依存句法语言模型对短语统计机器翻译性能的影响

董人菘[1] 王 华[1] 张晓钟[2] 余正涛[3] 张 涛[3]
(昆明理工大学冶金与能源工程学院 昆明 650093)[1] (昆明理工大学城市学院 昆明 650051)[2]
(昆明理工大学信息工程与自动化学院 昆明 650500)[3]

摘 要 为提高汉-英统计机器翻译的翻译效果,提出一个基于依存句法关系的语言模型,在较成熟的基于短语翻译的统计特征下,对解码产生的NBEST候选翻译结果进行进一步约束,重新计算得分,调整NBEST候选翻译序列,以得到最佳翻译。实验以"Pharaoh"为比较基准,以500句汉英句对为测试集,最终的实验结果表明,提出的基于依存句法关系的语言模型可以在一定程度上提高汉-英统计机器翻译最佳翻译的正确率。

关键词 依存句法,语言模型,机器翻译
中图法分类号 TP391.2 **文献标识码** A

Effects of Statistical Machine Translation with Language Model of Dependency Syntax Relationship

DONG Ren-song[1] WANG Hua[1] ZHANG Xiao-zhong[2] YU Zheng-tao[3] ZHANG Tao[3]
(Faculty of Metallurgical and Engineering, Kunming University of Science and Technology, Kunming 650093, China)[1]
(City College, Kunming University of Science and Technology, Kunming 650051, China)[2]
(School of Information and Automation, Kunming University of Science and Technology, Kunming 650500, China)[3]

Abstract In order to improve the the results of statistical machine translation in Chinese-English, the paper proposed a language model based on dependency syntax relationship. In the more mature features of phrase-based statistical translation, this new model can further constraint the result of the NBEST sequence by decoding, recalculate the NBEST sequence scores, and adjust the NBEST sequence to get a better translation. Experiments with a test set with baseline of "Pharaoh" in 500 English sentences and the final experimental results show that the proposed language model with dependency syntax relationship can improve the accuracy of Chinese-English's best statistical translation in some extent.

Keywords Dependency syntax, Language model, Machine translation

1 引言

近年来,句法信息被越来越多地引入到机器翻译的过程中。在形式化语法方面,最早的如吴德凯的ITG模型将同步语法用于统计机器翻译[1],Chiang提出的层次短语模型[2]与ITG类似,都是基于形式化语法的,但其在重排序方面比ITG更强。在语言学语法方面有Yamda等提出并发展的串到树模型[3,4],国内学者刘洋提出的树到串模型,Lin的基于路径的转换模型[5]及Quirk等人的基于依存句法分析的统计翻译模型[6],以及最近几年发展的森林到串的模型,都是基于句法的统计机器翻译模型,能够有效利用语言句法信息来对提高机器翻译的效果。

在语言学语法的有效使用方面,不论是依存句法还是短语句法,在对源语言或者目标语言进行句法分析后,句法信息只是在模板抽取过程中起作用,而且通常由于词的句法关系信息并没有包含在模板中,这导致在解码时,只能机械拼接模板,句法关系信息很难有效利用,对解码几乎没有贡献。因此,本文在相对成熟的基于短语的统计翻译方法的基础上,提出一种基于依存句法关系的语言模型,在解码后期产生NBEST候选翻译时,以基于依存句法关系的语言模型重新计算候选翻译的得分,改变NBEST候选翻译序列,在此基础上得到最佳翻译。

2 相关翻译方法及解码过程

对当前基于句法的统计机器翻译方法进行研究比较可以知道,不论是基于依存语法还是基于短语结构语法,在模板方面无论是短语-短语模型、树-串模型,还是森林-串模型,在训练时都是模板抽取,包括词汇化模板和非词汇化模板,训练时的句法关系信息,对于依存句法是节点之间的依存关系,这些节点之间的依存句法关系信息在抽取模板时起到很大的作用。

模板抽取完成后,在解码时多数采用直接翻译模型,如式(1):

$$\Pr(s_1^I \mid t_1^J)=\frac{\exp\left[\sum_{m=1}^{M}\lambda_m h_m(s_1^I,t_1^J)\right]}{\sum_{\tilde{e}_1^I}\exp\left[\sum_{m=1}^{M}\lambda_m h_m(\tilde{e}_1^I,t_1^J)\right]} \tag{1}$$

到稿日期:2013-05-20 返修日期:2013-08-25 本文受国家自然科学基金(61163022,61262041)资助。

董人菘(1978-),男,博士生,主要研究方向为智能控制;王 华(1965-),男,博士,教授,博士生导师,主要研究方向为冶金控制;张晓钟(1975-),女,硕士,主要研究方向为语言教学与翻译;余正涛(1970-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为机器翻译,E-mail:ztyu@hotmail.com;张 涛(1982-),男,硕士生,主要研究方向为机器翻译。

$h_m(s_1^I, t_1^I)$为特征函数，其中 $m=1,\cdots,M$，对于每个 $h_m(s_1^I, t_1^I)$，模型有相应的参数 λ_m。我们的目标是求解 λ_m 使汉语-英语互为翻译对的概率最大，即使 $h_m(s_1^I, t_1^I)$最大。式(1)中的分母部分对搜索结果没有影响，故我们的搜索模型为：

$$\hat{\alpha}=\arg\max_{\alpha}\{\sum_{m=1}^{M}\lambda_m h_m(s_1^I, t_1^I)\} \quad (2)$$

在搜索时，以特征函数为约束条件对模板拼接，计算翻译概率完成翻译，那么可以看出在解码过程中词之间的句法关系信息没有起到明显作用。

为了便于说明情况，以“Pharaoh”系统为比较基准，设计相同的特征函数，并在解码时以基于句法节点的语言模型重新计算 NBEST 候选翻译得分，以试图得到更好的最佳翻译结果。

3 基于依存句法关系的语言模型

3.1 依存语法

依存句法主要包括词类和依存关系，其中的依存关系描述两个词之间的关系，潜在地表达了一个词和另一个词结合的能力。在统计机器翻译领域中，依存语法具有一些适合于机器翻译的特性，如天然词汇化、更接近语义表达以及能更好地表达语言间的结构差异等[7]，而且由于依存关系比短语结构关系更加接近词之间本身的意义表达，因此本文采用依存句法分析来训练统计语言模型。

3.2 统计语言模型

统计语言模型在机器翻译、语言识别、信息检索、ORC 字符识别等领域均有广泛应用。其是针对某种语言建立的概率模型，使得正确词序列的概率值大于错误的词序列的概率值，对于词序列 $w_1,\cdots,w_m$ 的句子，其概率可以由式(3)计算。

$$p(w_1,\cdots,w_m)=p(w_1)\prod_{i=2}^{m}p(w_i|w_1,\cdots,w_{i-1}) \quad (3)$$

如果假设当前词的条件概率仅仅与前 $n-1$ 个词相关，则 $p(w_1,\cdots,w_m)=p(w_1)\prod_{i=2}^{m}p(w_i|w_{i-n},\cdots,w_{i-1})$，称为 Ngram 模型，本文训练的是 3 元语言模型。

3.3 依存语言模型训练

模型训练采用最大似然估计的方法，即

$p(w_i|w_{i-n},\cdots,w_{i-1})=\dfrac{C(w_{i-n}\cdots w_{i-1},w_i)}{C(w_{i-n}\cdots w_{i-1})}$，其中 $C(w_{i-n}\cdots w_{i-1},w_i)$为连续词汇 $w_{i-n}\cdots w_{i-1},w_i$ 在文档中出现的次数。

以 10 万句英文为训练集，以 Stanford 句法分析器为训练工具，对训练集中的每句英文进行句法分析，例如：

My dog also likes eating bananas.

句法分析后的依存树如下所示：

poss(dog－2,My－1)

nsubj(likes－4,dog－2)

advmod(likes－4,also－3)

xcomp(likes－4,eating－5)

dep(eating－5,bananas－6)

最左边的标记表示括号中词语之间的依存关系，词语后面的数字表示词语在句子中的位置。

将依存树中的词去除，按顺序保留句法树的词语之间的句法关系，从训练集中得到 10 万条经过依存句法分析后并抽取出来的只含有依存词之间句法关系的“句子”。对于上面的句子，即为

poss nsubj advmod xcomp dep

那么，以 Srilm 为语言模型的训练工具，即可训练 N 元基于依存句法关系的语言模型，本文 N 取语言模型中常见的 3。

4 解码

以 500 对汉英句对为测试集，在正常的解码过程产生 NBEST 候选翻译后，记录每个英文句子得分 *Pni*，对每个候选翻译进行依存句法分析，并得到依存句法节点序列，计算基于依存句法节点的语言模型得分 *DPni*，则 NBEST 候选翻译中第 i 个句子的最终得分如式(4)：

$$scorei=Pni+DPni \quad (4)$$

对 NBEST 候选翻译以新的 *scorei* 重新排序，最终输出最好的一个或者 N 个翻译结果。

5 实验结果

训练集为 10 万对汉英句对，测试集为随机选取的 500 句汉-英句对。训练集和测试集的情况如表 1 所列。

表 1 训练集和测试集情况

		中文	英文
训练集	句数	100000	
	词数	35374	37581
测试集	句数	500	
	词数	4015	4681

采用公用的 BLEU 评测方法，对翻译结果进行评分，以“Pharaoh”为对比实验系统，对每个句子产生的 NBEST 候选翻译输出最佳候选翻译，采用基于依存句法关系的语言模型约束后，训练集的最佳候选翻译的变化情况如表 2 所列。

表 2 增加句法约束后 NBEST 序列的变化情况

	句数
训练集	500
变化总量	23
正面变化	17
负面变化	6

在产生 NBEST 候选翻译后，采用基于依存句法关系的语言模型重新计算得分后，输出最佳候选翻译，BLEU2 测评得分的结果如表 3 所列。

表 3 采用 BLEU2 计算得分的结果

系统	BLEU4
“Pharaoh”	0.2312
“Pharaoh”＋基于依存句法关系的语言模型	0.2323

可以看出在 500 句的训练集中，大约有 4.6%的候选翻译出现变化，其中大约有 74%出现正面变化，16%的句子出现了负面的变化。综合来看，以句子为计算单位，并以 BLEU 测评标准计算得分，大约有 3.4%的句子翻译效果得到提升。

结束语 在已有翻译方法下，利用已有的汉语-英语双语平行语料库，提出一种基于依存句法节点的语言模型，并应用于解码阶段。从实验结果上看，提出的基于依存句法节点的语言模型可以在一定程度上提高翻译的准确率。下一步工作将拓展句法节点信息在翻译过程中的应用范围，试图在翻译方法上利用句法节点信息。

参 考 文 献

[1] Wu De-kai. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora[J]. Computational Linguistics, 1997, 23(3): 377-403

[2] Chiang D. A hierarchical phrase-based model for statistical machine translation[C]// Proceedings of ACL. 2005: 263-270

[3] Yamada K, Knight K. A syntax-based statistical translation model[C]// Proceedings of ACL. Toulose, France, 2001: 523-530

[4] Galley M, Hopkins M, Knight K, et al. What's in a translation rule? [C]// Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston, Massachusetts, 2004: 273-280

[5] Lin D, Cherry C. Word alignment with cohesion constraint[C]// Proceedings of the HLT-NAACL. Edmonton, Canada, 2003: 49-51

[6] Quirk C, Menezes A, Cherry C. Dependency treelet translation: Syntactically informed phrasal SMT[C]// Proceedings of ACL. Ann Arbor, Michigan, 2005: 271-279

[7] 熊德意. 基于括号转录语法和依存语法的统计机器翻译研究[D]. 北京: 中国科学院计算技术研究所, 2007

(上接第 58 页)

大的模块度的值，我们算法和 Brim 算法获得的模块度的值比较接近，但是我们不需要实现控制社区的个数。LPA 算法在求解次问题时，获得的最大模块度值为 0.5782。可以看出我们的算法要优于 LPA 算法。

结束语 本文提出了一个基于矩阵分解的二分网络社区挖掘算法。该算法首先将二分网络分为两个部分，每个部分尽可能保存完整的社区信息，然后分别对两个部分进行递归的拆分，直至不能拆分为止。在拆分的过程中，我们应用矩阵分解，使得到的分解能与网络的相关矩阵的行空间尽可能接近，以尽可能保持原图的社区信息。实验结果表明，该算法在不需任何额外参数的情况下，不但能较准确地识别实际网络的社区个数，而且可以获得很好的划分效果。

参 考 文 献

[1] Newman M E J. The structure and function of complex networks[J]. SIAM Rev., 2003(45): 16-256

[2] Strogatz S H. Exploring complex networks [J]. Nature, 2001 (410): 268-276

[3] Newman M E J. Scientific collaboration networks. I. network construction and fundamental results[J]. Physical Review E, 2001, 64: 016131

[4] Newman M E J. Scientific collaboration networks. II. shortest paths, weighted networks, and centrality[J]. Physical Review E, 2001, 64: 016132

[5] Le Blond S, Guillaume J L, LatapyM. C lustering in P2P exchanges and consequences on performances[C]// Castro M, Renesse R. Peer- to-Peer Systems IV. Berlin: Heidelberg, 2005: 193-204

[6] Watts D J, Strogatz S H. Collective dynamics of small world networks[J]. Nature, 1998, 393: 440-442

[7] 刘爱芬，付春花，张增平，等. 中国大陆电影网络的实证统计研究[J]. 复杂系统与复杂性科学, 2007, 4(3): 10-16

[8] Robins G, Alexander M. Small worlds among interlocking directors: network structure and distance in bipartite graphs[J]. Computational & Mathematical organization Theory, 2004, 10: 69-94

[9] Battiston S, Catanzaro M. Statistical properties of corporate board and director networks[J]. European Physics Journal B, 2004, 38: 345-352

[10] Ergun G. Human sexual contact network as a bipartite graph [J]. Physica A, 2002, 308: 483-488

[11] Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite networks[J]. Physical Review E, 2005, 72: 066107

[12] 陈文琴，陆君安，梁佳. 疾病基因网络的二分图投影分析[J]. 复杂系统与复杂性科学, 2009, 6(1): 13-19

[13] Lind P G, Gonzalez M C, Herrmann H J. Cycles and clustering in bipartite networks[J]. Physical Review E, 2005, 72: 056127

[14] Zhang P, Wang J L, Li X J, et al. Clustering coefficient and community structure of bipartite networks[J]. Physica A, 2008, 387: 6869-6875

[15] 吴亚晶，狄增如，等. 基于资源分布矩阵的二分网聚类方法[J]. 北京师范大学学报：自然科学版, 2010, 46(5): 643-646

[16] Michael J. Barber, Modularity and community detection in bipartite networks[J]. Phys. Rev. E 76, 066102, 2007

[17] Lehmann S, Schwartz M, Hansen L K. Biclique communities[J]. Phys. Rev. E, 2008, 78(1): 016108

[18] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks [J]. Phys. Rev. E, 2007, 76(3): 036106

[19] Liu X, Murata T. How does label propagation algorithm work in bipartite networks? [C]// Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '09. Washington, DC, USA, 2009

[20] Du N, Wang B, Wu B, et al. Overlapping Community Detection in Bipartite Networks[C]// WI. 2008: 176-179

[21] 王洋，狄增如，樊琪. 二分网络社团结构的比较性定义[J]. 复杂系统与复杂性科学, 2009, 6(4): 40-44

[22] Newman M E J. Modularity and Community Structure in Networks [J]. Proc Natl Acad Sci USA, 2006, 103(23): 8577- 8582

[23] Guimera R, Sales-Pardo M, Amaral L A. Module identification in bipartite and directed networks[J]. Physical Review E, 2007, 76: 036102

[24] Davis A, Gardner B B, Gardner M R. Deep South [M]. Chicago: The University of Chicago Press, 1941

[25] Scott J, Hughes M. The Anatomy of Scottish Capital: Scottish Companies and Scottish Capital [C] // CroomHelm. London, 1980: 1900-1979