软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2020,31(12):3772−3786 [doi: 10.13328/j.cnki.jos.005885]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 中文文本蕴含类型及语块识别方法研究*

于 东, 金天华, 谢婉莹, 张 艺, 荀恩东

(北京语言大学 信息科学学院,北京 100083)

通讯作者: 荀恩东, E-mail: edxun@126.com

**摘 要**: 文本蕴含识别(RTE)是判断两个句子语义是否具有蕴含关系的任务.近年来英文蕴含识别研究取得了较大发展,但主要是以类型判断为主,在数据中精确定位蕴含语块的研究比较少,蕴含类型识别的解释性较低.从中文文本蕴含识别(CNLI)数据中挑选 12 000 个中文蕴含句对,人工标注引起蕴含现象的语块,结合语块的语言学特征分析归纳了 7 种具体的蕴含类型.在此基础上,将中文蕴含识别任务转化为 7 分类的蕴含类型识别和蕴含语块边界-类型识别任务,在深度学习模型上达到 69.19%和 62.09%的准确率.实验结果表明,所提出的方法可以有效发现中文蕴含语块边界及与之对应的蕴含类型,为下一步研究提供了可靠的基准方法.

**关键词**: 文本蕴含识别;语块识别;蕴含类型;深度学习

**中图法分类号**: TP18

中文引用格式: 于东,金天华,谢婉莹,张艺,荀恩东.中文文本蕴含类型及语块识别方法研究.软件学报,2020,31(12):3772−3786. http://www.jos.org.cn/1000-9825/5885.htm

英文引用格式: Yu D, Jin TH, Xie WY, Zhang Y, Xun ED. Recognition method based on deep learning for Chinese textual entailment chunks and labels. Ruan Jian Xue Bao/Journal of Software, 2020,31(12):3772−3786 (in Chinese). http://www.jos.org.cn/1000-9825/5885.htm

## Recognition Method Based on Deep Learning for Chinese Textual Entailment Chunks and Labels

YU Dong, JIN Tian-Hua, XIE Wan-Ying, ZHANG Yi, XUN En-Dong

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

**Abstract**: Recognizing textual entailment (RTE) is a task to recognize whether two sentences have an entailment relationship. In recent years, RTE in English had made a great progress. The current researches are mainly based on type judgment, and pay less attention to locate the language chunks that lead to the entailment relationship. More over, it leads to a low interpretability of the RTE models. This study selects 12 000 Chinese entailment sentence pairs from the Chinese Natural Language Inference (CNLI) data and labeled chunks which lead to their entailment relationship. Then 7 entailment types are summarized considering Chinese linguistic features. On the basis, two tasks are proposed. One is to recognize the seven-category of entailment type for each entailment sentence pairs, another is to recognize the boundaries of the entailment chunks in it. The proposed deep learning based method reaches an accuracy of 69.19% and 62.09% in the two tasks. The experimental results show that proposed approaches can effectively identifying different types of entailment in Chinese and find the boundaries of the entailment chunks, which demonstrate that the proposed model provides a reliable benchmark for further research.

**Key words**: recognizing textual entailment; chunk labeling; deep learning

人工智能的发展离不开自然语言处理,而深度学习模型的进步,使得机器可以更容易地理解自然语言.自然语言处理很重要的一点就是实现文本的深度理解,进而在大量文本之间进行语义推理,促进阅读理解、问答系统、文本摘要等垂直任务的发展.

* 基金项目: 国家重点研发计划(2018YFB1005105)
Foundation item: National Key Research and Development Program of China (2018YFB1005105)
收稿时间: 2019-04-02; 修改时间: 2019-06-05; 采用时间: 2019-09-07

在这个过程中,文本蕴含识别(recognizing textual entailment,简称 RTE)是极为基础和重要的环节.文本蕴含是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).作为文本蕴含的基本任务,文本蕴含识别以语义理解为基础,判断两个句子之间语义关系.如果两个句子具有蕴含关系,那么这两个句子被称为蕴含句对.例如:

(1) *P*:一名男子与一名男孩说话. *H*:一名成年人与一名儿童说话.

例(1)中,“男子”与“成年人”是上下位词,“男孩”和“儿童”也是上下位词,所以 *P* 和 *H* 是由上下位词导致的蕴含现象.

(2) *P*:一位欣赏蝴蝶的年轻女孩. *H*:一个女孩很欣赏蝴蝶.

例(2)中,*P* 是由定中短语构成的陈述句,*H* 是将定中短语转变成主谓结构的陈述句,两句话的语义内容一致,但是句法结构不同,所以 *P* 和 *H* 是由句法变换导致的蕴含现象.

(3) *P*:一对年轻夫妇刚刚订婚. *H*:一对夫妇刚刚订婚.

例(3)中,*P* 中主语“夫妇”的定语修饰语“年轻”在 *H* 中被省略,所以 *P* 和 *H* 是由省略变换引起的蕴含现象.

从以上 3 组例句看到,蕴含关系取决于句子中标有下划线的语言成分.当两个句子中对应的语言成分具有蕴含关系,那么这两个句子就是蕴含的,反之则不然.这些语言成分被称为“语块”,这一概念最早来自美国心理学家、认知学家 Miller,她于 1956 年首次提出了记忆中“组块”,后被语言学家移植到语言领域[2,3].Wray[4]认为:语块是一个存储在大脑中的整体预制块,在使用时从记忆中被整块调用,而不是按照语法规则产出或分析的连续或非连续的由词汇构成的语串.在文本蕴含研究中,我们把导致蕴含关系的语块称为蕴含语块.蕴含语块介于词和句子之间,具有独立的语义和语用形式,蕴含语块之间的关系类型决定了蕴含句对之间的关系类型.

但是蕴含语块研究尚未得到广泛关注,很少有研究者尝试用现有模型发现蕴含语块,也很少有研究根据语块解释具体的蕴含类型和其中包含的推理机制.而语言学对于蕴含的研究集中于概念定义和逻辑命题证明[5,6],因此需要从大规模蕴含数据中标注出蕴含语块,分析其中的语言学特征,归纳形成一套较为系统的中文蕴含类型体系.该体系有利于直观描述蕴含的本质特征,加强人们对蕴含现象的理解,提升模型对文本蕴含识别的解释力.本文人工标注了 12 000 个中文蕴含句对,从词汇、句法、常识推理等 3 个角度归纳中文文本蕴含类型,并从 3 大类延伸出 7 个具体小类.

我们将文本蕴含识别任务细化为蕴含类型识别和蕴含语块-类型识别两个子任务.蕴含类型识别可以转化为分类任务.目前的蕴含类型识别受益于大规模数据集和深度神经网络模型,通常使用带有注意力机制的 LSTM(long short term memory)模型预测蕴含标签[7−9].现有针对蕴含语块识别的研究比较少,主要是用对齐的方法找出蕴含句对中相似部分[10],模型不需要理解句子的语义信息.另一方面,蕴含类型识别可以共享语块识别任务中得到的语义知识,但不能解决多种蕴含类型同时出现在一组蕴含句对里的情况.因此,我们提出蕴含语块-类型识别任务,它可以转化为序列标注任务,“位置-类型”为标签,标注出语块在句子中的位置信息和蕴含类型.我们用 ESIM[11],BERT[12]等模型作为基线,在标注数据上分别实现了 7 分类的蕴含类型识别任务和 17 分类的蕴含语块边界识别任务.

本文贡献在于以下 3 点.

(1) 数据方面,归纳了 7 个中文文本蕴含的类型,经过人工标注得到 12 000 条中文蕴含语块类型数据,为中文文本蕴含研究提供新的参考;

(2) 实验方面,将 ESIM 模型和 BERT 模型迁移到中文蕴含数据上,做了两个相关的任务,证明了带有注意力机制的模型在中文文本蕴含上是可行的(https://github.com/blcunlp/CTECL);

(3) 任务方面,提出一个同时预测蕴含语块边界和蕴含类型的新任务,扩展了文本蕴含研究内容,促进了蕴含研究的发展.

本文第 1 节介绍文本蕴含任务的相关工作.第 2 节介绍基于语言学特征的中文文本蕴含分类体系.第 3 节介绍蕴含语块类型的标注情况,并对数据进行类型和结构分析.第 4 节说明使用深度学习模型进行中文文本蕴含类型识别的算法,进行实验并分析结果.第 5 节说明使用深度学习模型进行中文文本蕴含语块-类型识别的算

法,进行实验并分析结果.第 6 节对全文内容进行总结.

## 1 相关工作

Dagan 等人[13]提出:自然语言中同一个浅层语义可以有多种不同的表达形式,这些表达形式之间存在一定的推理关系,即文本蕴含关系,如果找出不同表达之间的蕴含关系,那么就能够有效提升信息检索、文本摘要、机器翻译等应用的系统稳定性.经过 10 多年的研究发展,文本蕴含已成为了自然语言处理中一项基础且重要的任务,在数据制作、实验方法等方面取得了很大进展,有力地推动了语言理解和推理的发展.

现有的蕴含数据集仍以英文为主,早期有 RTE(recognizing textual entailment)评测[14−16]和 RITE(recognizing inference in text)评测[17−19]提供的小型数据集.随着深度神经网络的发展,SNLI[7],MNLI[20],QNLI[21]等由研究者专门为文本蕴含及其他推理任务制作的大规模数据集也被广泛应用.中文蕴含数据方面,早期有 RITE-2[18]和 RITE-3[19]提供的简体中文和繁体中文的蕴含数据.除此之外,CCL2018 的中文文本蕴含评测发布了包含 11 万条数据的中文自然语言推理数据集 CNLI[22].刘焕勇利用英汉对齐翻译的方法构建了 88 万的中文蕴含数据[23].以上数据集都提供了 2 分类标签“蕴含-非蕴含”或 3 分类标签“蕴含-中立-矛盾”,这些标签只能从整体上区分句子之间是蕴含还是矛盾关系,较为粗泛.想要深入研究蕴含类型以及导致蕴含的推理机制,需要对蕴含现象进行更为细致的分类.

Dagan 和 Glickan[14]从宏观角度把英语蕴含类型分成公理规则(axiom rule)、自反性(reflexivity)、单调性扩张(monotone extension)、限制性扩张(restrictive extension)、传递链(transitive chaining)等 5 类,这些概念较为抽象,在实际标注中很难操作.RITE-3[19]任务针对中文数据提出了 19 类蕴含现象和 9 类矛盾现象,任函[24]提出了 20 个面向汉语文本推理的语言现象标注类别,包含了同义词(近义词)、上下位词、时态、句法等.为建立专门的文本蕴含推理数据集,Bentivogli 等人[25]分析了 RTE-5 数据集中包含的语言学现象,提出了词汇、词汇-句法、句法、话语、推理等蕴含类型.其中,对推理现象的细粒度分类具有很好的参考意义,我们将有关数量、空间的推理知识和常识一起纳入本文的推理类型框架之下.在这些类型的基础上,结合实际标注情况,本文总结出了一套适用于中文蕴含的类型体系.

深度神经网络模型在文本蕴含关系识别上表现出很大的优势,是目前识别准确率最高的方法,具有极强的稳定性及可移植性.Bowman 等人[7]首次将循环神经网络(RNN)和 LSTM 神经网络用于文本蕴含识别,取得了不错的成绩.随后,Rocktäschel 等人[26−28]用两个 LSTM 模型分别对 *P* 和 *H* 建模,同时引入 Attention 机制,进一步提升了模型性能,较好地关注到了 *P_H* 的语义对应部分.Wang 等人[29]在 Rocktäschel 等人的工作基础上提出了 Mlstm 模型,重点关注 *P_H* 中语义匹配部分.Hickl 和 de Marneffe 等人[30]运用“对齐-过滤器”方法将命名实体和别的参数信息合并到表面文本中,他们使用人工标注数据来训练最大熵分类器,以确定表层语块的蕴含现象.Tsuchida[31]将基于词汇匹配的蕴含得分与基于深度学习的过滤机制结合起来,从词、短语和谓词获得蕴含信息.Camburu 等人[32]在 SNLI 数据中增加了对蕴含关系的人工解释,并将这些人工解释加入到模型训练过程中,改进了通用句子的表示方法,探索了模型的解释性问题.刘茂福等人[33]将文本间的蕴含关系转化为事件图之间的蕴含关系,联合多种特征,有效识别中文文本蕴含关系,强化了文本蕴含系统深层语义分析与推理能力.谭咏梅等人[34]使用基于卷积神经网络(CNN)与双向 LSTM(BiLSTM)的中文文本蕴含识别方法,避免了传统机器学习需要人工筛选大量特征以及使用多种自然语言处理工具造成的错误累计问题.金天华等人[35]在人工标注中文蕴含句对的基础上,将 Wang 的工作迁移到中文文本蕴含关系识别,进行了中文句法蕴含的语块单边界识别任务.

## 2 中文文本蕴含类型

通过分析文本蕴含数据的语言现象,可以发现隐藏在蕴含数据中的多种推理关系,在此基础上归纳出蕴含类型有助于深入研究蕴含的推理形式,更好地挖掘真实语料中的蕴含现象,并为蕴含生成提供重要的理论支持.我们参考前人的蕴含类型研究[14,24,25],结合实际语料,将中文文本蕴含分为词汇、句法结构、推理等 3 大类,其下还有 7 个小类.汇总见表 1.

**Table 1** Types of Chinese entailment
**表 1** 中文蕴含关系类型

| 类型 | | 例句 | |
|---|---|---|---|
| 词汇 | 上下位关系 | *P*:一个用推土机挖土的男人. | *H*:一个男人用重型施工设备挖土. |
| | 近义关系 | *P*:一个人坐在家门口读一本书. | *H*:一个人看书. |
| 句法结构 | 省略 | *P*:一个男人推着一辆装满购物袋的购物车. | *H*:那个男人推着购物车. |
| | 结构变化 | *P*:一只长耳狗在草地上奔跑. | *H*:一只狗跑在草地上. |
| 推理 | 位置信息 | *P*:人们在繁忙的人行道上打电话. | *H*:人们在户外打电话. |
| | 数量关系 | *P*:3 名女子和两名男子举着弦乐器站在一个楼梯下. | *H*:5 个人站在楼梯旁边. |
| | 常识 | *P*:一个年轻人坐在椅子上脸上盖着帽子. | *H*:一个年轻人坐在椅子上休息. |

### 2.1 词汇蕴含关系

词汇学是语言学的重要组成部分,词汇之间的语义关系早已成为蕴含关系的重点研究对象.上下位关系和近义关系对文本蕴含影响很大.

- 上下位关系(hyponymy)

这是语言学上一种不对称的词汇关系.上位词外延较大,内涵较小;下位词外延较小,内涵丰富.“水果-香蕉”就是一对具有上下位关系的词组.上下位关系在蕴含句对中十分常见.例如:

(1) *P*:一个用推土机挖土的男人. *H*:一个男人用重型施工设备挖土.

“重型施工设备”是“推土机”的上位词,能够概括“推土机”的特点,外延较少,因此,*P*,*H* 之间存在上下位蕴含.

- 同义关系(synonym)

反映了词语之间意义相同或相近的现象.除了词汇之间的同义现象外,表示相同语义的不同语法结构也可以归为同义关系.例如:

(2) *P*:一个人坐在家门口读一本书. *H*:一个人看书.

“读一本书”和“看书”都是动宾结构短语,两者具有同义关系.因此,*P* 和 *H* 之间存在同义蕴含.

### 2.2 句法蕴含关系

前人研究中不太重视句法蕴含,但我们在标注过程中发现,句法蕴含是经常在实际语料中出现的蕴含现象.句法蕴含可以分为省略和结构变化两个大类.

- 省略(ellipsis)

省略了定中结构短语或状中结构短语的修饰成分,但不改变中心语的蕴含情况.

(3) *P*:一个男人推着一辆装满购物袋的购物车. *H*:那个男人推着购物车.

“一辆装满购物袋的”是修饰“购物车”的定语,在 *P* 中被省略掉,因此,*P* 和 *H* 之间存在省略蕴含.

结构变化(struct-change):

句法结构的变化有多种,如“把”字句、“被字句”变换,句中主要成分被抽取出来单独成句,第一人称换第二、第三人称等.

(4) *P*:一只长耳狗在草地上奔跑. *H*:一只狗跑在草地上.

“在草地上奔跑”是状中结构短语,“跑在草地上”是动补结构短语,尽管两个短语的意义相同,但是两个的语法结构发生了变化,因此,*P* 和 *H* 之间存在结构变化引起的蕴含现象.

### 2.3 推理蕴含关系

推理蕴含是比较复杂的蕴含现象.我们在参考了 Bentivogli 等人之前的分类体系[25],将位置信息和数量关系归入推理蕴含,并将其余依靠背景知识的蕴含现象归为常识类.

- 位置信息(located-in)

表示地点的信息在 *h* 中被抽象为“外面”“户外”等反映地理位置或空间的信息,而句子中其他的成分没有矛盾关系,那么这两句话是蕴含的.空间地点信息的判断有时候需要借助背景知识的辅助.

(5) *P*:人们在繁忙的人行道上打电话. *H*:人们在户外打电话.

根据有关空间位置的背景知识,我们可以知道“在繁忙的人行道上”就是“在户外”,因此,*P* 和 *H* 之间存在位置信息引起的蕴含现象.

- 数量关系(quantity)

如果在 *P_h* 句子对里面有表示数量的词出现,且 *h* 句中的数量词反映了 *p* 中数量的和、差、积、商等数量关系,或者反映了抽象概括,例如用“们”、“一群”表示人物数量的概数.

(6) *P*:3 名女子和两名男子举着弦乐器站在一个楼梯下. *H*:5 个人站在楼梯旁边.

*P* 中的“3 名女子和两名男子”是有明确表述数量的词组,经过简单计算,可以得到一共有 5 个人的信息,与 *H* 中的“5 个人”一致,因此,*P* 和 *H* 之间存在数量关系引起的蕴含现象.

- 常识(common-sense)

常识蕴含中的推理形式复杂多样,现有研究难以清晰定义常识的概念.我们将所有需要复杂推理的蕴含句对都归为常识蕴含.

(7) *P*:一个年轻人坐在椅子上脸上盖着帽子. *H*:一个年轻人坐在椅子上休息.

根据常识信息,我们通常会认为“坐在椅子上,脸上盖着帽子”是“休息”的一种表现,所以从“脸上盖着帽子”推出“休息”,这是由常识引起的蕴含现象.

## 3 数据集构建及分析

### 3.1 数据标注

我们从中文自然语言推理语料库 CNLI 选取部分语义完整、结构清晰的蕴含语料,借助于开源语料标注平台 Brat(brat rapid annotation tool)对中文蕴含语料进行人工标注,形成一个规模为 12k 的中文蕴含语块类型数据集.数据集标注内容主要分为蕴含语块和蕴含关系类型两个部分.

- 蕴含语块指的是在蕴含句对 *P* 和 *H* 中具有蕴含关系的语块,需要分别在 *P* 和 *H* 中标注.蕴含语块可以是单个的词、多词短语、句子框架,甚至是在复句中的某一小句;
- 蕴含关系类型是两个语块之间的关系类型,由蕴含语块决定.关系类型的标注根据上文建立的中文蕴含类型体系进行标注.

标注员分别在蕴含句对 *P* 和 *H* 中标注出语块实体,然后选择这两个语块之间符合的蕴含关系,标注平台会实时保存标注信息.具体的标注示例如图 1 所示.

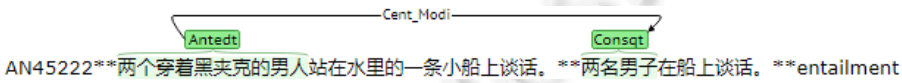



Fig.1 Annotation example by using Brat

图 1 Brat 标注示例

数据标注遵循以下原则.

(1) 语块成对出现,*P* 中的语块和 *H* 中的语块对应.例如,“两个穿着黑夹克的男人”和“两名男子”必须成对出现.省略类蕴含无法在 *H* 中找到与 *P* 对应的语块,用“*Null*”标记;

(2) 语块结构不固定,但是意义完整.语块的结构可以是词、短语、句中某些成分,但是其表达的意义必须是完整的.例如:“乘坐地铁”是动宾结构短语,“在地铁上”是介宾结构短语,两者结构并不相同,但是都具有“在地铁上”的意义,且这个意义是完整的,两者具有蕴含关系;

(3) 复合蕴含需要标出多组蕴含语块及对应的蕴含关系.复合蕴含指的是在同一组蕴含句对中存在多种蕴含现象,包括多组蕴含语块对应同一种蕴含关系以及多组蕴含语块对应多种蕴含关系.标注员应该尽可能地标注出句对中的蕴含现象.

本次数据标注员均具有专业的语言学背景,可以较好地区分蕴含句对中的语言学现象.为了控制语料标注

的一致性,我们对标注员进行了系统培训,同时开展试标注环节.在试标注期间,根据标注的实际情况调整标注原则.在正式标注期间,采用抽样复核的方式检查标注情况,最后的标注正确率达到90%以上.这个方法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.

### 3.2 数据分析

我们对所有标注数据做了统计分析,词汇、句法、推理这 3 个类型的比例关系如图 2.句法结构类蕴含关系是最多的,总共有 4 580 例,占 38.17%;推理关系其次,有 4 373 例,占比 36.43%;最少的是词汇关系类蕴含,有 3 047 例,占比 25.40%.

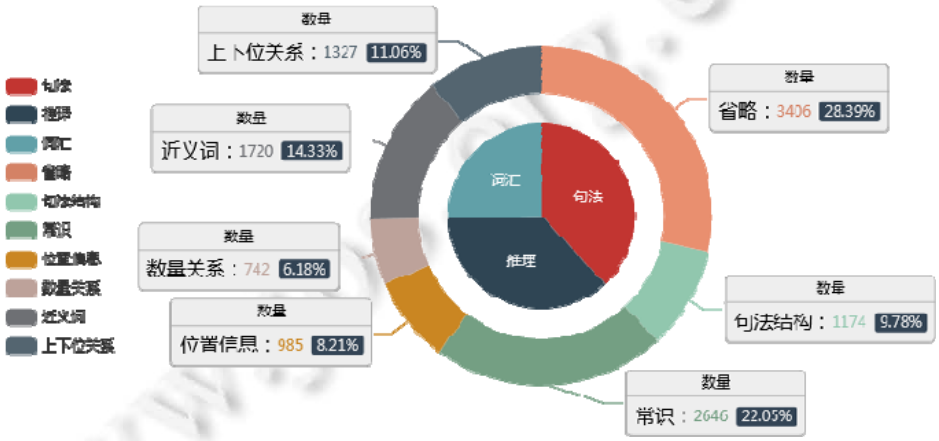



Fig.2 Chinese entailment type distribution

图 2 中文蕴含类型分布

在我们的标注数据中,句法结构类蕴含的占比最高,其次是推理类蕴含的占比,词汇类蕴含占比最少.这与金天华的工作有所出入.我们认为,数据量的扩大是主要原因.先前工作只标注了 4 000 个样例数据,而现在我们有 12 000 个样例,数据量扩大到 3 倍,意味着我们的文本更接近真实文本的情况.同时,数量结构、位置信息两类被加入推理大类中,推理的比例上升,词汇的比例就下降.

句法结构类占比最大,是可以从数据长度上证明的.统计发现:*P* 的平均长度是 18.3 个词,最大长度是 70 个词,出现频率最多的是 13 个词;*H* 的平均长度是 10.7 个词,最大长度是 62 个词,出现最多的长度是 9 个词.一般情况下,*H* 的长度比 *P* 小,这说明了以省略为代表的句法结构蕴含是非常普遍的.

对蕴含语块进行词性标注后统计分析,发现蕴含语块的结构类型十分丰富,不论是从 *P* 句得到的蕴含前件,还是从 *H* 句得到的蕴含后件,其类型数量都超出 1 000 种.蕴含语块出现频次排名前 3 的结构类型有:名词、数词+名词、动词+名词,如图 3 所示.

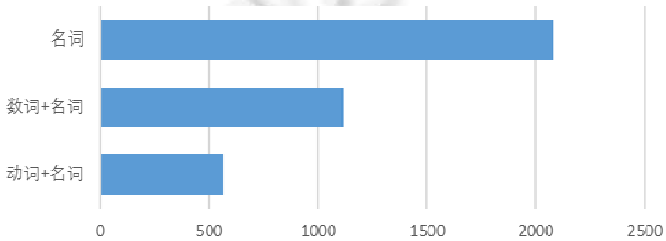



Fig.3 Top3 structures of entailment chunks

图 3 出现频次前 3 高的蕴含语块结构类型

说明名词、动词等结构类型在语块中分布广泛.名词具有极强的指称作用,在人们对事物的认知过程中起

着非常重要的作用.而动词+名、动词等结构往往代表了某一动作的发生,在人们对事件描述中有很重要的作用.所以,蕴含现象往往是在指称某一事物或者描述某一事件中发生的.

## 4 基于深度学习的中文蕴含类型识别

我们在上文的工作从词汇、句法、常识这 3 个角度归纳了 7 种中文文本蕴含类型.这些蕴含类型概括了中文蕴含数据的语言特征,将原本笼统的蕴含概念具体化,显示了蕴含现象内部的差异.中文蕴含类型识别就是以这 7 个蕴含类型作为任务标签,判断一组蕴含句对属于上文提到的哪种蕴含类型的分类任务.

虽然中文蕴含类型识别任务承袭自文本蕴含识别任务,但是两者具有明显不同.

- 文本蕴含识别判断的是两个句子是否具有蕴含关系,以“蕴含-矛盾-中立”为标签的 3 分类任务;
- 而中文蕴含类型识别任务是已知句对具有蕴含关系的情况下,进一步判断该句对属于何种蕴含类型,以中文文本蕴含类型为标签的 7 分类任务.

中文蕴含类型识别在原本的文本蕴含识别基础上更近一步,提高蕴含推理的解释性.在这一节,我们将用 ESIM[11]和 BERT[12]两个模型对中文文本蕴含类型进行分类.

### 4.1 基于ESIM的中文文本蕴含分类

Chen Q 在英文蕴含识别上提出了 ESIM 模型,采用了双向 LSTM 编码、注意力机制实现软对齐,兼顾句法信息的影响.ESIM 是文本蕴含领域的经典模型,我们对模型做了适当的调整,探索蕴含句信息和局部语块之间的语义关系,以便通过全局优化获得更好的语义表示框架.模型主要分为 4 层:输入编码层、局部推理建模层、推理组合层、池化层.为了验证蕴含语块的位置信息对 ESIM 模型识别蕴含类型的影响,我们设置了一组对照实验,将不包括蕴含语块位置信息的模型称为 ESIM1,包括蕴含语块位置信息的模型称为 ESIM2.两个模型的结构基本类似,仅在遮盖部分有所不同.以 ESIM2 为例,模型如图 4 所示.

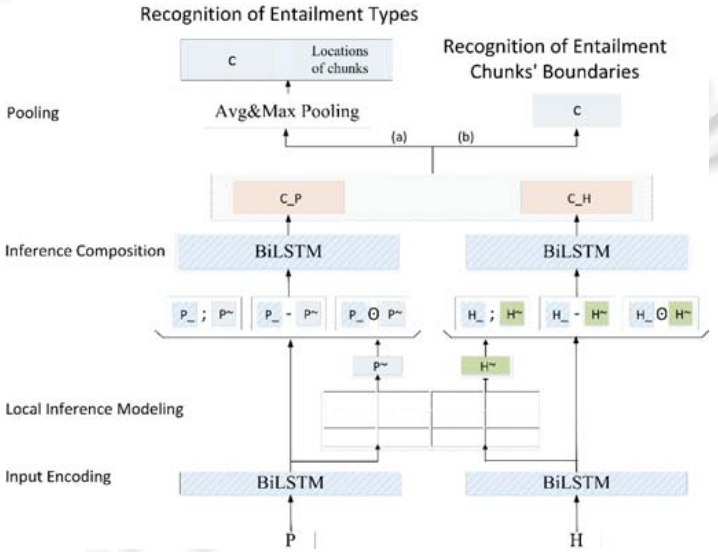



Fig.4 Structure of ESIM model

图 4 ESIM 模型结构图

在输入编码层,使用 BiLSTM 将蕴含前件 $P$ 和蕴含后件 $H$ 的语义信息表示为 $\overline{P}$ 和 $\overline{H}$ .

在局部推理层,使用注意力机制对 $\overline{P}$ 和 $\overline{H}$ 的词语及其上下文进行建模,得到局部推理关系 $\tilde{P}$ 和 $\tilde{H}$ .以 $P$ 为例,将得到的局部推理信息进行增强: $\overline{P}-\tilde{P},\overline{P}\times\tilde{P}$ ,与 $\tilde{P},\tilde{H}$ 组合得到全局与局部推理关系的组合表征.

同理,我们也对 $H$ 进行了上述操作.

在推理组合层,我们重新使用 BiLSTM 编码,把局部推理层的组合表示为 $C_P$,$C_H$.

在池化层,我们从时间维度上对 $C_P$,$C_H$进行平均池化和最大池化操作,将局部推理结果 $C_P$,$C_H$整合为全局推理关系 $C$,用 $C$ 表示模型对 $P$ 和 $H$ 推理关系的建模结果.

在 ESIM1 中,模型的输入中不包含蕴含位置 *P_pos* 和 *H_pos*,仅有蕴含句对的语义信息.我们将全局推理关系 $C$ 和未经遮盖操作的输入编码层输出 $\bar{P},\bar{H}$ 以及局部推理层输出 $C_P$,$C_H$ 组合起来,得到不包含蕴含语块位置信息的蕴含句对推理关系的最终表征 *final_C*,将其作为预测依据,对 7 个蕴含类型进行预测.公式如下:

$$final\_C = Concat\left\{C,\begin{bmatrix} C_P, C_H \\ \bar{P}, \bar{H} \end{bmatrix}\right\} \tag{1}$$

在 ESIM2 中,我们加入了遮盖层(masking),即图 4 中(a)线路中的蕴含边界位置.已知正确的蕴含位置 *P_pos*, *H_pos* 的情况下,使用遮盖操作来获取 $P$ 和 $H$ 中蕴含语块的语义信息.最后,遮盖操作作用于局部推理层的输出 $C_P$,$C_H$ 和输入编码层的输出 $\bar{P},\bar{H}$ ,得到只保留具有蕴含关系的语义片段,并将之与池化层输出 $C$ 组合起来,得到同时包含蕴含语块语义信息和蕴含句对推理关系的最终表征 *final_C′*.在预测层,*final_C′*作为预测依据,对 7 个蕴含类型做预测.公式如下:

$$final\_C' = Concat\left\{C,\begin{bmatrix} Mask(C_P, P_{\_pos}), Mask(C_H, H_{\_pos}) \\ Mask(\bar{P}, P_{\_pos}), Mask(\bar{H}, H_{\_pos}) \end{bmatrix}\right\} \tag{2}$$

### 4.2 基于BERT的文本蕴含分类

Devlin J 等人提出了利用双向 Transformer[36]的预训练语言模型 BERT,BERT 使用了 Transformer 的 Encoder 框架,有效利用了双向信息,在英文任务取得了很好的效果.同时,BERT 在大规模中文语料中进行了无监督预训练,并发布了基于中文的预训练语言模型.我们将其迁移至蕴含识别任务上,并根据中文文本蕴含数据的特点进行了适量微调.

如图 5 所示,BERT 在序列对文本分类任务中有独特的输入方式,输入的最开始添加起始标记[*CLS*],最末尾添加结束标记[*SEP*].我们依此调整了数据输入,蕴含前件和蕴含后件之间同样通过[*SEP*]来区分.*E* 表示输入的词向量,除此之外,每一个字都会添加一个表明当前句为蕴含前件或后件的向量,即 $E_P$ 或 $E_H$.而为了引入当前字的位置信息,第 $i$ 字还会添加一个位置嵌入 $E_i$.

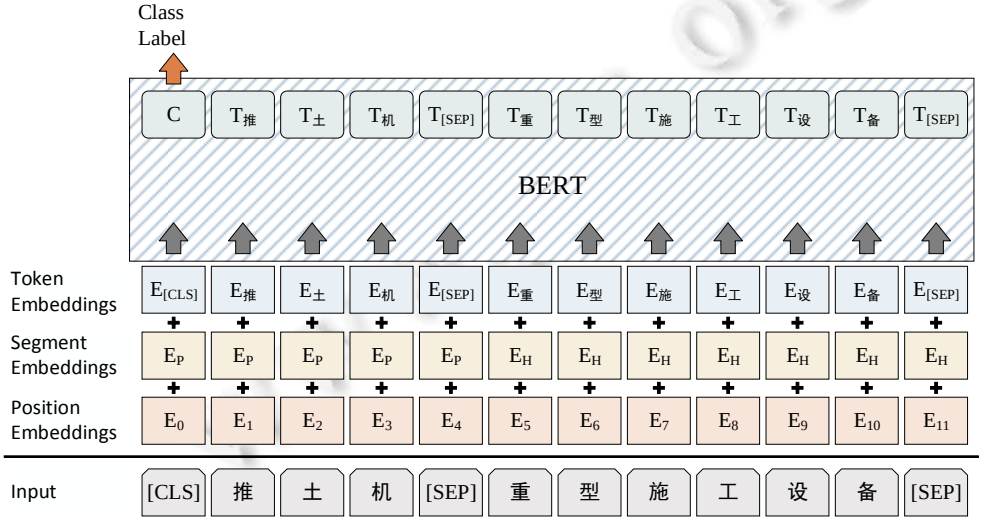



Fig.5 Input and output represantation in BERT

图 5 BERT 对分类任务的输入和输出表示

在输出部分,经过对输入层的调整后序列进入 BERT 模型进行训练,$T_i$ 表示第 $i$ 个字基于上下文的表示.而在

Transformer 最后一层的第一个词块[*CLS*]的输出能够得到整个序列的固定维度全局表征,我们将其作为分类层的输入,进行分类.

在实验过程中,我们尝试了多种数据输入方式,探究蕴含前件 *P* 和蕴含后件 *H* 以及蕴含语块 *P*′和 *H*′之间的语义关系.例如:

*P*:一个用推土机挖土的男人. *H*:一个男人用重型施工设备挖土.

在这组上下位词导致的蕴含句对中,蕴含前件语块 *P*′是“推土机”,蕴含后件语块 *H*′是“重型施工设备”.我们尝试的输入拼接方式如下:

(1) [*CLS*]+*P*+[*SEP*]+*H*+[*SEP*]:[*CLS*]+一个用推土机挖土的男人+[*SEP*]+一个男人用重型施工设备挖土+[*SEP*];

(2) [*CLS*]+*P*′+[*SEP*]+*H*′+[*SEP*]:[*CLS*]+推土机+[*SEP*]+重型施工设备+[*SEP*];

(3) [*CLS*]+*P*+*P*′+[*SEP*]+*H*+*H*′+[*SEP*]:[*CLS*]+一个用推土机挖土的男人+推土机+[*SEP*]+一个男人用重型施工设备挖土+重型施工设备+[*SEP*].

### 4.3 实验设计和结果分析

#### 4.3.1 实验设计

本文将人工标注的数据作为实验的数据集.经过数据预处理,剔除不符合要求的数据后,共得到 10 965 条蕴含句对.我们按照 8:1:1 的比例将数据随机分为训练集、验证集和测试集,其中,训练数据 8 771 条,验证数据和测试数据各有 1 097 条.

对于 ESIM 的系统,其中的超参数设置为 *hidden_units*=300,*embedding_size*=300.为了得到最好的结果,本系统采取 Early stopping 机制,轮数设置为 15.

对于 BERT 系统,我们根据不同的输入方式做了 3 组实验.

- BERT1 模型的输入为[*CLS*]+*P*+[*SEP*]+*H*+[*SEP*]的数据输入;
- BERT2 系统则采取[*CLS*]+*P*′+[*SEP*]+*H*′+[*SEP*]的数据输入格式;
- BERT3 的输入方式为[*CLS*]+*P*+*P*′+[*SEP*]+*H*+*H*′+[*SEP*].

对于这 3 个系统,超参数都设置为 *max_seq_length*=128,*train_batch_size*=16.为了以最少的运行时间得到最好的实验结果,我们设置 *learning_rate*=0.00003,微调轮数设置为 2.

为了评估文本提出的文本蕴含分类方法的有效性,实验采用准确率(accuracy)作为评价指标.准确率被定义为在指定的数据集上正确分类的样本数与总样本数之比,其值越高,效果越好.

#### 4.3.2 结果分析

由表 2 可以看出:ESIM2 在测试集上达到了 56.70%的准确率,比 ESIM1(43.57%)高 13.13%;同样地,BERT3(66.73%)比 BERT1(51.14%)在测试集的准确率高 15.59%.对比这两组实验,在同一个模型中包含蕴含语块位置信息的实验结果明显高于不包含语块位置的实验结果.这表明了蕴含语块位置信息可以提高模型的性能,有效预测蕴含关系类型.

**Table 2** Performances of various models on classification task

**表 2** 分类任务上各系统的性能

| 系统 | Train (%) | Dev (%) | Test (%) |
|---|---|---|---|
| ESIM1 | 48.85 | 47.95 | 43.57 |
| ESIM2 | 66.24 | 53.69 | 56.70 |
| BERT1 | 65.25 | 55.70 | 51.14 |
| BERT2 | 84.93 | 68.46 | **69.19** |
| BERT3 | 84.13 | 65.54 | 66.73 |

而对比 BERT2 和 BERT3,仅包含蕴含语块的 BERT2(69.19%)比同时包含蕴含句对和蕴含语块的 BERT3(66.73%)高 2.46%,差距较为明显.这证实了在同样拥有语块位置信息的基础上,非蕴含语块信息很可能会扰乱系统进行判别.

而对于都只有蕴含句对语义信息而没有蕴含语块信息的 ESIM1 和 BERT1,BERT1(51.14%)比 ESIM1(43.57%)高 7.57%.我们猜测:经过大规模语料的预训练后,BERT 能够学习到足够的基于上下文的表示,这在一些类别的分类上起到了辅助作用.

针对准确率最高的 BERT2 系统,我们观察其在测试集上的分类报告,见表 3.数量最少(仅 70 个)的数量关系类型却达到了最高的 *F*1 值 0.82,这意味着数量关系类型相对简单,模型能够充分地学习到相关知识来区分这个类别.而它的准确率仅为 0.73,远远低于召回率(0.93),这表明模型会更倾向于将数据预测为数量关系这一类别,这有可能会导致其他类型被错误预测为数量关系类型.对于同义关系这一类型,0.58 的准确率和 0.55 的召回率都不算特别高,说明蕴含中的同义关系相对较难,模型难以习得其分类依据.

**Table 3** Classification report of BERT2 on test set

**表 3** BERT2 在测试集上的分类报告

| | Precision | Recall | *F*1-score | Support |
|---|---|---|---|---|
| 省略 | **0.79** | 0.82 | 0.81 | 329 |
| 结构变化 | 0.64 | 0.55 | 0.59 | 110 |
| 常识 | 0.62 | 0.63 | 0.62 | 252 |
| 位置信息 | 0.74 | 0.70 | 0.72 | 91 |
| 数量关系 | 0.73 | **0.93** | **0.82** | 70 |
| 上下位关系 | 0.68 | 0.64 | 0.66 | 81 |
| 近义关系 | 0.58 | 0.55 | 0.56 | 164 |
| Avg/Total | 0.69 | 0.69 | 0.69 | 109 |

本实验的方法不足之处在于:一组蕴含句对可能含有多个不同类型的蕴含关系,容易在模型识别过程中产生干扰信息影响实验结果.BERT2 实验仅输入蕴含语块而放弃句对中非蕴含语块信息反而达到了实验的最优结果也证实了这一点.所以,下面我们将进行中文蕴含语块-类型识别的任务,进一步探究蕴含语块和蕴含类型之间的关系.

## 5 基于深度学习的中文蕴含语块-类型识别

上一节用深度学习模型做了中文蕴含类型识别实验,实际上是把蕴含语块作为有效输入来预测具体的蕴含类型.在实验过程中,我们发现模型对复合蕴含的识别效果较差.当输入只含有一组语块位置信息时,模型很难根据这一组语块判断出所有可能的蕴含类型标签.例如:

*P*:一个小男孩正在抚摸一只躺下的老虎. *H*:一个孩子正在爱抚一种危险的动物.

例句中有 5 组蕴含语块和 4 个蕴含关系类型:(1) “小男孩-孩子-上下位关系”;(2) “老虎-动物-上下位关系”;(3) “抚摸-爱抚-近义关系”;(4) “躺下的-*null*-省略”;(5) “老虎-危险的动物-常识推理”.

如果只给出“抚摸-爱抚”这一对语块,模型可以判断出这组语块具有近义关系,但是难以据此判断出其余 3 种的蕴含类型.

另一方面,在例句中,“上下位关系”对应了两组蕴含语块,“小男孩-孩子”和“老虎-动物”都具有上下位关系.已知例句含有上下位蕴含关系,模型可以在找出其中一组对应的语块,但是会遗漏另一组对应语块.

如果模型可以同时识别出蕴含语块和类型,这个问题就迎刃而解了.根据这个思路,我们提出两个解决方法,希望能够解决蕴含语块和蕴含类型之间的对偶性问题,进而提高蕴含类型识别准确率:首先,对复合蕴含句对进行扩充,使其在数据集中重复出现,每次出现只包含一组蕴含语块和蕴含类型的标注信息;其次,提出一个蕴含语块-类型识别的新任务,把蕴含语块识别和蕴含类型识别整合在一起,根据蕴含语块匹配蕴含类型.因为蕴含语块决定蕴含类型,所以蕴含语块-类型识别任务的逻辑应该是根据蕴含语块的语义信息去识别蕴含句对的具体类型,而不是在已知蕴含类型的情况下识别句对中的蕴含语块的边界.

蕴含类型识别本质是分类任务,而语块识别的本质是序列标注任务.识别语块就是在句子中找到语块边界,也就是找到语块在句中开始和结束的位置.我们用 IOB 标注法表示语块在句中的位置标签,用上文提到的 7 种中文蕴含关系类型作为类型标签.整合 IOB 位置标签和中文蕴含类型标签,得到 14 个形如“B-Cent_Modi”

“I-Cent_Modi”的位置-类型复合标签以及 PAD,UNK 和 *O* 等 3 个单独的位置标签.本任务总共有 17 个标签,可以在预测类别的同时预测语块边界.

### 5.1 基于ESIM的蕴含语块-类型识别方法

ESIM 模型也适用于蕴含语块-类型识别任务,在这个任务中,我们的输入与蕴含类型识别任务中 ESIM2 模型没有差别,如图 4 所示.经过了输入编码层,局部推理层和推理组合层后,我们得到了 $C_P$,$C_H$.由于语块-类型识别任务需要得到每个语块的信息,所以我们不做池化操作,如图 4(b)线路所示,直接将 $C_P$,$C_H$ 作为此任务的预测依据,整合语块的位置信息和蕴含类别,形成 17 个预测标签.

### 5.2 基于BERT的蕴含语块-类型识别方法

BERT 同样可以用于蕴含语块-类型识别任务,一般仅提供单序列标注的方法.然而,我们的蕴含语块边界识别研究是有关序列对的标注任务,所以我们将蕴含前件 *P* 和蕴含后件 *H* 用[*SEP*]进行连接,再作为输入进入模型.如图 6 所示,除了每个词块都有对应的标签之外,其他所需输入信息与蕴含类型识别任务没有差别.由于序列标注任务不能舍弃除蕴含语块外的其他部分(标签为“*O*”的部分),所以我们将整个蕴含句对的序列串作为输入.

在输出部分,对于序列中的每个词而言,都相当于一个分类任务,我们将词块 $i$ 的最终隐藏层的表示 $T_i$ 作为分类层的输入,使得对应输出的预测不受周围其他词块预测结果的影响.

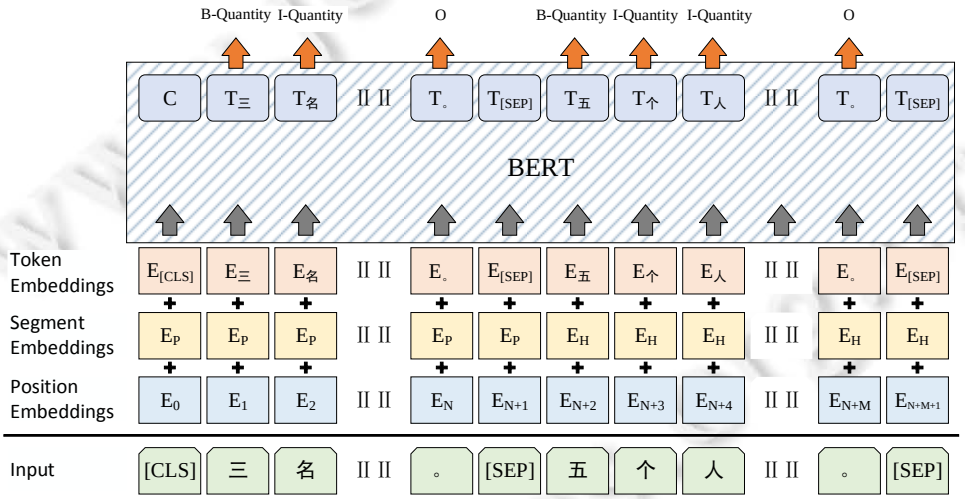



Fig.6 Input and output representation in bert for recognizing entailment chunks

图 6 BERT 对蕴含语块-类型识别任务的输入和输出表示

### 5.3 基于BERT+BiLSTM+CRF的蕴含语块-类型识别方法

双向 LSTM(bi-directional long short-term memory)[37,38]是循环神经网络的改进版,它是双向的长短时间记忆网络,非常适用于对时序文本进行建模.条件随机场(conditional random fields,简称 CRF)[39]是一个专门针对于序列标注任务的模型,在做归一化的时候考虑了全局的数据分布,做到了全局最优而不是局部最优.在过去的工作中,BiLSTM+CRF 广泛地应用于相关的命名实体识别工作中[40].命名实体识别和蕴含语块-类型识别两个任务具有相似之处:命名实体识别需要预测实体的头部和尾部,即实体的边界位置;蕴含语块-类型识别同样需要预测蕴含语块的边界,并且它们都需要预测句中每个词的标签,因此,我们也尝试将这种方法迁移到蕴含语块-类型识别任务上,将其作为实验的一部分.

BERT 在大规模语料上进行预训练,得到每个字符基于上下文的表征,这优于传统 word2vec 的表示方式,因此,我们尝试将 BERT 引入 BiLSTM+CRF 的模型中,代替底层 word2vec 部分,得到了 BERT+BiLSTM+CRF 模型.在编码部分,使用 BERT 的预训练语言模型将蕴含前件 *P* 和后件 *H* 进行编码,再将编码后拥有语义信息的嵌入

输入至 BiLSTM 层中,分别得到每个字嵌入的输出信息 $C_i$,最后将其输入至 CRF 层进行分类.在 CRF 层分类时,词块 *i* 对应的预测标签受周围其他词块预测结果的影响.

### 5.4 实验设计和结果分析

#### 5.4.1 实验设计

本组实验的训练数据有 8 771 条蕴含句对,验证数据和测试数据分别为 1 097 条.在本实验中共有 17 个“位置-类型”标签,句对中每个字都拥有对应的标签.

ESIM 系统的超参数设置为 *hidden_units*=300,*embedding_size*=300.为了充分训练并得到实验的有好结果,我们采取 Early stopping 机制,轮数设置为 15.

BERT 系统的数据输入格式为[*CLS*]+*P*+[*SEP*]+*H*+[*SEP*],其中,超参数设置为 *max_seq_length*=128,*train_batch_size*=16.由于 BERT 体量较大,所以为了以最少的训练时间得到最优效果,我们设置 *learning_rate*=0.00002,微调轮数为 3.

BiLSTM+CRF 系统的超参数设置为训练轮数 40,*hidden_dim*=300,*embedding_dim*=300,采用 Adam 优化器,学习率为 0.001.

BERT+BiLSTM+CRF 系统的超参数设置为 *max_seq_length*=128,*hidden_units*=1000.经过多次实验调整,这些参数能够得到本实验的最好效果.

为了评估本文提出的边界识别方法的有效性,实验采用 *F*1 值(*F*1-measure)作为评价指标.*F*1 值定义为在指定的数据集上准确率(*precision*)和召回率(*recall*)的综合平均,同时反映了两者的数值高低,*F*1 值越高,效果越好.

#### 5.4.2 结果分析

以上 4 个实验的输入均保持一致,仅在系统的模型结构部分不同.从表 4 看出,BERT(61.58%)在测试集上的表现优于 ESIM(52.80%).这证实了 BERT 模型在蕴含语块-类型识别任务上的有效性,其使用更加灵活,更能学习上下文的关系,对前后语块联系密切的边界标注任务十分友好.BERT+BiLSTM+CRF(62.06%)比 BiLSTM+CRF(50.92%)高 11.14%,提升显著.而 BERT+BiLSTM+CRF(62.06%)只比 BERT(61.58%)高 0.48%,提升效果并不明显.在有 BERT 的情况下,将句对作为整体输入,加上 Segment Label 以区分前后句,使得 BERT+BiLSTM+CRF 在句对标注任务上有更好的结果,然而 BiLSTM+CRF 在其中作用有限.虽然 BiLSTM+CRF 广泛应用在与命名实体识别相关的任务中,在单句标注任务上表现较好,但是本任务更关注句对之间的联合信息,BiLSTM+CRF 的实验效果在句对标注任务上不是很理想.

但是综合而言,这几个模型的表现并没有达到 65%,针对本任务的各个系统都有待提升.究其原因:一方面是因为标签过多,我们在这个任务上设置了 17 个标签,会降低模型学习效果;另一方面是数据分布不平均,部分类型的数据过于稀疏,模型难以学习到相关语义信息.

**Table 4** Performances of various models on recognizing entailment chunks

**表 4** 中文蕴含语块-类型识别任务上各系统的性能

| 系统 | Train (%) | Dev (%) | Test (%) |
|---|---|---|---|
| ESIM | 52.25 | 49.32 | 52.80 |
| BERT | 73.56 | 59.78 | 61.58 |
| BiLSTM+CRF | 87.90 | 49.71 | 50.92 |
| BERT+BiLSTM+CRF | 68.69 | 59.98 | **62.06** |

## 6 结 语

本文通过人工标注的方式获得了 12 000 条中文蕴含语块类型标注数据,通过分析蕴含语块的词汇、句法和语义特征,归纳出了 3 个大类(词汇蕴含、句法结构蕴含、推理蕴含)和 7 个小类(上下位关系、近义关系、省略、结构变化、位置信息、数量关系、常识)的中文蕴含关系类型.本文工作拓展了中文文本蕴含的研究对象,有利于挖掘蕴含中的推理知识,探索语义理解中的推理机制.

在标注数据的基础上,我们利用深度学习模型实现了中文蕴含类型识别任务,创新地提出了同时预测蕴含

语块和类型的中文蕴含语块-类型识别任务,探索了中文文本蕴含识别在新的任务形式上的可能性.实验结果表明:蕴含识别相关任务可以在基于大规模预训练数据的 BERT 模型上共享语义知识,有效预测句对中的符合蕴含现象的语块及其位置信息.该实验为小规模数据集上的中文文本蕴含识别任务提供了可靠的基线.

本文工作仍有待改进的地方.中文蕴含语块-类型识别有 17 个预测标签,每个标签需要同时预测语块位置信息和关系类型.标签数量多,内容复杂,预测结果比单纯预测类型的中文蕴含类型识别任务要低.在分析了蕴含类型识别实验结果后,我们发现模型难以学习蕴含数据中的近义关系,这启发我们在未来可以将外部知识加入模型中,提高预测准确率.词汇、句法结构等底层特征作为重要的模型输入,将会对模型性能产生重要影响,这些特征也将成为我们未来研究的重点关注对象.另一方面,中文蕴含类型和英文蕴含类型有部分重合,我们希望标注部分英文蕴含数据,做一组中英文蕴含识别的对比实验,比较深度学习模型在本文 3 个任务上的结果是否有差别.

**References**:


[1] Guo MS, Zhang Y, Liu T. Research advances and prospect of recognizing textual entailment and knowledge acquisition. Chinese Journal of Computers, 2017,40(4):889−910 (in Chinese with English abstract). http://cjc.ict.ac.cn/online/onlinepaper/gms-2017 45180721.pdf [doi: 10.11897/SP.J.1016.2017.00889]

[2] Li JM. An overview of the research on prefabricated chunks home and abroad. Shandong Foreign Language Teaching Journal, 2011, 32(5):17−23 (in Chinese with English abstract).

[3] Skehan P. A Cognitive Approach to Language Learning. Oxford: Oxford University Press, 1998.

[4] Wray A. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press, 2005.

[5] Russell B. Introduction to Mathematical Philosophy. North Chelmsford: Courier Corporation, 1993.

[6] Flew A. A Dictionary of Philosophy. London: Pan Book Ltd., 1979.

[7] Bowman SR, Angeli G, Potts C, *et al*. A large annotated corpus for learning natural language inference. arXiv preprint arXiv: 1508. 05326, 2015.

[8] Rocktäschel T, Grefenstette E, Hermann KM, *et al*. Reasoning about entailment with neural attention. arXiv preprint arXiv: 1509. 06664, 2015.

[9] Liu Y, Sun C, Lin L, *et al*. Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv:1605.09090, 2016.

[10] Sammons M, Vydiswaran VGV, Vieira T, *et al*. Relation alignment for textual entailment recognition. In: Proc. of the Text Analysis Conf. (TAC). 2009.

[11] Chen Q, Zhu X, Ling Z, *et al*. Enhanced lstm for natural language inference. arXiv preprint arXiv:1609.06038, 2016.

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

[13] Dagan I, Glickman O. Probabilistic textual entailment: Generic applied modeling of language variability. In: Proc. of the PASCAL Workshop on Learning Methods for Text Understanding and Mining. 2004. 26–29.

[14] Dagan I, Glickman O, Magnini B. The PASCAL recognising textual entailment challenge. In: Quiñonero-Candela, Joaquin, *et al*., eds. Proc. of the Int'l Conf. on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. Springer-Verlag, 2005. 177–190.

[15] Bar-Haim R, Dagan I, Dolan B, *et al*. The 2nd Pascal recognising textual entailment challenge. In: Proc. of the 2nd PASCAL Challenges Workshop on Recognising Textual Entailment. 2006,6(1):6.4.

[16] Giampiccolo D, Magnini B, Dagan I, *et al*. The 3rd Pascal recognizing textual entailment challenge. In: Proc. of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics, 2007. 1–9.

[17] Shima H, Kanayama H, Lee CW, *et al*. Overview of NTCIR-9 RITE: Recognizing inference in text. In: Proc. of the 9th NII Test Collection for Information Retrieval Workshop. 2011. 291–301.

[18] Watanabe Y, Miyao Y, Mizuno J, *et al*. Overview of the recognizing inference in text (RITE-2) at NTCIR-10. In: Proc. of the 10th NII Test Collection for Information Retrieval Workshop. 2013. 385–404.

[19] Matsuyoshi S, Miyao Y, Shibata T, *et al*. Overview of the NTCIR-11 recognizing inference in text and validation (RITE-VAL) task. In: Proc. of the 11th NII Test Collection for Information Retrieval Workshop. 2014. 223−232.

[20] Williams A, Nangia N, Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426, 2017.

[21] Demszky D, Guu K, Liang P. Transforming question answering datasets into natural language inference datasets. arXiv preprint arXiv:1809.02922, 2018.

[22] https://github.com/blcunlp/CNLI

[23] https://github.com/liuhuanyong/ChineseTextualInference

[24] Ren H. Research on annotation of linguistic phenomena for Chinese text reasoning. Journal of Henan Institute of Science and Technology, 2017,37(7):75−78 (in Chinese with English abstract).

[25] Bentivogli L, Cabrio E, Dagan I, *et al*. Building textual entailment specialized data sets: A methodology for isolating linguistic phenomena relevant to inference. In: Proc. of the LREC 2010. 2010.

[26] De Marneffe MC, Rafferty AN, Manning CD. Finding contradictions in text. In: Proc. of the HLT, Association for Computational Linguistics (ACL 2008). Columbus, 2008. 1039−1047.

[27] Iftene A. UAIC participation at RTE4. In: Proc. of the 1st Text Analysis Conf. (TAC). 2008. 35, 104, 105.

[28] MacCartney B, Manning CD. Natural logic and natural language inference. In: Proc. of the Computing Meaning. Dordrecht: Springer-Verlag, 2014. 129−147.

[29] Wang S, Jiang J. Learning natural language inference with LSTM. arXiv preprint arXiv:1512.08849, 2015.

[30] Sammons M, Vydiswaran VGV, Vieira T, *et al*. Relation alignment for textual entailment recognition. In: Proc. of the Text Analysis Conf. (TAC). 2009.

[31] Tsuchida M, Ishikawa K. IKOMA at TAC2011: A method for recognizing textual entailment using lexical-level and sentence structure-level features. In: Proc. of the Text Analysis Conf. (TAC). 2011.

[32] Blunsom P, Camburu OM, Lukasiewicz T, *et al*. e-SNLI: Natural language inference with natural language explanations. arXiv preprint arXiv: 1812.01193, 2018.

[33] Liu MF, Li Y, Ji DH. Event semantic feature based Chinese textual entailment recognition. Journal of Chinese Information Processing, 2013,27(5):129−136 (in Chinese with English abstract).

[34] Tan YM, Liu SW, Lv XQ. CNN and BiLSTM based Chinese textual entailment recognition. Journal of Chinese Information Processing, 2018,32(7):11−19 (in Chinese with English abstract).

[35] Jin TH, Jiang S, Yu D, *et al*. Chinese chunked-based heterogeneous entailment parser and boundary identification. Journal of Chinese Information Processing, 2019,33(2):17−25 (in Chinese with English abstract).

[36] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: Proc. of the Neural Information Processing Systems (NIPS). 2017. 5998−6008.

[37] Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans. on Signal Processing, 1997,45(11):2673−2681.

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

[39] Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the ICML. 2001. 282−289.

[40] Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360, 2016.

**附中文参考文献:**

[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.计算机学报,2017,40(4):889−910. http://cjc.ict.ac.cn/online/onlinepaper/gms-201745180721.pdf [doi: 10.11897/SP.J.1016.2017.00889]

[2] 李继民.国内外语块研究述评.山东外语教学,2011,32(5):17−23.

[24] 任函.面向汉语文本推理的语言现象标注规范研究.河南科技学院学报,2017,37(7):75−78.

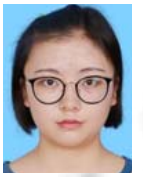
[33] 刘茂福,李妍,姬东鸿.基于事件语义特征的中文文本蕴含识别.中文信息学报,2013,27(5):129−136.

[34] 谭咏梅,刘姝雯,吕学强.基于 CNN 与双向 LSTM 的中文文本蕴含识别方法.中文信息学报,2018,32(7):11−19.
[35] 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.中文信息学报,2019,33(2):17−25.

**于东**(1982－),男,博士,副教授,主要研究领域为自然语言处理,人工智能.

**金天华**(1995－),女,硕士,主要研究领域为自然语言处理,人工智能.

**谢婉莹**(1997－),女,学士,主要研究领域为自然语言处理,人工智能.



**张艺**(1997－),女,学士,主要研究领域为自然语言处理,人工智能.

**荀恩东**(1967－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为自然语言处理,人工智能.