中文关系抽取的句级语言学特征探究

邢百西，赵继舜, 刘鹏远*
北京语言大学　信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
{xingbaixi,zhaojishun1997}@gmail.com,liupengyuan@blcu.edu.cn

摘要

神经网络模型近些年在关系抽取任务上已经展示出了很好的效果，然而我们对于特征提取的过程所知甚少，而这也进一步限制了深度神经网络模型在关系抽取任务上的进一步发展。当前已有研究工作对英文关系抽取的语言学特征进行探究，并且得到了一些规律。然而由于中文与西方语言之间明显的差异性，其所探究到的规律与解释性不适用于中文关系抽取。本文首次对中文关系抽取神经网络进行探究，采用了四个角度共13种探究任务，其中包含中文特有的分词探究任务。在两个关系抽取数据集上进行了实验，探究了中文关系抽取模型进行特征提取的规律。

关键词： 关系抽取 ； 探究任务

A Probe into the Sentence-level Linguistic Features of Chinese Relation Extraction

Baixi Xing, Jishun Zhao, Pengyuan Liu*
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center Print Media Language Branch
15 Xueyuan Road, Haidian District, Beijing, 100083, China
{xingbaixi,zhaojishun1997}@gmail.com,liupengyuan@blcu.edu.cn

Abstract

The neural network model has shown good results in relation extraction tasks in recent years. However, we know very little about the process of feature capture, which limits the further development of deep neural network models in relation extraction tasks. Current research work has explored the linguistic features of English relation extraction, and some rules have been obtained. However, due to the obvious differences between Chinese and Western languages, the laws and explanatory nature explored are not suitable for Chinese relationship extraction. This paper explores the Chinese relation extraction neural network for the first time, using a total of 13 types of exploration tasks from four perspectives, including Chinese-specific word segmentation exploration tasks. Experiments were carried out on two relation extraction data sets to explore the law of feature extraction in Chinese relation extraction model.

Keywords: Relation extraction , Probing task

* 通讯作者　Corresponding Author

第二十届中国计算语言学大会论文集，第643页-第654页，呼和浩特，中国，2021年8月13日至15日。

1 引言

信息抽取是自然语言处理的一项重要任务，它的基本目的之一是从原始的非结构化文本中提取有意义的结构化信息，以用于智能问答、检索等自然语言处理应用。信息抽取本身是一项庞大的任务，包括命名实体识别、关系抽取、事件抽取等子任务。关系抽取任务是信息抽取中的一项重要子任务，旨在预测句中给定实体对的语义关系，如图1所示

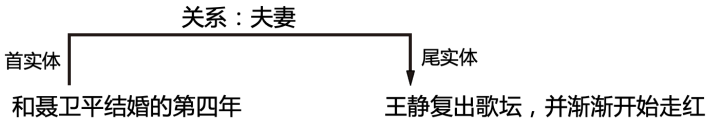

Figure 1: 一个关系实例，句中包含‘夫妻’关系在首实体‘聂卫平’和尾实体‘王静’

神经网络模型在关系抽取任务上已经展示出了很好的效果。研究者们使用神经网络去自动学习语言学特征 (Zeng et al., 2014; Lin et al., 2016; Zhou et al., 2016; Jiang et al., 2016)并且取得了很好的效果。大多数的神经网络模型使用一个编码器结构，去学习输入句子的向量表示，随后被传入一个分类层去预测关系标签。

这些模型的优秀表现说明了模型在学习向量表示的过程中，提取了一些语言和语义方面的特征，神经网络模型对关系进行分类时会依据这些特征进行分类 (Conneau et al., 2018)。然而，我们不清楚模型所学习到的特征。对神经网络黑箱内部处理特征过程的不可视性，限制了深度学习的进一步发展，越来越多的工作开始尝试对神经网络进行解释 (Karpathy et al., 2015; Kim et al., 2020)。

探究任务 (Adi et al., 2016; Alt et al., 2020)，或称诊断分类器，是一种分析模型的隐藏信息的方法。对于每一个探究任务，都对应着一个被训练的分类器，根据此分类器的表现来衡量该任务对应的特征被编码的过程。探究任务的选取通常是根据与原任务息息相关的下游任务。例如：一个用于训练关系抽取任务的编码器，直觉上应该同样编码了两个实体的关系类型，如果分类器正确地预测了实体类型，那么它表明了编码器仍然保留着实体类型特征，而这也直接影响了关系预测。这个方法的简单性使我们很容易地通过对下游任务直接进行探究，去指出模型所依赖的特征。

(Alt et al., 2020)提出了针对于英文关系抽取的探究任务，他们提出了多种探究任务去理解关系抽取模型所学习到的句子输入中对预测有帮助的特征，以便于更好的解释模型。虽然它们探究了英文关系抽取中，模型提取特征的规律性，但是由于中文与西方语言的差异性，这种规律能否迁移到中文关系抽取中是一个待探究的问题。

关系抽取领域中，中文与英文的不同主要体现在三个方面：首先，在中文文本中，字符与字符之间没有边界 (Li et al., 2008)。 (Li et al., 2019b)认为虽然对于神经关系抽取来说，没有必要去执行特征工程。但是大家仍然忽略了中文输入的不同粒度会对模型有重要影响。如：达尔文研究所有杜鹃。达尔文与杜鹃的关系会受到‘研究’和‘所’之间词的关系的影响，因此它提出了多粒度模型去整合字符级别和词级别信息。 (Li et al., 2019a)对比了基于字向量的神经网络和基于词向量的神经网络在语言模型，机器翻译等领域的表现。实验结果证明了由于数据稀疏性和未登录词，基于字向量的神经网络表现要更好。因此他们认为应该重新考虑基于深度学习的中文自然语言处理中的分词工作的必要性。虽然有这些工作去解决中文词语的边界问题，但是由于我们不知道神经网络内部的特征提取过程，因此我们无法得出确切的结论关于分词信息在神经网络中的贡献的规律。对于此中文特有的语义信息，仍然没有相关工作去进行探究分析。

其次，中文与其他西方语言在句法结构上非常不同。英语用词形变化表示词的句法功能；而汉语里的词没有表示句法功能的形态变化，词在句子里充当什么成分，主要靠语序来表示。例如，图2显示了同一句子的英文和中文的语法树如何不同。在中文判断关系更多依靠多样的句式所表达出的语义信息。这导致了同样的神经网络模型，在处理中文文本与英文文本时，所关注的句法特征与语义特征不同。

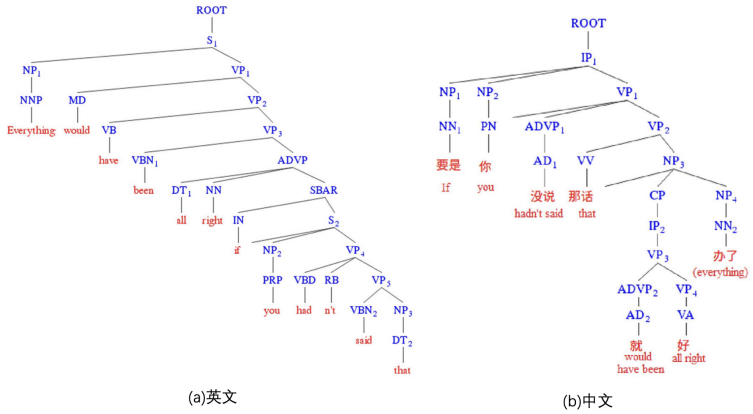

Figure 2: “要是你没说那话就好办了”在两种语言中的句法结构

最后，从语言类型上来看，汉语属于孤立语，大多数词由词根直接构成，缺乏词缀和词尾，并缺乏形态变化，句子成分之间的关系没有显性的语法标记；(黄伯荣and 廖序东, 2007)英语属于屈折语，其词由词根和附加成分组成，词根表示词汇意义，附加成分表示语法意义，句子成分之间的关系可以通过语法标记判断 (崔希亮, 2008)。早期的研究如 (Kambhatla, 2004)通过将词性信息加入到特征中，以便获取丰富的语义信息，然而这在中文中却无类似的工作。

基于上述原因，我们认为中文关系抽取有其独立的规律性，因此本文提出了中文关系抽取的探究任务，并进行了相关实验探究规律。与 (Alt et al., 2020)相似，本文关注一系列的关系抽取相关的语言学特征，从简单的表层信息(例如句子长度)，到句法信息(例如句法结构)和语义信息(例如实体类型)，还探究了特定于中文的分词语义信息。

本文贡献如下

- 首次进行了中文关系抽取探究任务，通过分析对深度模型在处理中文关系抽取过程中所做的特征提取工作有了全面的理解。
- 除了12种通用的探究任务，本文还提出了针对于中文的分词探究任务
- 本文比较了中文关系抽取中探究任务实验结果与英文的不同点，帮助理解了中文信息中更加重要的特征，同时也对关系抽取以外的领域提供了参考。

2 探究任务

本文参考了 (Alt et al., 2020)所提出的任务与设置，但是更加关注于中文关系抽取。因此对一些任务做了调整，如词性任务，以便于更加适合于探究中文。同时本文还提出了特定于中文的分词任务，以便于理解中文关系抽取过程中模型所关注的语义信息。探究任务的分类问题需要单句的句嵌入作为输入，而不是例如句嵌入+词嵌入，或多句表示等。这个过程与标准的关系抽取任务相符。

(Conneau et al., 2018)提出了针对于机器翻译的探究任务，这篇文章把机器翻译的下游任务分为三类，分别是表征信息，句法信息，和语义信息。分别对机器翻译密切相关的子任务如：主句-从句，合理词序等任务进行探究。但是不同的nlp任务所关注的信息各不相同，要为不同的nlp任务设计不同的探究任务。因此 (Alt et al., 2020)提出了针对于关系抽取的探究任务，设计了更加关注于关系抽取任务的下游探究任务。由于上文中提到的两种语言的种种不同，深度神经网络所关注的信息与捕获的特征也不相同，本文针对中文特性对部分神经网络提取的特征进行了探究。

表层信息 这些任务探究句子的编码器是否可以捕获到它们所编码的句子的表层特性，(Adi et al., 2016)提出了**句长任务**，预测一个句子中token的数量。本文把句长按照分布尽可能地均匀划分为若干类，同时把预测句长任务当作是一个多分类任务。由于关系抽取任务存在两个实体，直观上想要预测被预测关系的实体对之间词或字的数量。因此设计了**实体间长度任务**，预测实体间token的数量。与句长任务相类似，同样把长度均匀分为若干类当作分类任

务。 (Kambhatla, 2004; Surdeanu et al., 2011)在早期的特征工程中提到了实体间是否存在其他实体的特征，受此启发设计了**实体间是否存在其他实体任务**，这是一个二分类任务。此任务可以探究编码器对关系实体之间的上下文信息的编码程度。

句法信息 关系抽取任务的复杂性使得模型的表现非常依赖句法信息。许多关系抽取方法利用了依存树 (Bunescu and Mooney, 2005; Mintz et al., 2009)或词性信息 (Zhou et al., 2005)。本文因此包括了 (Conneau et al., 2018)在机器翻译探究中提出的**句法树深任务**，这个任务测试一个编码器是否可以储存句法结构。本文把句子中从根结点到任意叶子结点中最长的路径作为整句话的树深。将树深按照长度尽可能均匀分为若干组组。**最短依存路径任务**，探究模型是否学习到两实体间的最短依存路径可以很好的理解模型对于句法信息的处理。和句法树深任务相同，本文同样按照长度对语料进行标注。 (Chan and Roth, 2011)提出了所有的关系类型都可以被解释为几种受限的句法语义结构，通过对结构进行识别就可以很好的预测关系类型。同时在语义层面，许多关系类型是双向的，比如血缘关系中的顺序不同代表了完全不同的两种关系，实体出现的顺序会对实验结果产生重要的影响。因此本文包括了**实体顺序任务**，在此任务中，测试首实体是否先于尾实体出现。最后本文探究了**词性任务**，在关系抽取中，紧挨着实体左右侧的词语非常重要，这个任务可以测试编码器对实体的上下文环境是否敏感。我们共探究了四种词性任务，分别为首实体左侧词性，首实体右侧词性，尾实体左侧词性，尾实体右侧词性。

实体信息 实体是关系抽取任务的独有的特点， (Zhou et al., 2005)提出了将实体类型作为特征进行机器学习。受此启发本文提出了两个探究任务，**首实体类型**，和**尾实体类型**。实体类型信息对于关系抽取任务非常重要，举个例子，如果模型可以确定实体类别，那就可以缩小待选择的关系类别的数量，让关系抽取任务变得更加简单。此任务当作多元分类任务，分类数量为语料中出现的实体类别。

复杂语义信息 中文与英文有着很大的不同，其中非常重要的不同就是中文的词与词之间没有边界。在如今的预训练语言模型中，大量的工作都是基于词向量。 (Li et al., 2019a)认为应该重新思考分词工作的必要性。一是因为数据的稀疏性，他们统计了，发现48.7%的词语只出现了一次，二是因为即使最先进的模型的分词结果，仍然远远达不到完美，导致了误差的传递。第三，分词的意义在于，针对于标记的CWS数据集中存在多少附加语义信息。然而这个问题无法得到准确的回答。

上述工作只说明了分词工作目前仍然存在这些难以解决的问题，但是我们想要探究神经网络是否关注分词特征，因此本文设计了分词探究任务。分词任务作为一个复杂的自然语言处理任务，不同于上述简单分类任务，用现有的nlp工具进行标注难以避免有误差的存在，但此次实验主要探究规律性，所以暂时忽略误差的影响。

3 实验

3.1 训练过程

关系抽取模型通常遵循一个seq2vec方法，位于下层的编码器将输入编码为一个固定维度的向量表示，之后应用一个全联接的关系分类层。

如图三所示，探究任务的模型主要分为三个部分。首先是底层的编码器，此编码器被所要探究的领域与具体的探究任务共用。然后是两个分类器，一个用于关系抽取任务的分类，一个用于探究任务的分类。在图中，黑色的箭头表示数据的流向，而神经网络的颜色表示是否参与训练。

训练过程分为两个阶段，第一阶段为标准的关系抽取任务训练过程，此阶段我们希望得到一个被良好训练的模型，并且产生了良好的实验结果以证明模型的有效性。以此模型为基础进行下一步的探究任务。在这一个阶段中，编码器和关系抽取的分类器都参与训练。训练结束后，编码器与关系分类器的权重被固定。

第二个阶段为探究任务的训练过程，我们以在第一个阶段被训练好的编码器作为下层结构，之后接入探究任务的分类器。在此阶段只有探究任务的分类器参与训练，而下层的编码器不参与训练。这样探究任务的分类器的输入仍然是关系抽取任务所提取的特征。

3.2 编码结构

本篇论文我们共探究了两种神经网络中常用的编码结构CNN与LSTM。对于CNN网络，我们参照了 (Zeng et al., 2014)的网络结构，输入向量经过若干个不同尺寸的二维卷

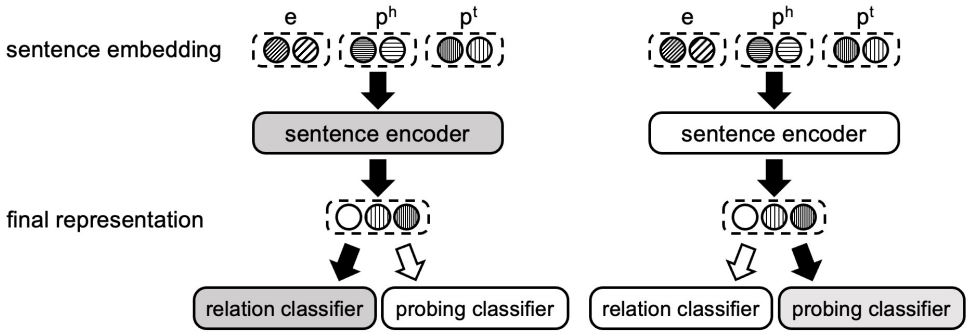

Figure 3: 训练过程，左侧为第一阶段，右侧为第二阶段，黑色的箭头表示了数据流的方向，黑色的网络结构表示参加了训练。e为句子的编码，p^h和p^t为首实体和尾实体的位置编码

积核，再经过一个最大池化层得到300维度的特征表示M_f。对于LSMT网络，我们采用双向LSTM神经网络，双向LSTM网络产生一个隐藏状态序列$\{h_t\}_{t=1,...,T}$，h_t是前向LSTM的h_t^f和反向LSTM的h_t^b的连结。采用双向LSTM可以同时根据词语的上下文环境进行编码。在探究实体左右位置相关的任务时可以保证两边的平衡。lstm的隐藏层维度设置为500。

3.3 语言知识

在关系抽取中添加额外的语法和语义特征作为输入可以提高模型表现(Zeng et al., 2014; Zhang et al., 2017)，然而最近几年，预训练语言模型的兴起极大地改变了这个过程。预训练语言模型在很多nlp任务上都取得了令人瞩目的表现，在本文中使用BERT作为神经网络的下层输入。

实体掩码被证实在关系抽取任务中会提高模型的表现，通过用实体类型和语法角色去代替实体提及。它限制了模型可以接受的实体信息，因此迫使模型更加关注语义信息。在本次实验中，由于中文的语法角色难以判断，因此本文用token中的'unused'去代替实体。

3.4 数据集与数据处理

本文采用了两个数据集，(Xu et al., 2017)提出的中文散文数据集和一个基于网络的人物关系抽取数据集。其中散文数据集语言较为规范，噪声较小，句子的结构工整；而人物关系数据集是经由网络抽取得到的，因此噪声较大，且句子之间的跨度较大。参考了两个数据集的不同，实验结果的共同性可以表明探究的规律对中文关系抽取的普遍适用性。散文中文数据集包含837篇中国文学文章中的9种关系，其中695篇用于训练，84篇用于测试，其余58篇用于验证。我们把文档级数据拆分为了句子级数据，经过去重处理共6881条训练数据。散文数据集噪声较小，人物关系数据集噪声较大。人物关系数据集包含10000条训练数据与1000条测试数据，共有11种关系类型和一种unknown关系，总共12种分类类型。为了得到所需要的句子的种种特征，需要先利用已有的工具对句子进行处理。本文使用Stanford corenlp包对语料进行标注，包括命名实体识别，词性标注等。

以数据集中的一句话为例:**音乐奇才三兄弟洪氏三兄弟在台湾歌坛都是名人，洪敬尧、洪荣宏、洪荣良都承袭了父亲洪一峰的音乐细胞。**其中，洪敬尧是标出的首实体，而洪一峰是标出的尾实体，以这句话为例解释本文的探究任务与数据处理。

对句长与句间任务，我们统计了语料库中句子长度的分布情况，尽可能的把它们均匀分为四类，对实体间长度也是进行了同样的操作。以句长为例，我们按照句长的分布，把长度分为0-35，35-45，45-55和大于55四类，对句长的探究任务做一个四分类任务。对于散文数据集，则按照0-30，30-45，45-70，大于70进行分类。

对实体间是否存在其他实体任务，因为是人物关系抽取数据集，因此我们只统计两实体间是否出现其他人物实体，是一个二分类任务。在这句话中，洪敬尧与洪一峰之间出现了被标注为PERSON的洪荣宏与洪荣良，因为我们把这句话关于存在实体任务标注为True。有序任务则是根据两实体在句中出现的先后顺序，进行分类。

对词性任务，中文的词性与英文的词性有很大不同，比如中文中出现了，DEC：从句"的"，DEG：修饰"的"，DER：得，DEV：地。以及标点在中文中也具备丰富的

语义关系。但是在本次实验中，我们没有对标点进行深入研究，只是把标点都归为punction。以这句话为例，首实体洪敬尧左侧是逗号，右侧是顿号，因此他的poslh和posrh是“PU”，尾实体洪一峰的左侧是父亲：名词，右侧是表示表示修饰的“的”，因此他的poslt是“NN”，posrt是“DEG”。

对于句法解析与依存树任务，我们用Stanford corenlp工具 (Manning et al., 2014)与spacy工具 (Neumann et al., 2019)对句子进行句法解析和依存分析。对于句法解析任务，我们把根结点到任意叶子结点的最长路径当作树深探究任务的标签。按照路径长度把它们尽量均匀地分为五类，0-5，6，7，8-9，10及以上。对于最短依存路径任务，我们用spacy工具获得两实体的最短依存路径。按照长度进行分类。

对于分词任务，我们应用jieba分词对噪声较小的散文语料进行标注，以此作为标准结果。

4 实验结果与分析

4.1 实验结果

Model	句长	句间	存在实体	句法树	依存路径	F1
Bert	0.427	0.365	0.785	0.316	0.513	0.753
Bert + mask	0.512	0.448	0.772	0.304	0.549	0.670
Bert + CNN	0.464	0.414	0.769	0.290	0.483	0.754
Bert + CNN + mask	**0.531**	0.389	0.772	0.298	0.489	**0.769**
Bert + LSTM	0.428	0.514	0.816	**0.402**	0.577	0.757
Bert + LSTM + mask	0.499	**0.518**	**0.827**	0.394	**0.590**	0.764

Table 1: 人物关系数据集，句子相关子任务的结果。在人物关系抽取任务中，没有包括实体类型任务。

Model	有序	posLH	posRH	posLT	posRT	F1
Bert	0.848	0.403	0.314	0.403	0.335	0.753
Bert + mask	**0.961**	0.473	0.355	0.467	0.374	0.670
Bert + CNN	0.573	0.282	0.198	0.264	0.190	0.754
Bert + CNN + mask	0.590	0.287	0.200	0.258	0.208	**0.769**
Bert + LSTM	0.785	**0.479**	0.386	**0.469**	**0.406**	0.757
Bert + LSTM + mask	**0.795**	0.455	**0.402**	0.457	0.400	0.764

Table 2: 人物关系数据集，实体相关子任务的结果。在人物关系抽取任务中，没有包括实体类型任务。

表1和表2展示了在人物关系抽取数据集上的探究任务的结果。

Model	首实体类型	尾实体类型	有序	存在实体	句法树	依存路径	F1
Bert	0.899	**0.881**	**0.885**	0.924	0.271	0.606	0.856
Bert + mask	0.889	0.873	0.868	0.910	0.256	0.585	0.864
Bert + CNN	0.770	0.667	0.673	0.816	0.265	0.554	0.854
Bert + CNN + mask	0.772	0.679	0.690	0.833	0.259	0.554	0.856
Bert + LSTM	0.892	0.876	0.869	0.922	**0.342**	0.619	0.861
Bert + LSTM + mask	**0.900**	0.869	0.876	**0.931**	0.338	**0.634**	**0.865**

Table 3: 散文数据集，探究任务的结果

表3和表4列出了散文数据集的进一步探究实验

对于所有分类探究任务的实验结果，本文都综合准确度与召回率计算F1值统计结果。表的最后一列为对应编码结构进行关系抽取任务的表现。

Model	句长	句间	posLH	posRH	posLT	posRT	F1
Bert	0.423	0.478	0.309	0.348	0.329	0.369	0.856
Bert + mask	0.383	0.457	0.309	0.339	0.319	0.341	0.864
Bert + CNN	0.577	0.399	0.171	0.260	0.213	0.197	0.854
Bert + CNN + mask	**0.621**	0.410	0.162	0.261	0.224	0.204	0.856
Bert + LSTM	0.457	**0.558**	0.370	**0.421**	**0.437**	0.379	0.861
Bert + LSTM + mask	0.469	0.546	**0.374**	0.395	0.401	**0.389**	**0.865**

Table 4: 散文数据集，探究任务的结果

Model	分词结果
LSTM+CRF分词训练	0.942
LSTM+CRF关系抽取	0.826

Table 5: 分词任务在散文数据集上的实验结果

表5为分词任务在散文数据集上的实验结果，其中LSTM+CRF关系抽取模型，lstm层不参与训练，因此它的输出为上一步模型所提取的特征。

4.2 规律分析

无论何种模型都可以准确的预测出实体类型，清晰地指出了在关系抽取任务中实体类型信息的重要性。对于句长任务，与实体间长度任务等表层信息，所有的编码器都无法进行准确的预测。有关长度的任务是一个简单任务，一个简单的模型再经过专门针对长度任务进行训练之后可以很好的预测长度，因此可以认为Bert+CNN+mask模型在人物关系数据集所取得的0.531的结果证明了句长信息对于关系抽取的不重要性。 (Conneau et al., 2018)在机器翻译的探究任务中也得到了同样的结论，对于越复杂的相关任务所训练出的模型，有关长度的探究任务表现越差，可以认为模型会捕捉更深层次的语义信息而遗忘表层信息。

虽然依存路径是一个复杂的任务，但是中文RE模型仍然可以捕获到一定的依存路径信息。指出了此特征对于中文关系抽取的帮助。实体相关的语义信息如实体间是否存在其他实体，和实体是否有序任务，模型都可以进行准确的预测，这和关系抽取任务主要依赖语义信息的直观判断相符。四种词性的中文探究任务表现不好，主要是因为中文的词性更加复杂，因此难以对关系抽取起到很明显的帮助。但是在人物关系数据集中，首尾实体左侧词性的重要性要比右侧词性大，这是由于中文中对实体进行表明修饰或动作的词语一般出现在名词之前。在散文数据集中，虽然有着文体与书写结构上的不同，但是尾实体左侧词性的重要性仍然是四种词性探究任务中最高的一个。可以说明句子被实体分成的三部分中重要性并不完全相同。

4.2.1 中文探究任务vs 英文探究任务

与 (Alt et al., 2020)的探究实验做对比，我们可以得出深度神经网络在处理中文关系抽取与英文关系抽取时所关注的不同信息。首先是众多词性任务，无论是何种编码结构都无法预测中文实体左右侧上下文的词性，而在英文的探究工作中，模型可以较好地预测实体左右侧单词的词性。这与我们对于两种语言不同点的分析相符。对于句法树深度探究任务，虽然模型都无法较好的预测，但是中文仍然比英文略有提高，英文依靠词形的变化表明词与词之间的联系，而中文依靠语义来分析句子中的联系。直观上认为英文更加容易对句法进行分析，实验结果与想象中不符，但是这也说明了中文关系抽取模型会捕捉到丰富的语义信息来帮助分析句法结构。对于实体间最短依存路径任务，中文比英文有了非常大的提升。依存关系的预测是一个复杂的nlp任务，虽然在本文中将路径长度按照范围进行分类简化了难度，但想要正确预测仍然需要依靠所捕捉到的复杂的语义信息。英文Bert + LSTM模型对于依存路径的预测F1值为0.351，而同样的Bert + LSTM模型在两个中文数据集上的F1值分别为0.590和0.634。不仅与英文数据集相对比模型预测依存路径的能力显著地提高了，在中文数据集中横向比较仍然可以看到表现的稳定性，说明中文的依存路径信息对关系抽取十分重要。而对于实体顺序和实体间是否存在其他实体等与语言种类关系较小的语义任务，两种模型都能进行很好的预测。

总体来看，中文关系抽取需要关注更多的语义特征，如依存路径信息与字间相互关联的语

义信息。这些特征对于中文关系抽取模型来说判断与提取相对而言更加复杂，但是模型却仍然保留了此类特征。而对于句法树与句长等句子整体的结构特征不敏感。此外，词性信息由于判定的复杂程度，以及在中文中对结构的帮助性会被语义信息所替代，对中文关系抽取帮助较小。中文关系抽取对整句的结构性特征不敏感。分词信息是会被模型所推断并加以利用的特征，但是并不是所有分词的重要性都相等，例如'不'，'少'等可以独立表示语义的字向量，在模型中会被当作独立的语义特征。

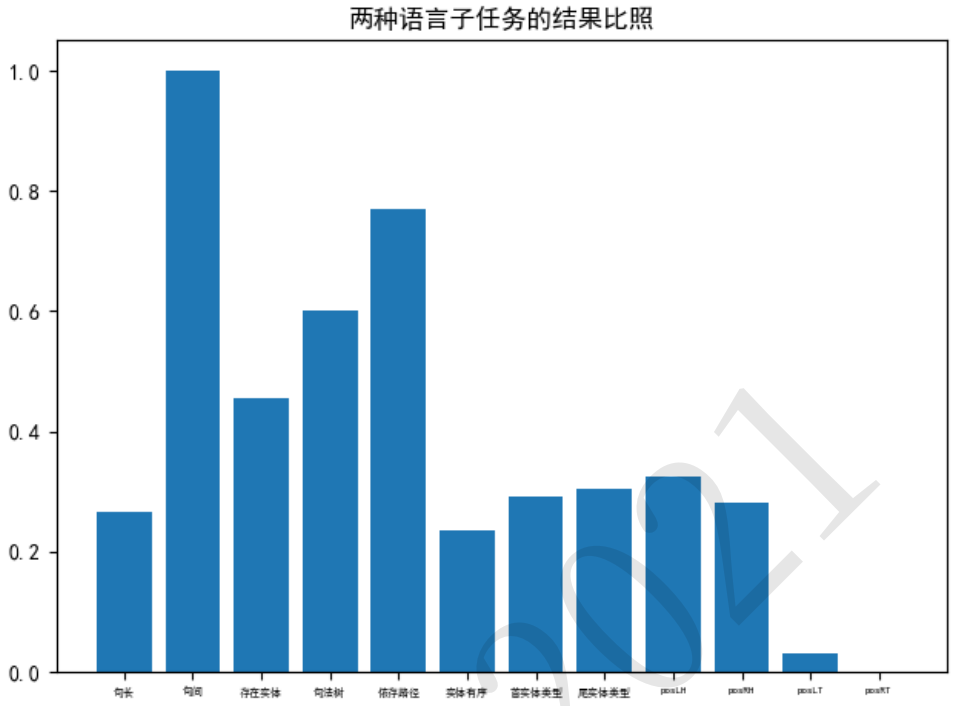

Figure 4: 两种语言比较结果

两种语言的比较结果如图4所示，我们采取平均探究表现较好的Bert+LSTM+Mask模型作为比较标准，将对应任务与关系抽取的结果比值作为此特征的重要程度，将中文与英文做对比以体现中英文中各特征的重要性区别。最后结果经过$x_{normalization} = \frac{x-Min}{Max-Min}$归一化处理，得分越接近一或零说明该子任务所代表的语义特征在中文与英文中的重要性差异越大，接近一偏向中文，接近零偏向英文。

4.2.2 中文分词信息

表5的实验结果证明了，即使是针对于关系抽取任务所训练的LSTM编码器，仍然可以对分词任务取得较好的结果。82.6%的准确率说明了基于字符的神经网络在处理中文文本时，同样会关注并保留大量的字间语义信息。在对分词结果进行错误分析时，发现即使存在着一些误分，但这些误分仍然有着很大的规律性。以一个分词实例为例。

原句：在中国近代史上，由于反抗外来侵略者，沿海有不少小村子名垂青史。

stanford分词结果：在 中国 近代史 上 ， 由于 反抗 外来 侵略者 ， 沿海 有 不少 小村子 名垂青史 。

关系抽取模型分词结果：在 中国 近代 史 上 ， 由 于 反抗 外来 侵略 者 ， 沿海 有 不 少 小 村子 名垂 青史 。

可以发现，模型的结果与标注规范虽然有一些不同，但是对于两字之间有着强烈联系的词语可以区分的很好，而对于如'不'，'小'等高频率词保留单字信息。这说明关系抽取模型在处理中文文本时，神经网络会依照字向量推断并保留字与字之间的联系。

4.2.3 编码结构

不同的编码结构会对探究任务产生影响。以句长任务为例，虽然各种编码结构都无法准确

预测句长任务，但CNN仍然取得了最好的效果，甚至比其他编码结构的表现要高出很多。同样地，对于词性任务，虽然各种编码结构都无法准确预测词性任务，但CNN对词性任务的预测取得了灾难性的表现，可以认为CNN完全忽略掉了词性信息。而LSTM网络结构则是在大多数任务上取得了最好的结果，这主要是因为探究任务的设计是根据人们对语义信息的理解而设计，LSTM神经网络更加专注于捕获具备上下文信息的语义特征。

一个有趣的发现是在关系抽取任务上表现更好的编码器，不一定在下游的探究任务上也表现的很好。以CNN+mask为例，虽然在关系抽取任务上取得了很好的效果，但是在大部分探究任务中都表现的不好。原因可能是设计的探究任务仍然没有体现出编码器所捕获的特征。编码器由于结构，尺寸等因素捕获了更加抽象的特征，这些特征对模型敏感而不易被人理解。

4.2.4 实体掩码

实体掩码被证实在关系抽取任务中会提高模型的表现，我们探究了增加实体掩码对探究任务的影响。关系抽取的实验结果表明再经过一个恰当的编码器结构之后，对实体进行掩码会帮助提高模型的表现。实体掩码同样会导致对探究任务得到不同的结果。一个很明显的例子就是实体类型任务。将输入句子中的实体进行掩码，本质上是隐藏了有关实体的信息，让模型更加关注上下文信息，然而模型仍然可以通过上下文信息推断出实体的类型。这也进一步说明了实体类型对关系抽取任务的重要性。

对于其他探究任务，对实体掩码也会影响结果。例如对于依存路径任务，增加掩码的模型探究结果会有明显的提高。这是因为对实体进行掩码会致使模型更加关注语义与结构性信息，而这也是关系抽取任务模型所依赖的特征。总的来说，对实体进行掩码可以提高关系抽取的准确率，也对基于上下文的语义特征如依存路径，实体间存在其他实体等探究任务的表现有帮助。

5 相关工作

早在神经网络兴起的初期，就有对神经网络进行解释性的探查与分析。(Karpathy et al., 2015)在2015年对LSTM的cells进行可视化，并且发现在LSTM中，不同的cell有着不同的功能。同时还分析了LSTM中三个门的饱和情况，与LSTM特性符合。这种对神经元进行分析的过程可以很客观的体现模型推理的过程，然而也存在一些问题，并非所有神经元都有明确的可被发现的规律，同时也无法通过一个可量化的指标来进行评判。

对神经网络的探究，现阶段已经有很多的工作，比如机器翻译 (Conneau et al., 2018)，语言模型 (Talmor et al., 2019; Roberts et al., 2020)，预训练 (Hewitt and Manning, 2019; Kim et al., 2020)等。(Shi et al., 2016)介绍了探究任务去预测神经翻译模型，是否可以捕获到句法特征。(Adi et al., 2016)也通过不同的探究任务，用于探究神经机器翻译模型中，编码器所捕获的特征。它们包括了长度任务，单词内容任务（去探究一个单词是否出现在句子中）以及词序任务。(Conneau et al., 2018)认为相关工作所提出的探究任务并不能全面的解释神经翻译模型，因此他设计了一个较为全面的探究实验，包括10个与机器翻译密切相关的子任务。(Alt et al., 2020)基于此工作重新设计了针对于关系抽取的探究任务，引入了实体信息等更加专注于关系抽取的探究任务。但是基于中文的语言特殊性，同样的语义探究任务会在中文语料上表现出独立的规律性，同时针对于中文更加特殊的分词信息，仍然没有相关的探究工作。因此我们在本文中针对于中文关系抽取进行了相关的探究工作。

6 总结

本文对关系抽取任务的下游探究任务进行了实验，共分析了13种探究任务以便于理解深度神经网络在进行关系抽取任务时，学习到的语言学特征。通过探究任务，我们更好地理解了深度神经网络的编码结构，在关系抽取任务的特征提取过程中所做的工作。本文比较了中文关系抽取中探究任务实验结果与英文的不同点，帮助理解了中文信息中更加重要的特征，同时也对关系抽取以外的领域提供了参考。

在未来的工作中，我们会设计更加具有代表性的探究任务。也会探究模型是否会推断外部知识，以便于理解外部知识对模型的帮助。

致谢

本文受北京市自然科学基金资助项目（4192057）资助。感谢匿名评审老师提出的修改建议。

参考文献

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing linguistic features of sentence-level representations in neural relation extraction. *arXiv preprint arXiv:2004.08134*.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. 2016. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1471–1480.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 178–181.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. *arXiv preprint arXiv:2002.00737*.

Wenjie Li, Peng Zhang, Furu Wei, Yuexian Hou, and Qin Lu. 2008. A novel feature-based approach to chinese entity relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 89–92.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019a. Is word segmentation necessary for deep learning of chinese representations? *arXiv preprint arXiv:1905.05526*.

Ziran Li, Ning Ding, Zhiyuan Liu, Haitao Zheng, and Ying Shen. 2019b. Chinese relation extraction with multi-grained information and external linguistic knowledge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4377–4386.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Mihai Surdeanu, David McClosky, Mason Smith, Andrey Gusev, and Christopher D Manning. 2011. Customizing an information extraction system to a new domain. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 2–10.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45.

GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*, pages 427–434.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.

崔希亮. 2008. 汉语作为第二语言的习得与认知研究. 汉语作为第二语言的习得与认知研究.

黄伯荣and 廖序东. 2007. 现代汉语教学与自学参考: 增订四版. 现代汉语教学与自学参考: 增订四版.

A 英文探究任务实验结果

	Type Head	Type Tail	Sent Len	Arg Dist	Arg Ord	Ent Exist	PosL Head	PosR Head	PosL Tail	PosR Tail	Tree Dep	SDP Dep	GR Head	GR Tail	F1 score
Majority vote	66.4	33.5	14.5	14.8	54.7	51.0	22.8	23.0	26.9	20.0	23.7	28.4	58.4	75.2	-
Length	66.4	33.5	**100.0**	13.8	54.8	59.4	18.6	24.7	26.9	20.1	**30.5**	29.6	58.4	75.2	-
ArgDist	66.4	33.5	16.5	**100.0**	54.7	77.5	14.9	23.0	26.9	19.8	23.8	35.3	58.4	75.2	-
BoE	77.7	47.6	61.1	22.6	97.3	66.5	33.7	41.5	32.5	36.3	29.8	31.0	66.3	77.4	39.4
CNN	94.0	85.8	47.6	88.1	98.8	**84.5**	70.7	76.1	84.0	86.5	28.5	44.0	78.0	88.6	55.9
+ ELMo	**97.0**	**90.2**	48.7	91.7	99.1	84.3	**76.1**	**81.2**	**86.6**	**90.1**	28.3	45.0	**82.8**	**91.9**	58.8
+ BERT ↓	95.9	88.8	44.7	46.0	93.8	79.9	64.7	74.4	80.8	88.4	29.4	41.0	77.7	90.0	59.7
+ BERT ↑	96.1	88.8	48.0	43.7	91.9	80.0	56.9	70.3	80.1	87.5	28.0	41.3	75.0	89.6	61.0
CNN ⊗	84.2	60.9	46.4	58.3	94.3	81.5	44.3	50.9	54.4	63.9	27.7	40.0	68.5	82.0	59.5
+ ELMo	82.8	69.8	47.4	75.6	98.1	82.9	54.2	60.2	65.4	77.3	28.7	42.4	71.9	85.0	61.7
+ BERT ↓	87.6	80.3	50.9	29.3	83.2	72.4	39.3	46.1	67.7	80.7	30.1	36.9	67.1	87.4	65.3
+ BERT ↑	87.2	79.3	50.6	25.3	78.3	69.8	39.6	42.9	59.9	77.5	30.3	35.1	65.6	86.9	66.1
Bi-LSTM	93.4	81.2	42.0	47.9	**99.4**	79.2	41.2	50.8	50.6	68.4	28.7	41.7	69.3	85.2	55.3
+ ELMo	96.4	89.6	27.9	47.0	97.9	80.9	47.8	52.5	67.2	72.6	25.2	42.8	72.1	90.0	61.8
+ BERT ↓	96.0	87.3	31.0	45.5	99.1	78.8	46.1	55.6	61.7	71.3	26.6	42.7	72.2	87.7	62.5
+ BERT ↑	96.0	87.7	28.6	45.3	97.7	80.4	48.0	50.9	61.4	67.4	25.1	42.3	70.8	87.0	63.1
Bi-LSTM ⊗	81.9	71.4	27.6	35.6	90.6	73.2	36.1	40.5	59.3	66.4	25.7	38.4	64.6	85.3	62.9
+ ELMo	82.8	50.7	30.6	19.7	73.4	65.0	32.0	35.9	37.9	41.8	28.0	32.2	63.0	79.5	64.1
+ BERT ↓	82.3	77.9	34.1	25.6	87.6	68.4	32.5	36.7	61.5	64.7	27.6	35.1	66.6	86.0	65.4
+ BERT ↑	81.7	79.6	30.2	21.3	81.1	67.0	30.6	33.8	55.9	55.1	27.3	34.2	64.1	84.9	66.1

Figure 5: 英文探究任务实验结果