基于组合的短语规则抽取方法

李强[†]，高永白[†‡]，肖桐[†]，张浩[†]，朱靖波[†]
[†]东北大学自然语言处理实验室　[‡]朝鲜平壤计算机技术大学
{liqiangneu,zhanghao1216}@gmail.com
{xiaotong,zhujingbo}@mail.neu.edu.cn
gaoyongbai@ics.neu.edu.cn

摘要

本文提出了一种基于组合的短语规则抽取方法。该方法受到句法规则抽取方法（Galley等，2006）的启发。该方法与已往方法的不同之处在于，目前最普遍使用的短语规则抽取方法将抽取所有与平行句对词对齐信息保持一致的规则。本文提出的方法则首先定义最小短语规则集，然后从最小短语规则集中组合出一个更加紧凑的、包含更多上下文信息的短语规则集。实验结果表明，与目前普遍使用的短语规则抽取方法相比，在保证翻译性能不降低的情况下，本文的方法抽取的短语规则集的规模比基准短语集减小56.6%。在进行短语规则组合时，同时考虑翻译性能与短语规则大小的情况下，组合两次的短语规则已基本达到最优。

1　简介

基于短语的统计机器翻译系统在机器翻译领域的性能表现出非常强的竞争力。基于短语的方法之所以有效很大一部分原因在于该方法依赖一个质量较高的短语规则集。在短语规则集中，每一个源语言短语被映射到一个或多个不同的目标语短语。在短语系统中，短语由句子中的一系列连续的单词构成，短语并没有语言学意义。目前，一些机器翻译领域研究人员（Och和Ney，2000；Deng等，2008）已经提出一些行之有效的短语规则抽取方法。在这些短语规则抽取方法中，Koehn等（2003）提出的启发式方法得到了广泛的应用。该抽取方法通过使用双语语料中每个句子对应的词对齐信息，抽取出所有与词对齐信息保持一致的短语规则。由于该规则抽取方法简单、易于实现，同时表现出非常优越的性能，所以在目前基于短语的统计机器翻译系统中得到了广泛的应用。我们知道，在使用（Koehn等，2003）的方法抽取短语规则的过程中，最终抽取出来的短语规则的数量与训练数据中单词的数量成二次方关系（Koehn，2008）。为了得到一个规模可控的短语规则集，通常的做法是对抽取的源语言及目标语言短语的长度加以限制。在多数的性能优异的机器翻译系统中，默认设置将抽取的源语和目标语短语所含单词个数的上限设置为7到10个词。例如，Moses[1]将抽取出来的短语的源语言端与目标语言端的长度限制为7个词。

Johnson等（2007）已经证明将短语规则集中的大部分规则删除并不会影响翻译系统的性能。为了减小短语规则集的大小，目前最普遍使用的方法是对现有的启发式规则抽取方法抽取出的短语规则进行过滤，从而减小短语规则集的大小。为了获得一个可操作的、规则数量合理的短语规则集，本文给出了另一种解决方案，即提出了基于组合的短语规则抽取方法[2]。该规则抽取方法的基本观点是，首先在双语语料中构造一个“最小短语规则”，然后通过组合最小短语规则来构造一个含有更多上下文信息、质量优良的短语规则集，在本文称之为“组合的短语规则”。在本文中，*n-composed*短语规则的含义是该规则可以由*1-n*个最小短语规则组成，即*(n-1)-composed*短语规则包含在*n-composed*规则之中。在基于组合的规则抽取方法中，规则集的大小是通过组合规则中*n*的取值进行控制的，即*n*取值越大，得到的规则集越大。这与以往

[1]http://www.statmt.org/moses/

[2]http://www.nlplab.com/members/liqiang/ext.usage.html，本文工作可以直接应用到NiuTrans开源系统中

规则抽取方法中通过限制源语与目标语短语所含单词的最大个数有所不同。经过NiuTrans开源系统[3]（Xiao等，2012）中的基于短语的统计机器翻译系统的验证，与Moses中默认设置的规则抽取方法相比，本文提出的基于组合规则抽取方法在保证翻译性能不降低的情况下，可以得到了一个更加简洁的短语规则集。当抽取*2-composed*短语规则时，本文的抽取方法得到的翻译规则的质量与Moses的默认规则集相当，同时短语规则集大小为Moses默认设置规则集的56.5%。实验结果同样表明，当随着组合最小短语规则次数的增多，翻译系统的性能与*2-composed*短语规则性能相比并没有显著的增长。在同时考虑系统翻译性能与短语规则集大小的情况下，*2-composed*短语规则已基本达到最优。

2 基准抽取方法

本文使用的基准短语规则抽取方法为（Koehn等，2003）中提出的规则抽取方法，该方法在性能优异的基于短语的统计机器翻译系统中得到了广泛使用，如Moses系统，NiuTrans系统。在Koehn等提出的短语规则模型中，短语规则必须满足一致性定义。

定义一：短语对（$\bar{s}$，$\bar{t}$）与词对齐信息保持一致，当且仅当$\bar{s}$中的所有单词在词对齐A中所对应的单词在$\bar{t}$范围之内，$\bar{t}$中的所有单词在词对齐A中所对应的单词在$\bar{s}$范围之内；与此同时，在$\bar{s}$与$\bar{t}$中，至少有一个单词对在词对齐A中。

在定义一中，$\bar{s}$表示源语短语，$\bar{t}$表示目标语短语。Chiang（2007）给出该定义的直观解释：给定一源语短语及目标语短语，在任意一端的短语中，至少有一个单词对应到另一端的短语中；同时，任意一端短语中的所有单词都不可对应到另一端短语之外。通过如上定义，在Koehn等提出的模型下的所有的短语规则都必须满足一致性的定义。我们可以根据如上定义直接从平行语料中抽取与词对齐信息保持一致的短语规则：首先在每一个句对中，从源语与目标语端循环查找所有短语，然后输出与词对齐信息保持一致的短语规则（Koehn，2008）。通过该方法进行短语规则集构造时，在规则抽取的过程中，需要设置抽取短语的所含单词的最大个数，这样才可避免得到规模不可控的短语规则集。

[3]http://www.nlplab.com/NiuPlan/NiuTrans.html

图1中右侧Baseline列表示用基准短语规则抽取方法从示例的含有词对齐信息的句对中抽取的短语规则。从抽取出的短语规则我们可以看出，这些规则均与词对齐保持一致。

3 组合规则抽取方法

前文所述，基准短语规则抽取方法有不可避免的问题，即在规则抽取过程中，短语长度需要进行机械的调试以获取最优的短语规则集。针对上述问题，本文提出了基于组合的规则抽取方法。该方法的动机来源于基于句法的规则抽取方法（Galley等，2006），Galley等提出的规则抽取方法通过组合较小的句法规则来生成较大、含有更多上下文信息、规模可控的句法规则。在本文的工作中，我们的思路与（Galley等，2006）方法相似。首先我们定义了最小短语规则集，之后通过一定方法组合该最小规则来生成更多质量较高、含有更多上下文信息的组合的短语规则集。

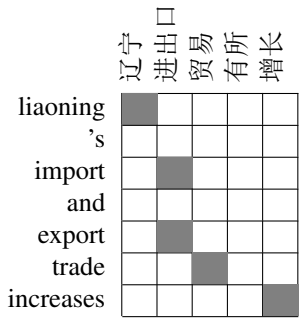
3.1 最小短语规则

在本文提出的基于组合的短语规则抽取方法中，首先关心的问题是什么样的规则才是最小短语规则。

定义二：最小短语规则就是在与词对齐信息保持一致的情况下，不能再被分解为两个或者更多的规则，最小规则集是翻译的最小单元，包含翻译所需的基本信息。

定义二给出最小短语规则的定义，最小规则集构成了一个最简洁的翻译模型。图1中右侧Minimal列表示用本文提出的短语规则抽取方法从示例的含有词对齐信息的句对中抽取的最小短语规则。在图1中所示的短语规则中，前五个规则符合本文对最小规则的定义。例如，（辽宁，*liaoning*）不可被分解为两个或两个以上的短语规则，所以该规则是最小短语规则。

需要注意的是，最小规则并不完全指代源语及目标语端短语只含有一个单词的短语规则。当词对齐为1对多或多对1的情况下，抽取出来的与词对齐保持一致的短语规则同样符合最小规则的定义。例如，（进出口，*import and export*）规则中，“进出口”在词对齐信息中相对的目标语单词为“*import*”和“*export*”，该规则与词对齐信息保持一致，是一个合理的短语规则，同时符合我们对最小规则的定义，在构造最小短语规则集时，我们将其加入最小规则集中。此外，如果与最小短语规则源



短语规则	Minimal	2-Composed	Baseline
(辽宁,liaoning)	yes	yes	yes
(辽宁,liaoning 's)	yes	yes	yes
(进出口,import and export)	yes	yes	yes
(进出口,'s import and export)	yes	yes	yes
(贸易,trade)	yes	yes	yes
(辽宁 进出口,liaoning 's import and export)	no	yes	yes
(进出口 贸易,'s import and export trade)	no	yes	yes
(辽宁 进出口 贸易,liaoning 's import and export trade)	no	no	yes
...			

图1. 词对齐数据中（左）抽取的短语规则（右）

语及目标语端相连的单词词对齐为空时，该最小规则可以向对空单词扩展，所构造的短语规则仍符合最小短语规则定义。例如，在（辽宁，*liaoning 's*）规则中，目标语单词*'s*出现在目标语短语的边缘，同时在词对齐信息中对空，该规则同样仅由一个最小短语规则（辽宁，*liaoning*）构成，所以该规则是最小短语规则。这种边缘对空单词的扩展对规则集的质量有非常大的影响，这在（Koehn等，2003）提出的短语规则方法中已经的到了验证，在本文的实验部分将对边缘单词对空情况进行详细的讨论。

3.2 组合短语规则

最小短语规则的定义符合人们的直觉，即在进行翻译时，总是希望使用的翻译规则尽量短小，同时翻译质量较高。然而，也正是由于最小短语规则仅含有翻译过程中使用的最基本的单词，最终构造最小短语规则集中丢失了大量的上下文信息，这些上下文信息是基于短语的统计机器翻译系统性能优异的关键因素之一（Koehn等，2003）。在极端的情况下，当抽取出的最小短语规则的源语及目标语端仅有一个单词时，翻译系统则退化到基于单词的翻译系统。为了提高短语规则的质量，使短语规则可包含更多的上下文信息，本文提出了通过组合最小规则来获取含有更多单词、更多上下文信息的短语规则的方法。

定义三：一条短语规则与词对齐信息保持一致，同时该短语规则由同一训练句对中的n个或小于n个的最小短语规则组合而成，我们称该类规则为*n-composed*短语规则。

从定义三中我们看出，*(n-1)-composed*短语规则包含在*n-composed*短语规则中。图1中右侧2-Composed 列表示用本文提出的组合短语规则抽取方法从示例的含有词对齐信息的句对中抽取的由两个或小于两个最小规则组合而成的组合短语规则。例如，（辽宁 进出口，*liaoning 's import and export*）由最小规则（辽宁，*liaoning 's*）与（进出口，*import and export*）组合而来，所以其为*2-composed*短语规则。

很明显，如果对组合短语规则中包含的最小短语规则的个数不加以限制时，本文提出的方法可抽取任意长度的短语规则。然而，在大多数情况下，将组合短语规则中包含最小短语规则的个数定义过大，并不会对构造出的短语规则集的质量有明显好的影响。这一问题将在第四部分实验中进行讨论。

注意，为了通用化，本文将最小短语规则定义为*1-composed*短语规则。

3.3 基于组合的短语规则抽取

通过对基准短语规则抽取算法进行简单修改，本文提出的基于组合的短语规则抽取方法非常易于实现。给定含有词对齐信息的双语平行语料，通过对*n-composed*中参数*n*进行合理设置，可通过如下方法抽取组合短语规则。

- 首先从给定的含有词对齐信息的双语平行语料中抽取最小短语规则集，将该规则集存放在名为*minimal*的哈希数据结构中。在进行组合时将通过*minimal*哈希结构判断一条规则是否是最小短语规则。

- 设置*n-composed*中参数*n*的值，构造组合短语规则。使用*minimal*检测所有可能的短语规则，判断该规则由几个最小短语规则组成，如果该规则由小于或等于*n*条*minimal*中最小短语规则组成，将其放入一个新的矩阵*composed*中。

- 输出*composed*中短语规则。

4 实验

本文将提出的基于组合的短语规则抽取方法应用到NiuTrans开源系统中的基于短语的翻译系统中，在NIST汉英翻译任务上，通过与基准短语抽取方法进行比较，评价该组合短语规则抽取方法对翻译系统性能影响。

4.1 基准方法

在本文的实验中，使用（Koehn等，2003）提出的基于短语的翻译框架作为基准翻译系统。基准系统中使用了开源系统Moses使用的所有标准的特征。此外，在本文的翻译系统中，集成了两个调序模型：Xiong（2006）提出的基于最大熵的词汇化调序模型与Galley和Manning（2008）提出的层次化短语调序模型。基准系统解码器使用束剪枝与立方剪枝技术（Huang和Chiang，2007）来加速解码，使用最小错误率训练来优化特征权重。默认的，调序最长距离设置为8，短语规则的源语端与目标语端包含单词个数限制为7（与Moses默认设置相同）。对于短语规则集来说，每一个源语短语根据短语翻译概率仅保留前30个翻译候选。

4.2 实验设置

本实验中使用的训练数据包含一百九十万条汉英双语句对，该训练数据来自于NIST MT 2008评测提供的大规模双语预料中NIST部分数据。首先，我们用GIZA++[4]工具对训练数据进行双向词对齐，之后用“*grow-diag-final-and*”启发性算法对双向词对齐结果进行对称化处理。此外，本实验中使用英语GIZAWORD的Xinhua部分[5] 和双语数据的目标语部分训练了一个5元语言模型。关于开发集和测试集，本文使用了NIST MT 2003的测试集（919句）作为权重调优的开发集，同时使用NIST MT 2004与NIST MT 2005的测试集（分别含有1788和1082个句子）作为评价系统翻译质量的测试集。翻译质量通过使用上下文不敏感的IBM版本的BLEU（Papineni等，2002）评价指标进行评价。

4.3 实验结果

表1表示基准抽取方法（Koehn，2003）与本文提出的组合规则抽取方法在不同组合值*n*设置下的实验结果，结果评价指标由BLEU值表示。从表1中“最小规则”行中我们可以看出，当仅抽取最小规则时，本文的方法将获得一个非常小的短语规则集，但由于最小规则集在抽取的过程中丢失了大量的上下文信息，所以在开发集及测试集上的平均翻译性能比基准系统降低1.37个BLEU点。当进行组合规则抽取时，我们可以得到包含更多上下文信息的短语规则集，同时BLEU值随规则数量的增多持续增长。例如，通过表1中“基准方法”与“2-Composed”方法进行比较，我们发现当抽取2-composed短语规则集时，可得到与基准方法相当的翻译性能，与此同时，2-Composed方法获得的短语规则集的大小比基准方法减小44.3%。通过实验进一步证明，当抽取3-Composed与4-Composed的短语规则时，开发集与测试集的平均BLEU值相比于基准系统与2-Composed方法都有一定的提高。在同时考虑翻译性能与短语规则大小的情况下，2-Composed短语规则的翻译性能与表1实验中的最高性能可比，同时短语规则大小却有了明显的下降，即2-Comopsed短语规则已基本达到最优。从表1的实验结果看出，本文提出的组合规则抽取方法可以有效的生成高质量的、紧凑的、同时含有较多上下文信息的短语规则集。

方法	规则数	MT03	MT04	MT05	均值
基准方法(k=7)	121.7M	37.74	36.12	36.13	36.66
最小规则	34.1M	36.07	34.77	35.04	35.29
2-Composed	67.8M	37.53	35.98	35.85	36.45
3-Composed	95.7M	37.71	36.26	36.26	36.74
4-Composed	112.6M	37.76	36.19	36.17	36.71

表1. 基准系统与组合方法在开发集（NIST MT 2003）及测试集（NIST MT 2004和NIST MT 2005）上的实验结果比较，其中每组实验结果通过5轮实验取平均值而来

在基准短语规则抽取方法中，当源语及目标语短语包含单词的最大个数设置为不同值时，可以有效的调整短语规则集的大小。图2比较了基准方法与组合方法在不同设置下的BLEU值。其中横轴表示为短语表的大小（单位百万），纵轴为BLEU值。图2中实线表示的是基准规则抽取方法中短语长度设置为不同值时的情况，在实线中实心方点表示的是具体的实验设置，如“length=3”表示的是在基准系统中短语规则的源语及目标语短语的最大长度均设置为3，其它与之类似。图2中虚线表示的

[4]http://code.google.com/p/giza-pp/

[5]LDC2003T05

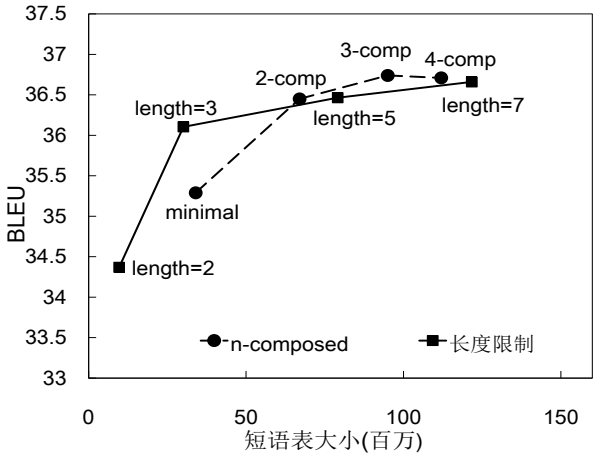


图2. 短语表不同大小对BLEU值的影响

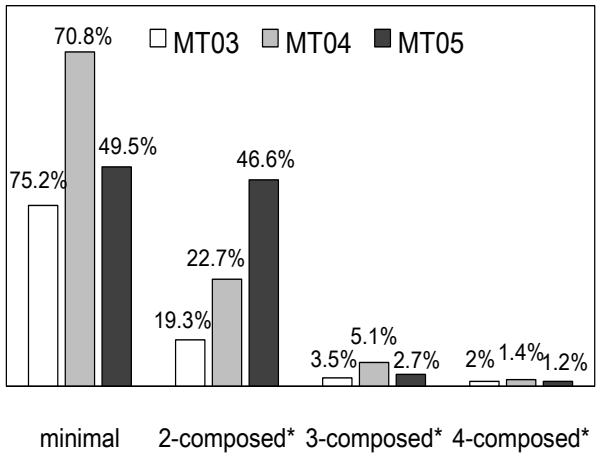


图3. 在30-best翻译结果中的组合规则使用比例情况

是基于组合的短语抽取方法中*n*设置为不同值时的情况。从图2中可以看出，本文提出的n-composed短语规则抽取方法中，当n≥2时，可获得与基准抽取方法相当的翻译性能；同时可以看出，本文提出组合短语规则抽取方法可更快的达到规则集大小与翻译系统的平衡。从此图中可观察到，仅使用最小规则集时，翻译性能比(≥2)-composed组合方法的值有很大程度的降低，这也从侧面说明了本文提出的基于组合短语抽取方法的有效性，同时说明含有更多上下文信息的短语规则对翻译系统的性能有非常大的影响。

本文对解码器使用最小短语规则及组合规则的比例情况进行了统计，该统计在开发集及测试集上的30-best翻译结果上进行。图3表示的是在开发集和测试集上的统计情况，其中*n-composed**表示仅由*n*个最小规则组合而成的组合规则。从图3中可以看出，解码器在使用短语规则进行翻译时，绝大多数情况下倾向于选择较短的规则（如*minimal*与*2-composed**）。由较多的最小短语规则构成的组合规则在翻译时则很少使用（如*4-composed**）。图3的实验结果同时解释了为什么表1中使用2-Composed组合规则可以取得较高性能。

在短语抽取过程中，在源语及目标语短语端的边缘对空单词进行空扩展时，同样会大大增加短语规则表的大小。表2中表示出在2-Composed的实验中，当边缘对空单词扩展词数进行限制时，抽取的短语规则表大小及BLEU值情况。其中*n-Unaligned*表示允许在短语规则的源语及目标语端进行*n*个对空单词的扩展，而∞-*Unaligned*表示不进行边缘对空单词个数进行限制的实验结果。从表2中第四行*2-Unaligned*的实验结果我们可以看出，通过在2-Composed组合短语规则中，边缘对空单词扩展限制为2个词时，抽取的短语规则集比基准方法的规则集减小56.6%，同时对翻译性能基本没有影响。当抽取短语规则时，如果不对边缘对空单词进行扩展，翻译系统的性能将会有大幅度的下降。从表2中第二行我们可以看出，在2-Composed情况下，不进行空扩展时性能比基准系统系统降低4.53个BLEU点。

方法	规则数	MT03	MT04	MT05	均值
基准方法(k=7)	121.7M	37.74	36.12	36.13	36.66
0-Unaligned	11.2M	33.02	31.47	31.90	32.13
1-Unaligned	35.8M	36.98	35.87	35.78	36.21
2-Unaligned	52.9M	37.68	36.05	35.96	36.56
∞-Unaligned	67.8M	37.53	35.98	35.85	36.45

表2. 短语抽取中边缘对空单词扩展的限制

在默认情况下，对于短语规则集来说，每一个源语言短语根据短语翻译概率仅保留前30个翻译候选，该方法称为直方图剪枝技术（Zens等，2012）。为验证本文规则抽取方法的有效性，本文将基准系统与本文提出的方法中的源语言短语对应的翻译候选个数*n*逐渐减小，即*n*=30、20、10、5分别进行实验，最终将不同方法在同一配置下的实验结果进行比较。具体实验结果如表3所示，从表3的实验结果中我们可以看出，虽然将同一源语言短语的对应的目标语短语个数逐渐减少，但是本文提出的规则抽取方法在每一种配置下均与基准系统可比，特别是当*n*=10时，基准规则抽取方法的翻译性能与基于组合的规则抽取方法相当。从该

实验结果中可以看出，虽然使用基于组合短语规则抽取方法极大的减小了短语规则集的大小，但是短语规则的质量并没有降低。Zens等（2012）比较了不同剪枝方法的有效性，在其论文中得出基于直方图剪枝技术的方法与基于绝对概率的和基于相对阈值的剪枝技术有相同的性能表现。由此可知，针对本文提出的短语规则抽取方法，同样可以应用后两种剪枝技术来获得更小的短语规则集。

	方法	规则数	MT03	MT04	MT05	均值
本文方法	n=30	52.9M	37.68	36.05	35.96	36.56
	n=20	50.4M	37.55	35.81	35.88	36.41
	n=10	45.6M	37.30	35.85	35.86	36.34
	n=5	38.9M	36.57	34.86	34.87	35.43
基准方法	n=30	121.7M	37.74	36.12	36.13	36.66
	n=20	118.1M	37.72	36.15	36.23	36.70
	n=10	110.0M	37.38	36.01	35.81	36.40
	n=5	97.8M	36.73	35.10	35.02	35.62

表3. 每个源语言短语对应*n*个翻译候选

本文将基于组合短语规则抽取方法简单的应用到NiuTrans开源系统中的基于层次短语的翻译系统中，同时与基准的层次短语系统进行比较。具体步骤如下：

- 1. 使用基准的短语规则抽取方法抽取源语言及目标语言短语长度分别为7，7的短语规则。
- 2. 使用基于组合的短语规则抽取方法抽取*2-composed*的短语规则，其中短语边缘对空单词扩展2个。
- 3. 抽取标准的层次短语规则[6]。
- 4. 将步骤1中规则与步骤3中规则进行组合，形成基准的层次短语翻译系统中使用的层次短语规则集。
- 5. 将步骤2中规则与步骤3中规则进行组合，形成层次短语翻译系统中使用的基于组合的层次短语规则集。

表4为本组实验结果，其中第一行“+基准方法”表示标准的层次短语规则集中加入基准的短语规则集；第二行“+组合方法”表示标准的层次短语规则集中计入基于组合的短语规则集。从表4的实验结果中可以看出，通过简单的将基于组合的短语规则抽取方法应用到层次短语系统中，在规则集规模比基准系统减小的同时，翻译系统的性能也下降了0.4个BLEU点。由于在标准的层次短语规则的抽取过程中是参考短语规则进行的，而在本实验中仅仅是在层次规则中加入了基于组合的短语规则，所以翻译性能的下降是可解释的。然而翻译系统BLEU值并没有过多的降低，这一现象同时说明基于组合的方法是有希望应用到基于层次短语的翻译系统中。由此可以推断，本文提出的短语规则抽取方法进行相应的修改，可以被应用到层次短语系统中。

[6]在这里，不包含变量的规则不在标准的层次短语规则之内

方法	规则数	MT03	MT04	MT05	均值
+基准方法	197.6+121.7M	37.60	36.08	35.97	36.55
+组合方法	197.6+52.9M	37.19	35.53	35.66	36.13

表4. 层次短语系统中使用基于组合的短语规则抽取方法与基准层次短语系统进行比较

5 相关工作

在基于短语的统计机器翻译中如何学习精简的短语规则集已经进行了长时间的研究。为了解决该问题，一些研究人员做了大量的研究工作。如，Johnson等（2007）将短语表过滤看作是显著性检验问题，通过该方法短语规则集的大小得到了有效的减小。Eck等（2007）使用解码器解码训练数据统计短语规则集中没有被使用的短语规则，通过过滤没有被使用的规则来有效的减小短语规则集的大小。Tomeh等（2009）将噪音作为短语规则集的评判标准，通过对短语规则集中噪音短语规则的识别及过滤来减小规则集的大小。Lv等（2007）通过使用训练数据中与测试集匹配最佳的句子来学习一个较小的短语规则集。Zens等（2012）系统的比较了不同规则过滤方法的有效性，同时提出了一种性能优异的基于熵的短语规则过滤方法。Ling等（2012）提出了一种基于熵的短语规则过滤方法，该方法通过度量长的短语规则与构成该规则的短语片段的在解码中的使用情况，进行短语规则的过滤。一般来说，上述方法都是通过对得到的初始短语规则表进行过滤来减小规则集的大小。这样，本文提出的方法抽取出来的短语规则集可作为上述方法的初始规则集，从而进行进一步的规则过滤来减小短语规则集的大小。在未来的工作中可以讨论上述方法在使用本文提出的短语抽取框架将取得怎么样的效果。

事实上，Mylonakis和Sima’an（2010）已经尝试着在他们的层次短语系统中使用最小规

则，但他们的实验结果表明仅使用最小规则时BLEU值有了明显的降低。然而，他们没有对最小规则进行组合来得到包含更多上下文信息的更大的规则。在本文中，简单的将基于组合的短语规则抽取方法应用到层次短语系统中，也并没有明显的降低系统的BLEU值，由此我们相信，本文提出的方法进行适当的修改，可被应用到层次短语系统中。相反，本文提出一套系统的框架来对最小规则进行组合以得到一个高质量的短语规则集，同时在真实的翻译任务中证明了该方法的有效性。

6 结论

本文提出了一套基于组合的短语规则抽取框架，通过使用本文提出的短语规则抽取方法，可以得到一个为基于短语的统计机器翻译系统服务的高质量、精简的短语规则集。通过与使用最广泛、性能表现优异的启发式短语抽取方法进行相比，在保证翻译性能不降低的情况下，本文提出的方法抽取的短语规则集比基准方法抽取的短语规则集减小56.5%。通过对实验结果的分析，我们发现，在某些数据集上，通过使用基于组合的短语抽取方法，我们可以获得BLEU值的提高。在本文中，同时通过大量的实验，对基于组合的短语规则抽取方法的有效性进行了合理的验证。

参考文献

David Chiang. 2007. *Hierarchical phrase-based translation*. Computational Linguistics 33(2):201-228.

Yonggang Deng, Jia Xu and Yuqing Gao. 2008. *Phrase Table Training For Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair?*. In *Proc. of ACL 2008*, pages 81-88.

Matthias Eck, Stephan Vogel and Alex Waibel. 2007. *Translation Model Pruning via Usage Statistics for Statistical Machine Translation*. In *Proc. of HLT-NAACL 2007*, pages 21-24.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, WeiWang and Ignacio Thayer. 2006. *Scalable Inference and Training of Context-Rich Syntactic Translation Models*. In *Proc. of COLING-ACL*, pages 961-968.

Michel Galley and Christopher D. Manning. 2008. *A Simple and Effective Hierarchical Phrase Reordering Model*. In *Proc. of EMNLP 2008*, pages 848-856.

Liang Huang and David Chiang. 2007. *Forest rescoring: Faster decoding with integrated language models*. In *Proc. of ACL 2007*, pages 144-151.

J Howard Johnson, Joel Martin, George Foster and Roland Kuhn. 2007. *Improving Translation Quality by Discarding Most of the Phrasetable*. In *Proc. of EMNLP-CoNLL 2007*, pages 967-975.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. *Statistical Phrase-Based Translation*. In *Proc. of HLT-NAACL 2003*, pages 127-133.

Philipp Koehn. 2008. *Statistical Machine Translation*. Cambridge, UK: Cambridge University Press.

Wang Ling, Joao Graca, Isabel Trancoso and Alan Black. 2012. *Entropy-based Pruning for Phrase-based Machine Translation*. In *Proc. of EMNLP-CoNLL 2012*, pages 962-971.

Yajuan Lv, Jin Huang and Qun Liu. 2007. *Improving Statistical Machine Translation Performance by Training Data Selection and Optimization*. In *Proc. of EMNLP-CoNLL 2007*, pages 343-350.

Markos Mylonakis and Khalil Sima'an. 2010. *Learning Probabilistic Synchronous CFGs for Phrase-based Translation*. In *Fourteenth Conference on Computational Natural Language Learning*, pages 117-125.

Franz Josef Och and Hermann Ney. 2000. *Improved statistical alignment models*. In *Proc. of ACL 2000*, pages 440-447.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu 2002 *BLEU: a Method for Automatic Evaluation of Machine Translation*. In *Proc. of ACL 2002*, pages 311-318.

Felipe Sánchez-Martínez and Andy Way. 2007. *Marker-based Filtering of Bilingual Phrase Pairs for SMT*. In *Proc. of EAMT-09*, pages 144-151.

Nadi Tomeh, Nicola Cancedda and Marc Dymetman. 2009. *Complexity-Based Phrase-Table Filtering for Statistical Machine Translation*. In *Machine Translation Summit XII*, pages 26-30.

Joern Wuebker, Arne Mauser and Hermann Ney. 2010. *Training Phrase Translation Models with Leaving-One-Out*. In *Proc. of ACL 2010*, pages 475-484.

Tong Xiao, Jingbo Zhu, Hao Zhang and Qiang Li. 2012. *NiuTrans: An Open Source Toolkit for Phrase-based and Syntax-based Machine Translation*. In *Proc. of ACL 2012 System Demonstrations*, pages 19-24.

Deyi Xiong, Qun Liu and Shouxun Lin. 2006. *Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation*. In *Proc. of ACL 2006*, pages 521-528.

Richard Zens, Daisy Stanton and Peng Xu. 2012. *A Systematic Comparison of Phrase Table Pruning Techniques*. In *Proc. of EMNLP-CoNLL 2012*, pages 972-983.