文章编号: 1003-0077(2011)02-0078-05

基于最大熵短语重排序模型的特征抽取算法改进

孙 萌[1,2], 姚建民[2], 吕雅娟[1], 姜文斌[1], 刘 群[1]

(1. 中国科学院 计算技术研究所 智能信息处理重点实验室, 北京 100190;
2. 苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘 要: 该文针对统计机器翻译中基于最大熵短语重排序模型特征抽取算法, 提出一种改进算法。该算法能够抽取出更多准确的短语重排序信息, 特别是逆序短语的特征信息, 解决了原算法中最大熵训练时特征数据不平衡的问题, 提高了翻译中短语重排序的准确率。以 NIST MT 05 作为汉语到英语翻译的测试集, 实验结果表明改进后的系统 BLEU 值比原系统提高 0.65%。

关键词: 最大熵; 特征抽取; 统计机器翻译; 重排序模型
中图分类号: TP391 文献标识码: A

An Improving Feature Extraction Algorithm for Maximum Entropy Based Phrase Reordering Model

SUN Meng[1,2], YAO Jianmin[2], LV Yajuan[1], JIANG Wenbin[1], LIU Qun[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. School of Computer Science & Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: This paper presents an improved feature extraction algorithm for maximum entropy based phrase reordering model. The algorithm can extract more accurate feature information of phrase reordering, particularly the feature of inverted phrases. It solves the problem of uneven distribution of feature information and increases the rate of correct translation. We use BLEU as a metric on Chinese-to-English translation, and the proposed algorithm obtains a relative improvement of 0.65% over baseline system.

Key words: maximum entropy; feature extraction; statistic machine translation; reordering model

1 引言

基于短语的统计机器翻译是当前机器翻译主流方法之一, 翻译的基本单元从词过渡到短语, 使得连续的词串在翻译过程中作为一个整体进行处理, 解决了词的上下文依赖问题。翻译的时候将输入的句子与短语词典进行匹配, 选择最好的短语划分, 同时将得到的短语译文重新排序, 得到最优的译文。

其中, 短语层次上重排序是基于短语机器翻译的一个重要研究问题。

许多系统(如 Pharaoh①, 丝路②) 采用扭曲模型概率调整目标语言短语之间的次序, 每个目标短语的扭曲概率可以根据当前目标短语的源语言短语的起始位置与前一个目标短语的源语言短语最后位置之间的距离计算。显然这种简单的基于惩罚长度的策略[1] 会影响短语重排序模型的正确率。将句法知识引入机器翻译系统, 可以有效地改进重排序的

① http://www.isi.edu/licensed-sw/pharaoh/
② http://www.nlp.org.cn/project/project.php?proj_id=14

收稿日期: 2010-08-11 定稿日期: 2010-11-30
基金项目: 国家自然科学基金资助项目(60873167, 60736014)

作者简介: 孙萌(1988—), 男, 硕士生, 主要研究方向为自然语言处理技术; 姚建民(1971—), 男, 博士, 教授, 主要研究方向为自然语言处理技术; 吕雅娟(1972—), 女, 博士, 副研究员, 主要研究方向为自然语言处理和机器翻译。

正确率[2-3]。其中 Wu[4] 提出的括号转录文法在机器翻译领域也得到了广泛的应用。但是由于括号转录文法并没有包含语言知识, 因而不能很好地预测两个相邻目标短语的组合次序。Xiong et al.[5] 在括号转录文法的基础上利用双语短语的边界单词作为特征进行最大熵训练得到重排序模型, 并通过计算相邻双语短语的特征获得在保序和逆序下的概率, 可以更好地预测相邻短语之间的次序, 从而有效地改善了翻译系统的翻译结果。

通过观察基于最大熵短语重排序模型进行最大熵训练的特征, 发现保序短语实例特征的数量远大于逆序短语实例特征的数量, 这是因为汉语和英语的语序大致相同。利用最大熵实现短语的重排序也可以视为一个分类问题, 即"保序类"和"逆序类", 而用以训练分类器的特征数据存在数据不平衡问题, 将有可能影响分类器的实际分类效果。例如, 选择 FBIS 作为训练语料, 基线特征抽取系统从中抽取 4 839 390 条特征实例, 其中保序特征实例占 82.7%, 而逆序特征实例仅占 17.3%。以所有特征实例中的 10 万句子作为对重排序模型的开放式测试集, 剩余数据作为最大熵训练集, 测试结果显示此重排序模型对保序特征的判断准确率 97.55%, 而对逆序特征的判断准确率仅为 72.03%。另外, 基于括号转录文法假设源语言端短语相邻则目标语言短语也相邻, 但是在实际的汉英句对中存在源语言短语相邻而目标语言短语不相邻的情况。针对以上情况, 本文从保序实例选取策略、引入组合特征以及加入新的短语次序三个方面改进最大熵的特征抽取算法, 以提高重排序模型的判断准确率, 最终达到提高翻译质量的效果。

2　基于最大熵短语重排序模型的统计机器翻译

Wu[4] 提出了一种基于括号转录文法的统计翻译模型。简化的括号转录文法仅包含以下两种规则:

$$R^l: A \rightarrow x/y$$
$$R^m: A \rightarrow [A_1, A_2] \mid < A_2, A_1 > \qquad (1)$$

其中 R^l 为词汇规则, 表示将源语言短语 x 翻译为目标语言短语 y。R^m 为合并规则, 源语言短语和目标语言短语的顺序可以表示为保序和逆序两种。在短语调序过程中, 可以为合并规则中的两种不同顺序设置先验的保序和逆序概率, 这种方法忽略了不同源语言—目标语言短语对之间的差异性。

Xiong et al.[5] 对以上括号转录文法模型的调序模型进行了改进, 提出了一个基于最大熵的括号转录文法的短语调序模型, 即运用最大熵模型进行短语的调序:

$$\Omega = p_\theta(o \mid A_1, A_2) = \frac{\exp\left(\sum_i \theta_i h_i(o, A_1, A_2)\right)}{\sum_o \exp\left(\sum_i \theta_i h_i(o, A_1, A_2)\right)} \qquad (2)$$

其中, h 为特征函数, θ 为特征权重, o 的取值为保序或逆序, 并且选取短语的尾词作为最大熵模型训练的特征。实验表明基于最大熵括号转录文法的短语调序模型的性能明显优于传统的基于扭曲的短语调序模型和基于括号转录语法的调序模型。但是, 从实验可以看出, 保序实例的数量要远高于逆序实例的数量, 可能会影响最大熵模型的性能。本文从重排序实例抽取算法和特征选择两方面切入, 旨在解决最大熵训练数据不平衡问题。在实验中, 将以采用基于最大熵调序模型的统计机器翻译系统 Bruin[5] 作为基线系统。

3　改进的重排序实例抽取算法

本文改进了最大熵短语重排序系统中重排序实例的抽取算法, 在实现上更加灵活简洁并且易于扩展, 可以满足实验中不同的抽取策略。

重排序实例抽取算法的输入是一个经过 GIZA++[①] 双向对齐的词语对齐矩阵, 输出是保序短语实例和逆序短语实例。

抽取算法首先遍历源语言端所有连续单词序列, 并抽取与此连续序列相对齐的目标语言最大跨度。然后过滤不满足对齐一致性的目标语言单词序列与源语言单词序列, 即依次反向扫描目标语言的跨度, 检查其对应的源语言跨度是否在原连续单词序列范围内。最后, 按照给定的不同抽取策略, 抽取出重排序实例。

3.1　变量定义

介绍重排序实例抽取算法之前, 首先定义与算法相关的变量。

(1) *alignset*

存放源语言到目标语言所有的对齐矩阵。

① http://code.google.com/p/giza-pp/

(2) *straightset*

存放目标语言短语保序次序实例的集合。

(3) *invertedset*

存放目标语言短语逆序次序实例的集合。

(4) *elseset*

存放源语言短语相邻目标语言短语不相邻的实例。

(5) *src_span[i,j]*

源语言从 i 到 j 的连续单词序列。

(6) *span[i,j]*

记录源语言 i 到 j 的连续单词序列以及对应目标语言的连续单词序列。

3.2 算法实现

本算法首先获得任意源语言 src_span[i,j] 对应的最大对齐矩阵 span[i,j], 然后过滤不合法的 span[i,j]。最后对重排序实例进行分类以及抽取实例特征。具体步骤见算法 1。

算法 **1** 改进的重排序实例抽取算法

```
1.  Input: 双语词对齐矩阵 A
2.  Initial(alignset, straightset, inwertedset, elseset);
3.  Foreach(src_span[i,j] ∈ s) do
4.      获取 src_span[i,j] 所对应的目标语言的对齐矩阵 span[i,j], 同时将 span[i,j] 存入 alignset 中;
5.  End for
6.  Foreach(span[i,j] in alignset)
7.      检查 span[i,j] 的对齐一致性, 删除不一致的 span[i,j];
8.  End for
9.  Foreach(span[i,j] in alignset)
10.     Foreach(i≤mid< j)
11.         if (span[i, mid], span[mid+ 1, j] 满足保序规则 S_i)
12.             straightset . push_back( span[i,j]);
13.         else if(span[i, mid], span[mid+ 1, j] 满足逆序规则 I_i)
14.             inwertedset . push_back( span[i,j]);
15.         else
16.             elseset . push_back(span[i,j]);
17.     End for
18. End for
```

算法第 9 行到 17 行, 描述的是改进的抽取实例算法的框架, 基于此框架可以方便制定各种抽取规则。其中第 10 行对抽取出来的双语词对齐矩阵, 检查是否可以将其拆分成两个相邻双语短语对, 并判断拆分后的相邻双语短语对的组合顺序。第 16 行, 本算法引入了一个新的分类, 即不相邻双语短语对。

3.3 重排序实例选择策略

基线系统采用了简单的方法控制重排序实例的数量, 即在保序实例中仅保留最小块, 对于逆序实例仅保留最大块。显然, 这样会损失一些短语边界特征, 并且保序实例的数量依然远超逆序实例的数量。这种特征数据的不平衡会影响最大熵重排序模型的判断准确率, 特别是对逆序实例特征的判断。以 10 万条实例进行开放式测试, 其中逆序实例数量为 17 286, 对逆序实例的测试精度仅为 72.03%。本文在 3.1 节提出的算法框架下, 对重排序实例选择策略依次进行以下 3 点尝试:

1) 为了解决最大熵训练过程中特征数据的不平衡, 最为直接的想法即是采取一定的选择策略直接限制保序实例的数量。相比基线系统选择保序实例中最小块, 本文采用随机算法选择保序实例数量, 避免了前种方法可能导致的长短语边界特征的缺失。

2) 在双语句子中会出现源语言短语相邻而目标语言短语不相邻现象, 针对这种情况, 本文在 1 的基础上增加一个新分类, 从一定程度上减轻特征数据的不平衡。抽取出来的实例, 如果不属于保序和逆序类, 即可将此实例归为一类。

3) 由于 giza+ + 对齐结果存在错误对齐, 对实例扩展未对齐词会提高短语特征抽取的召回率。这里定义保序、逆序规则 S_i、I_i, $i=\{0,1\}$; 其中当 $i=0$, 表示未对抽取实例进行未对齐词扩展; $i=1$, 表示对抽取实例进行未对齐词扩展。

4 特征抽取

从重排序实例中抽取特征, 以进行最大熵训练。重排序实例可以用< b_1, b_2 > 表示, 其中 $b=<c, e>$, c 代表源语言短语, e 代表目标语言短语, b_1 和 b_2 表示相邻或者不相邻短语。这里用 *c. h* 表示源语言短语的首单词, *c. t* 表示源语言短语的尾单词, 对于目标短语 e 也采用同样的定义。

基线系统考虑到特征抽取的规模, 仅利用重排

表 **1** 重排序实例的特征

尾词特征	$b_1.c.t$, $b_2.c.t$, $b_1.e.t$, $b_2.e.t$
首词特征	$b_1.c.h$, $b_2.c.h$, $b_1.e.h$, $b_2.e.h$
组合特征	$b_1.c.h\&b_2.c.h$, $b_1.e.t\&b_2.e.t$, $b_1.c.t\&b_2.c.t$, $b_1.e.h\&b_2.e.h$, $b_1.e.h\&b_2.e.t$, $b_1.e.t\&b_2.e.h$

序实例中的尾词。在特征抽取实验中, 除了以上四条尾词特征, 增加首词特征和组合特征。

由于汉语和英语语法结构的不同, 在汉语标点符号前后的短语或子句, 其对应的英语翻译有可能将此短语或子句逆序组合表达。基线系统的解码方法是, 如果在重排序窗口中搜索到标点符号, 则此窗口将不做逆序操作。此方法对于对称符号, 譬如“《》”“{}”是相当有效。但对“,”并不能以此简单判断。本文在增加重排序实例首词特征和组合特征的基础之上, 添加标点符号特征, 以进行最大熵训练。

5　实验结果及分析

实验中语言模型采用 N-gram 统计语言模型, 使用 LDC① 发布的 GigaWord 新华社部分作为训练英语语言模型的单语语料; 采用统计机器翻译领域公认的成熟开源语言模型训练工具 SRILM 进行 N-gram 语言模型的训练。实验采用规模为 518M 的四元语言模型。

基于重排序实例抽取算法, 我们设计了 7 个对比实验, 以对比不同特征抽取策略对最大熵训练的影响以及对最终翻译结果得分 BLEU 值的影响。选择 FBIS 作为训练语料, 抽取短语表以及重排序实例, 其中语料规模大约为 23.9 万句对。以 NIST MT 02 作为实验的开发集, NIST MT 05 作为测试集。

5.1　特征抽取策略对重排序结果的影响

选择重排序实例特征数据中的 10 万条记录作为最大熵重排序模型的开放测试集, 表 2 显示了从训练数据中抽取的重排序实例的规模、排序分类、各分类所占比例、测试精度和抽取的特征。其中, 测试精度为最大熵分类器正确判断样本的数量与测试集样本总数量的比值。

其中, 实验 1 是基线系统, 没有对保序实例的数量进行限制, 实验 2~ 7限制保序实例数量是逆序数量的 2 倍; 实验 2~ 4 抽取实例时没有对未对齐词进行扩展, 而实验 5~ 7 均进行未对齐词扩展; 实验 4、5 增加一个新的分类。

由于不同实验需要的特征不一致, 所以只能确定测试集的数量, 而不能确保测试集的内容的一致性, 因此不能简单的将最大熵重排序模型的测试精度高低反映为翻译性能的高低, 但仍然可以将最大熵重排序模型的测试精度作为一个参考指标。

表 2　重排序实例的规模、排序分类、测试精度和抽取的特征

	实例总量	保序 /%	逆序 /%	其他 /%	测试精度 /%	实验方案
1	4 839 390	82.7	17.3	–	93.9	尾词特征
2	1 467 337	65.1	34.9	–	86.5	+ 限制保序实例数量
3	1 467 337	65.1	34.9	–	92.8	+ 首词特征 + 组合特征
4	2 485 827	38.5	20.6	40.9	76.9	+ 第三类别
5	5 267 363	38.5	20.6	40.9	76.4	+ 扩展未对齐词
6	3 144 662	64.7	35.3	–	92.8	– 第三类别
7	3 144 662	64.7	35.3	–	91.5	+ 标点特征

从表 2 中可以看出, 实验 1 的测试精度达到最高值 93.9% , 实验 2 由于限制了保序实例的数量, 使得抽取出来的实例总量与实验 1 相比下降 70%, 导致最大熵训练的数据量不充足, 因此测试精度仅有 86.5%。考虑到在实例数量减少的情况下, 需要增加单个实例产生的特征数据量, 所以在试验 3 中, 对实例继续加入首词特征和组合特征, 测试精度达到 92.8%。但是源语言短语相邻, 并不表明目标语言短语相邻, 于是实验 4 引入第三类别, 即目标语言短语不相邻的情况。实验 4 的测试精度却下降至 76.9% , 这是因为新增的一个分类也增加了最大熵重排序模型判断的不确定性。实验 5 在实验 4 的基础上, 扩展未对齐词, 以增加实例的数量, 但是实验结果比实验 4 略低。实验 4 和实验 5 均是在实验 3 的基础上, 引入第三类而导致测试精度有较大下降, 从一定程度上说明第三类的引入不会提高最大熵模型判断准确率。因此, 本文设计实验 6, 在实验 3 的基础上扩展未对齐词; 实验 7, 在实验 6 的基础上引入标点符号特征。这两组实验的测试精度仅比实验 1 略低。

本文更关注特征抽取策略对于最大熵模型判断逆序实例的正确率, 图 1 显示最大熵重排序模型对测试集中保序子集(Mono) 和逆序子集(Invert) 的测试精度。

对测试集中的保序实例子集进行测试, 除了实验 4、5 因引入新分类而导致对保序特征判断的不确定性增大, 实验 2、3、6、7 与实验 1 的测试精度相差

① http://www.ldc.upenn.edu/

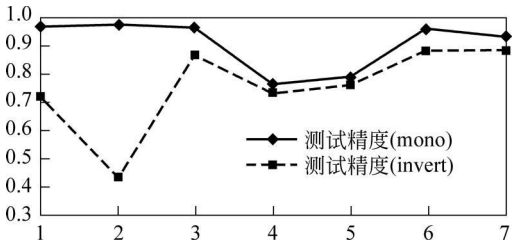

图 1 保序子集和逆序子集的测试精度

不超过 4%。观察测试集中的逆序实例子集的测试结果,实验 2 因为逆序特征的训练数据量较少,以致于对逆序实例子集的测试精度较低,而实验 3、4、5、6、7 均比实验 1 在逆序实例子集的精度高。其中,实验 6、7 的测试精度比实验 1 高达 16%。

从以上实验数据可以看出,本文提出的最大熵重排序模型特征抽取算法解决了由于特征数据不平衡导致最大熵模型对逆序特征判断不准确的情况。

5.2 翻译结果对比

在 NIST MT 05 上测试大小写敏感的 BLEU 值,图 2 显示 7 组用不同特征数据训练出来的最大熵重排序模型对最终翻译效果的影响。

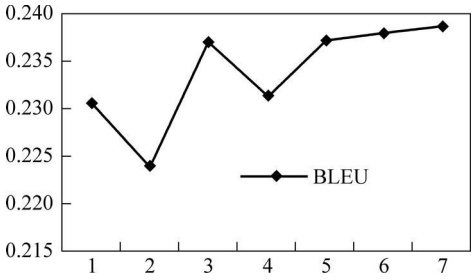

图 2 不同最大熵重排序模型对 BLEU 值影响

基线系统实验 1 的 BLEU 值为 0.228 3。从上图可以看到,除了实验 2 因特征训练数据过少以致最大熵重排序模型在翻译过程中性能有较大下降,实验 3、4、5、6 均基于实验 2 添加特征信息,并且在限制保序实例数量的同时重排序模型的性能均比基线系统高,其中实验 4 因引入“不相邻”分类翻译性能有所下降但是 BLEU 值仍高于基线系统,实验 7 加入标点特征,翻译的 BLEU 值达到最高值0.234 8。本文提出的重排序实例抽取以及特征抽取算法,通过限制保序实例数目和增加特征数量,可以显著提高重排序模型的性能从而提高翻译质量。

6 总结以及下一步工作

本文提出一种新的重排序实例抽取算法,并在此基础上加入新的特征,取得较好翻译效果。首先,通过限制保序实例的数目直接解决最大熵训练过程中的数据不平衡问题,由于特征信息过少而导致翻译性能下降。在此基础上,增加首词特征、组合特征翻译性能得到提高。其次,引入第三类短语组合顺序,即保序逆序之外的不相邻情况,虽然 BLEU 值有所下降但仍然高于基线系统。最后,本文在实验中尝试扩展对齐短语中的未对齐词,增加重排序实例特征数据量,翻译性能达到最好。

下一步工作我们将继续研究重排序实例特征对翻译性能的影响,重点在于融合句法知识特征,希望可以进一步提高翻译性能。此外,我们将深入探索基于括号转录语法框架下解码器的改进,以致可以处理源语言短语相邻而目标语言短语不相邻的情况。

参考文献

[1] Philipp Koehn . Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation Models [C]//Proceedings of the Sixth Conference of the Association for Machine Translation, Americas, 2004: 115-124.

[2] Kenji Yamada and Kevin Knight.. A Syntax-based Statistical Translation Model [C]//Proceedings of ACL, Toulouse, France, 2001: 523-530.

[3] David Chiang. A Hierarchical Phrase-based Model for Statistical Machine Translation [C]// Proceedings of ACL, Ann Arbor, Michigan, 2005: 263-270.

[4] Dekai Wu. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora [J]. Computational Linguistics, 1997, 23: 377-403.

[5] Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation[C]// Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 2006: 521-528.