基于知识库的交集型歧义字段切分系统

The System of Ambiguous Phrase Segmentation based on Knowledge Database

张培颖 李村合 (中国石油大学计算机与通信工程学院 东营 257061)

摘要:无论在自然语言处理还是在机器翻译中,中文自动分词都是一个重要的环节。其中歧义字段切分是中文自动分词研究中的一个"拦路虎",是影响分词系统切分精度的主要因素。能够正确切分某一类歧义字段的知识称为分词知识,所有分词知识的集合称为知识库或规则库。本文通过建立交集型歧义字段切分知识库,并采用知识学习的方法来丰富系统的知识,充分利用了知识库中积累的词的二元语法关系、语素构词、句法关系以及上下文等信息,提高了交集型歧义字段的切分精度。

关键词:知识库 规则库 交集型歧义字段 知识学习

1 引言

无论在自然语言处理还是在机器翻译中,中文自动分词都是一个重要的环节。其中歧义字段切分是中文自动分词研究中的一个"拦路虎"。所谓歧义字段,是指句中某个片段存在 2 种或 2 种以上的切分形式。只有歧义字段才能引起错误切分,歧义字段是影响分词系统切分精度的重要因素。

歧义字段分为两种基本类型:交集型歧义和组合型歧义。交集型歧义字段,占歧义字段总数的 85% ~ 90%。因此,如何解决好交集型歧义字段的切分问题,对于歧义字段的切分具有重要的意义。

2 系统的总体结构

我们设计的系统的总体结构如图 1 所示。中文字序列经过交集型歧义字段采集模块,识别出其中的交集型歧义字段;交集型歧义处理模块利用知识库里存放的各种分词知识进行歧义处理;最后生成分词结果。

交集型歧义字段的采集采用改进的逐词扫描法。对汉字串 $C_1C_2\cdots C_n$(C_i 为汉字)从 C_1 开始逐词进行正向最大匹配,找到第一个为词的位置 i,即 $C_1C_{i+1}\cdots C_m$ 为词;然后从 C_{i+1} 开始,再逐词进行正向最大匹配找到不成词的位置 j 且 $j > m$,即 C_iC_{i+1} 不成词。则 $C_iC_{i+1}C_{i+2}\cdots C_j$ 为交集型歧义字段。

交集型歧义处理模块,一方面通过提取知识库中的知识来进行交集型歧义字段的切分;另一方面,通过分词结果,可以采用知识学习的手段,以丰富知识库中的知识。

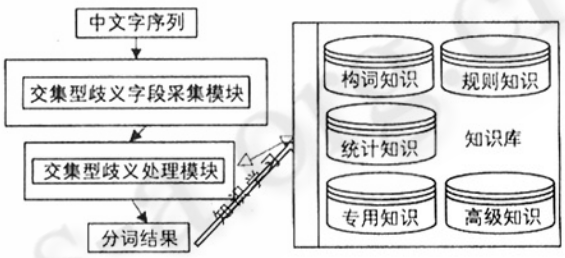

图 1 系统的总体结构图

3 知识库

能够正确切分某一类歧义字段的知识称为分词知识。所有分词知识的集合称为知识库或规则库。一直以来,规则方法都是歧义处理中的常用技术。通过分析从语料中采集的歧义现象,从语素构词、词的语法关系、词义搭配等多个角度进行分析,制定相应的构词规则和句法规则来切分歧义字段。此方法的明显优点是:解决歧义问题的方法,独立于自动分词方法,适用性强。对分词词典只要求能够提供必要的词项信息,故对自动分词的空间复杂度影响不大。再者采用这种方法便于在使用的过程中,通过知识学习的手段,以丰

富系统的知识，提高分词的精度。

3.1 分词知识的分类

目前所用的分词知识可以分为如下几类：

（1）构词知识。构词知识用于构造分词词典中没有的词，可以解决叠词的现象。例如："花花绿绿的世界"，按照汉语的构词法"花花绿绿"是一个词，但分词词典中不可能包括所有形如 AABB 的词，故词被错误地切分。构词知识能够构成形如 AA、AABB、AAB（AB 为词）、ABB（AB 为词）、ABAB（AB 为词）、AXAB（AB 为词）、AXA（A 为动词）、前缀词构成的词、后缀词构成的词等等，有了这些构词知识，这类词就可以正确切分。

（2）规则知识。从歧义字段形成的词与词之间的结构关系和词性关系出发，总结出一些规则来解决它们，部分规则如下：

规则 1：动词 + 名词 动宾类动词与其宾语同义的名词和独立成词频度很高的动词与双字以上的实意名词间的切分歧义，如："提/问题"、"办/事情"等。

规则 2：交集型歧义字段 $n_1 \cdots n_m$ 中，若 $n_1, n_2, \cdots, n_{m-1}$ 为同义或反义词，且 $n_1n_m, n_2n_m, \cdots, n_{m-1}n_m$ 分别成词，则 $n_1, n_2, \cdots, n_m$ 分别单独切分。如："初中级"、"高中低档"应分别切分为"初/中/级"、"高/中/低/档"。

规则 3：交集型歧义字段 $n_1 \cdots n_m$ 中，若 $n_2, \cdots, n_m$ 为同义或反义词，且 $n_1n_2, n_1n_3, \cdots, n_1n_m$ 分别成词，则 $n_1, n_2, \cdots, n_m$ 分别单独切分。如："关内外"应分别切分为"关/内/外"。

（3）专用知识。用于正确地解决 1 个字所形成的歧义字段的知识称为专用知识。

例如："把"字知识的描述如下：式中 W 为分词词典中词的集合，WD 为动词的集合。

$r =$ 把 $\alpha\beta \cap$ 把 $\alpha \in W \Rightarrow (\beta \in WD \rightarrow r1 =$ 把/α/$\beta) \cup (\beta \notin WD \cap \alpha\beta \in W \rightarrow r1 =$ 把/$\alpha\beta) \cup (\beta \notin WD \cap \alpha\beta \notin W \rightarrow r1 =$ 把 α/$\beta)$

利用此知识可以把"把头抬起来"正确切分为"把/头/抬/起来"，"把儿子给你"正确切分为"把/儿子/给/你"，"请拉好把手"正确切分为"请/拉/好/把手"等等。

（4）高级知识。必须通过上下文知识才能切分的歧义字段称为语用歧义字段。为了正确切分这类歧义字段而增加的语义知识、语用知识称为高级知识。

（5）统计知识。人之所以能对歧义字段做出正确切分，主要依赖字段中各种切分情况在实际生活中的使用频次做出判断的。因此，统计词出现的频次等信息对于歧义字段切分十分重要。

3.2 分词知识的选取

任何一个分词知识都是研究者通过对大量的歧义字段分析后提出的，它们都可以正确地处理某一类歧义字段，但同时可能会造成对另一类歧义字段的错误切分。因此在知识选取的时候，不能只考虑对某个歧义字段的正确切分，应该考虑对所有规则能处理的一类歧义字段的正确切分。

定义 1：某一个短语出现的频率定义为：

$$freq(phrase) = \frac{N}{S}$$

其中：S 为语料中短语的总数，N 为短语 phrase 出现的次数。

定义 2：设 W 为包括某一类歧义字段的短语集合，L 为 1 条分词知识，W_1 为 W 的子集，知识 L 对 W_1 中所有的短语都正确切分；W_2 为 W 的子集，知识 L 对 W2 中所有的短语都错误切分。则分词知识 L 的效率因子 I 定义为：

$$I = \frac{freq1}{freq2}$$

$$freq1 = freq(C_1) + freq(C_2) + \cdots + freq(C_i)$$

$$freq2 = freq(T_1) + freq(T_2) + \cdots + freq(T_j)$$

其中：$freq(C_i)$ 为短语 C_i 出现的频率，i 为 W_1 中所包含的短语数，j 为 W_2 中所包含的短语数，若 $I << 1$，则分词知识 L 可取，否则不可采用。

另一方面，随着分词知识的积累，分词知识之间可能存在相互冲突的现象，这时我们采取统计知识库中的信息对歧义字段进行切分，采用的切分公式为：

$$R = \sum_{i}^{n} Ln(freq(wi)) \times [len(wi)]^3$$

其中：n 为分出的词数，freq(wi) 为该词的词频，Ln 自然对数，len(wi) 该词所占的字符数，R 最终用来比较的权值。

对词频取自然对数后，使其数值差距明显缩小。取词长的 3 次方是因为可以有效地强化对词长较大的词优先选择的特点，同时使某些较短的却词频很高的词也有机会被选择。

（下转第 41 页）

(上接第 43 页)

下面通过具体的实例演示一下 R 值的计算过程。

取字符串“在意大利”来分。正向最大匹配算法分出的结果是“在意 - 大 - 利”,反向最大匹配算法分出的结果是“在 - 意大利”(词库中有“在意”和“意大利”)。其中,“在意”的词频是 543000,“意大利”的词频是 282000[6]。

前者 $R_1 = Ln(543000) \times 2^3 = 105.6$,

后者 $R_2 = Ln(282000) \times 3^3 = 338.8$。

因为 $R_2 > R_1$,所以结果取后者“在 - 意大利”。

3.3 知识库的结构

自动分词系统中所有的分词知识的集合称为知识库。并不是分词知识越多越好,由于知识之间的相互影响,知识的顺序不同,就可以有不同的切分结果。因此,为了使知识库中的所有知识之间不互相影响,知识库所有的知识应该是独立的,知识的调用应该是单个进行。知识库(规则库)应该是一个开放的集合,用户可以根据实际需要来进行调整、修改、增加等操作。

4 总结

本文通过建立交集型歧义字段知识库的方法来切分交集型歧义字段,充分利用了知识库中积累的汉语语法、句法、构词规则等信息,再者采用这种方法便于在使用的过程中,采用知识学习的手段,以丰富系统的知识,提高了分词的精度。

参考文献

1 黄河燕、李渝生,上下文相关汉语自动分词及词法预处理算法[J],应用科学学报,1999(6)。

2 温锁林,中文文本歧义字段切分技术[J],语文研究,2001(3)。

3 梁南元,汉语自动分词知识,中文信息学报,1990(2):29。

4 郑彦斌,书面汉语自动分词及歧义分析,河南师范大学学报,1997(4):25。

5 闫引堂、周晓强,交集型歧义字段切分方法研究,情报学报,2000(6)。…

6 张江,基于规则的分词方法,计算机与现代化,2005(4)。