一种错误敏感的词对齐评价方法[1]

黄书剑，奚宁，赵迎功，戴新宇，陈家骏

南京大学计算机软件新技术国家重点实验室，南京 210093
Email: {huangsj, xin, zhaoyg, daixy, chenjj}@nlp.nju.edu.cn

摘要： 对齐错误率（Alignment Error Rate，AER）是目前通用的词对齐评价标准。近年来的研究表明，AER 虽然在一定程度上能够反映词对齐的质量，但它与机器翻译最终结果 BLEU 得分的相关性并不好。本文针对基于短语的机器翻译系统（PBSMT）分析了 AER 可能存在的一些问题，并根据词对齐结果中可能存在的不同类型的错误,提出了一种错误敏感的词对齐评测方法 ESAER(Error-Sensitive Alignment Error Rate)。实验表明，本文提出的 ESAER 与 BLEU 的相关性要远远好于 AER。

关键字：统计机器翻译、词对齐、评价标准、AER、错误敏感

An Error-Sensitive Metric for Word Alignment in Phrase-based SMT

HUANG Shujian, XI Ning, ZHAO Yinggong, DAI Xinyu and CHEN Jiajun

State Key Laboratory for Novel Software Technology, Nanjing University, 210093
Email: {huangsj, xin, zhaoyg, daixy, chenjj}@nlp.nju.edu.cn

Abstract: *AER (Alignment Error Rate) is a widely used alignment quality measure. Recent study shows that the AER score is not well correlated with the BLEU score of the final translation result. In this paper, we analyze the possible reasons for this weak correlation in a phrase-based SMT environment. We also propose a new alignment quality measure ESAER (Error-Sensitive Alignment Error Rate) according to different alignment errors. Experimental result shows that ESAER gets a much higher correlation with BLEU score than AER.*

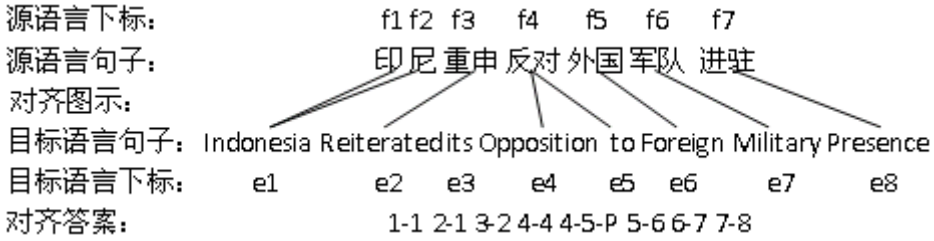
Keywords: *SMT, AER, Word Alignment, evaluation metric, error-sensitive*

1 引言

词对齐的基本任务是从句子对齐的双语平行语料中找到相互对应、互为翻译的词（如：图一）。在基于短语的统计机器翻译模型中[Koehn et al.，2003；Chiang，2005]，词对齐的质量直接影响到后续的词翻译概率计算、短语抽取、短语翻译概率计算、调序模型的训练等步骤的质量，从而很大程度上决定整个机器翻译系统的结果。因此，得到一个好的词对齐结果，

[1] 本文工作得到了国家 863 高科技项目资金（课题编号：2006AA010109）、国家自然科学基金（课题编号：60673043）、国家社科基金（课题编号： 07BYY051）以及南京大学研究生科研创新基金（课题编号：2008CL08）的资助。

是统计机器翻译取得良好结果的前提。



图一：平行的双语句对及其对齐

利用统计的方法，从大量的双语平行语料中自动得到词对齐信息的方案，从 1993 年 IBM Model[Brown et al.，1993]被提出以来就得到了广泛的认可。IBM Model1~5 本身就定义了词对齐概率的计算方法，1996 年，Vogel 等人提出了基于隐马尔可夫模型的词对齐方法[Vogel et al.，1996]。在这一阶段，词对齐仅仅是整个统计翻译模型的一个组成部分，其结果的评价主要是通过最终机器翻译系统的结果来衡量的。2003 年，Och 和 Ney 系统地比较了 IBM Model1-5 中的对齐模型以及基于隐马尔可夫的对齐模型的结果，并提出词对齐错误率（Alignment Error Rate，AER）这一标准，来单独衡量词对齐这一项任务的水平[Och et al.，2003]，词对齐的研究因而得以暂时脱离统计机器翻译模型这个整体，而成为一项单独的研究任务。近年来，各种提高词对齐水平的方法不断的被各国研究者所提出[Zhang et al.，2004；Liang et al.，2006；DeNero et al.，2007；Lacoste-Julien et al.，2007；等]。其中大部分的工作都将 AER 作为评价对齐质量提高的标准。Liu 等人的工作，还可以直接将 AER 作为优化的目标，采用判别式的方法来求得最优的对齐结果[Liu et al.，2005]。遗憾的是，这些研究工作虽然大都带来了 AER 的降低，却很少能显著地提高 BLEU 的水平。

2005 年以来，一些研究者开始对用 AER 来衡量词对齐质量这一做法提出质疑[Fraser et al.，2007；Ayan et al.，2006]。Fraser 等人认为，AER 中对准确率和召回率处以相同的权重实际上是有偏向的，并由此提出一种变权值的 F-measure 来衡量词对齐的质量[Fraser et al.，2007]。他们的实验表明这种变权值的 F-measure 与 BLEU 的相关性可以比 AER 有所提高。但是这种基于 F-measure 的方法仍然只是考虑的准确率和召回率两项指标，而没有考虑到词对齐错误对短语抽取乃至整个机器翻译系统的影响。另外，由于 F-measure 的不同权值是根据不同语料训练而得到的，使得这种方法在每次应用时都需要在新的语料进行参数调整，这也降低了该方法的可用性。

另一方面，考虑到词对齐与整个机器翻译流程的影响，研究者也开始试着从机器翻译其他任务的角度对词对齐提出不同的要求。在基于短语的统计机器翻译系统中，后续的工作是建立在基于词对齐的对齐短语抽取上的，Ayan 等人提出，需要用一致短语错误率（Consistent phrase error rate， CPER）来衡量词对齐的结果[Ayan et al.，2006]。但 CPER 每次计算时都要求先在多个对齐结果上进行短语抽取，其计算开销要远远大于 AER、F-measure 等指标，并不是一种高效的衡量方法。而且现有的短语的抽取方法是穷举所有与词对齐相一致的短语，这样的抽取过程有着一定的盲目性，在这样情况下得到的 CPER 是否可靠，还需要进一步的验证。

本文从短语抽取的角度分析了词对齐错误的不同类别、程度及其对短语抽取的影响，并以此制定了一个错误敏感的词对齐评价标准（Error-Sensitive Alignment Error Rate，ESAER），它根据词对齐错误的不同类别和不同程度，对词对齐结果进行不同的惩罚。相比于 Ayan 等人的工作而言，本文提出的方法仍然只是计算词对齐的信息，不需要进行短语抽取等额外步骤，因而仍然保持了 AER 等方法的高效性。相比 AER，ESAER 能取得更好的与 BLEU 的相关性。

本文后续部分结构如下：第二部分简述了目前通用的词对齐评价标准 AER 及其存在的

问题；第三部分分析了词对齐可能产生的错误，并在此基础上提出了一种错误敏感的词对齐评价方法；第四部分为相关的实验以及分析；第五部分总结了全文并展望了今后的工作。

2 对齐错误率及其问题分析

对齐错误率[Och et al.， 2003]是目前通用的词对齐质量评价标准，其计算要求一定数量的人工标记好的对齐结果作为标准答案。这些标记的对齐结果被分为两类，一类是确定的对齐（sure link），用集合 S 表示；另一类是可能的对齐（possible link，如：图一的例子中“反对-to”），用集合 P 表示（习惯上约定 $S \subseteq P$ ）。

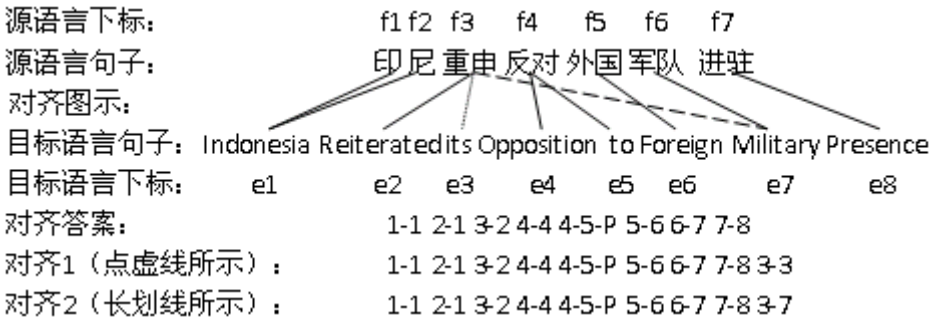
记需要评估的对齐结果为集合 $A=\{(j,a_j \mid a_j>0)\}$，则在上述条件下，结果 A 的对齐质量可以用改进的准确率和召回率来评价：

$$recall=\frac{|A\cap S|}{|S|},\ precision=\frac{|A\cap P|}{|A|} \qquad (1)$$

进一步的，结果 A 的对齐错误率可以定义为：

$$AER(S,P;A)=1-\frac{|A\cap S|+|A\cap P|}{|A|+|S|} \qquad (2)$$

由以上的描述可以看出，AER 的计算是以单个的对齐连接作为单位的，因而是非常简便而高效的。但是，恰恰由于 AER 仅仅把每一个对齐连接当作一个独立的单位来计算，而完全没有考虑这些对齐连接之间的相互影响，因而 AER 在预测短语抽取的质量方面有着明显的不足。



图二：两种影响不同的对齐错误

如图二的例子所示所示，假设由于平滑等原因发生了两个对齐错误：在对齐 1 中，“重申”被错误的与“its”对齐，而在对齐 2 中，“重申”被错误的与“Military”对齐。由上述 AER 的公式容易得到，这两种错误对齐的 AER 是相同的。但就短语抽取过程来说，显然对齐 1 中的错误影响要小得多，仅仅是造成了“重申-Reiterated”这一短语无法被抽取，而对后面其他短语的抽取不产生影响。对比而言，对齐 2 中的错误影响到了“军队-Military”、“外国军队-Foreign Military”、“军队进驻-Military Presence”、“外国军队进驻-Foreign Military Presence”等多个短语的抽取。显然这样的两个错误是必须被区分开来的。

本文采用对齐错误距离这一指标来衡量对齐错误的严重程度，并据此设计了一个错误敏感的词对齐评价标准。可以有效的区分这些对齐错误。

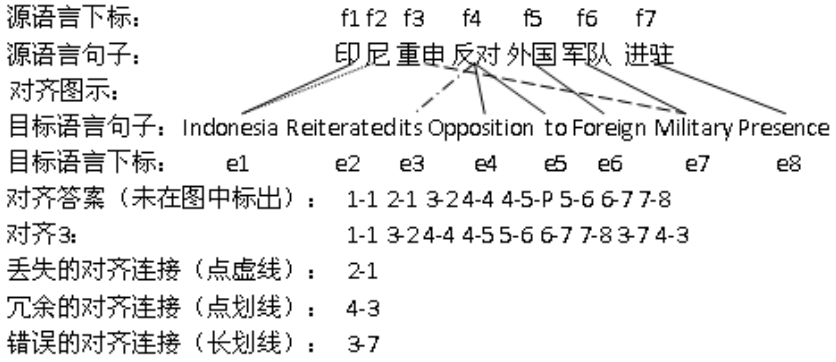
3 错误敏感的词对齐评价标准

3.1 直观的解释

本文提出的错误敏感的词对齐评价标准(ESAER)立足于对于不同类型、不同程度的词对齐错误进行不同的惩罚。

3.1.1 词对齐错误的类型

在类型上，词对齐的错误可以分为丢失对齐连接、冗余对齐连接和错误对齐连接三种类型（如图三所示）。对于短语抽取而言，丢失的词对齐连接只会导致个别短语无法被抽出，但不会影响到其它短语的抽取，因而其收到的惩罚应该相对较小；而冗余的对齐连接和错误的对齐连接虽然都会破坏其他短语的抽取，但由于被冗余对齐的词还保留了一部分正确的连接，因而也应该区别于错误的对齐连接而处理。对这三种不同的错误类型进行不同的惩罚，是本文工作的第一步。



图三：不同类型的词对齐错误

3.1.2 词对齐错误的严重程度

即使是同一种类型的错误，其惩罚也不应该相同，如在图二的例子中提到的两种错误就应该加以不同的惩罚。本文进一步提出应该以标准对齐位置与待判定的对齐位置之间的距离（对齐错误距离）为基础，来衡量不同程度的对齐错误所产生的影响。

3.2 形式化的定义

3.2.1 符号

按照惯例，我们用符号 S 表示双语句子的集合；用 f 和 e 表示源语言和目标语言的句子，f_j 和 e_i 表示源语言和目标语言中下标分别为 j 和 i 的词，m 和 l 表示源语言和目标语言的句子长度；源语言和目标语言的一组词对应关系称为一个对齐连接（alignment link），用 (e_i, f_j) 表示；F_j 表示与 e_i 对齐的一组源语言词；$\delta(\cdot)$ 为取指为 0 或 1 的指示函数；特别的，我们用上划线来注明所表示的符号与标准答案相关；此外，$d(\cdot,\cdot)$ 表示两个下标位置之间的距离函数；$mp(l,n)$ 表示丢失了 n 个对齐连接的惩罚；$rp(l,n)$ 表示冗余了 n 个对齐连接的惩罚；α,β,γ 表示实数值的惩罚因子，用以调节对不同类型的错误进行惩罚的力度。

3.2.2 词的对齐错误率

对于每一个目标语言的词e_i，按照其在标准答案中的对应定义其$ESAER(e_i)$如下：

- 若在标准答案中，有且仅有一个词（如$f_{\bar{j}}$）与e_i相对应；而在待评估的答案中：
  - 有且仅有一个词（如f_k）与e_i相对应：

$$ESAER(e_i) = ERROR(f_k, f_{\bar{j}}, l) = d(l, k, \bar{j}) \quad (3)$$

  - 没有词与e_i相对应（即缺少了一个对齐连接）：

$$ESAER(e_i) = ERROR(null, f_{\bar{j}}, l) = mp(l, 1) \quad (4)$$

  - 有一组词（如F_j）与e_i相对应：

$$\begin{aligned} ESAER(e_i) &= ERROR(F_i, f_{\bar{j}}, l) \quad (5) \\ &= rp(l, |F_i| - 1) + \min_{f_k \in F_i} ERROR(f_k, f_{\bar{j}}) \end{aligned}$$

- 若在标准答案中，没有词与e_i相对应；而在待评估的答案中：
  - 有一组词（如F_j）与e_i相对应：

$$ESAER(e_i) = ERROR(null, F_i, l) = rp(l, |F_i|) \quad (6)$$

- 若在标准答案中，有一组词（如$\bar{F}_j$）与e_i相对应；而在待评估的答案中：
  - 没有词与e_i相对应（即缺少多个对齐连接）：

$$ESAER(e_i) = ERROR(null, \bar{F}_j, l) = mp(l, |\bar{F}_j|) \quad (7)$$

  - 有一组词（如F_j）与e_i相对应：

$$\begin{aligned} &ESAER(e_i) = ERROR(F_j, \bar{F}_j, l) \quad (8) \\ &= \delta(|F_j| < |\bar{F}_j|)(\sum_{f_k \in F_j} \min_{f_m \in \bar{F}_j} ERROR(f_k, f_m) + mp(l, |\bar{F}_j| - |F_j|)) \\ &+ \delta(|F_j| \geq |\bar{F}_j|)(\sum_{f_m \in \bar{F}_j} \min_{f_k \in F_j} ERROR(f_k, f_m) + rp(l, |F_j| - |\bar{F}_j|)) \end{aligned}$$

3.2.3 整句及多句的对齐错误率

根据上述基于连接的错误函数，我们定义整句的错误率为目标语言词的错误率的均值（公式9）；定义多个句子的错误函数为每个句子错误率的均值（公式10）。

$$ESAER(e, f) = \frac{1}{m} \sum_{e_i \in e} ESAER(e_i) \quad (9)$$

$$ESAER(S) = \frac{1}{|S|}\sum_{(e,f)\in S} ESAER(e,f) \quad (10)$$

3.3 函数与参数设置

值得注意的是，上述形式化的定义中并没有对距离函数、丢失连接惩罚函数以及冗余连接惩罚函数做任何的限定。理论上讲，他们可以是任意实值函数，并且可以根据不同的翻译系统的实际情况选择不同的函数。此外，不同的惩罚因子的设置也带来了更大的自由性，可以使得ESAER在不同的错误类型中获得平衡。

为了简明起见，本文实验过程中采用了一组相对简单的设置，如下：

$$d(l,j,k) = \alpha * |j-k| \quad (11)$$

$$mp(l,n) = \beta * l * n \quad (12)$$

$$rp(l,n) = \gamma * l * n \quad (13)$$

$$\alpha = \beta = \gamma = 1 \quad (14)$$

4 实验及其分析

4.1 实验语料

在词对齐方面，本文采用LDC2003E14和LDC2005T10（CD1）作为双语对齐训练数据，其中部分数据经过句子对齐处理和繁简体转换等预处理。双语对齐的测试数据为2002年NIST机器翻译评测所使用的人工标注数据。

在机器翻译方面，我们采用了NIST2004机器翻译评测数据（LDC2006E43）的前1000句和全部NIST2005机器翻译评测的数据（LDC2006E38）分别作为开发和测试数据。具体细节如表一所示：

表一：实验所使用的数据

数据集合		LDC 编号	句子数量
分词	训练集	LDC2003E14	135,074
		LDC2005T10	153,037
	测试集	NIST2002	491
机器翻译	训练集	LDC2006E43	1,000
	测试集	LDC2006E38	1,082

4.2 实验系统

本文采用Giza++[Och et al.， 2003]作为对齐实验系统，实验了IBM Model1、HMM、IBM Model4三种模型。在机器翻译方面，我们选择了开源的Moses作为最终的机器翻译和评测平台。

4.3 实验过程和结果

为了得到质量不一但相对接近的多组对齐，我们用Giza++对训练数据进行了IBM Model1、HMM和IBM Model4三种对齐模型的训练（三个模型的迭代次数分别为5、5、3），并对每种模型的训练结果保留了两个方向的词对齐，以及两者取并集以及grow-diag-final [Och et al.， 2003]（下表中简写为GDF）之后的结果。对这样生成的12个不同的词对齐结果，

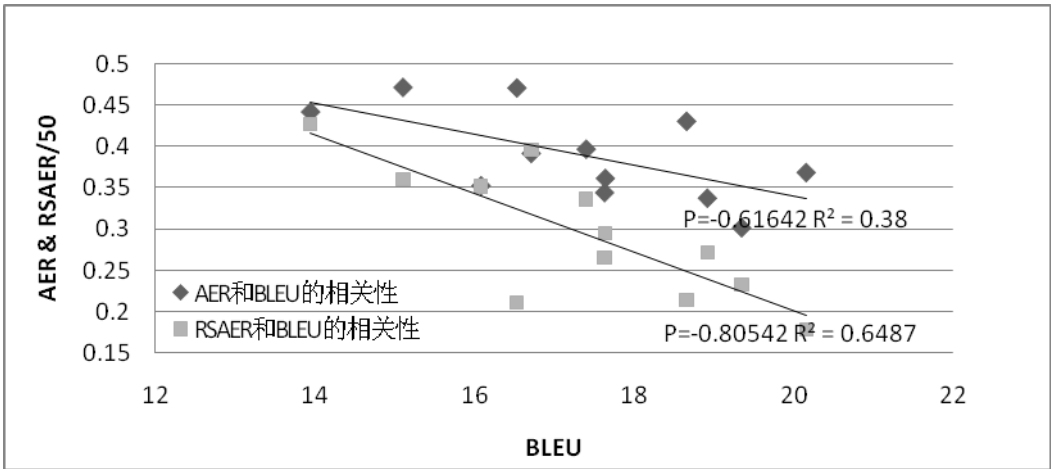
我们分别用Moses系统进行后续步骤如短语抽取和评分、调序模型学习、模型参数学习等，并翻译测试语料以得到最终的BLEU得分。由于Giza++的训练是一个非监督学习的过程，我们将双语对齐的训练数据和测试数据拼接在一起进行训练，以这样训练得到的测试数据上的结果作为评价本次训练的依据。根据这样的方法，我们单独计算了每组对齐结果的AER和ESAER（采用3.3中的设置）。为了便于比较，我们将ESAER的结果等比例缩小到与AER大致相同的范围（ESAER/50）。实验结果如表二所示。

表二：不同对齐结果的AER、ESAER与BLEU实验结果

（粗细下划线分别标出了当前标准下的最好结果和最坏结果）

Alignment		AER	ESAER	ESAER/50	BLEU
Model1	E2F	0.4705	10.5809	0.2116	16.53
	F2E	0.4417	21.3255	0.4265	13.94
	Union	0.4712	18.0103	0.3602	15.10
	GDF	0.3445	13.2756	0.2655	17.63
HMM	E2F	0.4304	10.7179	0.2144	18.66
	F2E	0.3917	19.7976	0.3960	16.71
	Union	0.3968	16.7958	0.3359	17.40
	GDF	0.3379	13.6103	0.2722	18.92
Model4	E2F	0.3687	8.9673	0.1793	20.16
	F2E	0.3528	17.5829	0.3517	16.08
	Union	0.3617	14.7509	0.2950	17.64
	GDF	0.3023	11.6513	0.2330	19.35

进一步，我们对上述实验结果进行了相关性分析，结果如图四所示。



图四：AER、ESAER与BLEU的相关性比较

（图中，P为Pearson相关性系数，R^2为Pearson系数平方值）

4.4 实验结果分析

从表二的实验结果中，我们可以看到：纵观全部12种对齐结果，对于BLEU表现最好的Model4E2F和BLEU表现最差的Model1F2E这两个对齐结果，AER并没有给出明显区别于其他对其结果的分数,但是ESAER很好的体现了这两种对齐与其他对齐的质量区别（分别用粗下划线和细下划线标出）。这说明了ESAER对不同的对齐质量有着很强的区分能力。

从单个模型的四种不同对齐结果上看，Och等人提出的grow-diag-final（GDF）的AER总是远远领先于其他的3种。但在实际的BLEU结果中，GDF并不总是最优的对齐，其最终

的翻译效果与E2F单向对齐的翻译效果相差不大。ESAER很好的反映了这一现象，在各模型的结果中，GDF和E2F的ESAER都领先于其他两种对齐，且这两种对齐的ESAER相差相对较小。

图四采用Pearson相关性系数定量的描述了AER、ESAER与BLEU的相关性。从图中可以看出，虽然AER和ESAER都与BLEU成的负相关关系，但AER的分布较为散乱，其R平方值仅有0.38，相关性较弱；而相对的，本文提出的ESAER除个别点以外，基本紧密的分布于趋势线附近，其R平方值达到0.65，相关性远远强于AER（Pearson相关性系数提高了30.7%）。

值得注意的是，ESAER对两个方向的对齐结果的评价有较大的差别，对E2F方向的词对齐结果比较偏好。这可能是由于ESAER对两种语言处理的不平等导致的。ESAER的计算过程实际是判别每一个英文词的对齐质量，而F2E方向的对齐是为每一个中文词寻找最可能的对齐，这一目标的不同可能是导致结果差别较大的主要原因。此外，F2E中存在的一对多现象，在以英文词为标准判断对齐质量时，会导致大量冗余链接的出现，这也造成了F2E方向的ESAER明显比E2F方向高。另一方面，相应的BLEU结果说明，不同方向的对齐对中英翻译系统的最后结果的影响确实有一定的差别（E2F方向对齐结果的翻译质量确实好于F2E方向）。这一观察也启发我们在后续工作中对对齐方向和翻译方向的关联做进一步的研究。

5 结论

本文从短语抽取的角度分析了AER可能存在的问题，并针对性的分析了词对齐的几种主要错误及其对短语抽取的影响。在此基础上，本文提出了一种错误敏感的词对齐评价标准，从短语对齐的角度对不同类型、不同程度的词对齐错误进行不同的惩罚。相对于以往工作而言，本文提出的方法仍然只是计算词对齐的信息，不需要进行短语抽取等额外步骤，因而仍然保持了AER等方法的高效性；同时，充分考虑了词对齐与短语对齐的关系，有效的克服了AER等方法在这个方面的缺陷。实验结果表明，在基于短语的统计机器翻译系统条件下，ESAER与BLEU的相关性要明显优于AER。

本文的后续工作主要着重于以下三个方面：

1．本文提出的词对齐评价标准还比较简单，对各种词对齐的错误只是进行了初步的划分，不同错误的惩罚函数和惩罚因子也都只是简单的进行了设置。是否能有一种相对较为通用的惩罚函数和惩罚因子的设置，使得ESAER对于各种翻译任务都适用，是本文今后工作的一个重点。

2．实验结果表明本文所提出的错误敏感的词对齐评价方法对不同方向的对齐结果惩罚并不相同的；而BLEU得分情况也表明，不同方向的词对齐结果确实对整个机器翻译系统的影响有不同的影响。今后的工作中，我们希望进一步通过理论和实验对对齐方向和翻译方向的关联进行分析。

3．本文的主要工作局限于基于短语的统计机器翻译领域，而目前，基于句法的统计机器翻译模型也能取得相当不错的效果，在有些方面甚至已经超过了基于短语的系统。采用的机器翻译模型不同，对对齐的要求自然也有变化，在后续工作中，我们拟对基于句法的机器翻译模型也做一定的研究，并努力提出与之相适应的词对齐评价标准。

参考文献

Ayan, N. F. & Dorr, B. J. 2006. Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT. *ACL,* 2006

Brown, P. F.; Pietra, S. D.; Pietra, V. J. D. & Mercer, R. L. The Mathematic of Statistical Machine Translation: Parameter Estimation
Computational Linguistics, 1993*, 19*, 263-311
Chiang, D. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. *ACL,* 2005
DeNero, J. & Klein, D. 2007. Tailoring Word Alignments to Syntactic Machine Translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics,* 2007, 17-24
Fraser, A. & Marcu, D. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics, MIT Press,* 2007*, 33*, 293-303
Koehn, P.; Och, F. J. & Marcu, D. 2003. Statistical Phrase-Based Translation. *HLT-NAACL,* 2003
Lacoste-Julien, S.; Taskar, B.; Klein, D. & Jordan, M. I. 2006. Word alignment via quadratic assignment. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics,* 2006, 112-119
Liang, P.; Taskar, B. & Klein, D. Moore, R. C. 2006. Bilmes, J. A.; Chu-Carroll, J. & Sanderson, M. *(ed.)* Alignment by Agreement. *HLT-NAACL, The Association for Computational Linguistics,* 2006
Liu, Y.; Liu, Q. & Lin, S. 2005. Log-linear models for word alignment. *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics,* 2005, 459-466
Och, F. J. & Ney, H. A systematic comparison of various statistical alignment models. *Computational Linguistics, MIT Press,* 2003*, 29*, 19-51
Vogel, S.; Ney, H. & Tillmann, C. 1996. HMM-Based Word Alignment in Statistical Translation. *COLING'96,* 836-841
Zhang, H. & Gildea, D. 2004. Syntax-based alignment: supervised or unsupervised? *COLING'04: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics,* 2004, 418