基于 CRF模型的组合型歧义消解研究

德鑫[1], 曲维光[1], 徐 涛[1], 董 宇[2]

(1. 南京师范大学 数学与计算机科学学院, 江苏 南京 210097; 2. 金陵科技学院 龙蟠学院, 江苏 南京 211169)

[摘要] 组合型歧义切分是汉语自动分词的难点之一. 为此, 利用 CRF(条件随机场)模型, 以歧义字段的上下文的词和词性建立特征模板, 进行歧义消解研究. 以 1998年半年《人民日报》为语料, 对常用的 10个组合歧义字段进行消歧, 平均消歧正确率达到 96. 35%, 取得了良好的效果. 实验表明, 利用该模型能有效提高消歧正确率.

[关键词] 中文自动分词, 组合歧义, CRF

[中图分类号] TP 311, TP 391. 12 [文献标识码] A [文章编号] 1672-1292(2008) 04-0073-04

Research of Disambiguating Combinational Ambiguity in Chinese Word Segmentation Based on CRF

Ding Dexin[1], Qu Weiguang[1], Xu Tao[1], Dong Yu[2]

(1. School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China;
2. Longpan School, Jinling Institute of Technology, Nanjing 211169, China)

Abstract Combinational ambiguity is one of the difficult points in Chinese word segmentation. Based on the CRF (Conditional Random Fields) model, this paper establishes feature template by the contextual words and part of speeches of the ambiguity word. 10 often-used ambiguity words are tested by using half of the 1998 "People's Daily" corpus, and the average accuracy is 96. 35%. The result of the experiment reveals that using the model is more effective for disambiguation.

Key words Chinese word segmentation, combinational ambiguity, CRF

自动分词是自然语言处理的基础, 后续深层的语言分析和理解都直接依赖于分词的结果[1]. 歧义切分又是影响分词系统切分精度的重要因素[2]. 歧义切分有两种基本类型: 交集型歧义和组合型歧义. 本文仅讨论组合型歧义切分字段.

定义 给定任意汉字串 AB 满足 (1) $AB \in W$、$A \in W$、$B \in W$、W 为词表; (2) 切分 …/AB/… 及 …/A/B/… 在真实文本中均出现, 则称 AB 为组合型歧义切分字段 (简称组合型歧义). 例如:

(a) 他的话常常博得观众会心的笑声.

(b) 如果这段能够成立的话, 那么我们就可以认定他就是嫌疑人.

上面例子中, "的话" 在不同的上下文中需要从分和从合.

迄今为止, 汉语分词歧义的研究多集中在交集型歧义, 针对组合型歧义的则较少. 大多数交集型歧义通常可根据字段内部的信息[2], 或以句法为主的局部上下文信息[3] 予以解决. 组合型歧义的处理策略则不相同, 往往须要考察更大范围的上下文[4].

1 相关研究

相对于交集型切分歧义, 组合型切分歧义的处理比较困难. 经过自然语言研究者几十年的努力, 取得了很大的进展, 提出了许多的解决方案.

收稿日期: 2008-06-18

基金项目: 国家自然科学基金(60773173)、国家"973"计划(2004CB318102)、江苏省社科基金(06JSBYY001、07YYB003)和国家社科基金(07BYY050)资助项目.

通讯联系人: 曲维光, 博士, 副教授, 研究方向: 计算语言学和人工智能. E-mail: wgqu@njnu.edu.cn

1.1 统计和规则相结合的方法

郑家恒等[5]对一些通常只有一种切分方式,只在个别情况下发生变化的字段采用基于统计信息的切分策略.具体方法是:建立基于枚举的歧义字段库,使得组合型歧义字段的切分通过数据库查询进行.计算值 freq = $nh/(nh+nf)$(其中 nh 为合形式的次数, nf 为分形式的次数, freq为切分形式为合形式的概率);设定阈值 a_1 和 a_2($a_1 > a_2$),当 freq > a_1 时,采用合的形式;当 freq < a_2 时采用分的形式.这种统计方法等同于选择概率最大的切分形式,小概率的切分形式将会被忽略.对于其它字段,利用组合型歧义字段与其前后相邻词语的词性有关的特性,采用基于词语/词性规则的切分策略.

1.2 基于向量空间模型的统计方法

肖云[6]和 Luo Xiao[7]针对组合型歧义字段的切分依赖于其上下文的句法和语义信息这一特点,提出"就问题本身的性质而言,组合型歧义切分字段的排歧问题是一个与词义消歧几乎等价的问题".遵循这一思路,以词义消歧中广泛使用的向量空间法为基本模型.即由训练得到歧义字段 W_0 的特征矩阵 $F_{分}(W_0)$ 和 $F_{合}(W_0)$.消歧时先得到 W_0 在当前句中的特征矩阵 $F(W_0)$,然后计算其与 $F_{分}(W_0)$ 和 $F_{合}(W_0)$ 的距离,取距离近者所对应的下角标作为切分结果.

1.3 基于上下文语境信息的方法

曲维光[8]提出了相对词频的概念,据此建立语境计算模型,利用歧义字段前后语境信息对组合型分词歧义进行消解.该模型不仅考虑了语境中存在的词频,而且考虑了语境中出现词语相对于整个语料词频的相对比率,用相对词频来模拟人们判断语境中出现词语对消歧的重要程度;同时又区分了语境的位置,将语境分为前语境和后语境,从而把前后语境出现的词语区分开来,提高了语境信息计算的准确性.

1.4 自组织的汉语组合型歧义消歧方法

冯素琴等[9]用人工校验后的分词语料提供的搭配实例作为组合歧义字段的初始搭配知识,提出使用搭配统计表的多元最大对数似然比进行消歧;继而根据实验确定了歧义字段的上下文窗口、窗口位置区分、权值估计等要素;在此基础上采用自组织方法自动扩充搭配集,使消歧信息趋于稳定.

由此可见,上述模型或方法对组合型歧义消解取得了一定的效果,本文尝试利用 CRF 模型进行歧义消解,试图取得更高的正确率.

2 基于条件随机场 *CRF* 的歧义消解模型

2.1 *CRF* 模型

条件随机场 CRF(Conditional Random Fields)[10],是一个在给定输入节点(观察值)条件下计算输出节点(标记)的条件概率的无向图模型,特别擅长处理序列标记问题.对于输入序列 x 和输出序列 y,可以定义一个线性的 CRF模型,形式如下:

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum \lambda_k f_k(y_{i-1}, y_i, x) + \sum \mu_k g_k(y_i, x) \right),$$

其中每个 $f_k(\cdot)$ 是观察序列 x 中位置为 i 和 $i-1$ 的输出节点的特征,每个 $g_k(\cdot)$ 是位置为 i 的输入节点和输出节点的特征,λ 和 μ 是特征函数的权重,Z 是归一化因子.作为一个无向图模型表现出比 HMM(隐马模型),MEMM(最大熵隐马模型)等有向图模型更好的效果.隐马模型一个最大的缺点就是由于其输出独立性假设,导致其不能考虑上下文的特征,限制了特征的选择,而最大熵隐马模型解决了这一问题,可以任意地选择特征,但由于其在每一节点都要进行归一化,所以只能找到局部的最优值,同时也带来了标记偏置的问题(label bias),即凡是训练语料中未出现的情况全都忽略掉,而条件随机场则很好地解决了这一问题,它并不在每一个节点进行归一化,而是所有特征进行全局归一化,具有表达元素长距离依赖性和交叠性特征的能力,能方便地在模型中包含领域知识,因此可以求得全局的最优值.

本文的实验使用的 CRF模型,具体采用了 TakuKudo编写的工具包"CRF++ 0.50"进行训练和测试(下载地址: http://crfpp.sourceforge.net/).

2.2 利用 CRF进行歧义消解

2.2.1 CRF 训练和测试语料的格式

我们把确定歧义字段的从合从分问题转化为序列标注问题,利用歧义字段上下文的词、词性来作为消

解的依据.

为了使用 CRF,训练和测试文件的格式必须符合要求. 一般说来,训练和测试文件必须包含多个 tokens,每个 token包含多个列. token的定义可根据具体的任务,如词、词性,等等,但最后一列是被 CRF训练用的标记. 每个 token必须在一行,各列之间用空格或制表符分开,一个 token序列组成一个句子. 句子跟句子之间用空行分开.

对于组合型歧义消解的任务,我们这样定义 token 包含 3列,分别是:词,词性,标记. 其中标记的定义是:对于句子中的歧义字段的其他词,标注为 *X*,对于歧义字段,若是从合的,该字段标注为 1,对于从分的,该字段标注为 2 对于歧义字段的词性,所有的词的从合从分形式一律标注为 *Y*. 例如句子:他 /*r* 的 /*u* 话 /*n* 常常 /*d* 博得 /*v* 观众 /*n* 会心 /*d* 的 /*u* 笑声 /*n*。/*w*. 该句转化为训练数据的格式如表 1

表 1 训练数据的格式

Table 1 Format of training data

词	他	的话	常常	博得	观众	会心	的	笑声	.
词性	*r*	*Y*	*d*	*v*	*n*	*d*	*u*	*a*	*w*
标记	*X*	1	*X*	*X*	*X*	*X*	*X*	*X*	*X*

测试语料的格式同训练语料格式基本相同,除了没有标记一列.

2.2.2 特征模板

由于 CRF是一个通用的序列标注工具,所以需要事先确定特征模板. 特征模板文件中的每一行代表了一个模板. 模板的基本格式是:% × [row, col],用于确定输出数据的一个 token 其中, row确定与当前 token的相对行数, col确定列的绝对位置. 考虑到歧义消解需要用到歧义字段上下文的词、词性等属性. 本文使用的 CRF特征模板及其意义解释如表 2

表 2 实验所用的 **CRF** 模板

Table 2 CRF template of the experiment

模板	意义
# Unigram	输出的一元文法
U_{01}:% × [- 2, 0]	左边第二个词
U_{02}:% × [- 1, 0]	左边第一个词
U_{03}:% × [0, 0]	该词
U_{04}:% × [1, 0]	右边第一个词
U_{05}:% × [2, 0]	右边第二个词
U_{06}:% × [- 1, 1]	左边第一个词性
U_{07}:% × [1, 1]	右边第一个词性
U_{08}:% × [- 2, 0] /% × [- 1, 0]	左边第二个词及左边第一个词
U_{09}:% × [1, 0] /% × [2, 0]	右边第一个词及右边第二个词
U_{10}:% × [- 1, 0] /% × [1, 0]	左边第一个词及右边第一个词
U_{11}:% × [- 2, 1] /% × [- 1, 1]	左边第二个词性及左边第一个词性
U_{12}:% × [1, 1] /% × [2, 1]	右边第一个词性及右边第二个词性
U_{13}:% × [- 1, 1] /% × [1, 1]	左边第一个词性及右边第一个词性

3 实验及分析

本文使用 1998年上半年《人民日报》的标准语料做实验语料,共计 1 300万字. 为了验证模型的效果以及便于和文献[11]比较,实验使用该文献中用过的 10个词进行消歧实验(由于本实验所用语料规模的限制,另外 4个词:“变为,并不,好的,不变”的从合或从分例句数为 0,正确率都是 100%,故不列出). 对其中的部分词(与其,上来,的话,同一,个人)的训练语料进行了人工校对. 为了更好地验证模型的性能,进行 10折交叉验证. 把半年语料平均分成 10份,9份做训练, 1份做测试,轮流循环 10次. 具体实验步骤如下:

(1)根据 10个歧义字段分别从 9份训练和 1份测试语料里抽取训练样本集和测试集.

(2)把 10个词的训练和测试文件转化成 CRF序列文件.

(3)分别训练和测试上述的训练样本集和测试集,得到测试结果.

(4)读取测试结果,统计各个词的消歧正确率,结果如表 3.

表 3 实验字段测试结果

Table 3 Results of experiment

词语	总的例句数	从合例句数	从分例句数	文献[11]正确率/%	本文正确率/%
个人	2 073	1 860	213	98.16	99.12
研究所	612	611	1	96.56	99.85
一起	1 614	1 406	208	94.70	97.50
与其	99	27	72	97.38	88.94
决不	355	354	1	95.46	99.70
一生	364	361	3	97.32	98.98
上来	357	124	233	97.81	93.69
的话	566	138	428	89.78	91.49
同一	163	158	5	96.69	94.69
及其	815	812	3	89.42	99.58
平均正确率				95.33	96.35

由于与文献使用的语料不同,测试的方法也不同,结果难以直接比较,但我们在大规模真实语料中取得了这样的平均正确率,结果是令人满意的.

4 结语

本文基于 CRF模型,对组合型歧义进行了研究,实验表明该模型具有较高的歧义消解正确率.本文仅利用了上下文的词和词性特征,今后,将继续研究是否可以利用其它一些信息作为歧义消解的特征.把 CRF模型应用到对兼类词的消歧及命名实体识别的工作,进一步提高语料标注的质量.

[参考文献] (References)

[1] 刘开瑛,由丽萍. 汉语框架语义知识库构建工程[C]. 北京: 清华大学出版社, 2006: 64-71.
Liu Kaiying, You Liping. On Chinese FrameNet Construction[C]. Beijing: Tsinghua University Press, 2006: 64-71. (in Chinese)

[2] 孙茂松,黄昌宁,邹嘉彦. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展, 1997, 34(5): 332-339.
Sun Maosong, Huang Changning, Benjamin K Tsou. Using character bigram for ambiguity resolution in chinese word segmentation[J]. Computer Research and Development, 1997, 34(5): 332-339. (in Chinese)

[3] 孙茂松,左正平. 消解中文三字长交集型分词歧义的算法[J]. 清华大学学报, 1999, 39(5): 101-103.
Sun Maosong, Zuo Zhengping. Algorithm for solving 3-character crossing ambiguities in Chinese word segmentation[J]. Tsinghua Univ (Sci & Tech), 39(5): 101-103. (in Chinese)

[4] 廉竹钧. 汉语组合型切分歧义字段消歧方法研究[D]. 北京: 北京语言文化大学, 2002.
Lian Zhujun. A Study on the Disambiguation of Combinatorial Ambiguities in Chinese Word Segmentation[D]. Beijing: Beijing Language and Culture University, 2002. (in Chinese)

[5] 郑家恒,吴芳芳. 多义型歧义字段切分研究[C]. 北京: 清华大学出版社, 1999: 129-134.
Zhang Jiaheng, Wu Fangfang. Research on Multi-sense Type Ambiguous Phrases Segmentation[C]. Beijing: Tsinghua University Press, 1999. 129-134. (in Chinese)

[6] 肖云,孙茂松,邹嘉彦. 利用上下文信息解决汉语自动分词中的组合型歧义[J]. 计算机工程与应用, 2001, 37(19): 87-81.
Xiao Yun, Sun Maosong, Benjamin K Tsou. Solving combinatorial ambiguity in Chinese word segmentation using contextual information[J]. Computer Engineering and Application, 2001, 37(19): 87-81. (in Chinese)

[7] Luo Xiao, Sun Maosong, Tsou B K. Covering ambiguity resolution in Chinese word segmentation based on contextual information[C] // Proceedings of the 19th International Conference on Computational Linguistics. Taiwan: [s.n.], 2002: 598-604.

(下转第 94页)

[参考文献](References)

[1] Pan L Y, Song H, Ma F Y. A macrocommittees method of combining multistrategy classifiers for heterogeneous ontology matching[C]// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2004, 3129: 672-677.

[2] Stoilos G, Stamou G, Kollias S. A string metric for ontology alignment[C]// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer-Verlag, 2005, 3729: 624-637.

[3] Euzenat J, Valtchev P. Similarity-based ontology alignment in owl-lite[C]// Proc of the European Conference on Artificial Intelligence. Valencia: IOS Press, 2004: 333-337.

[4] Castano S, Ferrara A, Montanelli S. Matching ontologies in open networked systems: techniques and applications[J]. Journal on Data Semantics V, Lecture Notes in Computer Science, 2006, 3870: 25-63.

[5] Liya Fan, Tianyuan Xiao. FCA-mapping: a method for ontology mapping[C]// Proc of the 4th European Semantic Web Conference. Berlin, Heidelberg: Springer-Verlag, 2007: 192-206.

[6] Godin R, Missaoui R, Alaoui H. Incremental concept formation algorithms based on galois (concept) lattices[J]. Computational Intelligence, 1995, 11(2): 246-267.

[7] Anna Formica. Concept similarity in formal concept analysis: an information content approach[J]. Knowledge-Based Systems Archive, 2008, 21(1): 80-87.

[责任编辑: 顾晓天]

(上接第 76页)

[8] 曲维光, 吉根林, 穗志方, 等. 基于语境信息的组合型分词歧义消解方法[J]. 计算机工程, 2006, 32(17): 74-76
Xiao Yun, Sun Maosong, Benjamin K Tsou. Solving combinatorial ambiguity in Chinese word segmentation using contextual information[J]. Computer Engineering and Application, 2001, 37(19): 87-81. (in Chinese)

[9] 冯素琴, 陈惠明. 一种自组织的汉语组合型歧义消歧方法[J]. 计算机工程与设计, 2007, 28(3): 737-749, 742.
Feng Suqin, Chen Huiming. AdaptiveChinese combinatorial ambiguities disambiguate method[J]. Computer Engineering and Design, 2007, 28(3): 737-749, 742. (in Chinese)

[10] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the 18th ICML. San Francisco: Mogan Koufmann, 2001: 282-289.

[11] 冯素琴, 陈惠明. 基于语境信息的汉语组合型歧义消歧方法[J]. 中文信息学报, 2007, 21(6): 13-16, 42.
Feng Suqin, Chen Huiming. Context-based approach to combinational ambiguity resolution in Chinese word segmentation[J]. Journal of Chinese Information Processing, 2007, 21(6): 13-16, 42. (in Chinese)

[责任编辑: 孙德泉]