词汇相似度约束的短语抽取

梁华参[1], 赵铁军[1], 薛永增[2], 孙加东[1]

(1. 哈尔滨工业大学 语言语音教育部-微软重点实验室, 哈尔滨 150001, hsliang@mtlab.hit.edu.cn;
2. 哈尔滨工业大学 媒体技术与艺术系, 哈尔滨 150001)

摘 要: 为克服传统的短语抽取方法对词对齐信息的依赖性强,抗噪声能力差这一缺陷,提出基于词汇相似度约束的短语抽取策略;在此框架下,提出了3种基于词汇相似度的约束方法:Dice系数、Phi平方系数和对数似然比.在IWSLT2004语料上进行的实验表明,3种基于词汇相似度的约束方法的翻译系统的BLEU评分均优于传统的翻译系统;其中基于对数似然比方法得到的翻译模型比基线系统Pharaoh的BLEU-4评分提高了15.14%.

关键词: 机器翻译;统计机器翻译;短语抽取;词汇相似度
中图分类号: TP391 **文献标志码:** A **文章编号:** 0367-6234(2010)05-0775-04

Phrase extraction based on constraints of word similarities

LIANG Hua-shen[1], ZHAO Tie-jun[1], XUE Yong-zeng[2], SUN Jia-dong[1]

(1. MOE-MS Key Lab of Natural Language Processing and Speech, Harbin Institute of Technology, Harbin 150001, China, hsliang@mtlab.hit.edu.cn; 2. Dept. of New Media and Art, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Aimed at the problem that the traditional phrase extraction method is strictly dependent on word alignments, and is not pruned to alignment errors, a loose phrase extraction method, which does not strictly depend on word alignments. In this method, constraints are posed on alignment points to avoid ill-formed phrase pairs. Three constraint strategies are proposed based on word similarities: Dice coefficient, Phi-square coefficient and log-likelihood ratio. Experiments were carried out on the corpus of IWSLT 2004. Results show that the BLEU scores of the best results of loose phrase extraction can be improved by 15.14%, compared with the baseline system Pharaoh.

Key words: machine translation; statistical machine translation; phrase extraction; word similarity

与传统的基于词的统计翻译模型相比,基于短语的模型有效利用了上下文关系来指导翻译过程,从而显著提高了翻译质量.王野翊[1]提出的基于结构的翻译模型,其实质是采用一个类似IBM词对齐模型2的方法来对齐双语短语,在此基础上再进行词一级的对齐.与此相类似的,Och[2]提出了对齐模板模型. Och将对齐短语泛化为基于词类的对齐模板,并采用了线性对数模型作为整体框架. Koehn[3]考虑了调序因素,提出了一个基于词对齐的短语翻译模型. Marcu等[4]采用联合概率代替条件概率,提出了基于短语和联合概率的翻译模型.张盈等[5]提出了短语对齐和切分相结合的短语等价对抽取方法.程葳[6]提出了双语语块的概念,并在此基础上建立了一个口语统计机器翻译系统. Vogel[7]分析比较了几种短语统计翻译模型,提出了一个混合模型.这些研究工作都是基于树串的统计机器翻译研究的基础[8-9].

本文在Koehn等人研究的基础上,针对短语等价对有效抽取问题,提出基于词汇相似度约束的短语抽取策略,来充分利用自动词对齐结果,并减小错误词对齐结果造成的精度损失.采用这种约束策略可以避免抽取到不完整的短语互译对.

收稿日期: 2009-06-08.
基金项目: 国家自然科学基金重点资助项目(60736014);国家高技术研究发展计划重点资助项目(2006AA010208).
作者简介: 梁华参(1982—),男,博士研究生;
赵铁军(1962—),男,教授,博士生导师.

1 基于短语的统计机器翻译系统框架

1.1 翻译模型

统计机器翻译中,翻译的任务就是在给定源语言句子 $\boldsymbol{f}$ 的条件下,搜索使得条件概率 $P(\boldsymbol{e}\mid\boldsymbol{f})$ 最大的目标语句子 $\hat{\boldsymbol{e}}$,作为翻译结果输出. 在对数线性模型下,条件概率 $P(\boldsymbol{e}\mid\boldsymbol{f})$ 通过一系列特征函数的线性组合来计算,即

$$\hat{\boldsymbol{e}} = \underset{\hat{e}}{\operatorname{argmax}} P(\boldsymbol{e}\mid\boldsymbol{f}) = \underset{\hat{e}}{\operatorname{argmax}} \sum_m \lambda_m h_m(\tilde{e},\tilde{f}). \quad (1)$$

式中:$(\tilde{e},\tilde{f})$ 为短语对,$h_m(\tilde{e},\tilde{f})$ 为特征函数,可以为 0 - 1 二值函数,也可以为实数函数. 当特征中涉及概率时,通常把特征函数定义为概率函数的对数形式,即 $h(\cdot) = \log p(\cdot)$.

基于短语的翻译模型把翻译过程,从传统的以词为单位的转换方式,转化为以短语为单位的转换方式. 在基于短语的翻译模型中,短语抽取方法中词对齐信息的利用对于翻译模型有直接影响.

1.2 对齐矩阵与重组

设:源语言和目标语言句子分别为 $\boldsymbol{f} = f_1\cdots f_m$,$\boldsymbol{e} = e_1\cdots e_n$,有下列定义:

定义 1(对齐点) 如果源语词 f_j 与目标语词 e_i 存在对应关系,则称 (j,i) 是一个连接,也称之为对齐点.

定义 2(对齐矩阵) 与句对 $(\boldsymbol{f},\boldsymbol{e})$ 对应的 $m\times n$ 阶的矩阵 $\boldsymbol{A}$ 被称作对齐矩阵.

$$\boldsymbol{A}(i,j) = \begin{cases} 1, & (i,j)\text{ 是一个连接;} \\ 0, & \text{其他.} \end{cases}$$

设源语言到目标语言的词对齐矩阵为 $\boldsymbol{A}_1$,相应的目标语言到源语言的词对齐矩阵为 $\boldsymbol{A}_2$,将两个方向上的词对齐结果中的连接重新进行组合得到的矩阵 $\boldsymbol{A}$ 称为词对齐重组矩阵.

双语词对齐的重组方法主要有:intersect, union, grow, grow-diag, grow-diag-final, grow-diag-final-and 等.

2 短语与词汇相似度约束

2.1 严格短语与非严格短语

定义 3 设:$\boldsymbol{f} = f_1\cdots f_m$,$\boldsymbol{e} = e_1\cdots e_n$ 分别为源语言和目标语言句子,a 是两个句子上的对齐,则短语互译对 $< e_{i_1}\cdots e_{i_m}, f_{j_1}\cdots f_{j_n} >$ 是与 a 一致的,当且仅当有下列条件成立:

1) $\forall j \neg\ \exists i'((i',j)\in a, i'\notin\{i_1,\cdots,i_m\}, j\in\{j_1,\cdots,j_n\}$;

2) $\forall i \neg\ \exists j'((i,j')\in a, i\in\{i_1,\cdots,i_m\}, j'\notin\{j_1,\cdots,j_n\}$;

3) $\exists k,l((i_k,j_l)\in a, 1\leqslant k\leqslant m, 1\leqslant l\leqslant n$.

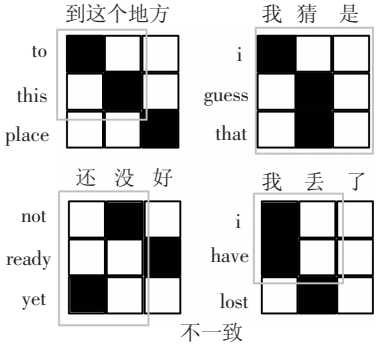
图 1(a)给出了几个一致与不一致的短语示例. 可以看出,这种短语抽取方法是严格按照词对齐进行的,因此本文称此类短语为严格短语.

由于严格短语完全符合词对齐限制,因此它的抗噪声能力不强,这在一定程度上影响了严格短语抽取的准确性. 本文尝试放宽一致性的条件,使得短语对中的词可以对齐到该短语之外. 只要这个词同时也和短语内的某个词对齐,也就是满足条件:

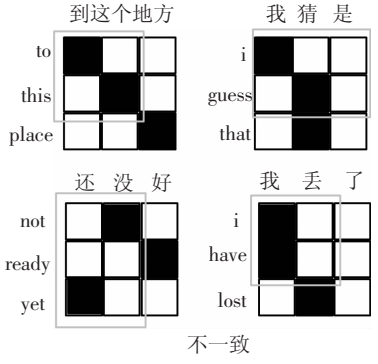
1) $\forall j((\exists i(i,j)\in a)\vee(\neg\ \exists i'(i',j)\in a))$, $i\in\{i_1,\cdots,i_m\}, i'\notin\{i_1,\cdots,i_m\}, j\in\{j_1,\cdots,j_n\}$;

2) $\forall i((\exists j(i,j)\in a)\vee(\neg\ \exists j'(i,j')\in a))$, $i\in\{i_1,\cdots,i_m\}, j\in\{j_1,\cdots,j_n\}, j'\notin\{j_1,\cdots,j_n\}$.

称这种短语为非严格短语,如图 1(b) 所示.



(a) 严格短语抽取



(b) 非严格短语抽取

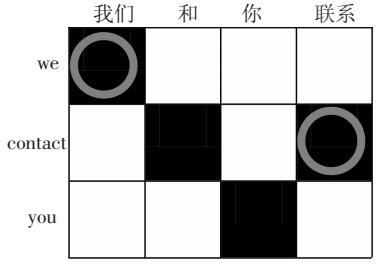
图 1 严格短语抽取与非严格短语抽取中的一致和不一致

2.2 约束短语

非严格短语抽取方法所需满足的条件过于宽泛,有时候会抽取到不完整的短语互译对. 例如,对于图 2 中的情形,因为在词对齐(黑框)中"we"同时对齐到"我们"和"联系",非严格短语抽取方法会抽取到错误短语互译对:"和 你 联系⇔we contact you".

本文尝试采用对对齐点进行约束的办法来避免这种情况,使得抽取到的短语互译对包含比较确定的互译词对,例如"我们⇔we"、"联系⇔contact"(灰圆圈),从而避免正确的互译词对在短语抽取中被拆开,以便抽取到正确的短语互译对. 即

增加条件.



无约束：<和你联系，we contact you>
有约束：<我们 和 你 联系，we contact you>

图 2　约束对短语抽取的影响

定义 4　称满足条件：

$\forall (i,j) \in \{(i,j) \mid \mathrm{sim}(e_i,f_j) \geqslant \theta,(i,j) \in a\}$，$((i \in \{i_1,\cdots,i_m\} \wedge j \in \{j_1,\cdots,j_n\}) \vee (i \notin \{i_1,\cdots,i_m\} \wedge j \notin \{j_1,\cdots,j_n\}))$.

的短语为 θ 约束短语，简称为约束短语，其中，$\mathrm{sim}(e_i,f_j)$ 是词汇相似度度量函数，θ 是阈值.

3　词汇相似度约束

给出 3 种相似度度量函数作为 $\mathrm{sim}(e_i,f_j)$：

1）Dice 系数（Dice Coefficient）.

设：$\#(e)$ 为目标语词 e 出现的频次，$\#(f)$ 为源语言词 f 出现的频次，$\#(e,f)$ 为 e 和 f 共现的频次，则这两个词的 Dice 系数定义为

$$\mathrm{Dice}(e,f) = \frac{2 \cdot \#(e,f)}{\#(e) + \#(f)}. \tag{2}$$

Dice 系数的值介于 [0,1] 之间. 数值越大表示两个词的相似度越高.

2）Phi 平方系数（Phi-Square Coefficient）.

在这种方法中，不仅要考察两个词同现的情况，还要考察两个词不同现的情况. 为此，对于每一个源语言词 f 和每一个目标语词 e，作联列表如表 1 所示.

表 1　联列表

源语言词	e	$\sim e$
f	a	b
$\sim f$	c	d

表 1 中 a 为同时包含目标语词 e 和源语言词 f 的句对数，b 为包含词 e，但不包含词 f 的句对数，c 为不包含词 e，但包含词 f 的句对数，$d = N - a - b - c$ 是不包含词 e 和 f 的句对数，N 为语料中句对总数.

Phi 平方系数 ϕ^2 是通过联列表来计算两个词的相似度的常用方法.

$$\phi^2(e,f) = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}. \tag{3}$$

ϕ^2 的值也介于 [0,1] 之间，值越大表示两个词之间的相似度越高.

3）对数似然比（Log Likelihood Ratio，LLR）.

通过联列表计算词汇相似度的另一种方法是对数似然比，又称为 G^2 - 统计量[10]，定义为

$$G^2(e,f) = 2\log \frac{B(a \mid a+b,p_1) \cdot B(c \mid c+d,p_2)}{B(a \mid a+b,p) \cdot B(c \mid c+d,p)}. \tag{4}$$

式中：$B(k \mid n,p) = \binom{n}{k} p^k (1-p)^{n-k}$ 为贝努利概率，各概率值通过最大似然估计得到：$p_1 = \frac{a}{a+b}$，$p_2 = \frac{c}{c+d}$，$p = \frac{a+c}{a+b+c+d}$，其中，a，b，c，d 即为表 1 中的各项. G^2 - 统计量的值越大表示两个词之间的相似度越高，该统计量比较适合小样本量的统计.

4　结果与讨论

4.1　严格短语模型和非严格短语模型的对比实验

在 IWSLT2004 汉英翻译数据集上测试并比较了 Koehn 的严格短语抽取方法和本文提出的非严格短语抽取方法. 其中训练集为 20 000 句汉英句对，测试集为 500 句汉语句子. 在这里对数据集略作处理：利用哈工大分词工具[11]对汉语部分重新进行了分词，英语部分则进行了切分. 严格短语抽取和解码方面采用了 Pharaoh 工具包. 在翻译结果中去掉了除“'”以外的所有标点符号，并且合并了类似“I 'll”这样的缩写. 翻译结果采用 BLEU 自动评价方法[12]进行评价，评价结果如表 2 所示.

表 2　严格短语抽取与非严格短语抽取的评价结果（BLEU）

对齐重组方法	严格短语抽取	非严格短语抽取
grow-diag-final-and	0.352 1	0.364 6
grow-diag-final	0.394 8	0.407 2
grow-diag	0.330 3	0.342 9
grow	0.278 4	0.299 4
intersect	0.283 8	0.283 8
union	0.377 5	0.410 6

从表 2 中可以看出，对齐重组方法的不同对最终翻译结果 BLEU 评分的影响较大. 但是非严格短语抽取的 BLEU 评分普遍好于严格短语抽取（两者基于 intersect 对齐的结果相同）. 这是因为非严格短语本身具有一定的抗噪声能力，从而减轻了对词对齐准确性的要求.

4.2　词汇相似度约束的影响

表 3 给出了不同约束策略对非严格短语抽取的影响. 从总体上看，应用约束后翻译结果普遍有所提高. Dice 系数约束对于 BLEU 评分提升幅度不大，效果不明显. Phi 平方系数约束在各种对齐重组方法下都能较为显著地提高 BLEU 评分，因

此是一个通用有效的约束策略. 对数似然比约束只对 union, grow, grow-diag 这些词对齐重组方法有效, 在 grow-diag-final, grow-diag-final-and 词对齐重组方法下 BLEU 评分有显著的降低.

表 3 不同约束策略下的评价结果(BLEU)

对齐重组方法	union	grow	grow-diag	grow-diag-final	grow-diag-final-and
无约束	0.410 6	0.299 4	0.342 9	0.407 2	0.364 6
Dice 系数	0.414 8	0.301 5	0.343 9	0.413 0	0.366 0
Phi 平方系数	0.418 5	0.302 6	0.347 7	0.420 6	0.367 3
对数似然比	0.424 7	0.374 9	0.421 9	0.347 7	0.301 1

Dice 系数仅仅考虑了双语词同现的情况, 没有考虑不同现的情况, 难以形成有效的约束, 效果不好. Phi 平方系数方法不仅考虑双语词同现的情况, 还考虑了双语词不同现的情况, 有利于避免间接共现这样的问题, 其约束效果比 Dice 系数方法要好. 而对数似然比方法虽然对于低频词有比较好的效果, 但是当应用于 grow-diag-final 和 grow-diag-final-and 词对齐重组方法时, 较易于仅将约束限制在新加入的对齐点上, 反而限制了短语抽取的有效性; 相反地, 当采用 union、grow、grow-diag 这些能够召回较多的词对齐点的重组方法时, 由于有较多的对齐点进行短语抽取, 限制减少, 其结果是 3 种约束策略中最好的.

5 结 论

1) 同样的对齐重组方法, 非严格短语模型的翻译评价结果好于严格短语模型.

2) 词汇相似度约束策略对于翻译结果的影响: Dice 系数约束策略效果不明显; Phi 平方系数约束策略普遍有效; 对数似然比约束策略虽然只对 union, grow, grow-diag 3 种词对齐重组方法有效, 但在这 3 种方法上的结果却是最好的.

参考文献:

[1] WANG Y. Grammar Inference and Statistical Machine Translation[D]. Pittsburgh: Carnegie Mellon University, 1998.

[2] OCH F J, NEY H. A systematic comparison of various statistical alignment models[J]. Computational Linguistics, 2003, 29(1):19 – 51.

[3] KOEHN P, OCH F J, MARCU D. Statistical phrase-based translation[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Morristown NJ: Association for Computational Linguistics, 2003: 48 – 54.

[4] MARCU D, WONG W. A phrase-based, joint probability model for statistical machine translation[C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Morristown NJ: Association for Computational Linguistics, 2002: 133 – 139.

[5] ZHANG Y, VOGEL S, WAIBEL A. An integrated phrase segmentation and alignment algorithm for statistical machine translation[C]//Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE′03). New York: IEEE Xplore, 2003: 567 – 573.

[6] 程崴. 限定领域内汉英口语的统计翻译方法研究[D]. 北京:中国科学院自动化研究所, 2003.

[7] VOGEL S, ZHANG Y, HUANG F, *et al.* The CMU statistical machine translation system[C]//Proceeding of the Ninth Machine Translation Summit. [S. l.]: [s. n.], 2003:110 – 117.

[8] MARCU D, WANG Wei, ECHIHABI A, *et al.* Spmt: Statistical machine translation with syntactified target language phrases[C]//Proceedings of the 2006 Conference on Empirical Methods in Natural Language. Morristown NJ: Association for Computational Linguistics, 2006: 44 – 52.

[9] WATANABE T, TSUKADA H, ISOZAKI H. Left-to-right target generation for hierarchical phrase-based translation[C]//Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Morristown NJ: Association for Computational Linguistics, 2006:777 – 784.

[10] DUNNING T. Accurate methods for the statistics of surprise and coincidence[J]. Computational Linguistics, 1993, 19(1): 61 – 74.

[11] 赵铁军, 吕雅娟, 于浩, 等. 提高汉语自动分词精度的多步处理策略[J]. 中文信息学报, 2001, 15(1): 13 – 18.

[12] PAPINENI K, ROUKOS S, WARD T, *et al.* BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Conference of the Association for Computational Linguistics (ACL – 02). Morristown NJ: Association for Computational Linguistics, 2002:311 – 318.

(编辑 张 红)