软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(12):3151−3161 [doi: 10.13328/j.cnki.jos.004816]

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

基于树形条件随机场的跨语言时态标注*

陈怡疆[1], 徐海波[1], 史晓东[2], 苏 畅[2]

[1](厦门大学 计算机科学系,福建 厦门 361005)

[2](厦门大学 智能科学与技术系,福建 厦门 361005)

通讯作者: 陈怡疆, E-mail: cyj@xmu.edu.cn, http://www.cs.xmu.edu.cn

摘 要: 提出时态树的概念和构造方法,从而将汉英时态转换问题转换为时态树标注的问题.而后,使用树形条件随机场为未标注时态树的结点标注英语时态.提出的特征函数的模板较好地满足了模型推断的需要.实验结果表明:与基于线性条件随机场模型的时态标注方法相比,基于时态树方法的准确率有大幅度的提高,说明使用时态树能够更好地表达子句间时态的依赖关系.

关键词: 汉英时态转换;时态树;未标注时态树;已标注时态树;树形条件随机场

中图法分类号: TP391

中文引用格式: 陈怡疆,徐海波,史晓东,苏畅.基于树形条件随机场的跨语言时态标注.软件学报,2015,26(12):3151−3161. http://www.jos.org.cn/1000-9825/4816.htm

英文引用格式: Chen YJ, Xu HB, Shi XD, Su C. Cross-Lingual tense tagging based on tree conditional random fields. Ruan Jian Xue Bao/Journal of Software, 2015,26(12):3151−3161 (in Chinese). http://www.jos.org.cn/1000-9825/4816.htm

Cross-Lingual Tense Tagging Based on Tree Conditional Random Fields

CHEN Yi-Jiang[1], XU Hai-Bo[1], SHI Xiao-Dong[2], SU Chang[2]

[1](Department of Computer Science, Xiamen University, Xiamen 361005, China)

[2](Department of Cognitive Science, Xiamen University, Xiamen 361005, China)

Abstract: In this paper, the concept of a tense tree and its construction method are proposed such that the issue of Chinese-English tense conversion is transformed into the issue of tagging a tense tree. Futher, tree-CRF is used to tag nodes of the untagged tense tree with English tenses. Templates of feature functions are suitable for the need of model inference. Experimental results show that the method of tree-based tense tagging is much better than linear-based tense tagging, and that tense trees can better express tense dependencies between clauses.

Key words: Chinese-English tense conversion; tense tree; untagged tense tree; tagged tense tree; tree-CRF

汉英机器翻译系统需要处理时态的问题[1],否则会影响翻译结果的准确度和流利度.例如,把“他在看书”翻译为过去进行时或现在进行时,句子表达的含义是不一样的.Gong 等人[2]使用 2005 NIST MT 数据检验出时态翻译对汉英机器翻译的结果有较大的影响.

对于英文动词的时态,“时”包括“现在”、“过去”、“将来”和“过去将来”,“态”包括“一般”、“完成”、“进行”和“完成进行”,其组合共有 16 种时态.中文句子没有显式的时态,动词都是原形,但有与时态相关的词,如“曾经、通常、将要”等副词和“了、着、过”等助词.龚千炎等人提出了具有汉语特点的一些时态种类[3].汉英时态翻译的核心部分是从汉语时态到英语时态的转换.

* 基金项目: 国家自然科学基金(61075058); 国家科技支撑计划(2012BAH14F03)

Foundation item: National Natural Science Foundation of China (61075058); National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2012BAH14F03)

收稿时间: 2014-03-23; 修改时间: 2014-08-27, 2014-12-01; 定稿时间: 2015-01-08

1 相关工作

人们提出了时态转换的多种解决方案,最常用的策略是采用手工规则的方法[4−6].例如[4]:if 有时间词为{“昨天”,“上周”,…} then 句子的“时”(时制,tense)标记为{过去时};if 句式为“已经”+V+…+“了” or V+补语且补语为{“完”,“好”,“掉”,“成”,…} then 句子的“态”(时体[6],aspect)为{实现态}.

马红妹[5]提出另一种手工规则的方法:

a) 将汉语的时制划分为 10 种,时体划分为 6 种,共有 54 种有效的组合,并把每种组合映射为英语的 16 种时态中的一种;

b) 在确定动词时态的时候,首先对语篇进行分析获取时间信息,用规则的方法确定汉语动词的时制和时体,然后,根据方法 a)中事先确定好的映射来确定该动词的英语时态.

由于自然语言错综复杂,基于手工规则(包括模板和映射)的方法往往过于粗糙和武断,规则的覆盖面和冲突也是问题.另外一种处理时态的策略是采用基于统计的方法[2,7−10].

Ye[7]将一个句子中的动词按照动词在句子中出现的先后顺序构造成动词向量$\langle v_1,v_2,\ldots,v_n\rangle$,每个动词分量 v_i 附上一些语法和语义特征,而后把确定每个动词时态的问题转化为序列标注问题,具体采用线性条件随机场的分类器.该方法的优点是通过统计机器学习的方法,覆盖复杂的语言现象,同时从整体上考虑各个动词时态的标注;缺点是没有准确抓住句子中主句和子句之间的时态的相互依赖关系;此外,没有动词的句子要另外处理等.

Murata[8,9]使用多种机器学习方法(*K* 临近算法、决策表、最大熵、支持向量机)来处理日英时态翻译,发现支持向量机的效果最好.

Gong 等人[2]提出基于 *N* 元模型的时态翻译模型,并与基于短语的统计机器翻译系统合成在一起.在该模型中,称一个英文句子中的某个动词的时态为句内时态(intra tense),称一个英文句子的主句中的动词的时态为句间时态(inter tense).通过构造句内时态序列语料库和句间时态序列语料库来训练各自的 *N* 元模型.而后,用这两个 *N* 元模型来修正翻译结果的概率值,*BLEU* 值从 28.30 提升到 28.92.但是,该模型没有利用到中文句子的时间词、时间副词和时间助词的信息,仅仅使用时态的 *N* 元模型来猜测可能的时态序列.

Gong 等人[10]进一步提出了基于分类的时态翻译模型,与基于短语的统计机器翻译系统合成在一起.首先,使用中文句子的单词和词性、时间词、历史时态信息和语篇的分类这些特征构造出 SVM 分类器来识别原文(中文)句子的英文时态 T_s,并使用译文(英文)的单词和词性特征构造出 SVM 分类器来识别译文(英文)句子的时态 T_g;其次,在翻译过程中,对输入的一个中文句子 *f*,使用特征函数 $F_1=(T_s==T_g)?1:0$ 和 $F_2=P(T_s|f)$来修正翻译的概率值.实验结果表明,*BLEU* 值从 28.30 提升到 29.27(提升了 0.97).该方法的不足在于:以句子的时态作为前提是不合适的.例如:句子“我们知道他吃了苹果.We know he ate the apple”中有两个英文时态,无论取哪个时态作为句子的时态,都会对翻译结果产生不良影响.

在本文中,我们提出了基于时态树[11]的英语时态标注算法.本文第 2 节分析影响和决定句子时态的因素,并提出时态树的概念及其构造算法.第 3 节提出使用树形条件随机场对时态树进行标注的算法.第 4 节对实验结果进行分析和讨论.第 5 节进行总结.

2 基于时态树的时态处理算法

2.1 影响句子时态的因素

首先考察影响动词时态的一些主要因素:

(1) 时态副词和助词

例如:已经、即将、曾经、VV+过、VV+着、VV+了,…

(2) 时间短语

例如:公元 1999 年、昨天、宣德 18 年、在 19 日之前、从 18 日到 21 日

(3) 动词的时相

动词的词义也内含时间信息,例如动词“击中”,该动词的词义表明该动作是一瞬间的、已经完成的.动词的时相(phase)体现动词词义的内在的时间特征,是确定动词时态的一个重要因素[3,12,13].例如:“导弹击中目标”,不大可能是将来时(特别在缺乏时间词的情况下).

(4) 主句与从句间的相互约束

复合句是由多个子句构成的,主句和从句之间存在着时态的相对约束.例如:当主句的时态是过去式时,宾语从句往往也是过去式的.再如,句子“当你来的时候,我把它给你”(when you come {came}, I will {would} give it to you.),虽然“来”是将来(或过去将来)发生的事,但是在这个从句中,“来”的时制必须是现在时(或过去时),不能是将来时;同时,主句中的“给”的时态也受到从句的时态约束,不可能是过去完成时等.

(5) 篇章内句子间的相互约束

例如:当语篇是在讲述过去的一件事时,整篇文章的句子之间相互约束,即,大多数是过去式.某个句子从表面上看没有表达过去的时间词,实际上是因为该句子的时间词在上几个句中已被提及而被省略.

(6) 语篇撰写的时间

例如:15 日我看书.如果撰写的时间是 12 日,该句子的时制是将来时;如果撰写的时间是 18 日,该句子的时制是过去时.

时态的确定是很困难的,远距离依赖,与句子的语义相关,时间词的省略,需要常识和逻辑推理的特点非常突出.

2.2 时态树及其构造算法

在已有的时态处理方法中,许多方法没有充分利用语法树的信息.我们认为,正确处理句子时态的一个前提必须是已获得该句子的语法树.在此前提下,我们提出了时态树[11]的概念,并提出了基于时态树的单句时态处理算法.

未标注时态树(untagged tense tree,简称 UTT):从语法树转化而来的压扁的、删减的树状结构,同时富含与时态相关的语法和语义信息.

已标注时态树(tagged tense tree,简称 TTT):在未标注时态树中,对子句 IP,CP 和动词(VA,VC,VE 和 VV)结点标注 16 种英语时态中的一种,其余结点标注为 empty,而构造出的新的树状结构.

未标注时态树和已标注时态树统称为时态树.对于任意一棵已标注时态树 TTT,存在唯一的未标注时态树 UTT 与之对应.

例如:句子“现在,热战和冷战虽已成为过去,但动荡和冲突仍看不到尽头”,图 1 为该句子的未标注时态树,图 2 为该句子的已标注时态树.

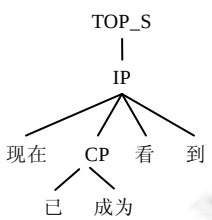


Fig.1 Untagged tense tree

图 1 未标注时态树

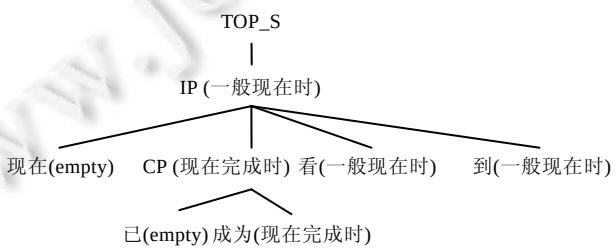


Fig.2 Tagged tense tree

图 2 已标注时态树

我们认为,子句(IP 和 CP)也可以有时态,该时态一般是该子句的主动词的时态.该方法一个意外的优点是:可以处理没有动词的句子时态.例如:“今天星期天”和“小明这个调皮蛋.”就没有动词.按已有的方法[7]仅对动词标注时态,则此时是空的动词序列,没有地方可以标注时态,但我们可以在 IP 结点上标注时态.

我们定义主动词为:在一个子句 IP(或 CP)中,其直接覆盖下的第 1 个动词被称为该 IP(或 CP)的主动词.例

如:“他去看电影”,“去”是主动词.又如:“你可以看电视”,情态动词“可以”是主动词.在图 2 中,“成为”是 CP 的主动词,但不是 IP 的主动词,因为“成为”被 CP 直接覆盖,而不被 IP 直接覆盖.

这样,我们就把时态处理转换成树 UTT 到树 TTT 的标注问题.

未标注时态树 UTT 的构造算法如下:

(1) 遍历中文句子的句法树,保留句法树中所有的 IP 和 CP(如果多个 CP 在纵向上形成父子关系链,则只保留最顶层的 CP),并生成对应的 IP 或 CP 结点,从而构造出未标注时态树 UTT 的骨架结构.此外,识别 IP 和 CP 是否在双引号中.此外,为未标注时态树 UTT 添加 TOP_S 结点作为根结点.

(2) 在构造 UTT 的骨架结构过程中,对时态树中的每个 IP 和 CP,扫描其直接覆盖的结点序列,并进行如下处理:

A. 识别动词,生成一个动词结点,并确定动词结点的属性.此外,为动词添加一个叶子结点 NULL_LEAF.此时,动词结点成为时态树的内部结点;

B. (用手工规则)识别结点序列中的时间短语,有可能的话,与语篇的写作时间作比较(过去、现在、将来),并生成一个时间短语结点作为所对应的 IP 或 CP 的叶子结点.作为宾语(的一部分)的时间词不加入到 UTT 中;

C. 识别与时间有关的副词,如“已经”,并生成一个时间副词结点表示该副词;

D. 识别与时间有关的助词,如“VV+了”,生成一个助词结点;

E. 识别逗号和分号,根据具体情况在必要时生成对应的分隔词结点;

(3) 为每个内部结点(TOP_S,IP,CP,动词结点)添加两个 STOP 结点,也就是说,该内部结点的第 1 个和最后一个孩子结点是 STOP 结点.这个步骤可以与上面两步同时进行.STOP 结点用来表示产生式的右手边的开始和结束.

在图 1 中,时间词“现在”对应一个时间短语结点,“看”、“到”、“成为”各自对应一个动词结点,“已”对应一个副词结点.我们不生成时间词“过去”所对应的时间短语结点,因为“过去”是宾语.时间短语作为宾语(的一部分)一般情况下不会影响时态,例如,“现在我们正思考过去.”判断时间短语是否是宾语(的一部分)的规则是:在语法树中,从该时间短语向上走,在碰到 IP 或 CP 或 TOP_S 的结点之前,如果碰到 VP 结点,那么认为该时间短语是宾语(的一部分).“但”、“仍”、“不”虽然是副词,但不是时间副词.时间副词是一个封闭的很小的有限词集,陆俭明[14]总结出了绝大部分的时间副词.“热战”、“和”、“冷战”、“虽”、“动荡”、“和”、“冲突”、“尽头”均不属于上述构造算法中的结点类型,所以不在 UTT 中产生对应的结点.

在未标注时态树 UTT 中,生成的各种结点都有各自的属性,具体见表 1.

Table 1 Attributes of nodes in untagged tense trees

表 1 未标注时态树中结点的属性

动词结点的属性	
动词单词本身	可以从语法树中获得
动词词性	可以从语法树中获得,4 种词性:VA|VC|VE|VV
是否是情态动词	情态动词集是一个封闭的很小的词集,为{可,可以,能,能够,该,应该,必须,…}
动词时相	可以从汉语词典中抽取出动词表,并为每个动词标注动词时相, 而后,只要查询该动词时相表就可以获得动词时相
是否在双引号中	可以采用规则的方法在语法树中进行判断
IP,CP 结点的属性	
结点的时间短语类型	如果 IP 或 CP 结点的孩子结点中有时间短语结点,则类型为“过去时”、“现在时”、 “将来时”、“不确定”中的一种,否则类型为“没有时间短语”
结点本身的标号(label)	为 IP 或 CP
结点在中文语法树中的父结点的标号(label)	
是否在双引号中	

Table 1 Attributes of nodes in untagged tense trees (continued)

表 1 未标注时态树中结点的属性(续)

时间短语结点的属性	
时间短语所表达的时制	具体类别为:过去时、现在时、将来时、不确定.例如,“昨天”的时制是过去时.
时间短语的时点或时段类型 事件发生时间与参照 时间的先后的类型	例如:“1998 年”是时点类型,“从 1998 年到 2000 年”是时段类型, 即:把事件发生时间与参照时间作比较,来确定事件发生时间与参照时间的先后
时间副词结点的属性	
时间副词本身	
时间副词的分类	时间副词的分类采用陆俭明[14]的标准
时间助词结点的属性	
助词本身	具体的内容有:了 1、了 2、着、过、起来、下去.如果“了”在动词之后, 则为“了 1”;如果“了”在名词之后,则为“了 2”
标点符号结点的属性	
标点符号本身	具体的内容有:逗号或分号
TOP_S 结点,STOP 结点和 NULL_LEAF 结点的属性	
无	除了结点本身的类型外,没有其他属性

我们采用手工规则的方法来识别时间短语,并根据时间短语的含义来确定时间短语结点的属性值,如:

node_rel_year→大前年|前年|去年|今年|明年|后年,

RelYear→RelCentury, node_abs_year|node_rel_year.

我们把规则中的非终结符看成是函数,自顶向下地进行调用.例如,RelYear 函数如果成功地调用 RelCentury 函数和 node_abs_year 函数,就返回 true 和处理的结果.

3 使用树形条件随机场处理时态树的标注

3.1 树形条件随机场

条件随机场不仅可以用在序列的标注上(例如对单词序列进行词性标注),也可以用在树(甚至图)的标注上[16,17].在实现过程中,我们具体采用树形条件随机场[15]来标注时态树.该过程可以假想为:对一棵未标注时态树 x,条件随机场枚举出所有可能的标注 y,从而构成许许多多的已标注时态树 $x[y]$,并计算每棵已标注时态树的条件概率 $P(y/x)$,然后取概率最大的树标注 y.在具体实现时,树形条件随机场采用了动态规划算法,所以并没有产生所有可能的标注.

在树形无向图 G 中,最大团 c 为一个结点和它的父结点构成的结点对(默认包含它们之间的边,即纵向边,下同).在 G 中,任意两个兄弟结点不在同一个最大团中,它们是条件独立的.但是实际问题中,它们往往是相互影响且不得不考虑的,如在语法树中,名词结点的前结点很可能是形容词结点而不可能是副词结点.所以我们假设,相邻的两个兄弟结点之间有边相连,相邻两个兄弟结点和它们的父结点组成一个最大团 c,最大团的重新定义主要体现在特征函数的定义和选择上.E^{ps} 表示父结点和子结点之间的边集合;E^{ss} 表示两个相邻兄弟结点之间的隐形边集合;c 表示当任意结点以及其右相邻兄弟结点,父结点的组成的最大团.重新定义势函数形式,得到:

$$p(y\mid x)=\frac{1}{Z(x)}\exp\left(\sum_{e\in\{E^{ss},E^{ps}\}}\sum_{k}u_k f_k(e,y_e,x)+\sum_{i}\sum_{k}s_k g_k(y_i,x)+\sum_{c}\sum_{k}v_k l_k(y_c,x)\right) \tag{1}$$

其中,$Z(x)$为归一化函数.我们的目标是:为观察序列 x 找到最优标注序列 y.在训练数据和测试数据中,特征函数 f_k,g_k,l_k 是固定的布尔函数,特征函数对应的权值 u_k,s_k,v_k 通过训练数据获得.我们从训练库中自动抽取特征函数.

3.2 特征函数的选择

根据树结构条件随机场最大团的特点,我们生成 4 类布尔特征函数(即公式(1)中的 4 种类型:结点本身的特征函数 $g_k(y_i,x)$、边 E^{ss} 和边 E^{ps} 的各自的特征函数 $f_k(e,y_e,x)$、最大团的特征函数 $l_k(y_c,x)$).布尔特征函数的形式为:if (条件) return 1 else return 0.条件是布尔表达式,在条件中除了使用未标注时态树 x 的任意信息之外,这 4 类特

征函数的条件中还必须分别包括:只包含当前结点的时态(*type*="*Current-Node*");包含当前结点和右邻兄弟结点的时态(*type*="*Current-Sibling*");包含当前结点和其父结点的时态(*type*="*Current-Parent*");包含当前结点以及其父结点、右邻兄弟结点的时态(*type*="*Current-Parent-Sibling*").

虽然从理论上来说,特征函数可以使用未标注时态树 x 的所有信息,但是未标注时态树 x 过于庞大,会造成数据极度稀疏,所以这 4 类特征函数使用未标注时态树 x 的部分信息.下面介绍如何选择未标注时态树 x 的信息的子集.

对于训练语料库中的每个已标注时态树 T,我们从根结点出发,依次向下为每个结点构造特征函数.在已标注时态树 T 中,构造结点 N 的特征函数使用的上下文信息除了结点自身的属性外还包括:

(1) 从已标注时态树 T 的根结点到结点 N 的路径信息以及路径上的结点的属性;

(2) 已标注时态树 T 中包含的时间短语结点信息以及与这些结点与结点 N 的位置关系;

(3) 已标注时态树 T 中包含的时间副词结点信息以及与这些结点与结点 N 的位置关系;

(4) 已标注时态树 T 中包含的时间助词结点信息以及与这些结点与结点 N 的位置关系.

因为时间短语结点、副词结点和助词结点在时态树中的影响范围是有限的,所以我们划分出 N 结点的上下文区域范围 $S0$~$S6$.将 N 结点的子结点集合称作区域 $S0$;将 N 结点的父结点和左兄弟结点集称作区域 $S1$,N 结点的右兄弟结点集称作区域 $S2$;将 N 结点的父结点称作 P 结点,P 结点的父结点和 P 结点的左兄弟结点集称作区域 $S3$,P 结点的右兄弟结点集称作区域 $S4$;将 N 结点的祖父结点称作 Q 结点,Q 结点的父结点 R 和 Q 结点的左兄弟结点集称作区域 $S5$,Q 结点的右兄弟结点集称作区域 $S6$.如图 3 所示.

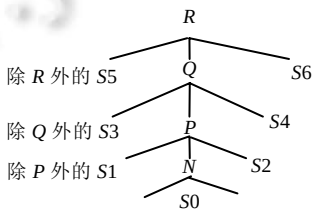


Fig.3 Contexts of node N

图 3 N 结点的上下文范围

$S0$~$S6$ 即本文中结点 N 可使用的最大上下文范围,在这个最大上下文范围内,按照构造特征函数时搜索上下文信息的范围从小到大确定了 4 种方案,即 $T0$,$T1$,$T2$,$T3$(见表 2).此外,我们定义结点 N 的直接可接触结点为:其父结点,子结点,兄弟结点的集合.

- $T0$:搜索的上下文信息来自结点 N 的直接可接触结点,即区域 $S0$,$S1$,$S2$;
- $T1$:搜索的上下文信息来自结点 N 和其父结点 P 的可接触结点,即区域 $S0$,$S1$,$S2$,$S3$,$S4$;
- $T2$:搜索的上下文信息来自结点 N 和其父结点 P 以及其祖父结点 M 的可接触结点,即区域 $S0$,$S1$,$S2$,$S3$,$S4$,$S5$,$S6$;
- $T3$:优先从结点 N 的可接触结点中寻找特定的上下文信息,其次是其父结点 P 的可接触结点,最后才是其祖父结点 Q 的可接触结点.即,区域优先顺序($S0$,$S1$,$S2$)>($S3$,$S4$)>($S5$,$S6$).例如,为结点 N 搜索已标注时态树 T 中的时间副词结点,我们优先从区域 $S0$,$S1$,$S2$ 中寻找,如果在这 3 个区域中包含有时间副词结点,我们即停止到区域 $S3$,$S4$,$S5$,$S6$ 中寻找时间副词结点.

Table 2 Search scopes of features

表 2 特征的搜索范围

$T0$	$S0$, $S1$, $S2$
$T1$	$S0$, $S1$, $S2$, $S3$, $S4$
$T2$	$S0$, $S1$, $S2$, $S3$, $S4$, $S5$, $S6$
$T3$	($S0$,$S1$,$S2$)>($S3$,$S4$)>($S5$,$S6$)

由于时间副词结点往往出现在动词结点的前面,且对于其左边的动词结点时态影响较小,所以在方案 *T*3 中,对结点 *N* 的上下文区域中同层中的时间副词结点的搜索也采用优先搜索策略,搜索优先级顺序为(*S*0,*S*1)>*S*2>*S*3>*S*4>*S*5>*S*6.时间助词结点则刚好与时间副词结点相反,常常出现在其修饰结点的后面,所以采用同时间副词结点相反方向的搜索策略(*S*0,*S*2)>*S*1>*S*4>*S*3>*S*6>*S*5.时间副词结点和时间助词结点在时态树中经常出现,但是很少出现同个分句中含有 3 个兄弟时间副词结点或 3 个兄弟时间助词结点的情况,所以在同一个搜索区域内我们按照搜索方向最多保存 2 个时间副词结点和 2 个时间助词结点,减少数据稀疏性,提高标注阶段的特征函数匹配度.

时间短语结点在语篇中包含量相对较少,但是影响却很重要,且很少出现同一个短句中包含两个时间短语结点的情况.我们注意到:如果同个分句中出现 2 个时间短语结点,则有可能它们在时间区域上相互矛盾,同时也增加标注的复杂性.所以我们将同个分句中第 1 个时间短语结点的属性直接包含到其父结点的属性中,在搜索过程中,直接从该结点的父结点,祖父结点,曾祖父结点的属性中获取时间短语结点信息.因此,结点 *N* 上下文区域中包括时间短语结点信息的区域为 *S*1,*S*3,*S*5.在 *T*3 方案中,优先搜索区域 *S*1 中 *N* 结点的父结点中包括的时间短语结点信息,没有搜索到才依次到 *S*3,*S*5 中搜索.

时间副词结点、时间助词结点、时间短语结点这 3 类结点的时态标注虽然固定为 *empty*,但是它们的父结点或者可能的右兄弟动词结点却是有时态信息的.在特征函数类型为 *Current-Sibling*,*Current-Parent*,*Current-Parent-Sibling* 时,对全局时态还是有较大影响的,所以还是需要生成其结点特征函数.

我们定义如下:

- 原子信息:*N* 结点的任意上下文区域(*S*0~*S*6)中的时间副词结点信息(包括有无该类结点、结点数目、结点属性)构成一条原子信息;*N* 结点的任意上下文区域(*S*0~*S*6)中的时间助词结点信息(包括有无该类结点、结点数目、结点属性)构成一条原子信息;*N* 结点的任意上下文区域(*S*1,*S*3,*S*5)中的时间短语结点信息(在其父结点中的时间短语属性)构成一条原子信息;
- 原子特征函数:结点 *N* 自身属性结合方案(*T*0,*T*1,*T*2,*T*3)区域内的任意一条原子信息按照 4 类特征函数的构成法,生成的特征函数统称原子特征函数;
- 复合特征函数:结点 *N* 自身属性结合方案(*T*0,*T*1,*T*2,*T*3)区域内的所有原子信息按照 4 类特征函数的构成法,生成的特征函数统称复合特征函数.

4 实验过程以及结果分析

4.1 实验数据

在条件随机场中,语料库的准确性对最后的结果至关重要,我们采用人工标注的方法标注中文树库 Chinese TreeBank6.0 (LDC2007T36).其中,chtb_0110.fid~chtb_0139.fid 的 30 个文件作为训练集,chtb_0140.fid~chtb_0144.fid 的 5 个文件作为开发集,chtb_0165.fid~chtb_0169.fid 的 5 个文件作为测试集.训练集有动词 1 504 个,共 2 859 个积极时态结点(动词结点和 IP,CP 结点).测试集有动词 237 个,共 430 个积极时态结点.

为了更好地分析标注结果,下面提供了测试集(chtb_0165.fid~chtb_0169.fid)在 3 种时态下的分布情况,见表 3 和表 4.需要注意的是:在测试集中,我们采用 16 种时态的人工标注,表 3 和表 4 是将 16 种时态转换为 3 种时态后的时态分布.

Table 3 Distribution of 3 tenses of verbs in the test set

表 3 测试集中的动词 3 种时态的分布

	Past	Present	Future	Total
数量(个)	75	144	18	237
百分比(%)	31.65	60.76	7.59	100.00

Table 4 Distribution of 3 tenses of IP, CP and verbs in the test set

表 4 测试集中的 IP,CP 和动词 3 种时态分布

	Past	Present	Future	Total
数量(个)	147	254	29	430
百分比(%)	34.19	59.07	6.74	100.00

在人工标注时态语料库时,时态语料库中 IP,CP 结点的时态可以使用如下的规则自底向上自动获取:

(1) 如果 IP(或 CP)结点有主动词,那么该 IP(或 CP)结点的时态为主动词的时态;

(2) 否则,如果 IP(或 CP)结点直接覆盖下的孩子结点中有 IP 结点,那么该 IP(或 CP)结点的时态为在这些 IP 子结点中出现次数最多的时态(如果次数最高的时态有多种,从左到右最先出现的优先);

(3) 否则,需要人工标注.

4.2 实验设置

实验开发使用的 XCRF(下载地址 https://gforge.inria.fr/projects/treecrf)[18,19]是用 Java 实现的一个通用的 CRF 工具包,用于标注线性或树结构的 XML 文件.我们使用准确率(accuracy)来衡量整体的标注效果.

$$Accuracy=\frac{\text{正确标注的个数}}{\text{所有标注的个数}}.$$

另外,我们使用精确率(precision)、召回率(recall)和 F 值(F_measure)来分析每一种时态的标注效果:

$$Precision=\frac{\text{标注为某个时态}t\text{的结点中正确标注的个数}}{\text{标注为某个时态}t\text{的结点的个数}},$$

$$Recall=\frac{\text{标注为某个时态}t\text{的结点中正确标注的个数}}{\text{正确标注中某个时态}t\text{的结点个数}},$$

$$F\_measure=\frac{2\times Precision\times Recall}{Precision+Recall}.$$

实验中的特征函数从训练语料库中自动抽取获得,按照 $T0$,$T1$,$T2$,$T3$ 的搜索规则依次搜索相应结点和区域,分别获得 3 类特征函数组合方案,获得的特征函数集依次编号为 $F1$~$F12$,具体含义见表 5.

Table 5 Combination of feature functions

表 5 特征函数的组合

$T0$	原子特征函数($F1$)	复合特征函数($F2$)	原子+复合特征函数($F3$)
$T1$	原子特征函数($F4$)	复合特征函数($F5$)	原子+复合特征函数($F6$)
$T2$	原子特征函数($F7$)	复合特征函数($F8$)	原子+复合特征函数($F9$)
$T3$	原子特征函数($F10$)	复合特征函数($F11$)	原子+复合特征函数($F12$)

按照 $F1$~$F12$ 的特征函数集组合策略,我们对于语料库完成自动抽取特征.

4.3 实验结果以及分析

由于受限于训练语料库的大小,在标注集中,很可能出现某个动词,IP,CP 结点没有任何符合条件的特征函数与之相匹配,也就是说,所有的特征函数取值均为 0.此时,结点时态标注为 *empty*,我们称其为空标注.对于出现这种情况有两种处理方法.

- 一是标注前的预标注,将全部待标注结点标注为训练语料库中出现概率最高的时态或者随机时态;
- 二是标注后再处理,将空标注的动词,IP,CP 结点时态标注为:它的前一个动词,IP,CP 结点(即当前结点的父结点或者兄弟结点)的时态,或者该时态树中出现频率最高的时态.本文实验中将空标注结点的时态标注为其前一个动词,IP,CP 结点的时态,这种方法可以最大提高标注的准确率.在实验 $F12$ 中,空标注大约占 2%左右.

4.3.1 标注为 16 种时态的实验结果

利用训练语料库来自动抽取特征函数和训练特征函数的权重,并对测试语料库中的文件进行标注(16 种时

态划分),标注结果分为全部的积极时态结点的标注准确率(称为全局准确率)和动词结点的标注准确率,见表 6.

Table 6 Experimental comparison of different combinations of feature functions (16 tenses)

表 6 特征函数不同组合的实验结果对比(16 种时态)

	全局准确率(%)	动词准确率(%)
*F*1	63.02	64.98
*F*2	64.65	64.98
*F*3	62.79	64.97
*F*4	62.79	64.14
*F*5	61.86	62.02
*F*6	62.79	64.14
*F*7	63.02	64.13
*F*8	62.79	66.24
*F*9	63.49	64.56
*F*10	61.62	62.44
*F*11	65.58	66.24
*F*12	61.40	62.45

从表 6 中我们可以看出:全局准确率总是低于动词的准确率,这主要是时间副词结点和时间助词结点往往修饰的是动词结点,对动词结点的影响更大,在以时间副词和时间助词为主要特征的情况下,动词结点准确率更高.IP,CP 结点作为动词结点的父结点或者更高层结点,在标注时选择主动词作为其时态,其孩子结点较多,受到的干扰较大.

在表 6 特征函数不同组合的实验结果对比(16 种时态)中,是在消除空标注之后统计出的全局准确率和动词准确率.倘若不消除空标注,从实验结果中有如下的结论:原子特征函数(*F*1,*F*4,*F*7,*F*10)的准确率相对较低,但是覆盖度大;复合特征函数(*F*2,*F*5,*F*8,*F*11)在能判断的情况下判断的准确度较高,但是复合特征函数中使用的条件很多,导致数据较稀疏,会出现较多空标注,覆盖度小.

4.3.2 与其他方法的实验结果的比较

由于 Gong[2,10]的方法没有提供时态转换方法的准确率、精确率和召回率,且没有标注所有的动词结点,所以无法与之比较.为了比较本文中的方法是否有效,我们将实验结果最优组 *F*11 中的 16 种时态转换成动词的 3 种时制(过去时、现在时、将来时)的精确率、召回率和 *F*-measure,并和 Ye[7]中的实验结果相比较.转换规则如下:

- {过去一般时,过去进行时,过去完成时,过去完成进行时,过去将来一般时,过去将来进行时,过去将来完成时,过去将来完成进行时,现在完成时,现在完成进行时}→过去时;
- {现在一般时,现在进行时}→现在时;
- {将来一般时,将来进行时,将来完成时,将来完成进行时}→将来时.

不仅动词有时态,IP,CP 也有时态,所以我们分别统计动词的标注结果与 IP,CP 和动词的标注结果,分别见表 7 和表 8.

表 7 是使用本方法时动词的最佳标注结果(*F*11)与 Ye 的最佳结果的对比.

Table 7 Experimental comparison of tagging verbs between *F*11 and Ye[7]

表 7 动词的 *F*11 与 Ye[7]的标注结果对比

	Ye[7]中,*Accuracy*=58.21%			*F*11 中,*Accuracy*=72.57%		
	Precision (%)	*Recall* (%)	*F*-measure(%)	*Precision* (%)	*Recall* (%)	*F*-measure (%)
过去时	67.57	79.55	72.10	69.35	57.33	62.77
现在时	42.50	27.48	32.07	74.56	87.50	80.51
将来时	29.66	25.56	21.56	50.00	16.67	25.00

从该表可以看出,本方法在整体准确率上有大幅度的提升.Ye[7]中对经常出现的现在时的精确率和召回率都很低,而 *F*11 大幅提升现在时的精确率和召回率.由于将来时在时态语料库和时态测试库中出现的概率往往较低,所以 Ye 和 *F*11 的 *F*-measure 均较低.*F*11 的将来时的精确率达到 50%,说明本方法可能还是能抓住决定将

来时的一些特征,然而训练语料库太小导致有些情形不能被学习到.

表 8 是在使用本方法时 IP,CP 和动词的最佳标注结果(*F*11).因为 Ye[7]的方法中没有使用 IP,CP 结点,所以不可比较.

Table 8 Results of tagging IP, CP and verbs (*F*11)

表 8 IP,CP 和动词的 *F*11 的标注结果

	*F*11 中 *Accuracy*=70.93%		
	Precision (%)	*Recall* (%)	*F*-measure (%)
过去时	67.26	51.70	58.46
现在时	72.55	87.40	79.29
将来时	63.63	24.14	35.00

从表 7 和表 8 可以看出,动词结点的标注结果要比全局结点(IP,CP 和动词)的标注结果好一些.

5 结束语

实验结果表明:基于时态树的汉英时态标注的动词准确率为 72.57%,要比基于序列的方法好得多(准确率为 58.21%),说明影响动词时态的因素具有层次性效果,线性条件下往往不能体现这种结构性特征.实验过程中,条件随机场表现较稳定,通过和 Ye[7]的实验进行对比,实验结果准确率的大幅提升说明,使用条件随机场方法处理时态树是一种有效的方法.

使用条件随机场处理时态树的自动标注,关键的步骤在于特征函数的选取.以上实验中,特征函数的选取有改进的空间,以后将考虑使用更精细的特征模板进行组合并与之比较.同时,语料库的规模也是限制实验结果的重要方面,今后将扩大人工标注集,采用不同规模的语料库进行实验.此外,本文提出的方法尚未利用到句子的上下文,在后续的工作中将研究基于语篇的汉英时态标注.

致谢 在此,我们向对本文的工作给予支持和建议的同行、老师和同学表示感谢,尤其感谢评审专家的评阅和指正.

References:

[1] Liu Q, Yu SW. Discussion on the difficulties of Chinese-English machine translation. In: Huang CL, ed. Proc. of the Int'l Conf. on Chinese Information Processing. Beijing: Tsinghua University Press, 1998. 507−514 (in Chinese).

[2] Gong Z, Zhang M, Tan C, Zhou G. *N*-Gram-Based tense models for statistical machine translation. In: Proc. of the EMNLP. East Stroudsburg: Association for Computational Linguistics (ACL), 2012. 276−285.

[3] Gong QY. Chinese Phase, Tense and Aspect. Beijing: China Commerce and Trade Press, 1995. 1−95 (in Chinese).

[4] Cheng JH, Dai XY, Chen JJ, Wang QX. Processing of tense and aspect in Chinese-English machine translation. Application Research of Computers, 2004,21(3):79−81 (in Chinese with English abstract).

[5] Ma HM, Wang T, Chen HW. Time phrase parsing of Chinese text and tense calculus. Journal of Computer Research and Development, 2002,39(10):1211−1220 (in Chinese with English abstract).

[6] Ma HM. Research on the representation and application of Chinese context in Chinese-English machine translation [Ph.D. Thesis]. Changsha: National University of Defense Technology, 2002 (in Chinese with English abstract).

[7] Ye Y, Zhang Z. Tense tagging for verbs in cross-lingual context: A case study. In: Proc. of the IJCNLP. Berlin, Heidelberg: Springer-Verlag, 2005. 885−895. [doi: 10.1007/11562214_77]

[8] Murata M, Uchimoto K, Ma Q, Isahara H. Using a support-vector machine for Japanese-to-English translation of tense, aspect, and modality. In: Proc. of the ACL Workshop on the Data-Driven Machine Translation. East Stroudsburg: Association for Computational Linguistics (ACL), 2001. 1−8. [doi: 10.3115/1118037.1118052]

[9] Murata M, Ma Q, Uchimoto K, Kanamaru T, Isahara H. Japanese-to-English translations of tense, aspect, and modality using machine-learning methods and comparison with machine-translation systems on market. Language Resources and Evaluation, 2006, 40(3):233−242. [doi: 10.1007/s10579-007-9022-z]

[10] Gong ZX, Zhang M, Tan CL, Zhou GD. Classifier-Based tense model for SMT. In: Proc. of the 24th Int'l Conf. on Computational Linguistics. 2012. 411−420.

[11] Chen YJ. Research on several key issues of Chinese-English machine translation [Ph.D. Thesis]. Xiamen: Xiamen University, 2011 (in Chinese with English abstract).

[12] Ye Y, Fossum V, Abney S. Latent features in automatic tense translation between Chinese and English. In: Hwee Tou Ng, ed. Proc. of the 5th SIGHAN Workshop on Chinese Language Processing. Australia: BPA Digital Press, 2006. 48−55.

[13] Olson M, Traum D, Dykema C, Weinberg A. Implicit cues for explicit generation: Using telicity as a cue for tense structure in a Chinese to English MT system. In: Proc. of the Machine Translation Summit VIII—Machine Translation in the Information Age. 2001. 259−264.

[14] Lu JM, Ma Z. Comments on Function Word of Modern Chinese. Beijing: Peking University Press, 1999. 106−141 (in Chinese).

[15] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. of the 18th Int'l Conf. on Machine Learning. 2001. 282−289.

[16] Tang J, Hong MC, Li JZ, Liang B. Tree-Structured conditional random fields for semantic annotation. In: Cruz I, ed. Proc. of the 5th Int'l Conf. of Semantic Web. Berlin, Heidelberg: Springer-Verlag, 2006. 640−653. [doi: 10.1007/11926078_46]

[17] Cohn T, Blunsom P. Semantic role labeling with tree conditional random fields. In: Proc. of the 9th Conf. on Computational Natural Language Learning. East Stroudsburg: Association for Computational Linguistics (ACL), 2005. 169−172.

[18] Jousse F, Gilleron R, Tellier I, Tommasi M. Conditional random fields for XML trees. In: Proc. of the ECML Workshop on Mining and Learning in Graphs. 2006. 525−533.

[19] Moreau E, Tellier I. The crotal SRL system: A generic tool based on tree-structured CRF. In: Proc. of the 19th Conf. on Computational Natural Language Learning. 2009. 91−96.

附中文参考文献:

[1] 刘群,俞士汶.汉英机器翻译的难点分析.见:中文信息处理国际会议论文集.北京:清华大学出版社,1998.507−514.

[3] 龚千炎.汉语的时相时制时态.北京:商务印书馆,1995.1−95.

[4] 程节华,戴新宇,陈家骏,王启祥.汉英机器翻译中时体态处理.计算机应用研究,2004,21(3):79−81.

[5] 马红妹,王挺,陈火旺.汉语篇章时间短语的分析与时制验算.计算机研究与发展,2002,39(10):1211−1220.

[6] 马红妹.汉英机器翻译中汉语上下文语境的表示与应用研究[博士学位论文].长沙:国防科学技术大学,2002.

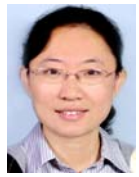
[11] 陈怡疆.汉英机器翻译若干关键问题的研究[博士学位论文].厦门:厦门大学,2011.

[14] 陆俭明,马真.现代汉语虚词散论.北京:北京大学出版社,1985.106−141.

陈怡疆(1972－),男,福建泉州人,博士,副教授,主要研究领域为自然语言处理,机器翻译,隐喻计算,文本信息抽取.

史晓东(1966－),男,博士,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.

徐海波(1988－),男,硕士生,主要研究领域为自然语言处理.



苏畅(1974－),女,博士,副教授,主要研究领域为自然语言处理,隐喻计算,人工智能.