文章编号：1003-0077(2011)03-0118-05

一种考虑对齐不一致的短语翻译概率估计方法

苏劲松[1,2]，刘 群[1]，吕雅娟[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. 中国科学院 研究生院，北京 100190)

摘 要：在统计机器翻译中，短语翻译概率特征对最终的翻译结果有着重大的影响。传统的估计方法只考虑了双语短语同时出现，满足对齐一致性的情况，而没有对其他情况进行统计，因而短语翻译概率的估计不够准确。该文中，我们修改了传统的短语概率计算公式，在估计概率的过程中充分地考虑短语的各种出现情况。多个测试集上的实验结果证明了我们方法的有效性。

关键词：统计机器翻译；对齐不一致；短语翻译概率

中图分类号：TP391　　**文献标识码**：A

A Phrase Translation Probability Estimation Method Considering Alignment Unconsistency

SU Jinsong[1,2], LIU Qun[1], LV Yajuan[1]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Beijing 100190, China;
2. Graduate University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The phrase translation probability features have great effect on the statistical machine translation. The traditional method has a deficiency in the estimation of phrase translation probability by just dealing with the phrases with consistent word alignments. In this paper, we modify the traditional formula to consider all occurrences of phrases in the corpus. The experimental results on the various test sets demonstrate the effectiveness of our method.

Key words: statistical machine translation; alignment unconsistency; phrase translation probability

1 引言

统计机器翻译近年来取得了巨大的发展，陆续出现了多种翻译模型[1]。从翻译基本单元来区分，模型主要可以分为下面三种：基于“词”的翻译模型[2]，基于“短语”的翻译模型[3-4]和基于“句法”的翻译模型[5-10]。后面两种模型可以说是当前统计机器翻译的主流模型，它们都采用了对数线性模型来融入多种特征。使用的特征主要包括翻译概率特征，语言模型分数特征，规则个数特征以及其他和具体模型相关的特征。在这些特征当中，短语① 翻译概率特征衡量了在已知源短语 f(目标短语 e)的情况下翻译成目标短语 e(源短语 f)的概率，该特征对机器翻译的最终结果有着巨大影响。

传统的短语翻译概率估计采用了最大似然估计的方法。已知源短语 f 和目标短语 e，短语翻译概率计算公式如下：

$$p(e \mid f) = \frac{count(f,e)}{\sum_{e'} count(f,e')} \quad (1)$$

$$p(f \mid e) = \frac{count(f,e)}{\sum_{f'} count(f',e)} \quad (2)$$

① 这里的短语即翻译规则，在层次短语中，即包括词汇化规则，也包括泛化规则。

收稿日期：2010-09-16　**定稿日期**：2010-12-21

基金项目：国家自然科学基金重点资助项目(60736014,60873167)

作者简介：苏劲松(1982—)，男，博士生，主要研究方向为自然语言处理，机器翻译；刘群(1966—)，男，博士，研究员，主要研究方向为自然语言处理，机器翻译；吕雅娟(1972—)，女，博士，副研究员，主要研究方向为自然语言处理，机器翻译。

从上面公式，我们可以清楚地看到，公式(1)和(2)只统计了源短语 f 和目标短语 e 满足对齐一致性的情况。然而，在语料库中，短语的分布情况却并非如此。对于某个源短语 f，并不一定存在满足对齐一致性的目标译文；同样，对于某个目标短语 e，也不一定存在满足对齐一致性的源译文。如图 1 所示，我们列举了两个例子来阐述传统估计方法的缺陷。

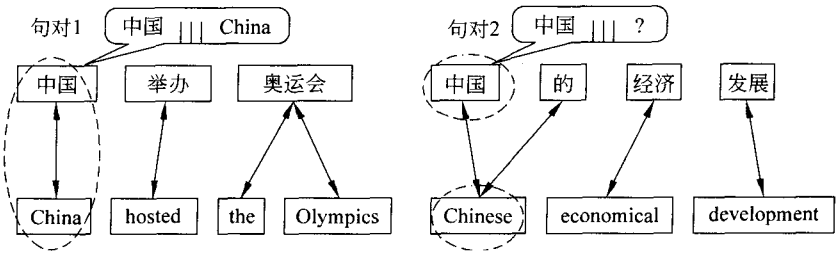

图 1　短语模型短语概率估计样例图

假设语料库中只含有以上两个句对(图 1)。在句对 1 中，我们可以抽取出双语短语("中国","China")，而在句对 2 中，源短语"中国"不存在满足对齐一致性的译文，因而不能抽取出合适的双语短语。根据传统的短语概率估计方法，我们可以得到 p("China"|"中国")=1。这里需要注意的是，在句对 2 中，"中国"并不是没有翻译，它和"的"合并在一起，翻译成目标短语"Chinese"。传统的短语概率估计方法并没有考虑这种情况，因而会过高地估计源短语"中国"翻译成目标短语"China"的概率。在翻译含有"中国 的"的新句子时，翻译模型会倾向于把所有出现的"中国"都翻译成"China"，而不选择由"中国 的"来翻译成"Chinese"。因而，我们需要对"源短语不满足对齐一致性译文"的情况赋予一定的概率，以使得翻译模型能够做出更好的源短语选择。

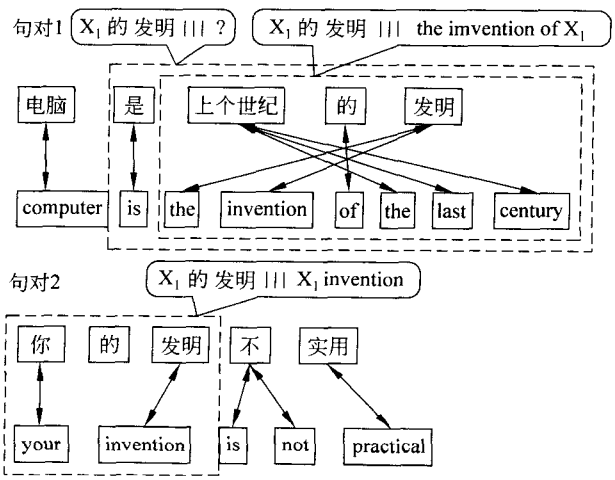

图 2　层次短语模型泛化规则概率估计样例图

再来看第二个例子。假设语料库中同样含有两个句对(图 2)，在层次短语模型中，根据传统的短语概率估计方法，我们可以得到 p("the invention of X_1"|"X_1 的 发明")=1/3。在此，我们先考虑句对 1。在窗口"上个世纪 的 发明"中，我们可以抽取出泛化规则"X_1 的 发明 ||| the invention of X_1"；而在窗口"是 上个世纪 的 发明"中，源泛化规则"X_1 的 发明"不存在满足对齐一致性的目标译文。与例 1 的道理相同，传统方法也会过高地估计了源泛化规则"X_1 的 发明"翻译成目标泛化规则"the invention of X_1"的概率。

基于以上分析，我们可以清楚地看到传统估计方法存在一定的缺陷。对此，本文对传统估计方法进行了改进，使得计算时能够充分考虑训练语料库中短语的各种情况。实验结果证明，我们的方法可以获得更为准确的短语概率估计，有效地提升了系统的翻译性能。

文章的组织结构如下：第 2 节详细介绍了我们的短语翻译概率估计改进方法。针对不同翻译模型，我们进行不同处理。特别地，针对层次短语模型的泛化规则，我们尝试了两种不同的翻译概率估计方法；第 3 节描述了实验设置和结果；第 4 节是文章的总结和展望。

2　考虑对齐不一致的短语翻译概率估计方法

根据前文分析，我们可以得知，传统的短语翻译概率 $p(e|f)$，可以看成是在已知源短语 f 存在满足对齐一致性的目标译文的情况下翻译为目标短语 e 的概率，同样 $p(f|e)$ 可以理解为在已知目标短语 e 存在满足对齐一致性的源译文的情况下翻译为源短语 f 的概率。在此，我们引入隐变量 b_{ue}(b_{uf})来表示给定源短语(目标短语)是否存在满足对齐一致性的目标短语(源短语)。传统估计方法可以表示为：

$$p(e \mid f, b_{we} = true) = \frac{count(f,e)}{\sum_{e'} count(f,e')} \quad (3)$$

$$p(f \mid e, b_{vf} = true) = \frac{count(f,e)}{\sum_{f'} count(f',e)} \quad (4)$$

在翻译过程中，对于源短语 f（目标短语 e），我们并不能确定它是否一定存在满足对齐一致性的目标译文（源译文）。因此可以说，传统方法估计的短语翻译概率和翻译过程并不对应。基于以上分析，我们对以上公式进行修改，修改后的公式如下：

$$p(e, b_{we} = true \mid f) = \frac{count(f,e)}{count(f)} \quad (5)$$

$$p(f, b_{vf} = true \mid e) = \frac{count(f,e)}{count(e)} \quad (6)$$

根据以上公式，在计算短语翻译概率的过程中，我们将考虑源短语 f（目标短语 e）的各种出现情况，然后进行频次累加，最后再进行归一化，以此来获得更为准确的概率估计。

然而，在不同模型中，频次累加的计算方式并不相同。对此，我们根据不同模型进行了不同处理。

2.1 短语模型

在短语模型中，频次累加的计算方式较为简单。根据文献[4-5]所描述的方法，若源短语 f（目标短语 e）存在满足对齐一致性的译文，它的频次累加值是1。而在我们的方法中，我们需要对不存在满足对齐一致性的译文的短语进行频次累加。在此，我们设置所有情况下短语的频次累加值都是1。对于公式(5)和(6)，我们不需要做特殊处理，直接统计每个句对中双语短语(f,e)，源短语 f 和目标短语 e 的出现次数。

考察图1中的例子，源短语“中国”总的出现次数为2，其中存在满足对齐一致性的次数为1次。根据公式(5)，我们可以得到 p(“China”|“中国”)=1/2。

2.2 层次短语模型

在层次短语中，翻译规则主要包括词汇化规则和泛化规则。词汇化规则的频次估计方法和短语模型中短语的估计方法是一样的。在此，我们也采用上述方法进行统计。而泛化规则的频次估计方法则稍微复杂一些，根据文献[5]描述的方法，在抽取泛化规则的过程中，当规则存在满足对齐一致性限制的译文时，它的频次累加值等于选定窗口可以抽取的泛化规则数目的倒数。例如：在图2的句对1中，从窗口“上个世纪 的 发明”，我们可以得到 *count*(X_1的 发明 ‖ | the invention of X_1)=1/6；在图2的句对2，从窗口“你 的 发明”，我们可以得到 *count*(X_1的 发明 ‖ | X_1 invention)=1/3。然而，当不存在满足对齐一致性的译文时，我们并没有办法正确计算对应的频次累加值。

下面我们采用两种方法来进行计算泛化规则的翻译概率：

方法1，我们采用与短语模型相类似的方法：对于源规则 f（目标规则 e），无论它是否存在满足对齐一致性的译文，我们都假设它的频次累加值是1。

考察图2中的例子，源泛化规则“X_1的 发明”在句对1的出现次数为3（分别在“电脑 是 上个世纪 的 发明”，“是 上个世纪 的 发明”和“上个世纪 的 发明”三个窗口中），在句对2中的出现次数为1（在“你 的 发明”窗口中）。根据公式(5)，我们可以得到 p(“the invention of X_1”|“X_1的 发明”)=1/(3+1)=1/4。

方法2，首先，我们假设“源规则 f 的译文是目标规则 e”和“源规则 f 存在满足对齐一致性的目标译文”是不相关的。然后，我们把概率 $p(e, b_{we}=true|f)$ 分解成为两个条件概率 $p(e|b_{we}=true, f)$ 和 $p(b_{we}=true|f)$ 的乘积。同理，我们也可以做类似假设，把 $p(f, b_{vf}=true|e)$ 进行分解。分解后的公式如下：

$$p(e, b_{we} = true \mid f) = p(e \mid b_{we} = true, f) \times p(b_{we} = true \mid f) \quad (7)$$

$$p(f, b_{vf} = true \mid e) = p(f \mid b_{vf} = true, e) \times p(b_{vf} = true \mid e) \quad (8)$$

对于 $p(b_{we}=true|f)$ 和 $p(b_{vf}=true|e)$，我们设定源短语 f（目标短语 e）的频次累加值都是1；而对于 $p(e|b_{we}=true, f)$ 和 $p(f|b_{we}=true, e)$，我们则仍然采用传统方法进行估计。

仍然考察图2中的例子。根据传统方法计算可得 p(“the invention of X_1”$|b_{we}=true$，“X_1的 发明”)=1/6/(1/6+1/3)=1/3；而源泛化规则“X_1的 发明”总的出现次数为3+1=4次，其中存在满足对齐一致性译文的次数为1+1=2次，可得 $p(b_{we}=true|$“X_1的 发明”)=1/2。根据公式(5)，我们可以最终得到 p(“the invention of X_1”，$b_{we}=true|$“X_1的 发明”)=1/3×1/2=1/6。

通过以上方法，我们可以获得两个新的短语翻译概率。在实际翻译过程中，我们把这两个新的翻译概率当作新的特征加入到翻译模型中，以此来提

高翻译模型的性能。

3 实验

3.1 实验设置

我们的实验设置如下：

1）训练语料：共含有来自 LDC① 的 1 548 447 个平行句对。该语料库一共含有中文词 42 334 463 个，英文词 48 152 996 个；

2）语言模型：采用 SRLIM② 训练的四元 GIGA XINHUA 语言模型；

3）开发集：NIST③ 02 评测集；测试集：NIST03，NIST05 评测集

4）解码器：Moses④，Bruin，Chiero；

5）译文评价：采用的评测指标是大小写不敏感的 BLEU-4[11]，使用的评测工具是 mteval-v11b.pl⑤。

对于解码器，我们采用最小错误率训练 MERT[12] 进行特征权重训练。解码中，我们设置候选翻译个数为 50，最终翻译 N-best 个数为 100。下面我们对所用解码器进行简单介绍。

Moses[3-4] 是著名的短语模型解码器，由 Philipp Koehn 开发。解码器基于对数线性模型，采用从左到右的方式进行解码。解码器提供多种了调序模型，实验中我们选用了 msd-fe 调序模型。

Bruin 是基于 BTG[6] 的短语模型解码器。该解码器基于对数线性模型，采用 CKY 方式进行解码。为了加快解码速度，解码器采用 Cube-Pruning 方法[13] 来减少搜索空间。

Chiero 是著名层次短语解码器 Hiero[5] 的 C++ 重实现版本。同样，该解码器也是基于对数线性模型，采用 CKY 方式进行解码，并采用 Cube-Pruning 方法来减少搜索空间。

3.2 实验结果

采取以上的实验设置，实验结果见表 1，表 2。

表 1　短语模型 实验结果

测试集	计算方法	Moses	Bruin
NIST03	传统短语翻译概率(公式(1)(2))	32.46	32.81
	＋ 新的短语翻译概率(公式(5)(6))	33.13	33.23
NIST05	传统短语翻译概率(公式(1)(2))	31.87	31.99
	＋ 新的短语翻译概率(公式(5)(6))	32.22	32.50

表 2　层次短语模型 实验结果

测试集	计算方法	Chiero
NIST03	传统短语翻译概率(公式(1)(2))	33.15
	＋ 新的短语翻译概率(方法 1，即公式(5)(6))	33.48
	＋ 新的短语翻译概率(方法 2，即公式(7)(8))	33.67
NIST05	传统短语翻译概率(公式(1)(2))	32.52
	＋ 新的短语翻译概率(方法 1，即公式(5)(6))	33.00
	＋ 新的短语翻译概率(方法 2，即公式(7)(8))	32.92

表 1 和表 2 分别列出在两类模型（短语模型和层次短语模型）上的实验结果。从上面的数据，我们可以清楚地看到，在不同的模型和测试集上，我们的方法都在一定程度上改进了翻译质量（在 NIST03 测试集上取得了 0.33～0.67 的提高，而在 NIST05 测试集上取得了 0.35～0.51 的提高）。因此，可以说我们的方法是有效的，获得了更为准确的短语概率估计。

4 总结与展望

本文对统计机器翻译中短语翻译概率的传统估计方法进行了分析，阐述了导致短语翻译概率估计不准确的主要原因就是传统方法只考虑了短语存在满足对齐一致性的译文的情况。对此，本文修改了传统计算公式，并针对不同翻译模型进行了不同处理。实验结果表明，我们的方法是有效的，在 NIST03 和 NIST05 测试集上，BLEU 值都有所提高。显然，我们的方法仍然存在继续改进的余地：

1）引入上下文来估计 $p(b_{te}=true|f,context)$ 和 $p(b_{vf}=true|e,context)$。

2）在我们的方法中，在估计 $p(b_{te}=true|f)$ 和 $p(b_{vf}=true|e)$ 时，短语的频次累加值简单设置成 1。在以后工作中，我们将引入调序图[14] 来获得更准确

① LDC 语料：http://www.ldc.upenn.edu

② SRILM 工具：http://www.speech.sri.com/projects/srilm/download.html

③ NIST 数据集：http://www.nist.gov/speech/tests/mt

④ Moses 解码器：http://www.statmt.org/moses/

⑤ BLEU 评测工具：http://www.nist.gov/speech/tests/mt/resources/scoring.hml

的频次累加值估计。

3）在层次短语模型中，我们采用了两种方法来计算翻译规则的频次累加值，效果一般。

今后我们将进行更为深入的研究，寻找更为合理的层次短语概率估计的方法。

参考文献

[1] 刘群. 统计机器翻译综述[J]. 中文信息学报，2003，17(4)：1-12.

[2] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. 1993. The mathematics of Statistical Machine Translation: Parameter Estimation[J]. Computational Linguistics. 1993,19:263-311.

[3] Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation[C]//Proc. of NAACL 2003:48-54.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation[C]//Proc. of ACL 2007, demonstration session. 2007: 177-180.

[5] David chiang. Hierarchical Phrase-Based Translation [J]. Computational Linguistics, 2007,33:201-288.

[6] Deyi Xiong, Qun Liu and Shouxun Lin. Maximum Entropy Based on Phrase Reordering Model for Statistical Machine Translation[C]//Proc. of ACL 2006: 521-528.

[7] Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. Scalable Inference and Training of Context-Rich Syntactic Translation Models[C]//Proc. of ACL 2006:961-968.

[8] Yang Liu, Qun Liu and Shouxun Lin. 2006. Tree-to-String Alignment Template for Statistical Machine Translation[C]//Proc. of ACL 2006, 2006:609-616.

[9] Haitao Mi, Liang Huang and Qun Liu. Forest-Based Translation[C]//Proc. of ACL 2008, 2008:192-199.

[10] Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan and Sheng Li. 2008. A Tree Sequence Alignment-based Tree-to-Tree Translation Model [C]//Proc. of ACL 2008, 2008:559-567.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation[C]//Proc. of ACL 2002, 2002:311-318.

[12] Franz Joseph Och. Minimum error rate training in statistical machine translation[C]//Proc. of ACL 2003, 2003:160-167.

[13] Liang Huang and David Chiang. Better k-best Parsing[C]//Proc. of IWPT 2005, 2005:53-64.

[14] Jinsong Su, Yang Liu, Yajuan Lv, Haitao Mi and Qun Liu. Learning Lexicalized Reordering Models from Reordering Graphs[C]//Proc. of ACL 2010: 12-16. short paper.

（上接第111页）

[7] 宗成庆，吴华，黄泰翼，等. 限定领域汉语口语对话语料分析[C]//计算语言学文集(全国第五届计算语言学联合学术会议论文集)，1999：115-122.

[8] 解国栋. 统计口语解析方法研究[D]. 中国科学院自动化所博士论文，2004.

[9] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996,22(1):39-71.

[10] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models[J]. Annals of Mathematical Statistics. 1972,43: 1470-1480.

[11] S. D. Pietra, V. D. Pietra and J. Lafferty. Inducing features of random fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997,19: 380-393.

[12] V. Vapnik. The nature of statistical learning theory [M]. Springer. 1995.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proc. of ICML, 2001: 282-289.

[14] *M. Eck and C. Hori. Overview of the IWSLT 2005* Evaluation Campaign[C]//Proceedings of IWSLT. 2005:11-32.

[15] 左云存，宗成庆. 基于语义分类树的汉语口语理解方法[J]. 中文信息学报，2006，20(2)：8-15.

[16] 解国栋，宗成庆，徐波. 面向中间语义表示格式的汉语口语解析方法[J]. 中文信息学报，2003，17(1)：1-6.