利用 1-m 词对齐信息改善统计机器翻译性能

陈如山，肖桐，朱靖波

东北大学自然语言处理实验室，沈阳，110004

E-mail: chenrs@ics.neu.edu.cn

摘 要： 词对齐是目前主流的统计机器翻译系统必备的模块，而 IBM 模型是词对齐最常用的模型。但是 IBM 模型不允许源语到目标语的一对多词对齐。这一限制在像汉英翻译这样频繁出现一对多对齐现象的任务中，影响了翻译系统的性能。针对这个问题，本文通过将目标语中满足一定条件的 bigram 合并，把一对多问题简化为一对一问题，进而改善词对齐的效果，最终达到提高统计机器翻译系统性能的目的。实验结果表明该方法能够在一定程度上提高汉英翻译的性能。

关键字： 机器翻译；词对齐；翻译对共现频率

Improve Statistical Machine translation performance using 1-m word alignment information

Chen Rushan, Xiao Tong, Zhu Jingbo

Natural Language Processing Lab, Shen Yang, 110004

E-mail: chenrs@ics.neu.edu.cn

Abstract: Word alignment is a necessary component of nowadays' machine translation system, and IBM Model is the most popular statistical word alignment model. But IBM Module doesn't allow one source word to be connected by multiple target words, when this restriction being applied to task like Chinese-English translation where the phenomenon occurs frequently, it will decrease the translation system's performance. Addressing this problem, this paper proposes a method which transforms some 1-m alignments to 1-1 alignments by merging bigrams which satisfy some conditions so as to improve the word alignment's effect, and finally improve the performance of the translation system. The results of the experiments show that this method does work.

Keyword: machine translation, word alignment, TCR

1 引言

统计机器翻译发展到现在，无论是基于短语的还是基于句法的，无一例外的用到了词对齐技术。所谓词对齐是指在源语和目标语构成的翻译对中找到词汇级的互译关系。Och[1]的工作已证明提高词对齐的性能能在一定程度上提高机器翻译系统的性能，因此，许多学者把研究工作集中在提高词对齐来改善统计机器翻译系统的性能[4][5][6][7]。

常用的词对齐模型包括 IBM 模型[2]，隐马尔科夫模型[3]等。为了简化词对齐问题，这两个模型都对词对齐进行了限制——一个源语言单词最多只能对齐到一个目标语单词。这个限制使得 IBM 模型和隐马尔科夫模型本身不能支持一对多和多对多的词对齐。对于这个问题，通常是通过源语和目标语的双向对齐来缓解由于模型的限制所产生的问题。但是像汉英翻译这样频繁出现一对多对齐现象的任务中，存在着大量的一个源语单词翻译成多个目标语单词的现象，比如："最

后”被翻译成“at last”，“后天”被翻译成“the day after tomorrow”，等等。大量一对多问题的存在，仍然会造成IBM词对齐模型性能的下降，从而导致翻译性能的下降。

针对这一问题最直接的想法莫过于直接修改模型，但这种做法的代价太大，我们希望能够通过简单有效的方式使IBM模型能够兼容源语到目标语一对多的对齐。Yanjun Ma[8]考虑利用bootstrapping技术来合并源语和目标语ngrams以简化词对齐，沿用这个思想，本文考虑利用反向对齐信息找出一个源语单词对齐到的多个连续的目标语单词的翻译对，然后通过合并这些连续的目标语单词来对目标语句子重新分词，从而使部分的一对多的问题转换为一对一的问题。此外由于在一对多对齐中，一对二对齐占有很大比重，它们对词对齐的影响较大，所以本文考虑只通过合并目标语的bigram来达到改进词对齐的目的。这样可以更好的发挥IBM等模型在一对一问题上的优势，进而提高词对齐和机器翻译的性能。针对如何选择连续的目标语单词进行合并，本文共提出了6种选择方法。它们分别使用了目标语bigram的词频，一些启发式规则以及源语到目标语翻译对共现频率等信息。实验结果表明充分合理地利用这些信息合并目标语bigram可以提高翻译系统的性能。

本文的第二部分简要说明了本文所用语料中对齐的分布，第三部分对合并目标语bigram的方法做了详细地说明，第四部分给出了实验结果并进行了分析。

2 一对多词对齐

在SSMT2007[1]的训练语料上，通过对GIZA++英汉对齐生成的维特比对齐文件进行统计，得到了如下数据：

表1 对齐统计数据

Tab.1 statistical data of word alignments

	全部	一对一	一对多	一对二	一对二（连续）
对齐数	8747460	4710265	1525753	1031213	776360
比率	100%	53.85%	17.44%	11.79%	8.88%

由此可见一个中文单词对到一个英文bigram在汉英词对齐中是非常普遍的，占全部对齐的8.88%，占一对多对齐的50.88%，该现象的大量存在必将降低IBM等模型在汉英词对齐应用中的性能。因此本文认为如果能找出其中高可信度的英文bigram并予以合并，将会对提高词对齐和机器翻译系统的性能有所帮助。

3 基于目标语bigram合并的词对齐

由于IBM等模型不具备处理一对多对齐的能力，因此本文就考虑利用反向对齐的信息找出一对多的翻译对，对其中部分的目标语bigram进行合并，把一对二问题转化为一对一问题以提高词对齐的精度，进而提高翻译系统的性能，为此本文提出了如下几种合并目标语bigram的方法：

3.1 朴素目标语bigram合并法

将在反向对齐中被对齐的所有目标语bigram进行合并。该方法不考虑一致性问题，同样的

[1] http://mitlab.hit.edu.cn/ssmt2007.html

目标语 bigram，可能在一个句子中合并，在另一个句子中不合并，这都取决于所在目标语句子对应翻译的对齐信息。

例如在反向对齐中存在如下对齐：

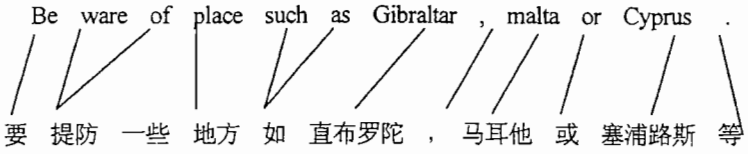

图 1 对齐实例 1

Fig. 1 alignment instance 1

则把 such 和 as 合并为一个目标语单词。

3.2 基于词频的目标语 bigram 合并法

将对齐次数高于一定阈值的 bigram 进行合并。该方法考虑了一致性问题，即一个 bigram 在是否合并必须在整个语料中是一致的。

若以 align(tbigram)表示目标语对齐次数，当出现例 1 所示词对齐时，align(such as)加 1，以 θ 表示阈值，则只有当 align(such as)大于 θ 时，才将 such as 这一 bigram 合并为一个单词，并且整个语料中的 such as 都将进行合并。

本文将词频阈值设定为 10。

3.3 基于源语到目标语翻译对共现频率的目标语 bigram 合并法

将源语到目标语翻译对共现频率高于一定阈值的 bigram 进行合并。bigram 是否合并在整个语料中是一致的。

翻译对共现频率由以下公式计算：

$$TCR = \frac{tran(tbigram, sword)}{cooc(tbigram, sword)} \tag{1}$$

其中：tbigram 表示目标语 bigram，sword 表示源语单词，tran(tbigram, sword)表示 tbigram 和 sword 构成的翻译对的出现次数，cooc(tbigram, sword)表示 tbigram 和 sword 的共现次数，TCR 表示翻译对共现频率。

当同时出现例 1 和如下所示对齐时：

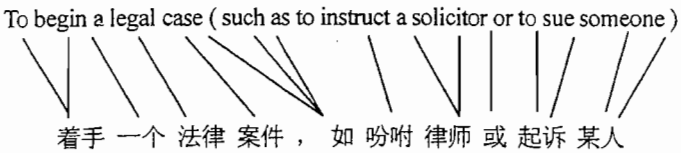

图 2 对齐实例 2

Fig. 2 alignment instance 2

公式中的 tran(such as ，如)加 1，cooc(such as, 如)加 2。

本文中将 TCR 阈值设定为 70%。

3.4 基于启发式规则的目标语 bigram 合并法

将满足一定启发式规则的bigram进行合并。Bigram是否合并在整个语料中也同样是一致的。这里的启发式规则是指人为地构造一些规则过滤掉本文认为没有意义的目标语 bigram。

例如存在一条启发性规则对所有包含as的bigram都不进行合并，则即便出现例1所示对齐，也不将 such as 进行合并。

3.5 基于上述标准组合的目标语 bigram 合并法

表2 组合方法

Tab.2 criteria combination

名称	组合方法
方法5	词频，源语到目标语翻译对共现频率的组合
方法6	词频，源语到目标语翻译对共现频率，启发式规则的组合

4 实验结果与分析

4.1 数据，Baseline 系统以及评测方法

本文使用 Moses 工具包搭建了一个 baseline 系统，训练语料和测试语料来自 SSMT 2007，其中训练语料包含 84 万句中英翻译句对，测试语料包含 1002 个中文句子，本文实验没有使用 MERT (Minimum Error Rate Training) 进行参数优化[2]，参数均使用系统默认值。语言模型的训练是在上述训练语料的英文语料上进行的。实验结果以机器翻译常用的指标 BLEU-4[9]值作为评价标准。

4.2 实验结果

Baseline 系统的性能为 19.48，下表为应用本文提出的 6 种方法后系统的 BLEU 值：

表3 实验结果（其中表格中的序号 1~6 与第 3 节中方法出现顺序一致）

Tab.3 experiment results (the numbers 1~6 in this table are consistent with the order of method shown in section 3)

序号	词频	TCR	启发式规则	BLEU
1	—	—	—	19.34
2	10	—	—	19.58
3	—	—		19.29
4	—	70%	—	19.76
5	10	70%	—	19.14
6	10	70%		19.51

由该表可以看出方法 4 的性能最好。

经统计，各方法的 bigram 连接情况如下表所示：

[2] 由于本文没有进行 MERT，因此没有使用 SSMT2007 开发集数据。

表 4 bigram 连接情况

Tab.4 bigram merging status

序号	被合并的 bigram 总数	被合并的 bigram 类别总数
1	776360	118834
2	1869062	8594
3	369273	78981
4	411411	81835
5	301711	6319
6	161105	3490

方法 1 对 bigram 的合并上下文不一致，同时合并的 bigram 总数很大，类别很多，上述两点造成训练语料出现了许多新的英文单词，使原本就不充分的数据更加稀疏，这也必将导致模型训练得不充分，因此导致了翻译系统性能的下降。

方法 2 相对于方法 1 加入了对词频的限制，这么做有利于得到可信度较高的 bigram，从表 4 可以看出这一限制使 bigram 类别数大大降低，达到了我们获取较高可信度的 bigram 的目的，从实验的结果可以看出，这一做法确实提高了系统的性能。但通过观察生成的新的英文语料，可以发现许多诸如“of the”，“’ s fine”这样的 bigram 被合并在一起，对于本文的方法来说，真正有意义的应该加以合并的应该是被源语大量对齐的那部分 bigram，bigram 的合并不应该是由于该目标语 bigram 的大量存在或词对齐的不确定性造成的，所以系统性能的提升不是那么明显。

针对方法 2 中合并大量无意义 bigram 的问题，方法 3 考虑用人为的启发式规则过滤掉部分频繁出现的 bigram，如带有“the”，“them”，“and”，“she”等的 bigram，通过对该方法的实验，本文发现单纯的使用启发式规则不仅不能带来性能的上升，反而使性能下降，究其原因，本文认为虽然经过启发式规则的过滤，但想单纯的依靠语言学的知识来改善词对齐是行不通的。

方法 4 不仅考虑了目标语，而且同时考虑了源语以及源语到目标语的对齐，这两个信息的加入很有助于发现本文认为有意义的那些 bigram，实验结果也同样证明了这一点，系统性能得到较大的提升。

最后本文考虑使用对不同指标进行组合的方式进行进一步的实验，实验结果显示这两种组合的方法并没有带来性能的提升，反而造成性能的下降，究其原因，本文认为标准与标准之间的关系十分复杂，简单地以逻辑与的方式对不同的标准进行组合并不能带来系统性能的提升。

5 结论及未来工作

本文提出了一种基于目标语 bigram 合并的改进词对齐的方法，利用包括词频，源语到目标语的翻译对共现频率，启发式规则等信息合并目标语 bigram，对目标语语料进行重新分词，并实验了 6 种 bigram 的合并方法，实验结果表明，充分合理地利用以上信息得到的新的目标语语料可以在一定程度上提高机器翻译系统的性能。

下一步的工作中，我们将尝试对实验中用到的参数进行优化，并考虑引入 bootstrapping 技术使本文的方法能够合并连续的 3 个或 4 个目标语单词，从而进一步提高词对齐和机器翻译的性能。

参考文献

[1] Franz Josef Och and Hermann Hey, A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics archive Volume 29, Issue 1, pp. 19-51, 2003

[2] Perter F. Brown, The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 1993

[3] Stephan Vogel, HMM-Based Word Alignment in Statistical Translation. Coling 1996

[4] Percy Liang, Alignment by Agreement. NAACL 2006

[5] Yonggang Deng, Guiding Statistical Word Alignment Models With Prior Knowledge. ACL 2007

[6] Phil Blunsom, Discriminative Word Alignment with Conditional Random Fields. ACL 2006

[7] John DeNero, Tailoring Word Alignments to Syntactic Machine Translation. ACL 2007

[8] Yanjun Ma, Bootstrapping Word Alignment via Word Packing. ACL 2007

[9] Kishore Papineni, Bleu: A Method for Automatic Evaluation of Machine Translation. IBM Technical Report 2001