文章编号：1003-0077（2011）00-0000-00

基于字符的中文分词、词性标注和依存句法分析联合模型*

郭振，张玉洁，苏晨，徐金安

（北京交通大学 计算机与信息技术学院，北京 100044）

摘要：目前，基于转移的中文分词、词性标注和依存句法分析联合模型存在两大问题：一是任务的融合方式有待改进；二是模型性能受限于全标注语料的规模。针对第一个问题，本文利用词语内部结构将基于词语的依存句法树扩展成了基于字符的依存句法树，采用转移策略，实现了基于字符的中文分词、词性标注和依存句法分析联合模型；依据序列标注的中文分词方法，将基于转移的中文分词处理方案重新设计为 4 种转移动作：Shift_S、Shift_B、Shift_M 和 Shift_E，同时能够将以往中文分词的研究成果融入联合模型。针对第二个问题，本文使用具有部分标注信息的语料，从中抽取字符串层面的 n-gram 特征和结构层面的依存子树特征融入联合模型，实现了半监督的中文分词、词性标注和依存句法分析联合模型。在宾州中文树库上的实验结果表明，本文的模型在中文分词、词性标注和依存分析任务上的 F1 值分别达到了 98.31%、94.84%和 81.71%，较单任务模型的结果分别提升了 0.92%、1.77%和 3.95%。其中，中文分词和词性标注在目前公布的研究结果中取得了最好成绩。

关键词：联合模型；中文分词和词性标注；依存句法分析；词语内部依存结构；半监督学习

中图分类号：TP391　　**文献标识码：**A

Character-level dependency model for joint word segmentation, POS tagging, and dependency parsing in Chinese

GUO Zhen, ZHANG Yujie, SU Chen, XU Jinan

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Recent work on joint word segmentation, POS tagging, and dependency parsing in Chinese has two key problems: the first is that word segmentation based on character and dependency parsing based on word were not combined well in the transition-based framework, and the second is that the joint model suffers from the insufficiency of annotated corpus. In order to resolve the first problem, we propose to transform the conventional word-based dependency tree into character-based dependency tree by using the internal structure of words and then propose a novel character-level joint model for the three tasks. For Chinese word segmentation, we design 4 transition actions: Shfit_S, Shift_B, Shift_M and Shift_E, through which the features used in previous research can also be integrated into the model. In order to resolve the second problem, we propose a novel semi-supervised joint model for exploiting n-gram feature and dependency subtree feature from partially-annotated corpus. Experimental results on the Chinese Treebank show that our joint model achieved 98.31%, 94.84% and 81.71% for Chinese word segmentation, POS tagging, and dependency parsing, respectively. Our model outperforms the pipeline model of the three tasks by 0.92%, 1.77% and 3.95%, respectively. Especially, the F1 value of word segmentation and POS tagging achieved the best result compared with those reported until now.

Key words: joint model； Chinese word segmentation and POS tagging； dependency parsing； word internal dependency structure；semi-supervised learning

---

* **收稿日期：**　　**定稿日期：**

基金项目：国家国际科技合作专项资助（2014DFA11350）；国家自然科学基金（61370130）；北京交通大学人才基金（KKRC11001532）

作者简介：郭振（1988—），男，硕士研究生，主要研究方向为依存句法分析和中文分词；张玉洁（1961—），女，教授，主要研究方向为自然语言处理和机器翻译；苏晨（1989—），男，硕士研究生，主要研究方向为机器翻译；徐金安(1970—)，男，副教授，研究方向为自然语言处理和机器翻译。

1 引言

中文分词、词性标注和句法分析是中文自然语言处理的三大基础任务，是一个中文句子被具体的自然语言处理应用（如机器翻译系统）使用之前，必须经过的处理步骤。以往的研究大都将中文分词、词性标注和句法分析看成独立的任务，任务的输入是人工标注的标准语料。但是单任务模型在实际应用中存在以下缺陷：

1）任务间的错误传递。例如在实际应用中，中文分词任务的输出直接作为词性标注任务的输入。此时，中文分词的错误会在词性标注任务中被放大，严重影响词性标注的精度。

2）多层次特征无法获取。例如某些词性标注歧义需要全局的句法信息才能得到消解，而传统的词性标注模型无法获取这些信息。

将多个任务融合到一个模型中同时处理的联合模型，是解决上述问题的一个有效方案。联合模型成为近年来研究的热点，许多有效的联合模型被提出来：中文分词与词性标注联合模型[1-2]；词性标注与依存句法分析联合模型[3-4]；中文分词、词性标注和基于词语的依存句法分析联合模型[5]；中文分词、词性标注和短语结构句法分析联合模型[6]。上述研究显示联合模型能使各任务的性能都得到不同程度的提高。

中文分词任务的输入是字符序列，而词性标注与句法结构分析的输入是词序列，解决好字符处理与词语处理之间的冲突是中文分词、词性标注和句法结构分析联合模型的关键。Hatori(2012)[5]假设词语内部字符之间有类似于句子中词语之间的依存关系，解码过程中每当一个词语构词成功后就假设词语内部之间的结构关系也建立完毕，从而在处理框架上统一了中文分词、词性标注和依存句法分析任务。但 Hatori(2012) [5]并没有真正利用词语内部之间的结构信息对联合模型进行改善。Zhang(2013) [6]认为构成词语的字符之间具有实际的语义结构，并对宾州中文树库 CTB5 的所有词语进行了结构标注，在此基础上实现了基于字符的中文分词、词性标注和短语结构句法分析联合模型。

联合模型的优点是可以同步处理多项任务，使各任务的中间结果可以相互利用，性能得到相互促进。然而用于联合模型的训练语料必须是在依存结构上经过人工标注的语料。而目前深加工的语料规模有限，难以满足训练高性能模型的需求。与此同时，大规模的生语料却相对容易获得，其中蕴含的知识也将有助于联合模型性能的提升。在以往的中文分词、词性标注和依存句法分析等单任务研究中，研究人员已经验证了利用生语料的半监督方法对各项任务的辅助作用[7-11]。如何从大规模生语料中抽取有价值的知识，融入到更复杂的联合模型中，是一个值得研究的新课题。

针对以上问题，本文做出了以下贡献：

- 将 Zhang(2013) [6]标注的词语内部结构转化为依存结构，将传统的基于词语的依存句法树扩展成了基于字符的依存句法树，在此基础上采用增量转移策略实现了真正意义上的基于字符的中文分词、词性标注和依存句法分析联合模型。
- 参考中文分词的序列标注思想，将中文分词的转移策略扩展为 4 种动作：Shfit_S、Shift_B、Shift_M 和 Shift_E。该扩展同时能够将以往中文分词研究中丰富而成熟的特征融入联合模型。
- 从大规模生语料中分别抽取了字符串层面的 n-gram 特征和结构层面的依存子树特征融入到联合模型中，首次实现了半监督的中文分词、词性标注和依存句法分析联合模型。
- 在 CTB5 的实验结果显示，本文的模型在中文分词、词性标注和依存分析任务上的 F1 值分别达到了 98.31%、94.84%和 81.71%，较单任务模型的分步处理结果分别提升了 0.92%、1.77%和 3.95%。其中，分词和词性标注在目前公布的结果中取得了最好成绩。

本文剩余部分组织结构如下：第二节介绍基于字符的中文分词、词性标注和依存句法分析联合模型；第三节介绍适用于联合模型的 n-gram 特征和依存子树特征的抽取和使用方法；第四节介绍评测实验；第五节对本文工作进行总结。

2 基于字符的中文分词、词性标注和依存句法分析联合模型

柱搜索和全局训练模型被应用于基于转移策略的自然语言处理框架，使得该框架在各项自然语言处理任务上取得了与其它经典模型同一水平的精度，并且保持了简单高效易于扩展的优势[12]。本文利用 Zhang(2013) [6]对 CTB5 的词语内部结构的标注信息，将基于词的依存句法树扩展成了基于字符的依存句法树。采用转移策略，实现了真正意义上的基于字符的中文分词、词性标注和依存句法分析联合模型。基于序列标注思想，重新设计了联合模型里中文分词部分的转移策略。该设计同时能够将以往中文分词研究中丰富而成熟的特征融入联合模型。模型用平均感知机算法进行全局训练，训练过程中采用参数提前更新策略[13]。训练和解码过程采用柱搜索算法实现。

2.1 基于字符的依存句法树

构成中文词语的汉字与构成英文单词的字母不同，单独的英文字母不能承载任何语义信息，而汉字属于表意文字，单独的汉字也承载了特定的语义信息。与词语通过相互影响产生语义修饰关系构成句子类似，构成词语的汉字之间也有特定的语义结构。例如“理发店”一词中，“理”和“发”通过动宾关系构成“理发”，“理发”作为定语修饰“店”构成词语“理发店”。这种汉字之间通过发生修饰关系构成词语的方式与词语构成句子的方式颇为类似。

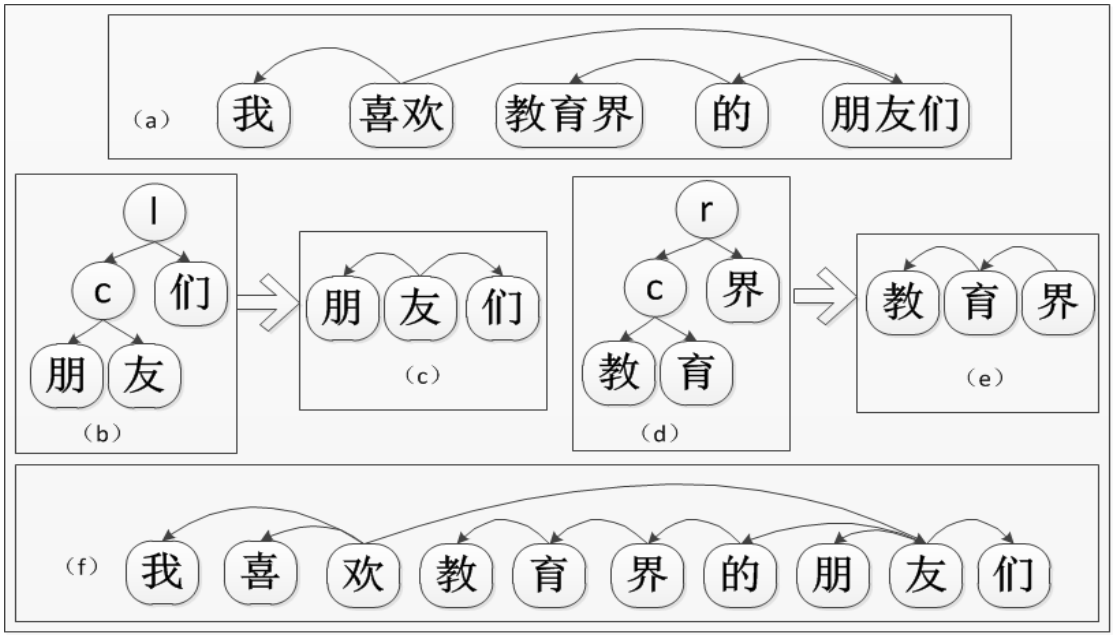

图 1　词语内部结构以及基于字符的依存句法树

Zhang(2013) [6]对宾州中文树库的所有词语进行了结构标注如图 1b 和图 1d 所示。“l”、“r”和“c”分别表示左边部分为右边部分的支配者，右边部分为左边部分的支配者以及左右两部分为并列关系。本文根据 Zhang(2013) [6]的标注信息将图 1b 和图 1d 的结构转换为图 1c 和图 1e 中的依存结构，转换时对于“c”我们选取右边部分为头结点。通过这种处理我们将图 1a 所示的基于词语的依存句法树转换成了图 1f 所示的基于字符的依存句法树。

基于字符的依存句法树比基于词语的依存句法树更适用于基于转移策略的中文分词、词性标注和依存句法分析联合模型。基于转移策略的联合模型将句子的分析过程分解为若干转移动作，为了减少搜索消耗，对于经历相同转移动作次数的候选结果，只保留模型评分较高的前 N 个结果。因此，模型要求经历相同转移动作次数的候选结果之间是可比的，即要求一个句子的所有候选结果从模型分析的初始状态到终止状态，恰好经历相同次数的转移动作。否则需要模型为此设计专门的中间结果对齐方案[5][14]。基于转移策略的联合模型将一个依存弧的建立设定为一个转移动作，而基于词的依存句法树中依存弧的个数随着分词结果的

变化而变化，使得经历相同转移动作次数的候选结果之间无法进行合理的竞争。基于字的依存句法树中依存弧的个数是固定的，即句子的字符数减 1，直接使上述难题得到解决。

另一方面，词语的内部结构信息有助于联合模型性能的提高。以中文分词为例，一个候选切分词语的内部结构越稳定合理，那么它真正成为词语的可能性越大。Zhang(2013) [6]的研究证明无论是根据特定规则强制构建的词语内部结构，还是人工标注的真实的词语内部结构，对于短语结构句法树的分析都有一定的辅助意义。Li(2011) [15]的研究表明，即使只使用词语的部分内部结构，也能提高中文依存句法分析的性能。

2.2 转移动作

基于转移策略的模型对输入句子从左到右进行处理，每次执行一个设定的转移动作，将句子从当前状态 T_i 转移到下一个状态 T_{i+1}。一个状态 T 包含一个栈 S={…S1,S0}和一个队列 Q={Q0,Q1…}，分别用来记录已经分析完成的部分结果，即依存子树，以及将要分析的字符。一个句子的初始状态 S 为空，Q 为句中所有字符；终止状态 S 为一棵完整的依存句法树，其中包含了中文分词与词性标注的结果，Q 为空。

为了将中文分词、词性标注和基于字符的依存句法分析三大任务融合到一个转移系统中，并且为了更加便捷的将以往单任务研究中成熟而丰富的特征加入到新的联合模型中，本文在前人研究的基础上重新设计了以下转移动作：

1） 中文分词和词性标注转移动作：

①SHIFT-B(t)：将队列 Q 的首元素作为非单字词的首字符移进栈顶，并给予词性 t。

②SHIFT-M：将队列 Q 的首元素作为非单字词的除首尾字符之外的字符移进栈顶。

③SHIFT-E：将队列 Q 的首元素作为非单字词的尾字符移进栈顶。

④SHIFT-S(t)：将队列 Q 的首元素作为单字词移进栈顶，并给予词性 t。

通过以上设计，基于转移策略的中文分词方法，统一到了将中文分词任务看做序列标注任务的处理框架下，同时使得以往基于序列标注思想的中文分词研究成果可以方便合理的融入到新的联合模型里。本文第 3 节提出的半监督的联合模型正是这一设计的有效利用，并且取得了显著效果。

2） 词语内部依存结构转移动作：

①REDUCE-SUBLEFT：栈 S 顶部的两个子树 S1 和 S0 出栈，建立依存关系 S1←S0（表示 S1 依存于 S0），将新形成的依存子树的根节点（即 S0）入栈。执行此动作的前提是子树 S1 和子树 S0 所包含的字符均属于同一个词语。

②REDUCE-SUBRIGHT：栈 S 顶部的两个子树 S1 和 S0 出栈，建立依存关系 S1→S0（表示 S0 依存于 S1），将新形成的依存子树的根节点（即 S1）入栈。执行此动作的前提是子树 S1 和子树 S0 所包含的字符均属于同一个词语。

词语内部依存关系的建立跟词语之间依存关系的建立类似，不同的是发生关系的元素类型不同，前者是字符后者是词语。

3） 词语之间依存结构转移动作：

①REDUCE-LEFT：栈 S 顶部的两个子树 S1 和 S0 出栈，建立依存关系 S1←S0，将新形成的依存子树的根节点（即 S0）入栈。执行此动作的前提是 S1 节点字符所属的词语和 S0 节点字符所属的词语是两个不同的词语，并且构词和词内依存结构分析均已完成。

②REDUCE-RIGHT：栈 S 顶部的两个子树 S1 和 S0 出栈，建立依存关系 S1→S0，将新形成的依存子树的根节点（即 S1）入栈。执行此动作的前提是 S1 节点字符所属的词语和 S0 节点字符所属的词语是两个不同的词语，并且构词和词内依存结构分析均已完成。

基于以上转移策略，一个字符数为 N 的句子，需要经过 2N-1 次状态转移即可完成从初始状态到终止状态的分析。

2.3 特征模板

本文使用的特征模板如表 1 所示。表 1 中的特征分为结构特征和序列特征两大类，分别表示依存子树的句法结构信息和中文分词与词性标注的上下文序列信息。句法结构信息包括基于词语的结构信息和基于字符的结构信息。

表 1 中的特征模板参考了 Hatori(2012)[5]的研究，本文对特征的使用阶段和使用方式进行了调整，以适用于 2.2 节所述的模型。P01-P20 主要抽取句法结构特征，在不同的转移动作中使用时，P01-P20 中的 w 会根据当时的环境选择代表一个完整的词或是一个词的一部分。W01-W20 是主要用来决定当前字符以什么方式参与词语的构成。T01-T05 被用来预测最新进入栈顶的词语的词性，只在 SHIFT-S(t)和 SHIFT-B(t)阶段使用。S01-S07 是本文新加入的基于字符的词语内部结构特征，与 P01-P20 一起辅助词语内部句法结构的分析。

3 半监督的中文分词、词性标注和依存句法分析联合模型

半监督的模型训练方法由于语料易得、方法简便高效等特点，广泛应用于各项自然语言处理任务。尤其在人工标注语料较少或专业领域资源匮乏的任务上，获得了显著效果。面对联合模型，半监督的方法遇到了新的机遇和挑战。对于中文分词、词性标注和依存句法分析联合模型，训练语料必须是经过人工中文分词标注、词性标注和基于字的依存句法结构标注的语料。而经过这样深层次人工标注的语料有限，难以满足训练高性能模型的需求。与此同时，不经过任何标注的完全生语料和只有部分标注信息的半生语料更容易获取，其中蕴含着丰富的信息可以用来提高联合模型的性能。由于联合模型中各任务的结果可以相互促进，使得生语料的加入可以同时促进多个任务性能的提升，这是单任务模型无法比拟的。但是如何将不同程度的生语料融入更加复杂的联合模型，是一个新的课题和挑战，需要专门研究。

表 1　基于字符的中文分词、词性标注与依存句法分析联合模型的特征模板

编号	特征	编号	特征
P01,P02,P03,P04	S0.w　S0.t　S1.w　S1.t	W01,W02	Q-1.w　Q-2.w-Q-1.w
P05,P06	S0.w-S0.t　S1.w-S1.t	W03	Q-1.w(Q-1 是单字词)
P07,P08	S0.w-S1.w　S0.t-S1.t	W04	Q-1.b-len(Q-1.w)
P09	S0.w-S0.t-S1.t	W05	Q-1.e-len(Q-1.w)
P10	S0.t-S1.w-S1.t	W06,W07	Q-1.e-C0　Q-1.b-Q-1.e
P11	S0.w-S1.w-S1.t	W08,W09	Q-1.w-C0　Q-2.e-Q-1.w
P12	S0.w-S0.t-S1.w	W10,W11	Q-1.b-C0　Q-2.e-Q-1.e
P13	S0.w-S0.t-S1.w-S1.t	W12	Q-2.w-len(Q-1.w)
P14	S1.t- S1.rc.t-S0.t	W13	Len(Q-2.w)-Q-1.w
P15	S1.t- S1.lc.t-S0.t	W14,W15	Q-1.w-Q-1.t　Q-2.t-Q-1.w
P16	S1.t- S1.rc.t-S0.w	W16	Q-1.t-Q-1.w-Q-2.e
P17	S1.t- S1.lc.t-S0.w	W17	Q-1.t-Q-1.w-C0
P18	S1.t-S0.t-S0.rc.t	W18,W19	Q-1.t-Q-1.e　Q-1.t-C0-C1
P19	S1.t-S0.w-S0.lc.t	W20	Q-1.t-Q-1.-C(C∈Q-1.w\e)
P20	S2.t- S1.t-S0.t		
T01,T02	Q-1.t-T0　Q-1.w-T0	C01,C02,C03,C04	S0.c　S0.ct　S1.c　S1.ct
T03,T04	Q-2.t-Q-1.t-T0　C0-T0	C05,C06	S0.lc.ct　S0.rc.ct
T05	Q-1.t-Q-1.e-C0-T0	C06,C07	S1.lc.ct　S1.rc.ct

注释：S0 和 S1 表示栈 S 顶部的第一个和第二个节点，Q-1 和 Q-2 表示相对于当前处理字符的前面第一个词和第二个词；w 表示词语，t 表示词性，c 表示字符；lc 和 rc 表示最左和最右子节点；C0 和 T0 表示当前处理字符及词性。P01-P13 用于所有转移动作，P14-P20 用于除“Shfit_M”和“Shfit_E”的所有转移动作；W01-W20 用于所有与分词相关的转移动作；T01-T05 用于“Shfit_S”和“Shfit_B”；C01-C07 用于与词内依存结构建立相关的转移动作。

本文从大规模生语料中抽取具有代表性的 n-gram 字符串特征和依存子树结构特征，研究生语料特征在联合模型中的使用方法，首次实现了基于字符的半监督中文分词、词性标注和依存句法分析联合模型，取得了显著的实验效果。图 2 为半监督联合模型的框架。

3.1 序列特征：n-gram 特征

本小节介绍用于联合模型的 n-gram 特征的抽取与使用。本文将 Wang(2011)[7]对完全生语料的处理方案移植到具有分词标注的语料上来，并根据抽取的信息为联合模型产生新的特征。

对于给定分词结果的一个句子 $S=C_0C_1\ldots C_n$，首先根据字符在词语中的位置对其进行标注，可以得到对应的标注序列 $T=T_0T_1\ldots T_n$。本文采用传统的四词位标注集[16]。然后，从句子中抽取 C_i、C_iC_{i+1}、$C_{i-1}C_iC_{i+1}$ 等不同长度的 n-gram 字符串，用 g 表示。对于一个特定的 g 如 C_iC_{i+1}，抽取与其相关的不同长度的标注串如 T_i、T_iT_{i+1}、T_{i+1} 等，用 seg 表示。这样就得到了一系列不同的（g，seg）。接下来，统计每种（g，seg）在语料中的出现频度 f（g，seg），就得到了一系列{g，seg，f（g，seg）}。然后，通过下面的方法将（g，seg）映射到不同的标签：出现频度前 10%的（g，seg），标签为 H；出现频度前 10%~30%的（g，seg），标签为 M；出现频度小于 30%的（g，seg），标签为 L。在此之前，先将出现频度小于 3 的（g，seg）去掉，本文认为这些（g，seg）的出现频度过低，不具有有效的统计意义。这样，就获得了一个{g，seg，lable}列表，新的 n-gram 特征就是基于这份列表产生的。

当联合模型执行与中文分词有关的转移动作时，即某个字符将要被移进分析状态的栈顶时，抽取与该字符相关的 n-gram 字符串 g，如果 g 存在于上文从生语料中抽取的列表中，就将对应的 seg 与 lable 连接起来，形成新的特征加入到模型中，即 n-gram 特征。沿用 Wang(2011)的做法，本文只采用 bi-gram 特征，在信息抽取和特征生成阶段，只关注 bi-gram C_iC_{i+1} 的相关信息，其中 C_i 为当前要处理的字符。

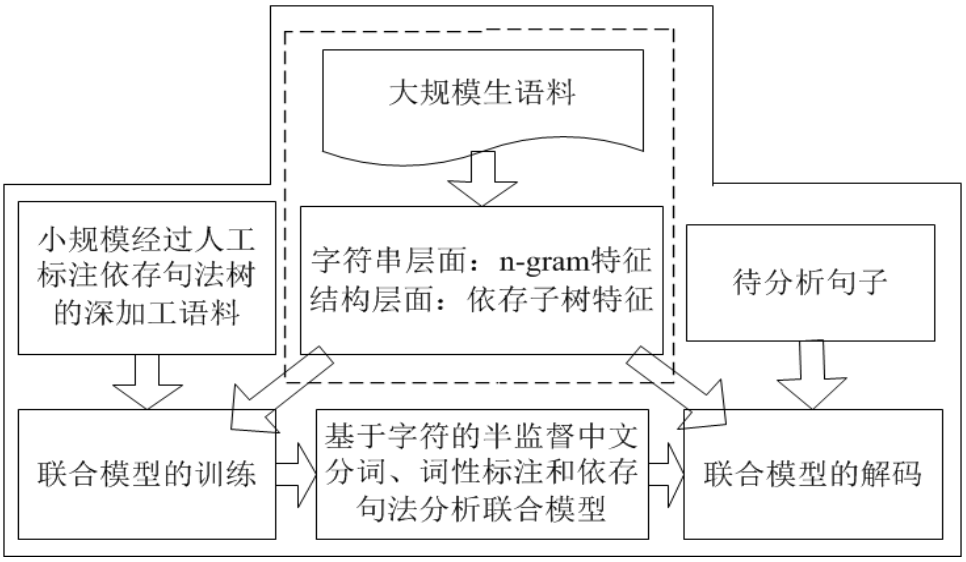

图 2　半监督的基于字符的中文分词、词性标注和依存句法分析联合模型框架

3.2 结构特征：依存子树特征

依存子树特征是指，从经过自动依存句法分析的生语料中抽取特定的依存子树，生成适用于有监督训练模型的特征。具有 2 个节点的依存子树和具有 3 个节点的依存子树使用最为广泛[10]。本文从生语料中抽取具有 2 个节点的依存子树生成用于联合模型的依存子树特征。

首先对大规模生语料进行自动依存句法分析。本文采用被广泛使用的单任务依存句法分析器，基于图模型的开源依存句法分析工具 MSTParser[1]。为了节约大规模生语料的预处理

[1] http://mstparser.sourceforge.net

时间，MSTParser 的训练和解码采用一阶模型，模型的训练语料与联合模型的训练语料相同。然后，从经过处理的大规模生语料中抽取具有 2 个节点的依存子树，即词语依存对：W1-W2-R/L。其中，词语 W1 和 W2 的顺序与它们在原句子中的顺序保持一致，R 和 L 分别表示右依存弧和左依存弧，即 W2 依存于 W1 和 W1 依存于 W2。统计所有依存子树 W1-W2-R/L 出现的频度 f（W1-W2-R/L），得到一系列{ W1-W2-R/L ，f（W1-W2-R/L）}。接下来，采用与 3.1 节类似的方法将（W1-W2-R/L）映射到不同的标签：出现频度前 10%的 W1-W2-R/L，标签为 H；出现频度前 10%~30%的 W1-W2-R/L，标签为 M；出现频度小于 30%的 W1-W2-R/L，标签为 L。在此之前，需要先将出现频度小于 3 的 W1-W2-R/L 去掉，理由如 3.1 所述。

当联合模型对当前状态的前两个栈顶元素进行依存关系决策时，为其生成两种依存子树 W1-W2-R 和 W1-W2-L，通过查询上文获得的依存子树信息表，获得相应的频度标签，将依存子树的依存弧方向和频度标签连接起来形成新的特征加入到联合模型中，例如“R-H”和“L-M”。

4 评测实验与结果分析

4.1 实验数据

标注语料采用宾州中文树库 CTB5，语料划分方案为：训练集 1-270 篇、400-931 篇和 1001-1151 篇；开发集 301-325 篇；测试集 271-300 篇[8]。训练集用于联合模型训练，开发集用于调参，测试集用于评测。用 PennMalt[2]将短语结构树转换为依存结构树。使用经过分词标注的人民日报（1998 年上半年）的数据作为具有部分标注信息的语料[3]，用于 n-gram 特征和依存子树特征的抽取。用基于条件随机场的词性标注模型对其进行词性标注，用基于图的依存句法分析模型对其进行依存句法分析。

对中文分词、词性标注和依存句法分析均采用准确率、召回率、综合性能指标 F1 值进行评测。对于依存句法分析，只有当具有依存关系的两个词语均被系统召回，并且依存弧的方向正确时，这个依存关系才被作为正确结果。遵循惯例，评测时与标点符号相关的依存关系不予考虑。

4.2 对比模型

根据第 2 节和第 3 节提出的方案，我们实现了一个基于字符的中文分词、词性标注和依存句法联合模型和三个半监督的联合模型。为了与单任务模型和部分任务的联合模型对比，实现了两套系统，并把它们作为 Baseline。细节如下：

- **SegTagDep：**本文提出的基于字符的中文分词、词性标注和依存句法分析联合模型。
- **SegTagDep+2-gram：**在 SegTagDep 中加入 3.1 节描述的 2-gram 特征。
- **SegTagDep+ subtree：**在 SegTagDep 中加入 3.2 节描述的依存子树特征。
- **SegTagDep+2-gram+subtree：**在 SegTagDep 中同时加入 3.1 节和 3.节 2 中描述的 2-gram 特征和依存子树特征。
- **CRF+MSTP：**基于条件随机场的中文分词和词性标注系统，使用开源工具 CRF++[4]，特征模板与 Wang（2011）[8]一致。基于图模型的中文依存句法分析[17]，采用开源工具 MSTParser，训练和解码采用二阶模型。
- **SegTag+MSTP：**采用 Zhang（2010）[1]的中文分词和词性标注联合模型。去除特殊规则处理，以保持与本文提出的联合模型一致。依存句法分析依然采用上面的 MSTParser。

4.3 实验结果

表 3 列出了上面六个系统的评测结果 F_1 值。表中第一行为本文提出的联合模型的性能。可以看出，联合模型在各项任务上均取得了好于基线系统的结果，其中在词性标注和依存分

2 http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html
3 http://www.icl.pku.edu.
4 http://crfpp.sourceforge.net/

析任务上的 F1 值，比仅在分词与词性标注任务上联合的基线系统提升了 0.68%和 1.66%。我们推断，词性标注性能的提升得益于句法特征的加入，使仅靠字符串特征无法消解的词性歧义得到解决；同时依存句法分析的性能，也随着更多正确词性被召回获得了提升。值得注意的是，联合模型的分词结果与最好的基线系统持平，我们分析主要因素是，目前系统的分词精度已经很高，分词性能的提高受限于标注语料的知识规模。我们随后提出的半监督模型通过使用更多的非全标注语料验证了这一点。

表 3　各模型在中文分词、词性标注和依存句法分析任务上的性能

模型	中文分词	词性标注	依存句法分析
SegTagDep	97.52	93.93	79.55
SegTagDep+2-gram	98.38	94.63	80.78
SegTagDep+subtree	97.74	94.25	80.40
SegTagDep+2-gram+subtree	**98.31**	**94.84**	**81.71**
CRF+MSTP	97.39	93.07	77.76
SegTag+MSTP	97.51	93.25	77.89
Hatori（2012）	98.26	94.64	--

下面观察生语料特征的加入对联合模型的影响。从表 3 的结果可以看出，从生语料中抽取的字符串层面的 2-gram 特征和结构层面的依存子树特征，都使联合模型在各项任务上的性能获得了不同程度的提高。2-gram 特征的加入使联合模型在中文分词、词性标注和依存句法分析的 F1 值分别提高了 0.86%、0.7%和 1.23%。依存子树特征的加入使联合模型在中文分词、词性标注和依存句法分析的 F1 值分别获得了 0.22%、0.32%和 0.85%的提升。由此我们认识到，由于联合模型中各任务的中间结果以特征形式及时反馈给其它任务，使得一个任务性能的提高会促进其它任务性能的提高，这使得来自生语料的特征信息在联合模型中获得的增益比在单任务模型中获得的增益更大。同时使用 2-gram 特征和依存子树的联合模型取得了在各任务上最好的性能：中文分词达到 98.31%，词性标注达到 94.84%，依存句法分析达到 81.711%，使联合模型在各项任务上分别获得了 0.79%、0.91%和 2.16%的性能提升，较单任务模型的分步处理结果分别提升了 0.92%、1.77%和 3.95%。

目前有关联合模型的研究报告来自 Hatori（2012）[5]，为了对比，我们将其评测结果列于表 3 最后一行。Hatori（2012）[5]的联合模型中加入了丰富的外部词典特征。可以看出，我们的模型在中文分词和词性标注任务上的性能优于 Hatori（2012）[5]的性能。由于 Hatori（2012）[5]并没有给出在本数据集上的依存句法测试结果，所以无法与其直接进行比较。

表 4　与以往中文分词与词性标注经典研究报告结果的比较

模型	中文分词	词性标注
Kruengkrai09	97.87	93.67
Zhang10	97.78	93.67
Sun11	98.17	94.02
Wang11	98.11	94.18
Hatori12	98.26	94.64
Zhang13	97.84	94.80
Our	98.31	94.84

表 4 列出了前人在中文分词与词性标注研究上获得的经典结果和本文获得的最好结果。“Kruengkrai09”是 Kruengkrai（2009）[2]实现的错误驱动模型；“Zhang10”是 Zhang（2010）[1]采用转移策略实现的中文分词与词性标注联合模型，并在训练与解码阶段对英文字符和阿拉伯数字采用了特殊的规则处理；“Sun11”是 Sun（2011）[18]的融合多个不同层次模型的处理方法，并且使用了词典信息；“Wang11”是 Wang（2011）[8]基于 CRF 实现的加入大规模

生语料的半监督模型；“Hatori12”是 Hatori（2012）[5]实现的中文分词、词性标注与基于词语的依存句法分析联合模型，并且加入了丰富的外部词典特征；“Zhang13”是 Zhang（2013）[6]实现的基于字符的中文分词、词性标注与短语结构句法分析联合模型。可以看出，本文的模型在中文分词和词性标注上取得了最佳的结果，显示了更大优势。

在本文投稿之后 ACL 发表了同样架构下的中文分词、词性标注和依存句法分析联合模型，在各项任务上的精度为 97.84%、94.33%和 82.14%[19]。本文的模型在依存句法分析的精度上略低于此文，我们将提高本文联合模型的精度作为今后工作的重点。

5 总结

本文利用词语内部结构信息，将基于词语的依存句法树扩展成了基于字符的依存句法树，采用转移策略提出并实现了真正意义上的基于字符的中文分词、词性标注和依存句法分析联合模型。在中文分词与词性标注部分，将序列标注思想与转移策略相结合，设计了 4 词位状态转移方案，使得以往中文分词的研究成果可以便捷的移植到联合模型中来。从大规模生语料中抽取字符串层次的 2-gram 特征和结构层次的依存子树特征，融入到新的联合模型中，首次实现了基于字符的半监督中文分词、词性标注和依存句法分析联合模型。实验结果显示，半监督的联合模型在各项任务上的性能均优于单任务模型和不同程度的联合模型，在中文分词、词性标注和依存句法分析方面 F1 值分别达到了 98.31%、94.84%和 81.71%，较单任务模型的分步处理结果分别提升了 0.92%、1.77%和 3.95%。作为今后的工作，我们一方面要进一步提高本文联合模型在中文依存句法分析任务上的精度，另一方面要优化模型实现方案，提高速度。

致谢 本文工作得到了中国科学院智能信息处理重点实验室，中科院计算所，北京，100190 的部分资助。

参考文献

[1] Zhang Y, Clark S. A fast decoder for joint word segmentation and POS-tagging using a single discriminative model[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 843-852.

[2] Kruengkrai C, Uchimoto K, Kazama J, et al. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging[C]//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: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 513-521.

[3] Hatori J, Matsuzaki T, Miyao Y, et al. Incremental Joint POS Tagging and Dependency Parsing in Chinese[C]//IJCNLP. 2011: 1216-1224.

[4] Li Z, Zhang M, Che W, et al. Joint models for Chinese POS tagging and dependency parsing[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1180-1191.

[5] 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: 1045-1053.

[6] Zhang M, Zhang Y, Che W, et al. Chinese parsing exploiting characters[C]//51st Annual Meeting of the Association for Computational Linguistics. 2013.

[7] Guo Z, Zhang Y, Su C, et al. Exploration of N-gram Features for the Domain Adaptation of Chinese Word

Segmentation[M]//Natural Language Processing and Chinese Computing. Springer Berlin Heidelberg, 2012: 121-131.

[8] Wang Y, Jun'ichi Kazama Y T, Tsuruoka Y, et al. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data[C]//IJCNLP. 2011: 309-317.

[9] Koo T, Carreras X, Collins M. Simple semi-supervised dependency parsing[J]. 2008.

[10] Chen W, Kazama J, Uchimoto K, et al. Improving dependency parsing with subtrees from auto-parsed data[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 570-579.

[11] Chen W, Kazama J, Torisawa K. Bitext dependency parsing with bilingual subtree constraints[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 21-29.

[12] Zhang Y, Nivre J. Analyzing the Effect of Global Learning and Beam-Search on Transition-Based Dependency Parsing[C]//COLING (Posters). 2012: 1391-1400.

[13] Collins M, Roark B. Incremental parsing with the perceptron algorithm[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004: 111.

[14] Zhu M, Zhang Y, Chen W, et al. Fast and Accurate Shift-Reduce Constituent Parsing[J].

[15] Li Z, Zhou G. Unified dependency parsing of Chinese morphological and syntactic structures[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1445-1454.

[16] Zhao H, Huang C N, Li M, et al. Effective tag set selection in Chinese word segmentation via conditional random field modeling[C]//Proceedings of PACLIC. 2006, 20: 87-94.

[17] McDonald R, Crammer K, Pereira F. Online large-margin training of dependency parsers[C]//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005: 91-98.

[18] Sun W. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1385-1394.

[19] Zhang M, Zhang Y, Che W, et al. Character-Level Chinese Dependency Parsing[C].//Association for Computational Linguistics. 2014.