一种应用句法分析技术的汉语自动分词模型

单玉秋

国家语言文字工作委员会语言文字应用研究所

An Automatic Chinese Word Segmentation Model By Applying Parsing Technology

Shan Yuqiu

the Institute of Applied Linguistics, the State Language Commission

本文意在探讨句法分析技术在自动分词任务中的作用。自动分词如果不利用更高级的语言知识，不可能获得很好的切分效果，因为切分歧义的解决不能仅仅依赖于简单的切分规则，句法甚至语义知识是必要的。句法分析技术是计算语言学研究很早的技术之一，它在利用自然语言知识上具有很好的系统性。如果将句法以及语义知识用于切词任务，可望得到更好的分词结果。本文提出的汉语切词模型考虑所有的切分可能，其最主要的特点是运用汉语句法等知识从各种切分可能中选择出合理的切分结果。该模型的切分分成两步：(1)对 $S=a_1a_2...a_i...a_n$ 进行处理，得到一个可能的词集 $Wset=\{W_p|W_p=a_i...a_j[a,b], (1 \leqslant i \leqslant j \leqslant n)$ 并且对任意 m,n， $W_m \neq W_n$， W_p 是一个分词单位,[a,b]是字串 W_p 的位置区间。$\}$；(2)处理 Wset，得到 $W'set=\{W_k|k=1,...,q,q \leqslant n\}$,其中，$|W_1...W_k|=|S|$ ($|W_1...W_k|$是词串 $W_1...W_k$ 的长度，$|S|$是句子 S 的长度)，且 $W_1...W_k$ 是 $a_1..a_n$ 的一个合适的切分。步骤 1 对字串 $a_1a_2...a_i...a_j...a_n$ 的处理，是从该字串中找出其中所有出现在词典中的分词单位（包括人名、地名以及在词典中出现的词汇等）.所有被发现的分词单位全部作为集合 Wset 中的元素。Wset 称为可能词集。步骤 2 则利用句法级以上的知识将 Wset 中的某些词选择出来，由它们组合成一个有顺序的合适的词序列。本文采用一个自由传播式句法分析网络技术得出合适的切分结果。短语结构形式描述的汉语句法规则采用相应的有向层次网络图表示。网络的词选择过程大致是这样：在网络运行之初，所有节点的状态为 no。可能词集中的所有元素进入网络的所有词类节点，这些信息便在网络中沿着边的方向进行传递。在规则类型节点处，对位置区间信息进行合并，并进行句法甚至语义分析；然后规则节点或直接传递原来信息或传递新生成的信息，并重新设置自己的状态；如此，只有当网络中最高层节点 s 达到稳定状态 ok 时，网络计算结束，s 节点处将输出合适的切分结果。如对句子“他马上来。”有：Wset={他[11],马[22],上[33],马上[23],上来[34],来[44]}。所有可能的分词结果有{他 11,马 22,上 33,来 44}，{他 11,马 22,上来 34}，{他 11,马上 23,来 44}。对于下面的句法子集而言,网络将输出正确的分词结果{他 11,马上 23,来 44}。句法规则子集为：Rs: s→np|vp|phnv; Rphnv: phnv→np+vp; Rnp1:np→rnp;Rnp2:np→npd;Rnp3:np→[rnz]+[a]+[usde]+ng;Rnp4:np→s+usde+np;Rvp1:vp→[d]+[va]+vgn+np+[utle];Rvp2:vp→[d]+[va]+vgo+[utle]; Rvp3:vp →[d]+[va]+vgj+[utle]+np+vp;Rvp4:vp→[d]+[va]+vgdo+[utle]+np+np。若向规则子集中增加一条规则 s→np+phnv，则通过在网络 Rs 规则节点处进行语义分析可排除{他 11，马 22，上来 34}这一切分可能，网络依然输出{他 11，马上 23，来 44}这一正确结果。