同编码器时，图中实线表示的阅读理解类模型准确率恒高于虚线表示的序列标注类模型，即 Bert-MRC、MatchLSTM-PrtNet 模型的表现分别优于 Bert-NER、NCRF 模型。经过分析，首先阅读理解模型的输出直接为文本片段的始末索引，避免了序列标注模型的无标注与语块不连续错误类型。其次，序列标注模型在分类层将句中每个字独立地映射在标签集合中，用标签代表矛盾语块边界，在进行预测时更强调每个字单独的类别。而抽取式阅读理解模型则将两个句子分别作为文本与问题，利用问题的信息在文本寻找相应的片段，强调了两个句子之间的信息交互和句子整体语义的理解。因此抽取式阅读理解模型更符合矛盾语块边界识别的思想与需求，实验效果更佳。

5.5 影响因素分析

本节从虚词、矛盾语块类型、句子长度三个角度研究了它们对矛盾语块边界识别正确率的影响，选取序列标注类与抽取式阅读理解类模型中性能较好的 Bert-NER 与 Bert-MRC 模型的实验结果进行分析。

虚词 在错误分析中我们发现了部分语块边界错开由边界附近的虚词导致。因此本文对 Bert-NER 和 Bert-MRC 模型的预测结果与真实结果之间的误差片段进行统计，并分别输出出现频率最高的前 5 个误差片段及其数量。

在 1000 条的测试集中，Bert-NER 模型与真实结果有 239 条误差片段，出现频率最高的片段分别为:“在”、“正在”、“正”、“上”和“站在”;Bert-MRC 模型有 269 条误差片段，出现频率最高的片段分别为:“在”、“正在”、“正”、“站在”和“坐在”。两个模型的误差片段中，出现频率最高的都是“在”“正在”“正”三个虚词，且绝大多数其他误差片段也包含虚词。这些虚词容易造成预测边界与真实边界位置错开的问题，降低了语块边界的正确率。

矛盾语块类型 本文对矛盾语块在不同类型上的正确率做出分析，如图 8。事件类型的语块正确率明显优于其他类型，这是因为该类型数据的数量占比超过半成，模型对其训练充分，正确率较高。反义和词义不兼容两种类型的数据量次之，正确率相比事件类型有所下降。在数字和反义类型上，Bert-MRC 模型明显优于 Bert-NER 模型，经过错误分析发现这是因为数字、否定类型的语块不仅包括数量词与否定词，还包括数量词或否定词修饰的片段。序列标注模型在对单字进行映射的时候，未将这些片段识别为矛盾语块内部的一部分，导致出错。两个模型在外部知识类型上的准确率均高于它们在反义、词义不兼容类别上

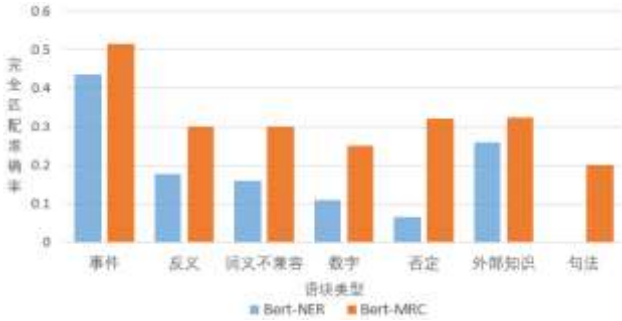

图 8 矛盾语块类型对模型正确率的影响

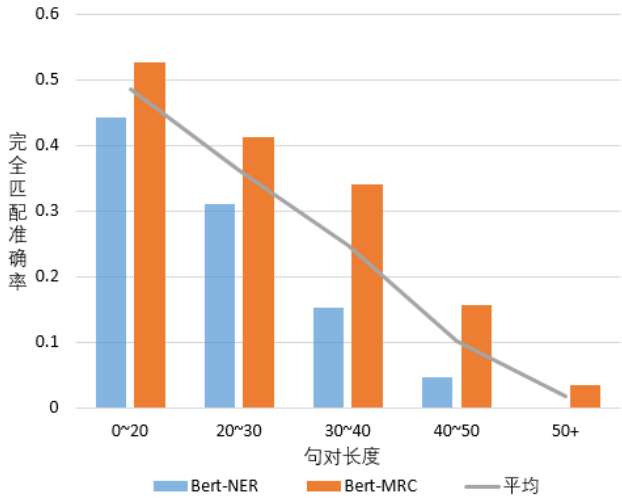

图 9 数据长度对模型正确率的影响

的正确率，说明 Bert 预训练语言模型包括了一定的外部知识，可以提高模型在该类型上的识别能力。模型在句法类别上的正确率很低且参考价值有限，因为句法类别的数据占比最少，导致训练不充分，且句法类型语块通常较长，对边界识别带来挑战。

句对长度 本文统计了句子长度与两个模型完全匹配正确率之间的关系，如图 9 所示。模型识别矛盾语块边界的性能和语料长度呈负相关。句长小于 20 的句子矛盾语块边界识别的正确率较高，平均达到 49%。随着句子长度的增加，句子数据量下降，且正确率也随之降低，当句子长度超过 50 时，完全匹配的正确率平均仅有 1.7%。这是因为当句子长度增加时，其句法结构和语义的复杂性也随之增加，增加了模型正确判断的难度，因此模型判别矛盾边界的正确率随之降低。

6 结论

本文基于文本矛盾的研究现状，提出矛盾语块的概念，将其归为 7 种类型并制定相应的标注规范。同时，对中文自然语言推断 CNLI 数据集矛盾类别的数据进行矛盾语块及类型的标注工作，构建了中文矛盾语块(CCB)数据集，并在该数据集上进行了矛盾语块边界识别的实验。实验基于序列标注与抽取式阅读理解两类模型，统一模型评价方式，比较了它们在该任务上的不同正确率，并分析矛盾语块边界识别效果的影响因素。本文工作检验了模型对矛盾语义的理解，并为文本矛盾的后续研究提供了基线模型。

参考文献

[1] De Marneffe M C, Rafferty A N, Manning C D. Finding contradictions in text[J]. Proceedings of ACL-08: HLT, 2008: 1039-1047.

[2] Badache I, Fournier S, Chifu A G. Contradiction in Reviews: is it Strong or Low? [C]//ECIR-Analysis of Broad Dynamic Topics over Social Media (BroDyn). 2018.

[3] Karimi H, Tang J. Learning Hierarchical Discourse-level Structure for Fake News Detection[J]. arXiv preprint arXiv:1903.07389, 2019.

[4] Lendvai P, Reichel U D. Contradiction detection for rumorous claims[J]. arXiv preprint arXiv:1611.02588, 2016.

[5] 刘茂福，王月，顾进广. 基于语义规则的中文矛盾关系识别方法[J]. 计算机工程与科学，2015，37(4)：806-812.

[6] Padó S, de Marneffe M C, MacCartney B, et al. Deciding Entailment and Contradiction with Stochastic and Edit Distance-based Alignment[C]//TAC. 2008.

[7] Gutiererz F, Dou D, Fickas S, et al. Online reasoning for ontology-based error detection in text[C]//OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2014: 562-579.

[8] Harabagiu S, Hickl A, Lacatusu F. Negation, contrast and contradiction in text processing[C]//AAAI. 2006, 6: 755-762.

[9] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[10] 金天华，姜姗，于东，等. 中文句法异构蕴含语块标注和边界识别研究[J]. 中文信息学报，33(2)：17-25.

[11] Ritter A, Downey D, Soderland S, et al. It's a contradiction---no, it's not: a case study using functional relations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 11-20.

[12] MacCartney B, Manning C D. Modeling semantic containment and exclusion in natural language

inference[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 2008: 521-528.
[13] Giampiccolo D, Magnini B, Dagan I, et al. The third pascal recognizing textual entailment challenge[C]//Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing. Association for Computational Linguistics, 2007: 1-9.
[14] Giampiccolo D, Dang H T, Magnini B, et al. The Fourth PASCAL Recognizing Textual Entailment Challenge[C]//TAC. 2008.
[15] Giampiccolo D, Magnini B, Dagan I, et al. The third pascal recognizing textual entailment challenge[C]//Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing. Association for Computational Linguistics, 2007: 1-9.
[16] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference[J]. arXiv preprint arXiv:1508.05326, 2015.
[17] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference[J]. arXiv preprint arXiv: 1704.05426, 2017.
[18] Lendvai P, Augenstein I, Bontcheva K, et al. Monolingual Social Media Datasets for Detecting Contradiction and Entailment[C]//LREC. 2016.
[19] Morante R, Blanco E. * SEM 2012 shared task: Resolving the scope and focus of negation[C]//* SEM 2012: The First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). 2012, 1: 265-274.
[20] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. COLING 2014.
[21] Liu M, Wang L, Nie L, et al. Event graph based contradiction recognition from big data collection[J]. Neurocomputing, 2016, 181: 64-75.
[22] Marques R. Detecting Contradictions in News Quotations[D]. Master's thesis, IST, University of Lisbon, November, 2015.
[23] Li L, Qin B, Liu T. Contradiction detection with contradiction-specific word embedding[J]. Algorithms, 2017, 10(2): 59.
[24] Hashimoto C, Torisawa K, De Saeger S, et al. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 619-630.
[25] Shih C, Lee C, Tsai R T, et al. Validating contradiction in texts using online co-mention pattern checking[J]. ACM Transactions on Asian Language Information Processing (TALIP), 2012, 11(4): 17.
[26] MillerGA. Wordnet: A lexical database for english[J]. Commun.ACM.
[27] Zadrozny W, Hematialam H, Garbayo L. Towards semantic modeling of contradictions and disagreements: A case study of medical guidelines[J]. arXiv preprint arXiv:1708.00850, 2017.
[28] Stenetorp P, Pyysalo S, Topić G, et al. BRAT: a web-based tool for NLP-assisted text annotation[C]//Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012: 102-107.
[29] Yang J, Zhang Y. Ncrf++: An open-source neural sequence labeling toolkit[J]. arXiv preprint arXiv:1806.05626, 2018.
[30] Lafferty J D, McCallumA, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] // [S.l.:s.n.], 2001.
[31] 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.
[32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[33] WangS, JiangJ. Machine comprehension using match-lstm and answer pointer[M]. [S.l.:s.n.], 2016.
[34] Vinyals O, Fortunato M, Jaitly N. Pointer networks[C]//Advances in Neural Information Processing Systems. 2015: 2692-2700.
[35] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text[J]. arXiv preprint arXiv:1606.05250, 2016.
[36] Ramshaw L A, Marcus M P. Text chunking using transformation-based learning[M]//Natural language processing using very large corpora. Springer, Dordrecht, 1999: 157-176.

李博涵（1996—），硕士研究生，主要研究领域为自然语言处理。
E-mail：hcina@163.com

姜姗（1995—），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_blcu@126.com