doi:10.3772/j.issn.2095-915x.2022.03.005

基于图的汉语字级别依存分析联合模型

开放科学（资源服务）标识码（OSID）

汪凯 梁宇腾 张玉洁 徐金安 陈钰枫

北京交通大学计算机与信息技术学院 北京 100044

摘要： [目的 / 意义] 汉语分词、词性标注和依存句法分析作为汉语自然语言处理的三大基本任务发挥着至关重要的作用。基于转移的三个任务联合模型曾经取得最好精度，但是随着神经网络和计算能力的发展，具有全局信息建模能力的图模型，在单任务和两个任务上已经超过转移模型。如何在基于图模型下联合三个任务，进一步提升精度成为新的挑战。[方法 / 过程] 本文提出一种基于图的汉语分词、词性标注和依存句法分析的联合模型，通过设计统一的字级别标签实现三个任务的联合，并采用预训练语言模型融合上下文信息的字表示方法和基于双仿射注意力机制的评分函数。本文也设计了联合模型的解法算法用于三个任务的解码。[结果 / 结论] 实验结果表明，本文词性标注任务的引入方式可以建模词性与分词以及词性与依存句法分析之间的关系，从而带来其他两个任务上精度的提升。与目前精度最好的 Yan[1] 工作相比，在三个任务上都取得最好精度。

关键词： 依存分析；联合模型；词性标注；汉语分词

中图分类号： G35；TP391

A Joint Model for Graph-based Chinese Character Level Dependency Parsing

WANG Kai LIANG Yuteng ZHANG Yujie XU Jinan CHEN Yufeng

Beijing Jiaotong University, Beijing 100044, China

Abstract: [Objective/Significance] Chinese word segmentation, POS tagging and dependency parsing play a vital role. The three-task transition-based joint model has achieved the best accuracy, but with the development of neural networks and computing capabilities, the graph-based model with global information modeling capabilities has surpassed the transition-based model in single-task and two-tasks. How to combine the three tasks based on the graph-based framework to further improve the accuracy has become a new challenge. [Methods/Process] This paper proposes a joint model of three tasks based on graph-based framework. The combination is realized by designing unified character-level tags, and the character context representation method

基金项目 国家自然科学基金 (61876198, 61976016)。

作者简介 汪凯（1997-），硕士，研究方向为自然语言处理、依存句法分析；梁宇腾（1998-），硕士，研究方向为自然语言处理、依存句法分析、复述；张玉洁（1961-），教授，研究方向为自然语言处理、机器翻译、文本大数据处理，E-mail：yjzhang@bjtu.edu.cn；徐金安（1970-），教授，研究方向为自然语言处理、机器翻译；陈钰枫（1981-），教授，研究方向为自然语言处理、机器翻译。

引用格式 汪凯 , 梁宇腾 , 张玉洁 , 等 . 基于图的汉语字级别依存分析联合模型 [J]. 情报工程 , 2022, 8(3): 68-80.

based on pre-training language model (e.g. BERT). The scoring function implemented by the biaffine attention mechanism. This paper also designs the solution algorithm of the joint model for the decoding of the three tasks. [Results /Conclusions] The experimental results show that, the introduction of the POS tagging can better model relationship between part-of-speech and word segmentation, as well as between part-of-speech and dependency parsing, so as to improve the accuracy of the other two tasks. Compared with the Yan work[1], the best performance is achieved on the three tasks.

Keywords: Dependency parsing; joint model; POS tagging; Chinese word segmentation

引言

随着计算机的普及和网络技术的发展，电子化的科技文献、专利和网络文本数据与日俱增，成为人们获取信息的主要来源之一和情报加工的重要素材[2]。面向大规模电子化文本数据的信息收集及情报加工的自动化处理方式也因此进入情报研究的视野，出现多文档摘要、事件抽取、热点跟踪与发现，以及知识图谱构建等面向文本的大数据分析任务[3-5]。以自然语言描述的文本数据属于非结构化数据，其处理方式与结构化数据不同，需要对自然语言句子进行语言分析，在语义理解的基础上才能获取正确的信息和有意义的情报，语言分析作为语义理解的基础技术显得尤为重要。汉语文本数据的分析技术包括分词、词性标注和依存句法分析，它们的性能直接关系到上述大数据分析任务的工程应用效果。

分词、词性标注和依存句法分析三个任务一直是汉语信息处理研究的核心课题[6-7]，早期三个任务的串行处理方式是在分词之后进行词性标注，然后再进行依存句法分析。之后研究人员提出联合处理方式，以避免串行方式中的错误传播，同时加强三个任务之间中间结果的相互利用，在三个任务的性能上取得显著提升[8]。目前存在两种主流的联合处理方法，一种是基于转移框架的联合模型，另一种是基于图框架的联合模型。早期的联合模型主要采用基于转移的框架[9]，但是转移模型预测时只能利用局部信息，导致长距离依存弧的分析精度难以提升[10-12]；而基于图框架的联合模型具有全局信息的建模能力，可使长距离依存弧得到有效处理，成为主流框架。

基于图框架的联合模型在包括汉语在内等多种语言上为依存句法分析带来性能上的提升[13-15]，其中包括词性标注与依存句法分析的联合模型[16]，表明联合词性标注可以提升依存分析的精度；也包括汉语分词和依存句法分析的联合模型[1]，在两个任务上取得目前最好性能。在基于图框架下如何联合三个任务，进一步提升精度成为新的挑战。为此，本文首次提出在图框架下三个任务的联合模型。

本文剩余部分的组织结构如下：第一节介绍相关研究；第二节介绍基于图的三个任务联合模型的框架；第三节描述三个任务联合模型中的核心技术；第四节介绍实验部分；第五节为结语，对本文研究进行总结。

1 相关研究

汉语的分词、词性标注和依存句法分析三个任务的联合处理方式是缓解串行处理方式中

doi:10.3772/j.issn.2095-915x.2022.03.005

错误传播问题的重要手段。早期，研究人员主要关注基于转移框架的联合模型，取得了诸多研究成果。转移框架是为了依存句法分析提出的一种解析方式，具体包括一个缓存区（Buffer，用于存放输入句的单词序列）、一个栈（Stack，用于存放单词间的依存关系）以及几个动作（用于缓存区与栈之间的数据操作），比如“移进（进栈）”，“左归约”和“右归约”等动作，然后通过预测一个移进规约转移动作序列构建一棵依存句法树，将依存分析问题建模为寻找最优动作序列的问题。由于每步决策只能利用栈顶和缓冲区内部分单词的信息，因此是一种局部信息建模的方式。Hatori 等 [9] 在 2012 年首次提出基于转移框架的三个任务联合模型，以依存句法分析的转移框架为基础，将分词与词性标注两个任务分别转化为两种特殊的字级别的依存句法分析任务。具体的，为分词与词性标注两个任务分别设计分词动作与预测词性动作并加入到原先依存句法分析的动作集合中，在汉字执行规约操作时进行分词和词性的预测，从而在汉字级别实现三个任务的统一与联合。Zhang 等 [10] 在 2014 年手工标注了汉语单词内部汉字之间的依存关系类型，并且定义了单词内部的特征模板，通过结合词内和词间的依存特征提升了依存句法分析的精度。随着深度学习的发展，Kurita 等 [11] 在 2017 年首次将神经网络的方法应用于三个任务的联合模型，该模型结合了原有特征模板和基于深度学习的预训练字向量、词向量，并采用前馈神经网络进行转移动作的预测，在分词和词性标注的精度上得到大幅提升。

随着计算机计算能力的大幅提升，具有全局信息建模能力的图框架受到研究人员的关注，随之提出的一些模型在精度上也超过了基于转移框架的模型。图框架设置点和边来分别表示单词和单词之间的依存关系构建一个有向图，将依存句法分析转换成在有向图中寻找得分最大生成树的问题 [17]。与转移框架相比，图框架能够同时考虑所有单词的信息，使得在长距离依存弧的预测上也具有优势。Yan 等 [1] 在 2020 年首次在基于图的框架下将汉语分词与依存句法分析进行联合，在两个任务上的精度超过基于转移框架模型的最好精度。

考虑先前研究中基于转移框架的联合模型的优势以及近年基于图框架的依存句法分析的发展趋势，我们尝试实现基于图框架的三个任务的联合方式，意图利用图框架全局信息的建模能力和三个任务的相互学习机制，提升三个任务的精度。本文一方面探索如何在基于图的框架下进行三个任务的转换，另一方面探究词性标注任务引入对性能的影响及原因。

2 基于图的三个任务联合模型的框架

依存句法分析任务的输入是一个带有词性标注的单词序列，输出为依存句法树，通过树结构刻画句子内部单词之间的组合或修饰关系。形式化地输入句子可以表示为：$x=\$_0 w_1, ..., w_n$，对应的词性序列为 $p_0 p_1, ..., p_n$，其中 w_i 表示输入句子的第 i 个词，p_i 为 w_i 的词性，$\$_0$ 为人工设置的伪词，作为依存句法树的根节点，用于指明句子中的核心动词。一棵符合依存文法的句法树由多条依存弧构成，可以表示为：$t = \{(h,m,l); 0 \le h \le n, 0 < m \le n, l \in L\}$，其中（$h,m,l$）

表示从核心节点（Head）w_h 到依存节点（Dependent）w_m 的一条依存弧，标签 l 表示依存弧上的关系类型，L 是所有依存关系类型的集合，同时我们采用 $w_m \overset{l}{\curvearrowleft} w_h$ 进行标记。

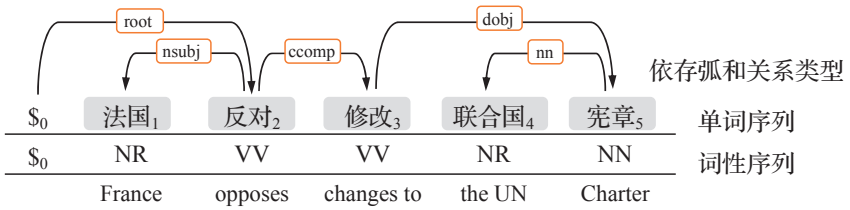

图 1 依存树示例

图 1 表示输入句子“法国 $_1$ 反对 $_2$ 修改 $_3$ 联合国 $_4$ 宪章 $_5$”的依存树。“反对”是句子中的核心动词，从词“反对 $_2$”到“法国 $_1$”的依存弧表示“反对 $_2$”是核心节点，“法国 $_1$”是依存节点，它们之间的依存关系类型为“nsubj”，表示名词作为动词的主语（noun as subject），可以标记为“法国 $_1$ $\overset{nsubj}{\curvearrowleft}$ 反对 $_2$”。

由于汉语分词是汉字级别的任务，而词性标注和依存句法分析却是单词级别的任务，为了将分词、词性标注和依存句法进行联合，首先需要将三个任务转换成统一的处理方式。本文的解决方式是将单词级别的词性标注和依存句法分析都转换成汉字级别的任务，为此本文首先设计相应的标签转换方式，将分词标签和词性标签与依存句法标签进行融合，形成三个任务的字级别标签，借此可以将三个任务都统一成汉字级别的依存句法分析任务。其次，在基于图框架的依存句法分析中，只需将图中的节点由单词替换成汉字，并借助现有图框架的分析方法可以在汉字级别实现实现三个任务的联合。

联合模型的整体框架如图 2 所示，输入是汉字序列，输出是分词、词性标注、依存句法分析三个任务的分析结果。联合模型主要包含三大部分，分别是汉字表示方法、评分函数以及解码部分。字表示方法负责将汉字序列转换成一组包含上下文信息的语义向量；评分函数部分负责对依存弧、依存关系类型进行评分；最后的解码算法负责将模型输出的字级别标签解码成三个任务的结果。下面将对联合模型所涉及的核心技术进行详细描述。

3 核心技术概述

3.1 三个任务的字级别标签设计

为了实现三个任务的联合，我们首先需要将分词和词性标注转换为依存句法分析任务。我们可以将分词任务看作词内字之间的依存句法分析。首先将每个单词转换成一个特殊的子树，单词中的字以其右邻的字作为核心节点，设置新的依存关系类型“app”作为词内字之间的依存关系类型。接着需要将词级别的依存句法分析转换为字级别的依存句法分析，我们可以将单词之间的依存弧转化为两个单词中最后一个字之间的依存弧，即单词最后一个字的核心节点为该单词核心节点词的最后一个字，依存关系类型为原来单词之间的依存关系类型。以图 1 中的例子说明，其中单词“联合国”中

doi:10.3772/j.issn.2095-915x.2022.03.005

的“联”与“合”以其右邻字“合”与“国”为其核心节点，并获得依存关系类型“app”[1]，可表示为“联$^{app}_{\curvearrowleft}$合$^{app}_{\curvearrowleft}$国”；而单词之间的依存弧“联合国$^{nn}_{\curvearrowleft}$宪章”转换为两个词最后的字“国”与“章”之间的依存弧，可表示为“国$^{nn}_{\curvearrowleft}$章”。图 1 中的例子原先是词级别，经过转换后得到字级别结果如图 3 所示，如此实现分词和依存句法分析两项任务的联合。

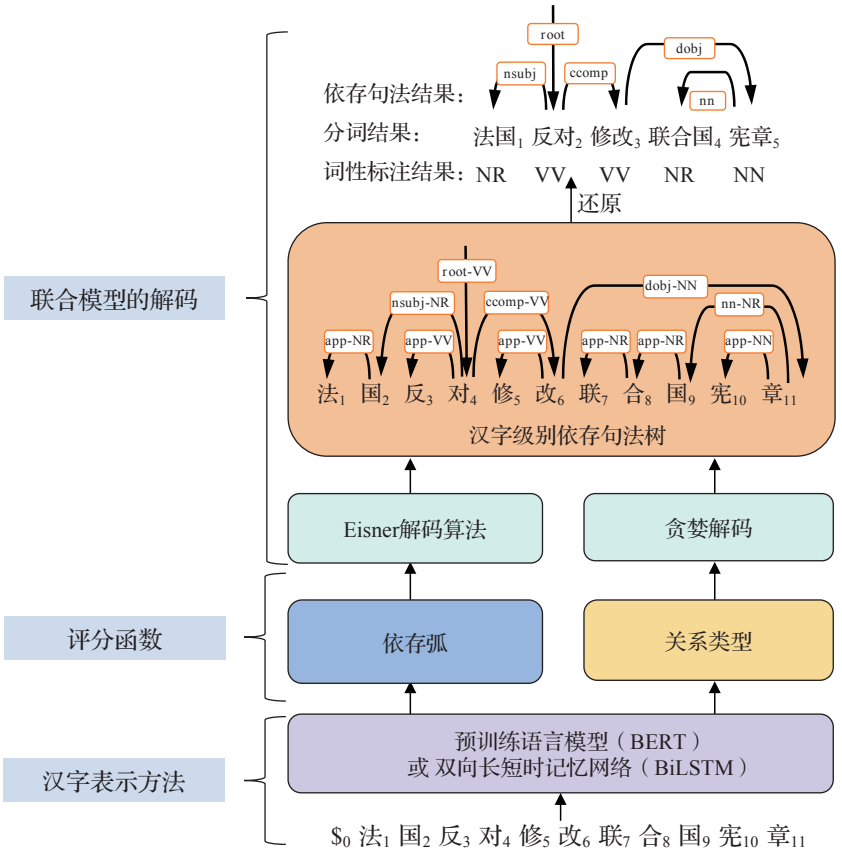

图 2 基于图框架的三个任务联合模型

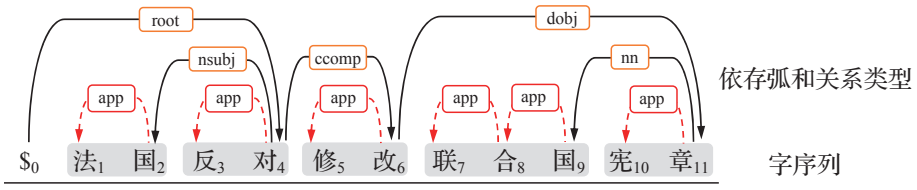

图 3 联合分词和依存句法分析的字级别依存树

为了将词性标注任务与分词和依存句法分析一起融合到图框架中，本文设计一种基于字的词性标注方式。首先，我们可以将词性标注转换为对于字的词性标注。我们认为构成词的每个字具有与该词相同的词性，如此对于词性为 p_i 的单词 w_i，可以将构成 w_i 的所有字的词性设置成 p_i。为了与分词和依存句法分析任务联合，我们将每个字的词性拼接到与其核心节点之间的关系类型上。例如，“联合国”的词性为“NR”（Proper Nouns），构成该词的三

个字的词性标签都设置为“NR”，前两个字的词性标签“NR”与“联$\overset{app}{\curvearrowright}$合$\overset{app}{\curvearrowright}$国”中的关系类型“app”进行拼接，得到“联$\overset{app-NR}{\curvearrowright}$合$\overset{app-NR}{\curvearrowright}$国”，最后一个字的“NR”与“国$\overset{nn}{\curvearrowright}$章”中的关系类型“nn”进行拼接，得到“国$\overset{nn-NR}{\curvearrowright}$章”。图 2 中的例子经过转换后的结果如图 4 所示，如此将词性标注任务引入到分词和依存句法分析的联合模型中，从而实现三项任务的联合。如此我们可以在现有基于图的依存分析框架下实现三个任务的联合。

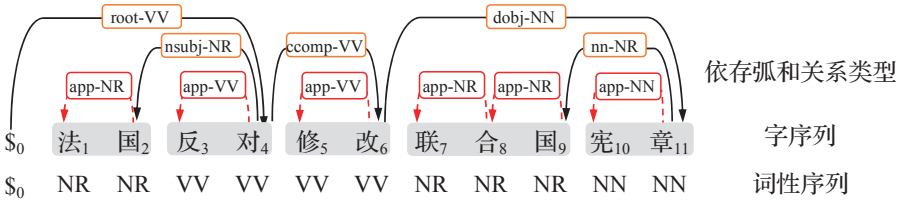

图 4 转换后能融合三个任务的字级别依存树

3.2 汉字表示方法

经过 3.1 节介绍的方法，已将三个任务的标签都标注在字级别依存树上，如此只需要针对汉字序列，让模型预测出汉字级别依存树，即可通过还原的方式恢复出三个任务的分析结果。对于输入句 $x=\$_0\, c_1\, c_2, ..., c_m$，其中 c_i 表示第 i 个字。字的表示方法将离散的字序列映射为语义空间中上下文相关的一组向量。本文设计两种融合上下文信息的表示方法，一种是基于双向 LSTM[18]（BiLSTM）的表示方法，另一种是基于 BERT（Bidrectional Encoder Representations from Transformers）[19] 的表示方法。

当表示方法采用 BiLSTM 时，首先使用传统的 ngram 对短距离的上下文信息进行建模，具体是使用单字（unigram）、双字（bigram）和三字（trigram）[1] 特征。每个字表示为 $\boldsymbol{e}_i=\boldsymbol{e}_{ci} \oplus \boldsymbol{e}_{cici+1} \oplus \boldsymbol{e}_{cici+1ci+2}$，其中⊕为向量拼接操作符，$\boldsymbol{e}_{ci}$、$\boldsymbol{e}_{cici+1}$ 和 $\boldsymbol{e}_{cici+1ci+2}$ 为单字 c_i、双字 c_ic_{i+1} 和三字 $c_ic_{i+1}c_{i+2}$ 的嵌入表示。本文使用三层 BiLSTM 对长距离上下文信息的建模。融合上下文信息后第 i 个字 c_i 的表示用 LSTM 最后一层输出 $\boldsymbol{h}_i$ 表示，如公式（1）所示。

$$\boldsymbol{h}_i = \text{BiLSTM}(\boldsymbol{e}_i, \theta_{\text{bilstm}}) \quad (1)$$

其中 θ_{bilstm} 为 BiLSTM 模型参数。

当表示方法采用 BERT 时，使用 BERT 的最后一层输出作为融合上下文信息后字的表示，为了保持后续统一，我们仍使用 $\boldsymbol{h}_i$ 来表示字 c_i。

$$\boldsymbol{h}_i = \text{BERT}(x, \theta_{\text{bert}}) \quad (2)$$

其中 θ_{bert} 为 BERT 模型的参数。

3.3 依存弧与依存关系类型评分函数

上面我们已经将三个任务的联合转换为字级别的依存句法分析任务，所以图模型上的节点与句子中每个字相对应，节点之间的弧与字级别依存弧相对应。本文采用基于双仿射注意力 [20] 的依存弧评分函数。由于每个字既可以作为依存节点又可作为核心节点，我们用两个多层感知机 $\text{MLPs}^{\text{arc-head}}$ 与 $\text{MLPs}^{\text{arc-dep}}$ 分别学习 c_i 节点作为核心节点和依存节点的特征表示，一个学习 c_i 作为核心节点的表示（arc-head），另

doi:10.3772/j.issn.2095-915x.2022.03.005

一个学习 c_i 作为依存节点的表示（arc-dep），分别如公式（3）和公式（4）所示。

$$\boldsymbol{r}_i^{\text{arc-head}} = \text{MLPs}^{\text{arc-head}}(\boldsymbol{h}_i) \quad (3)$$

$$\boldsymbol{r}_i^{\text{arc-dep}} = \text{MLPs}^{\text{arc-dep}}(\boldsymbol{h}_i) \quad (4)$$

在此基础上，利用双仿射注意力建模所有依存弧的评分，即计算任意两个节点 c_i 和 c_j 之间构成依存弧 $c_j \curvearrowleft c_i$ 和 $c_i \curvearrowleft c_j$ 的评分，以 $c_j \curvearrowleft c_i$ 为例，其构成依存弧评分计算如公式（5）所示。

$$s_{ij}^{\text{arc}} = \boldsymbol{r}_j^{\text{arc-dep}} \boldsymbol{W}^{\text{arc}} \boldsymbol{r}_i^{\text{arc-head}} + u^{\text{arc}^{\text{T}}} \boldsymbol{r}_i^{\text{arc-head}} \quad (5)$$

其中，$\boldsymbol{W}^{\text{arc}} \in \mathbb{R}^{d \times d}$ 是 2 维的权重矩阵，$u^{\text{arc}} \in \mathbb{R}^{d}$ 为权重向量。对于依存关系类型的预测，与依存弧的预测过程相似，也为每个字设置两个向量表示，一个作为依存关系类型的核心节点的表示（label-head），另一个作为依存关系类型的依存节点的表示（label-dep），分别如公式（6），（7）所示。依存弧 $c_j \curvearrowleft c_i$ 所有可能的依存关系类型评分如公式（9）所示。

$$\boldsymbol{r}_i^{\text{label-head}} = \text{MLPs}^{\text{label-head}}(\boldsymbol{h}_i) \quad (6)$$

$$\boldsymbol{r}_i^{\text{label-dep}} = \text{MLPs}^{\text{label-dep}}(\boldsymbol{h}_i) \quad (7)$$

$$\boldsymbol{r}_{ij}^{\text{label}} = \boldsymbol{r}_i^{\text{label-head}} \oplus \boldsymbol{r}_j^{\text{label-dep}} \quad (8)$$

$$\boldsymbol{s}_{ij}^{\text{label}} = \boldsymbol{r}_i^{\text{label-head}} \boldsymbol{U}^{\text{label}} \boldsymbol{r}_j^{\text{label-dep}} + \boldsymbol{W}^{\text{label}} \boldsymbol{r}_{ij}^{\text{label}} + \boldsymbol{u}^{\text{label}} \quad (9)$$

其中 $\boldsymbol{U}^{\text{label}} \in \mathbb{R}^{K \times p \times p}$ 是 3 维的权重矩阵，$\boldsymbol{W}^{\text{label}} \in \mathbb{R}^{K \times 2p}$ 是 2 维的权重矩阵，$\boldsymbol{u}^{\text{label}} \in \mathbb{R}^{K}$ 是偏置向量，K 为不同依存关系类型的数目。在预测时，依存弧 $c_j \curvearrowleft c_i$ 的依存关系类型被设置成得分最高的关系类型。

3.4 联合模型解码算法

本文设计的解码算法主要分为两个步骤：第一步是预测字级别依存树和字级别依存关系类型；第二步将字级别依存树和字级别依存关系类型转换成词级别的预测结果。

第一步预测方法如下：采用 Eisner 算法 [17] 从所有符合依存文法的依存树中找到得分最高的字级别依存树，即得到字级别依存树中的所有依存弧；然后使用贪婪解码 [20] 来获得依存弧上的字级别关系类型。

第二步转换步骤如下。

（1）分词预测：利用第一步得到的字级别依存关系类型进行分词预测。若当前字与其核心节点之间的字级别依存关系类型为“app”开头，则表示这两个字属于同一个词；若非“app”开头，则表示当前字属于词的最后一个字，处于词的边界。如图 5 所示，字 c_1 和 c_2 与其核心节点的依存关系类型分别是“app-NR”和“app-NN”，都是以“app”开头，字 c_3 与其核心节点 c_5 的字级别依存关系类型为“nsubj-NR”，不再以“app”开头，则说明 c_3 是词的最后一个字，前三个字 $c_1c_2c_3$ 构成一个词。

（2）词级别词性预测：针对步骤（1）得到的分词结果，根据词内的字级别依存关系类型拆分出字级别词性，然后使用投票机制得出预测词的词性，若出现的最多次数相同，则随机选择一个。例如图 4 中 $c_1c_2c_3$ 构成一个词，三个字与其核心节点的字级别依存关系类型分别为“app-NR”，“app-NN”以及“nsubj-NR”。根据分隔符“-”进行拆分得到字级别词性分别为“NR”，“NN”和“NR”。这三个词性中出现次数最多的词性为“NR”，所以词 $c_1c_2c_3$ 的词性预测为“NR”。

（3）词级别依存弧和依存关系类型预测：根据步骤（1）（2）得到的分词结果与词性，根据词中最后一个字的字级别依存弧和字级别

依存关系类型，可以得出词级别依存弧和依存关系类型。如图 5 所示的两个词 $c_1c_2c_3$ 和 c_4c_5，两个词中最后一个字分别为 c_3 和 c_5 之间的依存关系为 $c_3^{nsubj-NR}c_5$，从而得到词 $c_1c_2c_3$ 为依存节点、词 c_4c_5 为核心节点，类型为“nsubj”的依存句法。

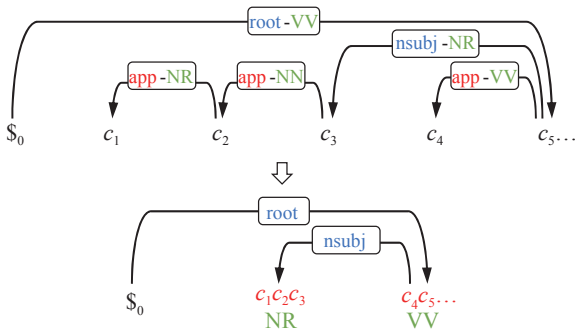

图 5 字级别结果转换为词级别

我们为上述模型的训练目标涉及两个损失函数 L_1 和 L_2，L_1 用于学习依存弧，L_2 用于学习依存关系类型，并都采用交叉熵损失 [21] 函数。学习依存弧时，交叉熵损失 L_1 最大化正确依存弧的评分，以依存弧 $c_j \curvearrowleft c_i$ 为例，损失由公式（10）计算。

$$L_1 = -\log \frac{e^{s_{ij}^{arc}}}{\sum_{0 \le k \le m} e^{s_{kj}^{arc}}} \quad (10)$$

学习依存关系类型时，在正确依存弧的基础上，交叉熵损失 L_2 最大化该弧正确关系类型的评分，如公式（11）所示。

$$L_2 = -\log \frac{e^{s_{ij}^{label}}}{\sum_{k \in L} s_{ij}^{k}} \quad (11)$$

其中 L 为所有依存关系类型的集合。本文提出的联合模型设计到两个目标函数 L_1 与 L_2，我们采用线性加权的方式融合两个目标函数，如公式（12）所示。

$$L=\omega L_1+(1-\omega)L_2 \quad (12)$$

4 实验

4.1 实验数据与评价方法

本文在 CTB5、CTB7 和 CoNLL09 三个汉语公开数据集上进行实验。对于 CTB5 和 CTB7，数据划分遵循已有工作 [11]。分词、词性标注和依存句法分析的评测指标均采用准确率、召回率、综合性能指标 F1 值。同时为了方便对比实验分析，在部分实验中也采用了依存句法评价指标 UAS（Unlabeled Attachment Score）和 LAS（Labeled Attachment Score）。

4.2 超参数设置

本文提出模型的超参数设置如下：多层感知机（$MLPs^{arc\text{-}head}$、$MLPs^{arc\text{-}dep}$、$MLPs^{label\text{-}head}$、$MLPs^{label\text{-}dep}$）的层数都设置为 1 层，激活函数设置为 leakyReLU，前两个的输出维度 d 设置为 500，后两个的输出维度 p 设置为 100。在 BiLSTM 和所有的 MLP 中加入 dropout 层，dropout 率设置为 0.33；损失函数公式（13）中的 ω 设置为 0.5，batchsize 设置为 128，双向 LSTM 的维度设置为 400。本文采用 Adam 优化算法更新模型参数。此外，当表示方法采用双向 LSTM 时，学习率设置为 0.002 训练 100 轮；当表示方法采用 BERT 时，学习率设置为 0.00002 训练 15 轮。我们在开发集上验证模型，并且保留开发集上依存句法 F1 值最好的模型，用于最后在测试集上进行评测。

4.3 与已有模型的对比结果

本文与已有的基于转移和基于图的代表性工作进行比较，比较结果列在表 1 中，分别

doi:10.3772/j.issn.2095-915x.2022.03.005

用 T 和 G 表示基于转移和基于图的模型。在基于转移的模型中，我们从已有工作中选取了在 CTB5 和 CTB7 上表现最好的模型进行对比，其中 Kurita 等人 [11] 的模型在分词和词性标注任务上的精度最好，模型 Zhang14 STD[10] 在 CTB7 数据集的依存句法分析任务上精度最好，模型 Zhang14 EAG[10] 在 CTB5 数据集上的依存句法分析任务精度最好。从中可以看出，基于转移的这些模型在不同数据集上和不同任务上所具有的优势不同。同时，我们注意到没有一种方法可以在两个数据集的三个任务上同时达到最好精度，具有不稳定性；特别是模型 Kurita17 在分词和词性标注任务上的精度提升并没有带来依存句法分析任务上相应的精度提升。关于基于图的模型，我们将 Yan 的工作 [1] 列在表 1 中进行对比，从中看出无论是 Yan20 还是本文的模型，与基于转移的模型相比，在 CTB5 数据集上分词和依存句法分析至少提升 0.11% 和 6.16%，在 CTB7 数据集上分词和依存句法分析至少提升 0.78% 和 6.17%。基于图的模型在三个任务上的精度提升显示图模型在全局信息建模上的优势。

表 1 联合模型实验结果

模型	类型	CTB5			CTB7			CoNLL09		
		$F1_{seg}$	$F1_{pos}$	$F1_{udep}$	$F1_{seg}$	$F1_{pos}$	$F1_{udep}$	$F1_{seg}$	$F1_{pos}$	$F1_{udep}$
Zhang14 STD	T	97.67	94.28	81.63	95.53	90.75	75.63	-	-	-
Zhang14 EAG	T	97.76	94.36	81.70	95.39	90.56	75.56	-	-	-
Kurita17	T	98.37	94.83	81.42	95.86	90.91	74.04	-	-	-
Yan20	G	98.48	-	87.86	96.64	-	81.80	-	-	-
Ours	G	**98.49**	**95.88**	**87.95**	**96.78**	**93.03**	**82.11**	**96.75**	**93.07**	**84.65**
Yan20 (BERT)	G	98.46	-	89.58	97.06	-	85.06	-	-	-
Ours (BERT)	G	**98.70**	**96.72**	**90.59**	**97.27**	**94.26**	**85.94**	**97.12**	**93.17**	**88.17**

与 Yan20 相比，我们的模型（Ours）在引入词性信息后，在 CTB5 和 CTB7 两个数据集上分词和依存句法分析精度都有提升，尤其在 CTB7 数据集上分词和依存句法分析分别提升 0.14% 和 0.31%。Yan 没有进行词性标注，所以我们与基于转移的工作中在词性标注上表现最好的模型 Kurita17 相比，在 CTB5 和 CTB7 上词性标注的精度分别提升 1.00% 和 2.10%。上面的分析结果表明本文模型在所有报道的三个任务上的精度达到最好。我们分析三个任务上精度的同时提升是在图框架中引入词性标注任务带来的结果，词性信息引入一方面带来分词精度提升，从而使依存句法分析精度提升；另一方面词性信息对依存句法分析本身也有直接帮助。我们会在后面进行详细分析。

当表示方法采用 BERT 时，本文模型 Ours（BERT）与 Ours 相比在三个任务上精度均有提升，说明使用预训练模型 BERT 相较于 BiLSTM 会带来性能上的提升。与同样使用预训练模型的 Yan20（BERT）相比，在 CTB7 上的分

词和依存句法精度分别提升 0.21% 和 0.88%，说明在使用预训练模型的基础上引入词性标注任务依旧有助于提升分词和依存句法分析的精度，同时也说明 BERT 在预训练过程中学习词性信息并不充分，本文联合词性标注任务可以弥补并进一步提升精度。

4.4 引入词性标注任务的效果

本文在 Yan 的工作上基于图的框架下引入词性标注任务，下面分别分析其在分词和依存句法分析上的效果。

4.4.1 词性信息对未登录词的作用

未登录词是指出现在测试集而未出现在训练集中的词，又称为集外词（out of vocabulary, OOV），实际应用中未登录词的出现是普遍现象，会严重影响分词的精度，因此对于未登录词的处理能力是衡量模型的重要指标。我们通过未登录词的召回率这一指标来衡量模型对未登录词的处理能力。我们比较了引入词性标注任务的模型 Ours（BERT）和没有引入的模型 Yan20（BERT）对未登录词的召回能力，比较结果列于表 2。可以看出引入词性标注任务的模型在三个数据集上的未登录词召回率均有提升，并且我们进一步分析发现登录词的召回率并没有下降，由此得出结论，分词精度的提升来源于词性标注任务对未登录词处理能力的提升。我们进一步分析 Ours（BERT）与 Yan20（BERT）的差异，具体的对前者能正确分词的登录词，但后者未能正确分词的未登录词进行统计，并按照未登录词的分类 [22] 进行分析，结果列于表 3。我们发现词性信息的引入在各种类型的未登录词上都起到了改善的效果，其中在“普通生词”上的改善最为明显，其次是“成语”，在改好的结果中占比分别达到 50.5% 和 19.2%。我们分析其原因为汉语中构成词的字之间遵循一定模式，本文引入词性标注任务并将其转换为字上词性标签的预测方式，能够建模字的词性与分词之间的关系，从而帮助包括未登录词在内的分词精度提升。

表 2 不同数据集上未登录词的召回率

模型	CTB5	CTB7	CoNLL09
Yan20 (BERT)	86.11	86.30	89.43
Ours (BERT)	**86.70**	**86.90**	**89.69**

表 3 分词正确类型统计

正确类型		正确数	比例/%		例子
命名实体	人名	15	5.4	13.0	罗伯特·雷
	地名	10	3.6		台南县市
	组织机构名	4	1.4		县支行
	时间和数字	7	2.5		一千年
专业术语		30	10.9		给排水
成语		53	19.2		肝胆相照
普通生词		139	50.5		开矿
英语单词		17	6.1		Chinese
合计		275	100		

4.4.2 词性信息对依存句法的影响

为了进一步分析词性标注任务的引入对依存句法分析的影响，我们与没有词性标注任务的 Yan 工作进行对比。我们首先根据构成依存弧的头节点和依存节点的词性模式对依存弧进行分类，选取其中占比很大的头节点为动词的依存弧，然后在这些依存弧上比较两个模型的错误率，比较结果按占比从高到低排序，列在表 4 中。词性模式 $X \sim Y$ 表示从词性为 X 的单

doi:10.3772/j.issn.2095-915x.2022.03.005

词到词性为 *Y* 的单词的依存弧。我们发现本文模型在这些词性模式上 VV↷NN 的依存弧的错误率都有所降低。例如在 CTB7 数据集上以模板的依存弧数量高达到 23049，本文所提模型 Ours（BERT）预测的错误率降低 0.26%；在 CoNLL09 数据集上，错误率降低 0.34%。我们分析其原因：由于构成依存弧的头节点和依存节点在词性上遵循一定模式，以上面表 4 中显示的模式为例说明，一方面有些词性的单词不大可能作为“VV”的依存节点，比如“JJ”与“DEG”等；另一方面能够作为其依存节点的词性的单词按照其数量也有大小差异。所以当本文引入词性标注任务并将其转换为字之间的依存关系类型的预测时，能够建模词性模式与依存弧构建之间的关系，从而帮助依存弧的预测。

表 **4** 不同词性标签模板下解析错误率对比

X↷*Y*	CTB7			CoNLL09		
		Yan20 (BERT)	Ours (BERT)		Yan20 (BERT)	Ours (BERT)
	数量	错误率	错误率 (±)	数量	错误率	错误率(±)
VV↷NN	23049	12.31	12.05(-0.26)	7250	12.63	12.29(-0.34)
VV↷AD	13967	13.07	12.63(-0.44)	3638	10.94	10.25(-0.69)
VV↷P	6872	11.66	11.31(-0.35)	2127	10.30	9.12(-1.18)
VV↷NR	4047	9.27	8.90(-0.37)	1340	9.93	8.36(-1.57)

表 **5** 汉语分词正确情况下依存分析性能对比

模型	CTB5		CTB7		CoNLL09	
	UAS	LAS	UAS	LAS	UAS	LAS
Yan20 (BERT)	92.39	89.74	88.51	84.92	90.50	87.25
Ours (BERT)	**92.61**	**89.96**	**88.79**	**85.12**	**91.58**	**88.60**

上面的实验都是在自动分词的基础上进行的实验结果分析，我们分析引入词性标注任务可以带来分词精度的提升，从而提升依存句法分析精度。为了分析引入词性标注任务对依存句法分析的直接影响，我们在分词都正确的基础上进一步比较本文模型与 Yan 的工作。为了使用金标分词，我们在联合模型解码时，选择让模型 Yan20（BERT）和 Ours（BERT）读入金标分词，之后再进行词性标签和依存句法预测。如此在两个模型分词都正确的情况下，进行依存句法分析结果的对比，并将结果列在表 5 中。由于两个模型分词都是正确的，我们采用更为常见的词级别依存句法分析指标 UAS 和 LAS，分别评测依存弧和依存关系类型预测结果的召回率，通常依存关系类型的预测是更难的任务，LAS 的指标可以指示模型在复杂任务上的性能。从上面表中的结果可以看出，在三个数据集上本文模型的 UAS 和 LAS 都超过了 Yan20（BERT），说明词性信息对依存句法分析精度的提升有直接帮助。

5 结语

本文首先在三个任务上提出了基于图的联

合模型，将汉语分词，词性标记和依存分析融合到统一的框架中，从而能够充分利用三个任务之间的共享知识并减少错误传播。同时本文词性标注任务的引入方式可以建模词性与分词以及词性与依存句法分析之间的关系，从而带来两个任务上精度的提升。本文在三个汉语数据集上进行实验验证和详细分析，发现引入词性信息的确有助于提升汉语分词和依存句法分析的精度。

参 考 文 献

[1] He Y, Qiu X P, Huang X. A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing[J]. Transactions of the Association for Computational Linguistics, 2020, 8(2):78-92.

[2] 宗成庆 , 夏睿 , 张家俊 . 文本数据挖掘 [M]. 北京 : 清华大学出版社 , 2020.

[3] Guo D Y, Tang D Y. Syntax-Enhanced Pre-trained Model[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021

[4] Tian Y H, Chen G M. Dependency-driven Relation Extraction with Attentive Graph Convolutional Networks[C]. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. 2021

[5] Roller S, KielA D, Nickel M. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora[C]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018.

[6] 李正华 . 汉语依存句法分析关键技术研究 [D]. 哈尔滨 : 哈尔滨工业大学 , 2013.

[7] 张宇 . 基于树形条件随机场的高阶句法分析 [D]. 苏州 : 苏州大学 , 2021.

[8] 郭振 , 张玉洁 , 苏晨 , 等 . 基于字符的中文分词、词性标注和依存句法分析联合模型 [J]. 中文信息学报 , 2014, 28(6):1-8.

[9] Hatori J, Matsuzaki T, Miyao Y, et al. Incremental joint approach to word segmentation, POS tagging, and dependency parsing in Chinese[C]. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. Association for Computational Linguistics, 2012.

[10] Zhang M S, Yu Z, Che W X, et al. Character-level Chinese dependency parsing[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 1326-1336.

[11] Kurita S, Kawahara D, Kurohashi S. Neural Joint Model for Transition-based Chinese Syntactic Analysis[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017.

[12] Shi T, Liang H, LEE L. Fast(er) Exact Decoding and Global Training for Transition-Based Dependency Parsing via a Minimal Feature Set[C]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

[13] Mcdonald R T, Crammer K, Pereira F. Online large-margin training of dependency parsers[C]. ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA. Association for Computational Linguistics. 2005.

[14] Mcdonald R T, Pereira F. Online Learning of Approximate Dependency Parsing Algorithms[C]. OFEACl. 2006.

[15] Carreras X. Experiments with a Higher-Order Projective Dependency Parser[C]. EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic. 2007.

doi:10.3772/j.issn.2095-915x.2022.03.005

[16] Zhou H, Li Z. Is POS Tagging Necessary or Even Helpful for Neural Dependency Parsing?[C]. Natural Language Processing and Chinese Computing. NLPCC. 2020.

[17] Eisner J. Bilexical Grammars and their Cubic-Time Parsing Algorithms[M]. Springer Netherlands, 2000.

[18] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. Arxiv preprint, 2018. https://arxiv.org/pdf/1810.04805.pdf

[20] Dozat T, Manning C D. Deep Biaffine Attention for Neural Dependency Parsing[J]. Arxiv preprint, 2016. https://arxiv.org/pdf/1611.01734.pdf

[21] Ji T, Wu Y, Lan M. Graph-based Dependency Parsing with Graph Neural Networks[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[22] 宗成庆 . 统计自然语言处理 [M]. 北京 : 清华大学出版社 , 2013:1-10.