语义分析中谓词标识的特征工程

汪红林[1,2],王红玲[1,2],周国栋[1,2]
WANG Hong-lin[1,2],WANG Hong-ling[1,2],ZHOU Guo-dong[1,2]

1.苏州大学 计算机科学与技术学院,江苏 苏州 215006
2.江苏省计算机信息处理技术重点实验室,江苏 苏州 215006
1.School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China
2.Jiangsu Provincial Key Laboratory of Computer Information Processing Technology,Suzhou,Jiangsu 215006,China
E-mail:064227065055@suda.edu.cn

WANG Hong-lin,WANG Hong-ling,ZHOU Guo-dong.Feature engineering for predicate identification and classification in semantic analysis.Computer Engineering and Applications,2010,46(9):134-137.

Abstract: Predicate is the most important component in a sentence,which greatly influences the identification of the semantic analysis.The performance of predicate identification and classification relies on lots of features,but how to combine those features is more important.This paper picks out 7 basic features and over 30 new features with different combinations.By adding useful combinations of the features into the baseline system with the maximum entropy classifier,it improves by 5% of *F*1-score(from 84.7% up to 89.8%) on predicate identification and also gains about 2% increase of *F*1-score(from 80.3% up to 82.1%) on predicate classification.It shows that those new features and the combination of them can much improve the performance of the system.
Key words: predicate identification and predicate classification;semantic analysis;feature engineering;maximum entropy classifier

摘 要:谓词是句子中的最重要的成分,它的正确与否对语义分析的影响非常大。而众多的特征直接影响到谓词标识的性能,如何组织这些特征显得尤为重要。选取了7个基本特征和30多个新特征以及它们的组合,使用最大熵分类器,在基本特征的基础上通过增加有利特征的方法,使得谓词标注的*F*1值增长了约5%(由84.7%增加到89.8%),词义识别的*F*1值增长了约2%(由80.3%增加到82.1%),结果表明,这些新特征及其组合大大提高了性能。
关键词:谓词标注和词义识别;语义分析;特征工程;最大熵分类器
DOI:10.3778/j.issn.1002-8331.2010.09.038 **文章编号**:1002-8331(2010)09-0134-04 **文献标识码**:A **中图分类号**:TP18

1 引言

语义分析就是根据句子的句法结构和句中每个实词的词义,推导出能够反映句子意义的某种形式化表示。对句子进行正确的语义分析,一直是从事自然语言理解研究的学者们追求的主要目标。随着自然语言处理基础技术,如:中文分词、词性标注、句法分析、机器学习等的逐步成熟,以及语义分析在问答系统、信息抽取、机器翻译等领域的广泛应用,使得其越来越受到重视。

现在大多数的语义分析是以"谓词-论元"的结构驱动的,所以谓词的标识是进行语义分析的前提,它的正确与否直接影响到语义分析的性能。谓词可以是动词和名词,给定一个CoNLL2008 shared task WSJ语料的实例(1):

Meanwhile,overall evidence(.01) on the economy remains (.01)fairly clouded. (1)

实例(1)所构成的依存关系树如图1。

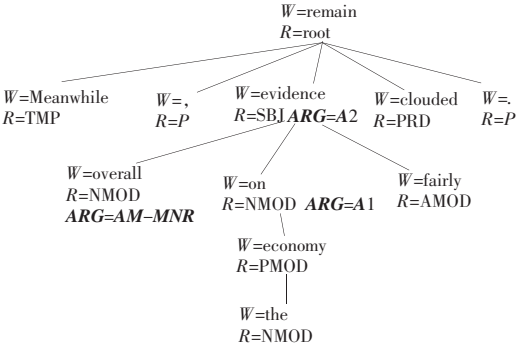

图1 实例(1)对应的依存树

图1中,用黑体字表示各依存关系承担的角色,*W*表示单词,*R*表示依存关系。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60673041);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z147);高等院校博士学科点专项科研基金(the China Specialized Research Fund for the Doctoral Program of Higher Education under Grant No.20060285008)。
作者简介:汪红林(1985-),男,硕士研究生,主要研究方向:自然语言处理;王红玲(1975-),女,博士研究生,主要研究方向:中文信息处理;周国栋(1967-),男,教授,博士生导师,主要研究方向:自然语言处理、信息抽取等。
收稿日期:2008-09-26 **修回日期**:2009-01-17

在实例(1)中，有两个谓词分别是：名词evidence 和动词remains，其后面括号里表示的是它们的词义，一般单词都是多义词，在语料库中对每个单词都有个词义排序，evidence(.01)表示在实例(1)中 evidence 是第一个词义，查词网得到第一个词义是“迹象，证据”。谓词标识的目的就是首先要找出句子中的谓词，然后再对所找到的谓词进行词义判别，最终确定它的意义。

2　相关工作

目前，谓词的标识并没有成为一项很大的任务，很多都是基于规则的标识方法，如：Lluís[1]等将句子中所有的动词都当作是谓词，除了助动词和被动语态中的谓词，对于名词而言，他将训练集中出现过的名词性谓词作为一个标准来参考，凡测试集中有与其中某个词相同的名词都作为谓词，没有出现过的名词不作为谓词，这样会遗漏很多名词，也会标错很多名词，性能不太好。

Wang[2]等也是将句子中所有的动词都当作谓词，通过一个词表，查表后将有可能充当动词的名词也当作谓词，和 Lluís[1]等取得的性能差不多。

Yuret[3]等在谓词标注阶段使用基于特征的方法，使用了一些基本特征，但在词义识别阶段却用的是基于规则的方法，它首先统计训练集中的单词词义数目，使用了出现频率最高的词义，如果单词没有在训练集中出现，则使用“.01”。

此外，基于统计的方法有：Ciaramita[4]等用了 7 种特征，如：谓词原型、词性、孩子结点数目等，在谓词标注和词义识别都用了同样的特征，在测试集(该文的测试集均指的是 CoNLL2008 shared task 提供的 WSJ 测试集)上得到的 $F1$ 值分别为 71.99%/70.11%。

Che[5]等和 Ciaramita[4]等利用了一部分相同的特征，但是特征数目更多，例如：加入了很多有关兄弟结点的信息，还有句法成分的信息，并且谓词的标注和词义识别使用了不同的特征，效果明显提高。

Morante[6]等使用了比较少的特征，但是将名词性谓词和动词性谓词进行了分开标注，在开发集(该文的测试集均指的是 CoNLL2008 shared task 提供的 WSJ 测试集)上的精确率分别是 $P(N)$=89.8%，$P(V)$=95.9%。

Watanabe[7]等使用了 CRF 分类器的方法，将谓词标注和词义识别分开进行，只使用了该文的几个基本特征，最终的 $F1$ 值是 79.02%。

3　系统描述

本实验分为两步：谓词标注和词义识别，后者在前者所标注的谓词的基础上继续进行词义的分配，利用最大熵分类器进行分类，实验数据分别为：CoNLL2008 shared task(http://www.yr-bcn.es/conll2008/)提供的 WSJ 训练集 39 280 句，其中谓词数有 185 138 个，WSJ 测试集 2 400 句，其中含有的谓词数为 10 818 个。

3.1　特征选取

共使用了 30 多个特征，除了相同的 7 个基本特征之外，此两步所用的特征不完全相同。参考了 Ciaramita[4]等和 Che[5]等的工作，最终的性能是基本特征加上有利特征得到的。假设实例(1)所构成的依存树中，当前单词为 remains，现将各特征列举如下，特征如果不存在就用 NULL 代替。

3.1.1　基本特征

依存关系：指当前单词与其父亲单词的关系，实例(1)中本特征为：root。

单词本身：指当前单词本身，实例(1)中本特征为：remains。

单词词性：指当前单词词性，如果有 gold 词性，则使用 gold 词性，否则用自动分析的词性。实例(1)中本特征为：VBZ(gold 词性)。

单词原型：指当前单词的原型，实例(1)中本特征为：remain。

中心词本身：指当前结点的中心词，实例(1)中本特征为：NULL。

中心词词性：指当前结点的中心词词性，实例(1)中本特征为：NULL。

中心词原型：指当前结点的中心词原型，实例(1)中本特征为：NULL。

3.1.2　谓词标注的特征

分离词单词原型：指当前词如果是连接词，如“Atlanta-based”，则将此单词分为 3 个单词分开处理：“Atlanta”，“-”，“based”。此处窗口大小为 1，假设当前结点编号为 i，则 FI1 包含的单词原型有：i+(i−1)，i−1，i，i+1，i+(i+1)。实例(1)中本特征为：(remain+economy)economy remain fairly(remain+fairly)。

分离词单词本身：意义同上，实例(1)中本特征为：(remains+economy)economy remains fairly(remains+fairly)。

预测的词性：指当前结点自动分析的词性，窗口大小为 2，假设当前结点编号为 i，包含的单词词性有：(i−2)+(i−1)，(i−1)+i，i−1，i，i+1，i+(i+1)，(i+1)+(i+2)。实例(1)中本特征为：DT+NN NN+VBZ NN VBZ RB VBZ+RB RB+VBN。

单词形式：指当前单词的变现形式。如：单词“Brazil”表示为“Xx*”，窗口大小为 2，假设当前结点编号为 i，包含的单词有：(i−2)+(i−1)，(i−1)+i，i−1，i，i+1，i+(i+1)，(i+1)+(i+2)。实例(1)中本特征为：x*+x* x*+x* x* x* x* x*+x* x*+x*。

孩子数：指当前结点的单词数目。实例(1)中本特征为：5。

孩子相关特征：指当前结点孩子的分离词单词原型、预测词性、依存关系、当前结点的分离词单词原型+当前结点孩子的分离词单词原型，当前结点的预测的词性+当前结点孩子的预测的词性。实例(1)中本特征为：meanwhile RB TMP meanwhile+remain RB+VBZ，P_，+remain P+VBZ evidence NN SBJ evidence+remain NN+VBZ cloud VBN PRD cloud+remain VBN+VBZ.P_.+remain P+VBZ。

父子位置差：指当前结点孩子与其本身的编号之差。例如：当前结点编号为 8，它的 5 个孩子结点的编号分别为：1，2，4，10，11。所以实例(1)中本特征为：−7 −6 −4 2 3。

成分首单词本身：指当前结点的句法成分的第一个单词。实例(1)中本特征为:overall。

成分首单词词性:指当前结点的句法成分的第一个单词词性。实例(1)中本特征为:JJ。

成分首单词原型:指当前结点的句法成分的第一个单词原型。实例(1)中本特征为:overall。

成分末单词本身:指当前结点的句法成分的最后一个单词本身。实例(1)中本特征为:fairly。

成分末单词词性:指当前结点的句法成分的最后一个单词词性。实例(1)中本特征为:RB。

成分末单词原型:指当前结点的句法成分的最后一个单词原型。实例(1)中本特征为:fairly。

成分词性链:指当前结点的句法成分所包含的单词的词性(除了第一个和最后一个单词词性)。实例(1)中本特征为:DT。

孩子词性链:指当前结点的孩子结点的词性链。实例(1)中本特征为:RB+,+NN+VBN+.。

孩子词性链(*N*):指当前结点的孩子结点的词性链(遇到重复词性则只能用一次)。实例(1)中本特征为:RB+,+NN+VBN+.。

孩子依存关系:指当前结点的孩子结点的依存关系链。实例(1)中本特征为:TMP+P+PRD+SBJ+P。

孩子依存关系(*N*):指当前结点的孩子结点的依存关系链(遇到重复关系则只能用一次)。实例(1)中本特征为:TMP+PRD+SBJ+P。

兄弟依存关系链:指当前结点的兄弟结点的依存关系链。实例(1)中本特征为:NULL。

兄弟依存关系链(*N*):指当前结点的兄弟结点的依存关系链(遇到重复的依存关系只能取一次)。实例(1)中本特征为:NULL。

兄弟词性链:指当前结点的兄弟结点的词性链。实例(1)中本特征为:NULL。

兄弟词性链(*N*):指当前结点的兄弟结点的词性链(遇到重复词性则只取一次)。实例(1)中本特征为:NULL。

动词语态:如果当前结点是动词,并且其中心词是"be"或者"get",或者其关系类型是"APPO",则语态为"passive",否则是"active"。如果当前结点是名词,则语态为"NULL"。实例(1)中本特征为:active。

除此之外,还有很多的组合特征:分离词单词原型+分离词单词本身,分离词单词本身+孩子数,分离词单词本身+孩子依存关系,分离词单词本身+孩子依存关系,孩子依存关系+兄弟依存关系链,兄弟依存关系链+兄弟词性链。

3.1.3 词义识别的特征

训练集中共有962个多义谓词,占整个训练集的14%,除了使用谓词标注的相关特征以外，还开发了另外的新的特征,它们分别为:

句中所有单词:本句中所有的单词组合,实例(1)中本特征为:Meanwhile+,+overall+evidence+on+the+economy+remains+fairly+clouded+。

单词本身(*P*):意义同上,但是在单词之前加上一标记,"L_"表示此单词在谓词之前,"R_"表示在谓词之后,"T_"表示谓词本身。实例(1)中本特征为:L_Meanwhile + L_,+ L_overall+ L_evidence+ L_on+ L_the+ L_economy+ T_remains+R_fairly+ R_clouded + R_。

单词词性(*P*):本句中所有单词词性组合,但是在单词词性之前加上一标记,"L_"表示此单词在谓词之前,"R_"表示在谓词之后,"T_"表示谓词本身。实例(1)中本特征为:L_RB+L_JJ+L_NN+L_IN+L_DT+L_NN+T_VBZ+R_RB+R_JJ+R_。

窗口为11的单词:当前单词的前后5个单词组合(包含当前单词本身)。实例(1)中本特征为:over+evidence+on+the+economy+remains+fairly+clouded+。

3.2 实验结果与分析

3.2.1 基础系统

无论是谓词标注还是词义识别,基础系统都是基于上述7个基本特征的,然后逐次测试各个特征的性能,选取有用的特征。基础系统在测试集上的*P*/*R*/*F*为:88.4%/81.2%/84.7%。表1列举了谓词标注中的单个特征的性能,黑体字表示性能增长明显的特征。

表1 谓词标注基础特征上分别增加不同特征后的性能表 (%)

特征	*P*	*R*	*F*	特征	*P*	*R*	*F*
+分离词单词原型	**92.5**	**84.5**	**88.3**	**+孩子词性链**	**92.5**	**84.1**	**88.1**
+分离词单词本身	**89.8**	**82.6**	**86.1**	**+孩子词性链(*N*)**	**92.5**	**84.1**	**88.1**
+预测的词性	**91.5**	**83.8**	**87.5**	**+孩子依存关系**	**90.9**	**83.1**	**86.8**
+单词形式	88.7	81.6	85.1	**+孩子依存关系(*N*)**	**90.9**	**83.1**	**86.8**
+孩子数	89.6	82.1	85.7	+兄弟依存关系链	88.4	80.8	84.4
+孩子相关特征	**93.1**	**84.6**	**88.6**	+兄弟依存关系链(*N*)	88.4	80.9	84.4
+父子位置差	89.7	82.4	85.9	+兄弟词性链	88.5	81.1	84.6
+成分首单词本身	89.1	82.0	85.4	+兄弟词性链(*N*)	88.6	81.2	84.7
+成分首单词词性	89.1	81.9	85.4	+动词语态	88.4	81.2	84.6
+成分首单词原型	89.2	82.1	85.6	+成分末单词原型	90.0	81.8	85.8
+成分末单词本身	89.9	82.0	85.8	+成分词性链	90.4	82.3	86.1
+成分末单词词性	90.0	82.5	86.1				

组合特征	*P*	*R*	*F*
+分离词单词原型+分离词单词本身	88.9	81.1	84.8
+分离词单词本身+孩子数	89.1	81.2	85.0
+分离词单词本身+孩子依存关系	90.0	82.1	85.9
+分离词单词本身+孩子依存关系(*N*)	**90.0**	**82.8**	**86.5**
+孩子依存关系+兄弟依存关系链	**90.5**	**82.8**	**86.5**

表1中,几乎所有的单个特征都会增加性能,最高增加了3.9%(孩子相关特征),只有少数几个特征使性能下降(兄弟词性链等),可以看出,它们是:分离词单词原型,分离词单词本身,孩子相关特征,孩子词性链,孩子词性链(*N*),孩子依存关系,孩子依存关系(*N*),分离词单词本身+孩子依存关系(*N*),孩子依存关系+兄弟依存关系链。

据统计,训练集中有近4 000个连字(如:Atlanta-based),这些单词都必须分开处理,而分离词单词原型和分离词单词本身分别表明了这些分离词的特征,故效果明显。一般中心词前后的几个词都是修饰词,修饰词对中心词的标注影响也很大,预测的词性恰好显示了这一特性。和预测的词性相比,孩子相关特征更加准确地表达了修饰词(孩子结点)的信息,故性能提高得最明显。其他几个特征也都具有相似特性。

词义识别基础系统的 *P*/*R*/*F*:91.6/71.3/80.3,表 2 列举了词义识别中分别加入单个特征的性能。

表 2 词义识别基础特征上分别增加不同特征后的性能表 (%)

特征	*P*	*R*	F	特征	*P*	*R*	*F*
+分离词单词原型	**93.7**	**72.9**	**82.0**	+孩子词性链	**91.9**	**71.6**	**80.5**
+分离词单词本身	91.8	71.4	80.4	+孩子词性链(*N*)	**91.9**	**71.6**	**80.5**
+预测的词性	91.9	71.5	80.4	+孩子依存关系	**92.2**	**71.8**	**80.8**
+单词形式	91.7	71.3	80.3	+孩子依存关系(*N*)	**92.2**	**71.8**	**80.8**
+孩子数	91.7	71.3	80.3	+兄弟依存关系链	91.6	71.3	80.2
+孩子相关特征	**93.1**	**72.5**	**81.6**	+兄弟依存关系链(*N*)	91.6	71.3	80.2
+父子位置差	91.6	71.3	80.3	+兄弟词性链	91.6	71.3	80.2
+成分首单词本身	**91.8**	**71.5**	**80.4**	+兄弟词性链(*N*)	91.6	71.3	80.2
+成分首单词词性	**91.8**	**71.5**	**80.4**	+动词语态	91.8	71.4	80.4
+成分首单词原型	91.6	71.3	80.2	+句中所有单词	91.6	71.0	80.1
+成分末单词本身	91.6	71.3	80.2	+单词本身(*P*)	91.6	71.2	80.2
+成分末单词词性	91.6	71.3	80.2	+单词词性(*P*)	91.0	71.1	80.0
+成分末单词原型	91.6	71.3	80.2	+窗口为 11 的单词	**91.7**	**71.5**	**80.4**
+成分词性链	91.6	71.3	80.2	+孩子词性链+动词语态	**91.8**	**71.5**	**80.4**

从表 2 可以看出,大多数的特征都没有起到提高性能的作用,很多特征反而会使得性能下降(如:句中所有单词等),究其原因,词义和句子的其他结构没有关系,只与谓词本身的几个特性有关系,如谓词的词性,谓词依存关系等等,所以最终在词义识别中并没有采用更多的特征,只保留了一些基本的特征和少数几个使性能提高的特征。

3.2.2 最佳系统

通过对基础系统的不断测试,最终谓词标注选择的特征有:7 个基本特征、分离词单词原型、分离词单词本身、预测的词性、孩子数、孩子相关特征、成分首单词本身、成分首单词词性、成分首单词原型、成分末单词本身、成分末单词词性、成分末单词原型、孩子数、孩子词性链(*N*)、孩子依存关系、孩子依存关系(*N*)、兄弟依存关系链、兄弟依存关系链(*N*)、兄弟词性链、兄弟词性链(*N*)、动词语态、分离词单词原型+分离词单词本身、分离词单词本身+孩子数、分离词单词本身+孩子依存关系、分离词单词本身+孩子依存关系(*N*)、孩子依存关系+兄弟依存关系链。而词义识别选择的特征有:7 个基本特征、成分词性链、孩子词性链、孩子依存关系、兄弟词性链、动词语态、窗口为 11 的单词。表 3 列举了最佳系统的性能。

表 3 谓词标注和词义识别的最佳性能表 (%)

类别	*P*	*R*	*F*1
谓词标注(PI)	94.1	85.9	89.8
词义识别(PC)	93.8	73.0	82.1

作为 CONLL2008 shared task 中的一项子任务,参加了开放测试,并取得了较好的结果,与 Ciaramita[4]等(PI 和 PC 的 *F*1 值分别为:84.87%和 78.94%)相比,性能有了很大的提高,关键还是特征的作用,因为除了使用与 Ciaramita[4]等相同的特征外,还开发了很多新的特征。Watanabe[7]等在谓词标识过程中用到了 CRF 分类器和 K 邻近算法,最终获得 PI 和 PC 的平均 *F*1 值是:79.02%,还是比较低,再次说明了特征的重要性。

4 结论和展望

选取了依存分析和句法分析中常用的近 30 个特征进行了谓词标识实验,结果证明,随着有利特征的增加性能明显提高,PI 的 *F*1 值增加了 5 个百分点,PC 的 *F*1 值也增加了 2 个百分点,此外,实验表明特征的组合也很重要。未来,还要加入更多的规则,如:被动语态的动词肯定不是谓词,作为修饰词的动词的过去分词也不是谓词等等,从 Lluís[1]等可以看出纯粹利用规则也取得了不错的性能。此外,还要将名词和动词分开处理,因为大部分的动词都充当谓词,而名词则反之。相信结合这两点,性能还会提高不少。

参考文献:

[1] Lluís X,Màrquez L.A joint model for parsing syntactic and semantic dependencies[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:188-192.

[2] Wang Hong-ling,Wang Hong-lin,Zhou Guo-dong.Dependency tree-based SRL with proper pruning and extensive feature engineering[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:253-257.

[3] Yuret D,Yatbaz M A,Ural A E.Discriminative vs generative approaches in semantic role labeling[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:223-227.

[4] Ciaramita M,Attardi G,Dell'Orletta F.DeSRL:A linear-time semantic role labeling system[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:258-262.

[5] Che Wan-xiang,Li Zheng-hua,Hu Yu-xuan,et al.A cascaded syntactic and semantic dependency parsing system[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:238-242.

[6] Morante R,Daelemans W,Van Asch V.A combined memory-based semantic role labeler of English[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:208-212.

[7] Watanabe Y,Iwatate M,Asahara M,et al.A pipeline approach for syntactic and semantic dependency parsing[C]//Proceedings of the 12th Conference on Computational Natural Language Learning,Manchester,2008:228-232.

[8] Gildea D,Jurafsky D.Automatic labeling of semantic roles[J].Computational Linguistics,2002,28(3):245-288.

[9] Gildea D,Palmer M.The necessity of syntactic parsing for predicate argument recognition[C]//Proceedings of ACL-2002,Philadelphia,PA,2002:239-246.

[10] Surdeanu M,Harabagiu S,Williams J,et al.Using predicate-argument structures for information extraction[C]//Proceedings of ACL-2003,Sapporo,Japan,2003.