语义分析和结构化语言模型*

李明琴[1+], 李涓子[2], 王作英[1], 陆大绘[1]

[1](清华大学 电子工程系,北京 100084)

[2](清华大学 计算机科学与技术系,北京 100084)

Semantic Analysis and Structured Language Models

LI Ming-Qin[1+], LI Juan-Zi[2], WANG Zuo-Ying[1], LU Da-Jin[1]

[1](Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62781704, E-mail: lmq@thsp.ee.tsinghua.edu.cn, http://www.tsinghua.edu.cn

Received 2004-05-14; Accepted 2004-09-07

Li MQ, Li JZ, Wang ZY, Lu DJ. Semantic analysis and structured language models. *Journal of Software*, 2005,16(9):1523−1533. DOI: 10.1360/jos161523

Abstract: An integrated semantic analysis system is presented, and the structured language models are proposed based on it. The semantic analysis system can automatically tag semantic class for each word and analyze the semantic dependency structure between words with the precision of 90.85% and 75.84% respectively. In order to describe sentence structure and long-distance dependency, two kinds of structured language models are examined and analyzed. Finally, these two language models are evaluated on the task of Chinese speech recognition. Experiments show that the best semantic structured language model—headword trigram model—achieves 0.8% absolute error reduction and 8% relative error reduction over the trigram model.

Key words: semantic analysis; dependency analysis; language model; speech recognition

摘 要: 提出了一个语义分析集成系统,并在此基础上构建了结构化的语言模型.该语义分析集成系统能够自动分析句子中各个词的词义以及词之间的语义依存关系,达到 90.85%的词义标注正确率和 75.84%的语义依存结构标注正确率.为了描述语言的结构信息和长距离依存关系,研究并分析了两种基于语义结构的语言模型.最后,在中文语音识别任务上测试两类语言模型的性能.与三元语言模型相比,性能最好的语义结构语言模型——中心词三元模型,使绝对字错误率下降 0.8%,相对错误率下降 8%.

关键词: 语义分析;依存分析;语言模型;语音识别

中图法分类号: TP18 文献标识码: A

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2001AA114071 (国家高技术研究发展计划(863))

作者简介: 李明琴(1977−),女,吉林和龙人,博士生,主要研究领域为语音识别,语言模型,中文语义分析;李涓子(1964−),女,博士,副教授,CCF 高级会员,主要研究领域为自然语言理解,中文信息处理,网络中的知识挖掘和管理;王作英(1935−),男,博士,教授,博士生导师,主要研究领域为语音识别;陆大绘(1928−),男,教授,博士生导师,主要研究领域为语音识别.

虽然 N 元语法模型广泛地应用于语音识别、中文输入法等任务中[1],并取得了较好的效果,但是它本身仍然存在着局限性,即它只能描述句子中词之间的线性关系,无法利用自然语言中丰富的语法、语义结构.因此,许多学者一直致力于研究如何将自然语言结构信息融入到语言模型中.

目前,改进语言模型的方法大致可以分为两类.一类是利用浅层语法、语义结构信息的语言模型,它们只引入较少的语法、语义知识,不定义显性的语法结构.如基于词类的语言模型[2]、基于 Trigger 的语言模型[3]、潜在语义分析模型[4]、Skipping 语言模型[5]等.另一类是基于语法或者语义结构的语言模型,这类模型从分析句子语法、语义结构入手构建语言模型,它们更多地利用了语言的结构信息.如结构语言模型[6,7]、基于自上而下句法分析器的语言模型[8]、无监督学习的依存结构模型[9]等.

本文在早期语义依存分析研究工作[10]的基础上,实现了一个完整的中文语义分析系统,并且构建了两类基于语义依存结构的语言模型.文中的语义分析系统能够自动分析句子中各个词的词义和句子中词之间的语义依存关系,达到 90.85%的词义标注正确率和 75.84%的语义依存结构标注正确率.在语义分析的基础上,本文提出了两类结构化的语言模型——最优标注句子模型和中心词三元模型,并详细讨论了两类模型的特点.最后,在中文语音识别任务上,测试两类语言模型的性能.与三元语言模型相比,两类结构化语言模型都降低了字错误率,其中性能最好的语言模型使绝对字错误率下降 0.8%,相对错误率下降 8%.

本文第 1 节首先介绍本文语义分析的语言学基础,然后描述整个语义分析系统的结构和统计模型.在此基础上,第 2 节提出两类基于语义结构的语言模型.第 3 节测试并分析语义分析系统和结构语言模型的性能.第 4 节给出与其他相近工作的比较.最后一节对所做的工作进行讨论和总结.

1 语义分析

1.1 语言学基础

理解一个句子需要理解句子中每个词的意义和句子中词之间的关系.

学术界一般认为词的意义(简称词义)是将字符串映射到某个概念上.在本语义分析系统中,采用了《同义词词林》[11]中定义的词义体系,将词映射到 1 343 个词义类上.

句子中词的关系主要体现了句法学、语义学约束,在语言理解中发挥着重要的作用,我们用语义依存语法[12]来描述.依存语法认为,词之间的关系是有方向的,通常是一个词支配另一个词,这种支配与被支配的关系就称作依存关系.依存关系既可以是句中词与词之间的句法关系,也可以是语义关系.本文中的语义依存语法主要关注语义,它定义了 59 个语义关系,占整个关系标注集(共 70 个关系)的绝大多数.

在语义依存语法中,支配词又称为被支配词的中心词.中心词通常可以表现它所在短语的主要语法、语义特征,例如,动词、名词短语中的中心词是动词、名词,方位词短语的中心词是地点名词.短语间的支配、被支配关系由短语的中心词间的支配、被支配关系表示.以图 1 中的句子为例,其中修饰词用带箭头的实线指向中心词,中心词用粗实线相连.“杨 博士”(空格表示词的边界)是句子的主语,“重视”是谓语动词,“杨 博士”是动作“重视”的经验者.在语义依存语法中,表示为“博士”依存于“重视”,它们的语义关系为“经验者/Experiencer”.同时,用“杨”和“博士”之间的“限定”关系表示“重视”的经验者是某个姓杨的博士,而不是其他博士.

语义依存语法不仅描述句子主干的词间关系,而且描述句子的短语成分内部的细节结构,例如,在短语“其 发明 成果 的 推广 使用”中,“其 发明 成果 的”修饰“推广 使用”,指出被“推广 使用”的内容为“其 发明 成果”,因此,在语义依存语法中表示为“成果”和“推广”之间的“内容/content”关系,“推广”是中心词.

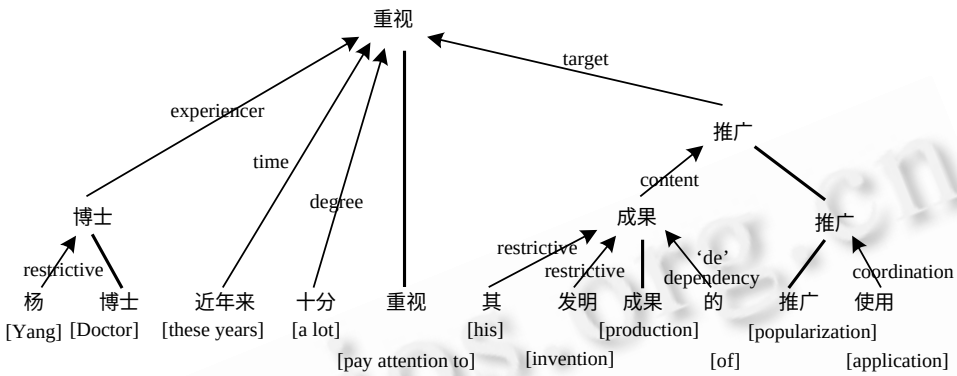
严格地说,对于长度为 N 的句子 $W=\{w_1,w_2,...,w_N\}$,语义依存语法定义了一个语义依存关系表 $SRL=\{SR(1),SR(2),...,SR(N)\}$,其中 $SR(i)=(H_i,R_i)$.SR(semantic relation)表示句子中第 i 个词的中心词为第 H_i 个词,它们之间的语义关系为 R_i.如果句子中第 j 个词是整个句子的中心词,那么 $SR(j)=(H_j,R_j)=(-1,$‘kernel word’).例句的语义依存关系表如图 2 所示,它与图 1 的语义依存树是完全一致的.

句子(w_i/s_i): 杨/Dd15 博士/Ae13 近年来/Ca11 十分/Ka01 重视/Gb21 其/Aa04 发明/Hc05 成果/Da14 的/Kd01 推广/Ie01 使用/Hj28

English: These years, Doctor Yang pays a lot of attention to the popularization and application of his invention.

(a) The sentence tagged with semantic classes

(a) 带有语义标注的句子



(b) The semantic dependency tree

(b) 语义依存树

Fig.1 Semantic dependency structure of a Chinese sentence

图 1 句子的语义依存结构

句子(w_i/s_i): 杨/Dd15 博士/Ae13 近年来/Ca11 十分/Ka01 重视/Gb21 其/Aa04 发明/Hc05 成果/Da14 的/Kd01 推广/Ie01 使用/Hj28

English: These years, Doctor Yang pays a lot of attention to the popularization and application of his invention.

(a) The sentence tagged with semantic classes

(a) 带有语义标注的句子

-1		SR(*i*)		
Modifier (*i*)		HeadWord (H_i)		Semantic relation (R_i)
Index	Word	Index	Word	
1	杨/Yang	2	博士/Doctor	限定/Restrictive
2	博士/Doctor	5	重视/pay attention to	经验者/Experiencer
3	近年来/these years	5	重视/pay attention to	时间/Time
4	十分/a lot	5	重视/pay attention to	程度/Degree
5	重视/pay attention to	-1	-1	核心成分/Kernel word
6	其/his	8	成果/production	限定/Restrictive
7	发明/invention	8	成果/production	限定/Restrictive
8	成果/production	10	推广/popularization	内容/Content
9	的/of	8	成果/product	'的'字依存/'De' dependency
10	推广/popularization	5	重视/pay attention to	目标/Target
11	使用/application	10	推广/popularization	连接/Coordination

(b) Semantic dependency relation list

(b) 语义依存关系表

Fig.2 Semantic dependency relation list of a Chinese sentence

图 2 句子的语义依存关系表

1.2 语义分析模型

1.2.1 基本模型

令句子 $W=\{w_1,w_2,...,w_N\}$ 相应于语义类串 $S=\{s_1,s_2,...,s_N\}$ 和语义依存树 $T=SRL=\{SR(1),SR(2),...,SR(N)\}$,语义分析模型 $P(T,S,W)$ 可以按从左向右的顺序,依次预测每个词的词义、词、依存结构,如式(1)所示.

$$P(T,S,W)=\prod_{k=1}^{N}\{P(s_k\mid T_{k-1},S_{k-1},W_{k-1})P(w_k\mid T_{k-1},S_k,W_{k-1})P(T_{k-1}^{k}\mid T_{k-1},S_k,W_k)\} \quad (1)$$

其中, W_{k-1} 为前 $k-1$ 个词对应的词串, S_{k-1} 为前 $k-1$ 个词对应的词义串, T_{k-1} 为前 $k-1$ 个词生成的部分语义依存结构, T_{k-1}^k 为分析第 k 个词时新增加的部分语义依存结构, $T_k = T_{k-1} \cup T_{k-1}^k$.

式(1)中模型参数空间太大,很难准确估计模型参数.因此,为了充分而有效地估计模型参数,我们选择鉴别力强的特征作为模型的历史等价类:

$$P(s_k \mid T_{k-1}, S_{k-1}, W_{k-1}) = P(s_k \mid s_{k-1}, s_{k-2}) \tag{2}$$

$$P(w_k \mid T_{k-1}, S_k, W_{k-1}) = P(w_k \mid s_k, w_{k-1}) \tag{3}$$

上面两式分别被称为语义预测模型和词预测模型.语义依存关系分析模型 $P(T_{k-1}^k \mid T_{k-1}, S_k, W_k)$ 将在第 1.3 节详细讨论.

1.2.2 对偶模型

语义分析模型 $P(T,S,W)$ 的分解顺序可以调换一下,即先预测词,再预测词义,最后预测语义依存关系.该模型为式(1)的对偶模型.

$$P_c(T,S,W) = \prod_{k=1}^{N} \{P_c(w_k \mid T_{k-1}, S_{k-1}, W_{k-1}) P_c(s_k \mid T_{k-1}, S_{k-1}, W_k) P(T_{k-1}^k \mid T_{k-1}, S_k, W_k)\} \tag{4}$$

它的历史等价类定义为

$$P_c(w_k \mid T_{k-1}, S_{k-1}, W_{k-1}) = P(w_k \mid w_{k-1}, w_{k-2}) \tag{5}$$

$$P_c(s_k \mid T_{k-1}, S_{k-1}, W_k) = P(s_k \mid w_k, s_{k-1}) \tag{6}$$

为了简便起见,称上一节中的模型为 SWT 模型,本节中的对偶模型为 WST 模型.

1.3 语义依存关系分析模型

在描述语义依存关系模型之前,我们先说明语义依存树分析过程.语义依存树的分析过程可以用二叉语义依存树来描写,如图 3 所示,其中修饰词用带箭头的实线指向中心词,中心词用粗实线相连.首先,合并互相依存的相邻词,生成一个新结点,同时在新结点上标注两个词的合并关系和中心词(为了方便起见,称为新结点 i 的合并操作 r_i 和中心词 hw_i 及中心词语义类 hs_i).然后,新结点与其他相邻的结点合并,生成更大的结点.重复上述合并操作,直至生成一个包含全句的结点.合并操作一般写成{“依存关系”,“左合并”/“右合并”}的形式,其中“依存关系”可以是语义依存语法中定义的 70 个关系中的任意一个.“左合并”/“右合并”表示中心词的来源,即“左合并”表示新结点的中心词继承左子结点的中心词;反之,为“右合并”.特别地,定义空合并操作(NULL)表示叶子结点 k 的合并操作 r_k = NULL.

在某些语义依存树中,左、右相邻的结点都直接依存于中间结点,例如,图 1 中“发明”和“的”同时依存于“成果”,此时,它们的合并顺序是任意的,例如,图 3 中“成果”先合并“发明”,但是先合并“的”也是允许的.因此,同一个语义依存树可以对应多个不同的二叉语义依存树,然而,一个二叉语义树则只能对应唯一一个语义依存树.式(1)中的树 T 可以是多叉语义依存树或者二叉语义依存树.此后,文本中的 T 特指二叉语义依存树.

式(1)中的语义依存分析模型 $P(T_{k-1}^k \mid T_{k-1}, S_k, W_k)$ 可以定义为生成部分依存结构 T_{k-1}^k 的合并操作概率乘积:

$$P(T_{k-1}^k \mid T_{k-1}, S_k, W_k) = \prod_{i=1}^{\tau_k} P(q_i \mid T_{k-1}, S_k, W_k, q_1, ..., q_{i-1}), \tag{7}$$

其中, q_i 为第 i 个合并操作. τ_k 表示分析第 k 个词时新增加的合并操作数,即 $|T_{k-1}^k|$.例如,在图 3 中分析词“成果”时,将新增加 2 个合并操作,分别为“发明”和“成果”之间的操作(*Restriction*,right)、“其”和“成果”之间的操作(*Restriction*,right).

由于模型 $P(q_i \mid T_{k-1}, S_k, W_k, q_1, ..., q_{i-1})$ 中包含的参数非常多,不易直接估计模型概率.因此,我们假设操作 q_i 只与被合并左、右子结点 l_i, g_i 有关,并且 q_i 只与左、右子结点 l_i, g_i 的合并操作和中心词有关——$P(q_i \mid r_{l_i}, r_{g_i})$ 和 $P(q_i \mid \langle hw_{l_i}, hs_{l_i} \rangle, \langle hw_{g_i}, hs_{g_i} \rangle)$.定义合并操作的得分为

$$Score(q_i \mid T_{k-1}, S_k, W_k, q_1, ..., q_{i-1}) = (1-\lambda) \log P(q_i \mid \langle w_{l_i}, s_{l_i} \rangle, \langle w_{g_i}, s_{g_i} \rangle) + \lambda \log P(q_i \mid r_{l_i}, r_{g_i}) \tag{8}$$

其中, λ 为加权系数,在实验中 $\lambda = 0.3$.关于模型 $P(q_i \mid r_{l_i}, r_{g_i})$ 和 $P(q_i \mid \langle hw_{l_i}, hs_{l_i} \rangle, \langle hw_{g_i}, hs_{g_i} \rangle)$ 的训练和平滑问题已

经在文献[10]中做了详细的讨论,在此不再赘述.

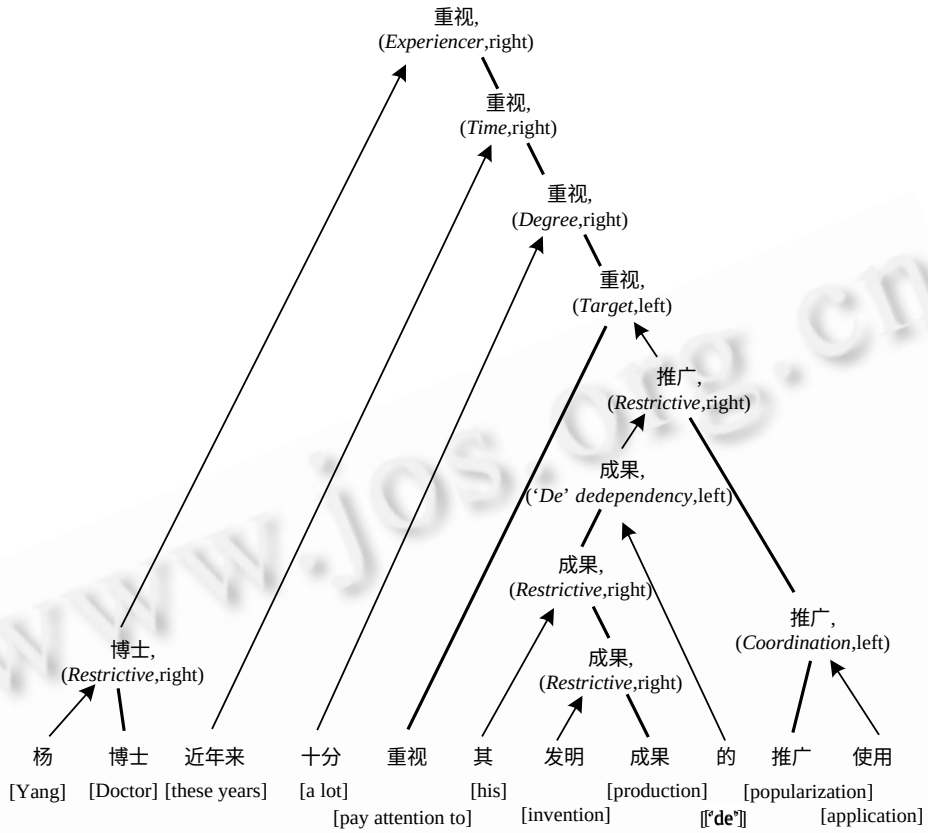


Fig.3 One of the binary semantic parse trees for the semantic dependency tree in Fig.1

图 3 图 1 中语义依存树对应的多个二叉语义依存树之一

2 基于语义依存结构的语言模型

本文尝试了两种结构化语言模型,以便描述句子中的结构信息和长距离依存关系.一种是由语义分析模型的概率直接推导句子概率,这种方法很直观,但是语言模型的性能直接受到语义分析模型的限制.另一种是从语义结构中提取特征估计句子的概率,例如结构中心词三元模型.

2.1 模型1:最优标注句子模型

句子 W 的概率为词串 W 与所有可能的语义串 S 和语义依存树 T 的联合概率的边际分布.

$$P_1(W)=\sum_{T,S}P(T,S,W) \tag{9}$$

由于语义分析过程[10]中使用了剪枝技术,使得一些可能的语义类串和语义依存树被剪掉,所以式(9)只能改写为

$$P_1(W)\approx\sum_{i=1}^{Pr}P(T^{(i)},S^{(i)},W) \tag{10}$$

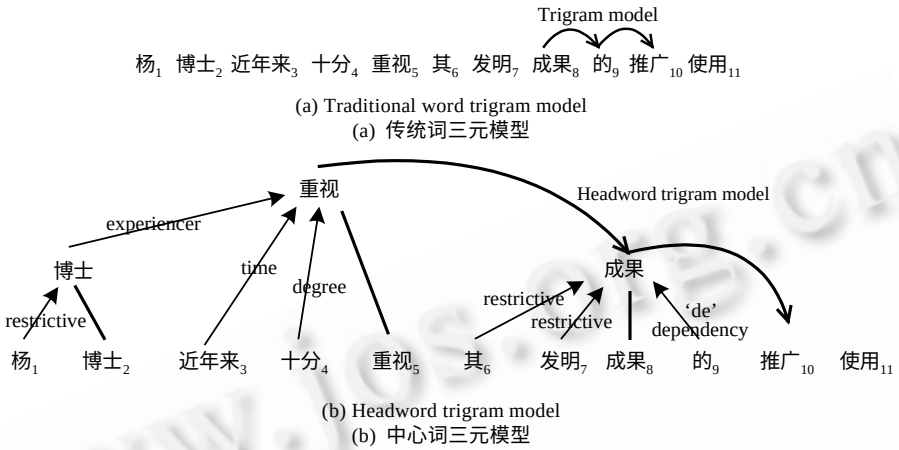
其中, Pr 为剪枝后保留的路径数.实际上,在句子整体模型(式(10))中最优标注的概率占 $P_1(W)$ 的主要部分,所以通常以最优标注的概率近似句子整体模型,即令 $Pr=1$,

$$P_1(W)\approx P(T^*,S^*,W) \tag{11}$$

其中, $T^*,S^*=\arg\max_{T,S}P(T,S,W)$.后面的实验也表明,最优标注句子模型(式(11))的性能和句子整体模型(式(10))是基本相当的.

2.2 模型2:中心词三元模型

中心词通常可以代表短语的主要语法、语义特征,它被认为具有较强的预测能力[6].而且,中心词预测结构通常跳过部分助词和副词,使长距离约束发挥作用.例如在图 4 中,前 9 个词构成的部分依存结构突出中心词“重视”和“成果”,我们直觉认为它们将比传统词三元模型中的条件“成果”和“的”的预测能力更强.



(a) Traditional word trigram model
(a) 传统词三元模型

(b) Headword trigram model
(b) 中心词三元模型

Fig.4

图 4

对于给定的词串,语义依存分析器能够给出多个依存结构候选和相应的得分.语言模型可以利用全部的候选结构预测下一个词,

$$P_2(w_k \mid W_{k-1}) = \sum_{T_{k-1},S_{k-1}} P_{WP}(w_k \mid T_{k-1},S_{k-1},W_{k-1})P(T_{k-1},S_{k-1},W_{k-1}) \tag{12}$$

这里的词预测模型 P_{WP} 可以完全不同于分析器模型(式(1)、式(4)中的词预测模型).本文选择了前 $k-1$ 个词生成的语义依存结构中突现出来的最后两个中心词 $h_{-1}^{k-1}, h_{-2}^{k-1}$ 来预测下一个词,即用 k 位置前面两个中心词来预测下一个词.我们称 P_{WP} 为中心词三元模型:

$$P_{WP}(w_k \mid T_{k-1},S_{k-1},W_{k-1}) = P(w_k \mid h_{-1}^{k-1}, h_{-2}^{k-1}) \tag{13}$$

同样,由于语义分析中的剪枝问题,模型 2 在实际运算中需改写为

$$P_2(w_k \mid W_{k-1}) = \sum_{i=1}^{Pr_{k-1}} P_{WP}(w_k \mid T_{k-1}^{(i)},S_{k-1}^{(i)},W_{k-1})\rho(T_{k-1}^{(i)},S_{k-1}^{(i)},W_{k-1}) \tag{14}$$

其中, $\rho(T_{k-1}^{(i)},S_{k-1}^{(i)},W_{k-1})$ 是不同部分语义依存结构的权重,是归一化后的候选依存结构得分,

$$\rho(T_{k-1}^{(i)},S_{k-1}^{(i)},W_{k-1}) = \frac{P_N(T_{k-1}^{(i)},S_{k-1}^{(i)},W_{k-1})}{\sum_{j=1}^{Pr_{k-1}} P_N(T_{k-1}^{(j)},S_{k-1}^{(j)},W_{k-1})} \tag{15}$$

这里,部分语义依存结构得分 $P_N(T_{k-1},S_{k-1},W_{k-1})$ 没有采用类似于 Chelba 的计算方法[6](如式(16)):

$$P(T_{k-1},S_{k-1},W_{k-1}) = \prod_{j=1}^{k-1}\{P(s_j \mid T_{j-1},S_{j-1},W_{j-1})P(w_j \mid T_{j-1},S_j,W_{j-1})P(T_{j-1}^j \mid T_{j-1},S_j,W_j)\} \tag{16}$$

其中, $P(T_{j-1}^j \mid T_{k-1},S_k,W_k) = \prod_{i=1}^{\tau_j} P(q_i \mid T_{j-1},S_j,W_j,q_1,...,q_{i-1})$.

因为对应相同词串 W_{k-1} 的不同部分语义依存结构 $(T_{k-1},S_{k-1},W_{k-1})$ 可能包括了不同数目的词、词义类预测操作和语义依存合并操作,例如式(16)中,部分依存结构中将包括 $k-1$ 个词、词义类预测操作和 $\sum_{j=1}^{k-1}\tau_j$ 个合并操作,这将导致包含较少预测操作、合并操作的部分依存结构概率比包含较多操作的概率大很多,使中心词三元模型

倾向于结构简单的部分依存结构.因此,我们将部分依存结构的得分按照词、词义类预测和合并操作数目进行归一化,即首先计算分别计算预测概率、合并概率的几何平均值,再以它们的乘积作为归一化的概率 P_N .最后,按照式(15)计算归一化的部分依存结构得分.

$$P_N(T_{k-1},S_{k-1},W_{k-1})=\left\{\prod_{j=1}^{k-1}\left[P(s_j\mid T_{j-1},S_{j-1},W_{j-1})P(w_j\mid T_{j-1},S_j,W_{j-1})\right]\right\}^{1/k-1}\cdot\left\{\prod_{j=1}^{k-1}P(T_{j-1}^{j}\mid T_{j-1},S_J,W_j)\right\}^{1/\sum_{j=1}^{k-1}\tau_j} \quad (17)$$

式(17)能够有效地降低简单的语义依存结构(即包含较少合并操作的语义依存结构)的权重,使中心词三元模型可以发现更远距离的历史.第 4 节中的比较实验也表明,新的归一化方法(式(17))不仅明显增加了平均历史的长度,而且提高了语言模型的性能.

3 实验与分析

3.1 语义分析系统

在训练和测试语义分析系统时,我们使用了语义依存网络数据库(semantic dependency net,简称 SDN)[12]和 T9394 人民日报数据库.SDN 数据库共包括约 1 百万个词,全部数据带有词义类标注和语义依存结构标注.T9394 人民日报数据库包括 1993 年 1 月~1994 年 12 月《人民日报》全部的新闻文本,共约 2 千万个词.在实验中,将 SDN 数据库分为两部分,一部分含 17 万个词(约合 2 万个句子)用于测试,其余 SDN 数据库和全部 T9394 数据库用于训练.

本实验模型训练过程如下:首先,用 SDN 训练数据部分训练语义标注模型[13]和语义依存关系分析模型[10];然后,再用语义标注器和语义依存标注器自动标注 T9394 数据库,生成带有词义和语义依存结构标注的数据库 T9394-SDN;最后,用 T9394-SDN 数据库训练各种词预测模型和词义预测模型.

我们定义了词义类正确率,以评价词义标注的性能.

定义 **1**(词义标注正确率 **SCR**). $SCR=\dfrac{\text{标注正确词义的词数}}{\text{总词数}}$.

为了评价语义依存关系标注的性能,我们首先定义了两个概念:语义依存关系正确和语义依存结构正确.如果语义依存关系三元对(修饰词 w_i ,中心词 w_{H_i} ,关系 R_i)中各个元素都正确,则为该语义依存关系正确.然而,语义依存结构正确只需要三元对(修饰词 w_i ,中心词 w_{H_i} ,关系 R_i)中修饰词与中心词正确.由此,我们进一步定义:

定义 **2**(语义依存关系正确率 **SRCR**). $SRCR=\dfrac{\text{标注正确语义依存关系的词数}}{\text{总词数}}$.

定义 **3**(语义依存结构正确率 **SSCR**). $SSCR=\dfrac{\text{找到正确中心词的词数}}{\text{总词数}}$.

在表 1 中,第 1 行为的词义类标注器(Sem class tagger)的标注正确率.第 2 行为语义依存关系分析器(Dep parser)的标注正确率[10],Dep parser 以带有正确词义类标注的句子为输入,词义标注正确率可以看作 100%,它只表现语义依存关系分析模型(式(7)和(8))的性能.第 3 行和第 4 行是基于 SWT 模型(式(1))及基于 WST 模型(式(4))的语义分析器正确率.

Table 1 Results of semantic analysis system

表 **1** 语义分析系统的性能

	SCR (%)	SRCR (%)	SSCR (%)
Sem class tagger	91.41	–	–
Dep parser	100	67.25	76.87
SWT-Parser	90.85	66.50	75.84
WST-Parser	90.20	66.32	75.66

从表 1 可以看出,SWT 模型标注正确率普遍略优于 WST 模型.SWT 模型中的词及词义类预测模型类似于词义标注中的隐含马尔科夫模型[13].与 WST 模型相比,SWT 模型更准确地描述了词义类的接续三元模型

$P(s_i \mid s_{i-2}, s_{i-2})$,使 SWT 模型的词义标注正确率略高于 WST 模型,从而使语义依存关系标注正确率也相应地较高,并且在语言模型的实验(第 3.3 节)中,SWT 与 WST 模型也有类似的表现.

实验还表明,词义类的自动标注对语义依存关系和语义依存结构标注正确率影响很小,这主要是因为语义依存关系分析模型是一个基于词的模型,词义类信息主要用于平滑,词义类标注正确率对依存分析的影响不大.

3.2 语义结构语言模型

语言模型的实验是在清华大学语音识别实验室的 THSP 数据集上进行的,THSP 数据集共包括 11 组数据,其中 6 组为女声,5 组为男声,每组数据包括 120 个摘自《人民日报》的句子,句子平均长度为 25 个字.

多数文献[6,7]表明,将结构化语言模型与词三元语言模型插值会取得更好的效果.本实验也采用类似的方法.同时,针对两类结构化语言模型的特点选用了不同的插值方法.

最优标注句子模型以句子为单位给出模型概率,因此只能在句子级插值:

$$P_{I1}(W) = \lambda P_1(W) + (1-\lambda) P_T(W) \tag{18}$$

其中,词三元模型的句子概率 $P_T(W)$ 为

$$P_T(W) = P(w_1)P(w_2 \mid w_1)\prod_{k=3}^{N} P(w_k \mid w_{k-2}, w_{k-1}) \tag{19}$$

中心词三元模型则是以词为单位给出模型概率,可以直接在词级插值:

$$P_{I2}(w_k \mid W_{k-1}) = \lambda P_2(w_k \mid W_{k-1}) + (1-\lambda) P_T(w_k \mid w_{k-2}, w_{k-1}) \tag{20}$$

3.2.1 中心词三元模型复杂度

语言模型的性能通常可以用测试集的复杂度(PPL)来评价.对于给定语言模型,复杂度越小,语言模型越接近客观存在的语言模型,模型的质量也就越好.模型复杂度的极小值为语言自身的熵.

$$PPL = \exp\left(-\frac{1}{K}\sum_{i=0}^{K-1}\log P_M(w_i / w_1,..., w_{i-1})\right).$$

使用复杂度评价语言模型时,要求模型是完备的,即 $\sum_{w_i} P(w_i \mid w_1,..., w_{i-1}) = 1$.但是,由于最优标注句子模型的近似假设破坏了模型的完备性,因此这里只计算中心词三元模型的复杂度.

本节实验采用与第 3.1 节相同的测试集.当 SWT-中心词三元模型(式(20))中的 λ 取不同值时,模型复杂度见表 2.实验表明,中心词三元模型描述了句子的结构信息和长距离的词依存信息,显著降低了语言模型的复杂度.当插值系数 $\lambda = 0.4$ 时,模型复杂度比 Trigram 绝对下降 28.11%,相对下降 10.17%.

Table 2 Perplexity of headword trigram model under different interpolation weights

表2 不同插值系数下中心词三元模型复杂度

λ	0	0.2	0.4	0.6	0.8	1
PPL	276.31	250.34	248.20	253.32	267.11	314.44

3.2.2 中心词三元模型的历史长度

中心词三元模型作为长距离语言模型,使用了长度超过 2 的历史.我们采用与文献[6]类似的模型历史长度分布定义(如式(21)),令 $D(T_{k-1})$ 为部分语义依存结构暴露出来的中心词 h_{-2} 距当前词 w_k 跨越的长度.

$$P(d) = \sum_{k=1}^{N}\sum_{i=1}^{\Pr_{k-1}}\delta(D(T_{k-1}^{(i)}), d)\rho(T_{k-1}^{(i)}, S_{k-1}^{(i)}, W_{k-1}) \tag{21}$$

其中,$\rho(T_{k-1}^{(i)}, S_{k-1}^{(i)}, W_{k-1})$ 如式(15)定义.$\delta(x, y)$ 为指示函数.当参数 $x = y$ 时,$\delta(x, y) = 1$;否则,$\delta(x, y) = 0$.

实验中采用了 THSP 测试集,中心词三元模型的历史长度分布如图 5 所示.其中 41%以上的情况历史长度超过 2,平均长度为 2.733 637,这表明中心词三元模型中包括了很多长距离约束.

3.2.3 语音识别中的实验结果

在语音识别实验的声学层中,采用 45 维 MFCC 和能量特征,其中包括 14 维 MFCC、1 维归一化能量和它们的一阶和二阶差分,同时采用 DDBHMM(duration distribution based hidden Markov model)[14]声学模型,并假设段长为均匀分布.语言层中,首先用词三元语言模型(trigram)搜索声学层生成的拼音图,得到 100 个最优句子;然

后,用基于语义依存结构的语言模型再次打分、排序,得到最优识别结果.本实验中的语义分析系统与第 3.1 节中的完全相同.表 3 为基线系统(即 trigram)的第 1 条最优路径(1-best)和前 n 条最优路径(n-best)字错误率.

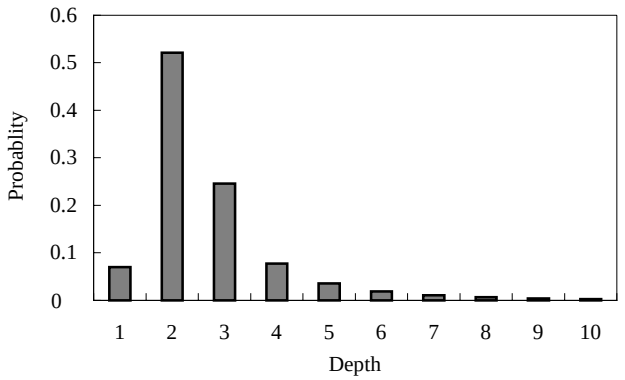


Fig.5 Depth distribution of headword trigram model

图 5 中心词三元模型历史长度概率分布

Table 3 Chinese character error rates of 1-best and n-best paths of baseline (%)

表 **3** 基线系统 1-best 和 n-best 路径字错误率 (%)

1-best	5-best	20-best	100-best
10.20	7.66	6.24	5.08

表 4 中列出了 3 类基于语义依存结构的语言模型(句子整体模型(FSDM)、最优标注句子模型(BSDM)和中心词三元模型(HTM))的实验结果.实验表明,不同的语义依存分析器、不同的结构语言模型都不同程度地降低了语音识别的字错误率.

Table 4 Results of all semantic structure language models (%)

表 **4** 各种语义结构语言模型的识别结果 (%)

	CER (relative error reduction)		
	FSDM	BSDM	HTM
SWT-Parser	9.56 (6.22)	9.40 (7.80)	9.39 (7.94)
WST-Parser	9.87 (3.19)	9.89 (2.99)	9.37 (8.14)

从句子整体模型(FSDM)和最优标注句子模型(BSDM)的比较来看,以最优标注概率近似全部标注概率和的假设是可行的,它基本不影响字误识率.基于 SWT-Parser 和 WST-Parser 的最优标注句子模型分别使字误识率下降 7.80%和 2.99%.

从最优标注句子模型(BSDM)和中心词三元模型(HTM)的比较来看,中心词三元模型性能优于最优标注句子模型.我们认为其中的原因主要包括 3 个方面.第一,最优标注句子模型直接由语义分析模型概率推导句子概率,使语言模型的性能直接受到语义分析模型性能(语义关系正确率只有 67%)的限制.然后,中心词三元模型主要利用了语言的结构信息.语义分析器的结构正确率明显高于关系正确率,因此中心词三元模型减小了语义依存分析器精度有限对语言模型的影响.第二,中心词三元模型使用了约 2 千万词的训练数据,而最优标注句子模型只使用了约 1 百万个词的数据训练语义依存分析器.第三,中心词三元模型中考虑了多个语义结构候选,对多个候选结构对应的中心词三元模型取概率加权和.因此,中心词三元模型比最优标注句子模型对于性能有限的中文语义分析器来讲性能略好,并且具有更强的鲁棒性.

总之,实验表明语义结构模型描述了句子的结构信息和长距离间词的依存关系,提高了语音识别的性能.其中,中心词三元模型相对于基线系统绝对字错误率下降 0.8%,相对错误率下降 8.14%.另一方面,语义结构模型还能够分析被识别句子的结构和语义关系,有助于对句子意义的理解.

4 与相近工作的比较

本文提出的语义分析系统实现了非特定领域词义排歧和语义依存关系分析的一体化模型.在非特定领域的中文处理中,周明实现了自动句法分析器[15],它的句法分析正确率为 67.7%.与句法分析相比,语义分析的复杂度更大,然而我们的标注正确率基本相当.如果在周明的句法分析之后再进行语义分析,则很难达到与本文相当的语义分析正确率.

在近几年的改进语言模型中,本文的语义结构语言模型与文献[6,7]的语言模型比较相近,Chelba 模型可以看作是一种 WST 中心词三元模型.而它们的区别主要在于:

1. 依存结构的分析方法不同

首先,依存结构的获取方法不同.Chelba 的依存结构是从 Upenn Treebank 句法树按照规则转化得到的,而我们的语义依存结构完全是针对中文的特点定义的,我们的语义结构分析器是从根据语义依存语法手工标注的语义依存网络数据库上学习得到的.其次,我们的依存分析统计模型也不同.

2. 语言模型形式不同

本文提出并分析了更多形式的语义分析模型和结构化语言模型.在语义分析模型中,我们分析了不同预测顺序的两个对偶模型——SWT 模型和 WST 模型.在结构化语言模型中,分别分析了句子整体模型、最优标注句子模型和中心词三元模型.如果只考虑模型总体形式,不考虑依存结构分析方法及部分语义依存结构得分归一化方法,那么 Chelba 模型可以看作是一个 WST 中心词三元模型.

3. 中心词三元模型中部分语义依存结构得分归一化方法不同

本文修正了中心词三元模型中部分语义依存结构得分归一化方法(式(17)),对部分依存结构的得分按照词、词义类预测和合并操作数目进行归一化.改进前后对测试集的语言模型历史长度分布如图 6 所示,在测试集上 Chelba 模型历史的平均长度为 2.208 588,而改进后 WST 中心词三元模型历史的平均长度为 2.733 637.在中文语音识别实验中,采用 Chelba 形式的语言模型的字错误率为 10.08%,而改进后 WST 中心词三元模型字错误率为 9.37%.

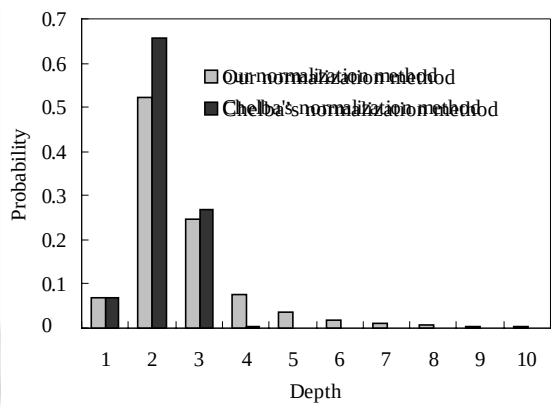


Fig.6 Depth distribution comparison between headword trigram models with and without our normalization method

图 6 归一化前后中心词三元模型历史长度概率分布比较

5 讨论与总结

本文在早期语义依存分析工作[10]的基础上,实现了完整的中文语义分析,并构建了两类基于语义结构的语言模型.文中的语义分析系统能够自动分析句子中各个词的词义以及句子中词之间的语义依存关系,达到90.85%的词义标注正确率和 75.84%的语义依存结构标注正确率.在语义分析的基础上,本文研究了两类不同的语言模型——最优标注句子模型和中心词三元模型.由于中心词三元模型回避了语义依存分析模型精度有限

的问题,同时利用多个候选结构信息,使得语言模型具有更高的可靠性.中心词三元模型比传统词三元模型复杂度相对下降 10.17%.在语音识别任务上,中心词三元模型比传统词三元语言模型,字错误率绝对下降 0.8%,相对下降 8%.

目前,语义依存分析及其结构语言模型的研究仍处于初级阶段,还有很多有待改进和解决的问题.今后的工作至少要着重解决如下 3 个问题:(1) 如何提高语义分析系统的性能;(2) 如何用统计的方法更有效地描述语言学知识,例如,如何从语义结构中选取更优的特征来预测词,如何设计 P_{WP} 模型;(3) 针对目前自然语言处理水平非常有限的现状,语言模型如何利用这些包含噪音的语言学信息,如何减少这些噪音的干扰.

References:

[1] Jelinek F. Self-Organized language modeling for speech recognition. In: Waibel A, Lee KF, eds. Readings in Speech Recognition. San Mateo: Morgan Kaufmann Publishers, 1990. 450−506.

[2] Brown PF, DellaPietra VJ, DeSouza PV, Lai JC, Mercer RL. Class-Based *n*-gram models of natural language. Computational Linguistics, 1992,18(4):467−479.

[3] Lau R, Rosenfeld R, Roukos S. Trigger-Based language models: A maximum entropy approach. In: Sullivan BJ, ed. Proc. of the Int'l Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol II. 1993. 45−48.

[4] Bellegarda JR. A multi-span language modeling framework for large vocabulary speech recognition. IEEE Trans. on Speech Audio Processing, 1998,6(5):456−467.

[5] Gao JF, Suzuki H, Wen Y. Exploring headword dependency and predictive clustering for language modeling. In: Hajic J, Matsumoto Y, eds. Proc. of the Empirical Methods in Natural Language Processing (EMNLP). 2002. 248−256.

[6] Chelba C. Exploiting syntactic structure for natural language modeling [Ph.D. Thesis]. Johns Hopkins University, 2000.

[7] Xu P, Chelba C, Jelinek F. A study on rich syntactic dependencies for structured language modeling. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). ACL, 2002. 191−199.

[8] Roark B. Probabilistic top-down parsing and language modeling. Computational Linguistics, 2001,27(2):249−276.

[9] Gao JF, Suzuki H. Unsupervised learning of dependency structure for language modeling. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL). ACL, 2003. 7−12. http://research.microsoft.com/~jfgao/paper/dlm-ACL03.pdf

[10] Li MQ, Li JZ, Wang ZY, Lu DJ. A statistical model for parsing semantic dependency relations in a Chinese sentence. Chinese Journal of Computers, 2004,27(12):1679−1687 (in Chinese with English abstract).

[11] Mei JJ, Zhu YM, Gao YQ, Yin HX. Tongyici Cilin (Dictionary of Synonymous Words). Shanghai: Shanghai Cishu Publisher, 1983 (in Chinese).

[12] Li MQ, Li JZ, Dong ZD, Wang ZY, Lu DJ. Building a large Chinese corpus annotated with semantic dependency. In: Ma Q, Xia F, eds. Proc. of the 2nd SIGHAN Workshop on Chinese Language Processing. 2003. 84−91.

[13] Zhang JP. A study of language model and understanding algorithm for large vocabulary spontaneous speech recognition [PH.D. Thesis]. Beijing: Department of Electronic Engineering, Tsinghua University, 1999 (in Chinese with English abstract).

[14] Wang ZY, Xiao X. Duration distribution based HMM speech recognition models. Chinese Journal of Electronics, 2004,32(1):46−49 (in Chinese with English abstract).

[15] Zhou M. A block based dependency parser for unrestricted Chinese text. In: Proc. of the 2nd Chinese Language Processing Workshop. 2000. 78−84. http://research.microsoft.com/china/papers/Robust_Dependency_Parser_Chinese_Text.pdf

附中文参考文献:

[10] 李明琴,李涓子,王作英,陆大绘.中文语义依存关系分析的统计模型.计算机学报,2004,27(12):1679−1687.

[11] 梅家驹,竺一鸣,高蕴琦,殷鸿翔.同义词词林.上海:上海辞书出版社,1983.

[13] 张建平.大词汇量连续语音识别中的语言模型和理解算法的研究[博士学位论文].北京:清华大学电子工程系,1999.

[14] 王作英,肖熙.基于段长分布的 HMM 语音识别模型.电子学报,2004,32(1):46−49.