用n元语法统计确定短语的边界*

Determining Phrase Boundaries by N-gram Statistics

郭志立 苑春法 黄昌宁

清华大学计算机系，北京 100084

在自底向上的句法分析算法中，找出句子里相对独立的片段，并把片段内的词结合起来构成某种语法功能的短语，是减少后续工作量的重要环节。一个模式（在文本中出现多次的语言片段）在句中是否构成相对独立的短语，有赖于它所处的上下文环境。但汉语中各种词在句子中的功能非常灵活，很难确定什么样的模式就是一个短语。我们设想在一定长度范围内对一个模式的左右边界进行扩展，考察扩展后的模式频率和扩展进来的词性标记与原模式之间的互信息，根据频率的下降梯度及互信息之间的对比来确定短语的边界。

以某一 n元模式为核心，在 n+1 元语法的统计结果中查找这个 n元模式，并记录这些 n+1 元模式各自的频率。然后在 n+2元语法中扩展各n+1元模式，在n+3元语法中扩展各n+2元模式……以此递归地对原模式在长度上进行扩展，模式的频率随词数的增加而逐渐降低。根据频率下降的梯度和扩展而来的词与原模式之间的互信息可以发现一些短语的边界。考察模式频率下降的梯度，可以发现有些模式在扩展到一定长度后，频率出现跳跃式的陡然下降，这个陡降的位置多是扩充后的长模式边界，即扩充后的长模式构成短语。为了简化计算，定义$\Delta(k)=\log_{10}(R_{n+k}\ /\ R_n)$，其中$R_{n+k}$、$R_n$分别表示第k次扩展后n+k元模式的频率和作为核的n元模式的频率，用它作为评价频率下降梯度的指标。以模式“a+的+ng”向左的一次扩展为例，观察Δ(k)随k的变化形势，可以看出“a+的+ng”向左扩展了“dd”(程度副词)以后，再次向左扩展遇到“ng”时，频率发生急剧下降。以此认为，模式“dd a 的 ng …”在其左边词性为“ng”的情况下构成相对独立的短语。

对于作为扩展核心的模式本身，仅仅根据模式向左(右)第一次扩展时不同词性标记的频率下降幅度是不能确定模式自身的左(右)边界的。我们用向左(右)扩展的第一个词性标记X_{11}与模式本身K的互信息 $I(X_{11}) = \log\ P(X_{11} \mid K) / P(X_{11})$ 作为衡量词性标记X_{11}与模式 K结合紧密程度的标准。以模式“mx+qng+ng”为例，可以看到rm(代词“每”)、maf(前助数词)与“mx+qng+ng”结合较紧，形成更长的短语，而频率较高的vgn、usde等都可以做为“mx+qng+ng”的左边界。这跟依存关系的标注标准相符合。

我们对160万词的语料进行了统计，并且用上述方法详细寻找了“mx +量词(包括量词“个”qng、个体量词qni、种类量词qnk等)+ng”、“a+的+ng”、“vg+的+ng”、“f+的+ng”、“ng+的+vg”等五类模式扩展后的短语边界，然后以标注了依存关系的735个句子中的短语切分作为评测标准。实验表明这个方法是可行的。

---

*国家自然科学基金支持项目 № 69373036