一种改进词语对齐的新方法*

罗维[1,2]，吉宗诚[1,2]，吕雅娟[1]，刘群[1]

[1](中国科学院计算技术研究所，中国科学院智能信息处理重点实验室，北京，100190)；

[2](中国科学院研究生院，北京，100049)

E-mail: {luowei, jizongcheng, lvyajuan, liuqun}@ict.ac.cn

摘　要： 词语对齐是目前主流的统计机器翻译系统的基本模块，GIZA++是词语对齐最常用的工具，但是GIZA++的对齐结果仍然存在不足。本文从约束双语命名实体之间的对齐的角度出发，提出了一种改进词语对齐结果的方法：首先，识别双语命名实体；其次，用标记替换双语命名实体；再次，用GIZA++重新进行词语对齐；最后，将标记还原为原始的命名实体。实验表明，针对基于短语的机器翻译系统，该方法提高了NIST值，并且该方法还有改进空间。

关键词： 词语对齐；命名实体；统计机器翻译

A Novel Approach to Improve Word Alignment

Luo Wei[1,2], Ji Zongcheng[1,2], Lv Yajuan[1], Liu Qun[1]

[1](Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190);

[2](Graduate University of Chinese Academy of Sciences, Beijing 100049)

E-mail: {luowei, jizongcheng, lvyajuan, liuqun}@ict.ac.cn

Abstract: Word alignment is a fundamental problem for statistic machine translation (SMT) systems. However, modern alignment methods are not reliable enough to yield high quality alignments for SMT, especially for distantly-related language pairs such as Chinese-English. We propose an approach to improve word alignment by restricting the alignment between named entities of language pairs. Our experiments show that our approach achieves improvement in NIST score.

Keywords: Word Alignment, Named Entity, Statistical Machine Translation

1　引言

机器翻译（Machine Translation, MT）是指依靠计算机来实现不同自然语言之间的自动翻译。词语对齐在机器翻译领域中占有重要的地位，它是基于短语的翻译模型和基于句法的翻译模型抽短语和规则的基础，而且词语对齐的质量影响着这些模型的解码效果[1]。此外，词语对齐在语义消歧、双语词典的建设等方面都起着重要作用[2]。

* 本文受国家自然科学基金重点项目《融合语言知识与统计模型的机器翻译方法研究》（批准号 60736014）以及863重点项目《面向跨语言搜索的机器翻译关键技术研究》（课题编号 2006AA010108）资助。

然而，通过分析现有的词语对齐工具的对齐结果，发现在处理双语命名实体之间的对齐时表现较差。本文针对这个问题，提出一种尝试改进词语对齐效果的方法，它尝试约束双语两端命名实体之间的对齐。实验结果表明，该方法对基于短语的翻译模型 Moses[3]并不十分有效，但是通过分析实验结果，我们发现该方法的性能仍然有一定的提升空间。

2 背景

自从上世纪 90 年代初，Peter Brown 等学者提出了基于噪声信道思想的统计机器翻译模型[4]以来，在不足 20 年中，基于统计方法的机器翻译已经成为机器翻译领域的主流方法，其中的研究成果包括基于词的翻译模型、基于短语的翻译模型以及基于句法的翻译模型。IBM 提出的 5 个模型[4]是基于词的翻译模型，而从另一个角度看，这些模型都是词语对齐模型。

Franz Och 等学者开发出了基于 IBM 提出的 5 个模型的词语对齐开源工具 GIZA++[2]，大大推动了机器翻译领域的研究。由于 GIZA++所实现的 IBM 模型是建立在一对多的模型基础上，所以需要在汉英和英汉这两个方向分别用 GIZA++来做词语对齐，然后通过一种启发式的方法来合并两次词语对齐结果[1]。基于短语的翻译模型和基于句法的翻译模型，会从词语对齐的语料库中抽取双语短语。其要求双语短语必须与词对齐相容，即短语对中的词语只能对齐到短语对中的词语，不能对齐到短语对之外的词语。该要求一般称为对齐一致性[5]。

3 词语对齐的改进方案

从抽取短语的过程可以看出，词语对齐的质量会直接影响到短语对齐的质量。通过人工查看词语对齐结果的方式，我们发现由于命名实体的存在，导致句子中出现某一端的命名实体的词对到另一端的非命名实体的词上，或者某一端的非命名实体的词对到另一端的命名实体的词上，也就是说，命名实体的存在影响了词语对齐的效果。图 1 给出一例。

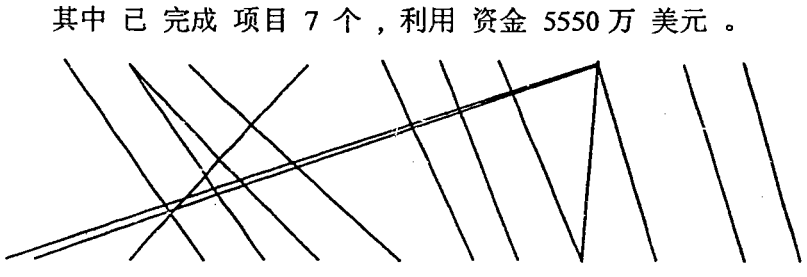

图 1 命名实体的存在影响词语对齐

从图 1 可以看出，汉语端的命名实体“5550 万”除对齐到英文端的“55.5 million”外，还对齐到“so far”，其中对齐到“5550 万”对齐到“so far”是多余而且错误的，这就是命名实体中的词对齐到非命名实体的词上。因此考虑从命名实体下手，通过约束命名实体只能对应到命名实体上，不允许命名实体与非命名实体的对应，以期改进对齐结果。由于我们所做的实验是汉英

方向的机器翻译，所以在下文的论述中，源语言指汉语，而目标语言指英语。当然本文谈到的方法思路具有一般性，可以应用其他的语言翻译任务中。

3.1 方案流程

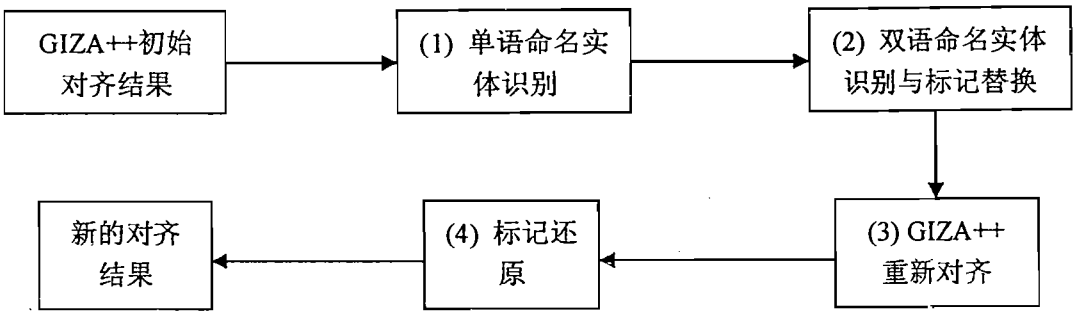

图 2 针对 GIZA++对齐结果的改进方案

图 2 给出了改进方案的实施流程。它从 GIZA++计算出的初始对齐结果出发，通过一系列的步骤处理，最终得到改进后的对齐结果。这时，就可以应用新的对齐结果来抽取短语和规则。下面将在 3.2 到 3.5 小节中，介绍各步骤的详细内容。

3.2 单语命名实体识别

命名实体可以划分为三大类名称词汇：实体类（包括人名、机构名、地名）、时间类（包括时间、日期）和数字类（包括数量、数码、序数词和货币等）[6]。根据需要，我们设置了 5 种命名实体类别：人名、机构名、地名、数词和时间词。具体来说，系统在双语端是这样完成命名实体识别的：

(1) 中文端：首先使用 ICTCLAS[1]对汉语句子进行分词和词性标记，另外 ICTCLAS 自身还带有命名实体识别模块，借助该模块，可以对汉语句子标上命名实体。

(2) 英文端：使用 Stanford NER[2]直接来识别命名实体。

3.3 双语命名实体识别与标记替换

双语命名实体是指来自两种不同语言的互译命名实体，而双语命名实体的识别对于跨语言信息检索和机器翻译等自然语言处理领域都是非常有用的[7]。标记替换是指将双语端对应的命名实体用相应的标记替换掉。可见，双语命名实体识别是建立在单语命名实体识别的基础上，同时它又为标记替换做准备。

由于命名实体划分为人名、机构名、地名、时间词和数词等 5 类，所以需要针对不同类型的命名实体采用不同的策略。我们设计的策略如下：

(1) 实体类的处理

对人名（包括中国人名和外国人名）、地名（包括中国地名和外国地名）和机构名，由于其用字灵活，且可以由更小的语言单元拼接而成，所以采用查词典和基于规则相结合的方式来翻译。

[1] http://ictclas.org/
[2] http://nlp.stanford.edu/software/CRF-NER.shtml

(2) 时间类和数字类的处理

由于时间、数字命名实体的构造形式简单，命名规则有规律可循，因此，针对时间词和数词，采用基于规则的方法来处理。

在系统中，用 PERSON、ORGANIZATION、LOCATION、TIME 和 NUM 分别代表识别出的人名、机构名、地名、时间词和数词。在 3.3.1 和 3.3.2 中将详细介绍双语命名实体识别及其替换原则。

3.3.1 双语命名实体识别和替换原则

为了实现双语命名实体识别和替换，有两种不同的方法：

(1) *"基于命名实体翻译的替换"* 操作（简称为 *"翻译替换"* 操作，下同）：将源端的命名实体翻译为目标端的命名实体，和将目标端出现的命名实体翻译为源端的命名实体。在实现的系统中，考虑到只做汉英方向的命名实体的翻译替换，已能得到较高的双语识别和替换效果，因此在系统中暂时没有实现英汉方向的命名实体的翻译替换。

(2) *"基于原始对齐信息的替换"* 操作（简称为 *"对齐替换"* 操作，下同）：根据初始的对齐信息，将满足对齐一致性的双语命名实体用相应的标记替换掉。

这两种不同的办法是直观而且是简洁的。当然，在实现上，对同一个句子，不同的命名实体应该用不同的标记来替换。图 3 给出一例来说明这个问题。系统识别出汉语端的 "8300 万 美元" 和 "600 万美元" 为数词，那么依次用"###NUM1###"、"###NUM2###"替换，识别出 "1999 年" 为时间词，那么用"###TIME1###"替换。而对英文端出现的与汉语端相对应的命名实体，系统将以相同的标签替换。

【原始句对信息】
【汉】管理 和 行政 支出 为 8300 万 美元 ，比 1999 年 少 600 万 美元 。
【英】expenditure for management and administration amounted to $ 83 - million , $ 6 million less than in 1999 .
【命名实体用对应的标记替换后的句对结果】
【汉】管理 和 行政 支出 为 ###NUM1### , 比 ###TIME1### 少 ###NUM2### 。
【英】expenditure for management and administration amounted to ###NUM1###, ###NUM2### less than in ###TIME1### .

图 3 标记替换结果

我们分析并提出如下替换命名实体的原则：

(1) 先进行 "翻译替换" 操作，而后进行 "标记替换" 操作。这是由于 "翻译替换" 由于是根据规则或者词典来实施对命名实体的翻译，具有较高的可信度，所以先进行 "翻译替换" 操作。

(2) 由于命名实体可以分为若干类，所以应该针对不同类型的命名实体，采取不同的翻译策略。但是，系统都会碰到一个汉语端命名实体能有多种英文的翻译结果，对这种情况，系统并不从中挑选一个最佳的翻译结果，而是保留所有可能的翻译结果，并到英文端的句子中做逐一匹配。一旦有一个翻译结果匹配上，那么结束。

(3) 对汉语端中表示一位数字的命名实体（包括汉语数词 "一"、"二"、"三" 等，1 位阿拉伯数字以及 1 位罗马数字等）不进行 "翻译替换" 操作。这是因为语料中经常会出现如图 4 中的

现象，如果仍然使用“翻译替换”操作，那么会造成标记替换错误，影响GIZA++的重新对齐。在图4中，汉语端第一个“8”，在英文端是没有对应的词语。这时如果依旧采用“翻译替换”的策略，那么会把英文端本该对应到汉语端“8 岁”的”eight”错误地对应到汉语端第一个的“8”。这是由语言应用的灵活性造成的，所以有必要考虑，对表示一位数字的命名实体不进行“翻译替换”操作。

【原始句对信息】
【汉】北京 春季 长跑 比赛 的 线路 分为 5 公里 和 10 公里 两 个 大 组 , 按 参赛者 的 年龄 又 分为 8 个 组别 , 参赛者 从 8 岁 到 80 岁 都 可以 报名 。
【英】the race has two events -- five kilometers and 10 km , with age groups set for entrants ranging from eight up to 80 years old in age .

图4 不适合对表示一位数字的命名实体进行翻译替换操作

3.3.2 翻译替换与命名实体翻译的区别

在3.3.1一节中详细地介绍了双语命名实体识别和标记替换的内容。但是需要强调的是，“翻译替换”工作与命名实体的翻译工作是不同的。具体表现在：

(1) 从目标上讲，“翻译替换”操作是针对机器翻译的词语对齐这个特定的目标，所以并不要求对系统去识别并翻译所有的命名实体，而是要考虑对齐正确率这个指标的前提下，恰当地选择命名实体来翻译，比如对汉语端中表示一位数字的内容，系统就不进行“翻译替换”操作。

(2) 从过程上讲，系统有时还需要处理语法上不正确的命名实体（这类问题的产生原因可能是语料预处理时，汉语分词引入的错误等）。图5给出了语料中经常出现的一类现象。

【原始句子信息】
【汉】当 建厂 工程 完成 至 百分之八十时

图5 语料库中分词错误举例，该错误影响词语对齐

图5所示句子中的“百分之八十时”，是分词错误，应该是“百分之八十 时”，同时它还会造成了GIZA++的对齐错误。因为“百分之八十时”不是一个正确的命名实体，所以一般的命名实体翻译系统是不考虑翻译这类词语。但是从提高对齐效果的角度上看，如果通过规则，依旧把“百分之八十时”翻译为“eighty percent”及若干其他等价的翻译结果(如”80 percent”等)，那么，只要通过“翻译替换”操作在英文端匹配上一种翻译结果，系统就能保证GIZA++重新对齐的结果中不会出现命名实体与非命名实体词语之间的对齐。

3.4 GIZA++重新对齐

在用标记替换命名实体后的双语句对上，重新使用GIZA++工具进行双向对齐，即可得到新的对齐文件。系统将约束双语端的标签之间是相互对齐的。举例来说，汉语端的“NUM1”标签只能对齐到英文端的“NUM1”，汉语端的“TIME2”只能对齐到英文端的“TIME2”；不允许出现汉语端的”NUM1”对齐到英文端的“NUM2”，也不允许其对齐到英文端的“TIME1”。

需要注意的是，新的对齐文件是含有替换命名实体所用的标签，所以还需要进行“标记还原”一步（参见3.5一节）才能得到真正的对齐文件。

3.5 标记还原

在3.4一节得到的对齐文件的基础上，通过扫描原始双语句对，可以将标签还原为其对应的命名实体，这样即可得到真正的对齐文件。随后可以抽取短语和规则。

4 实验结果以及分析

由于缺乏人工标注的词语对齐语料，难以用对齐错误率(alignment error rate, AER)[2]这个指标来检验词语对齐改进方案的效果。然而考虑到从对齐文件中抽取出的双语短语表或者规则是解码器的基本组成部分，所以，考虑检查抽取出的短语表或者规则是否能提升解码器的解码效果。

我们使用开源短语模型Moses作为解码器；训练语料采用FBIS语料（约23万平行句对），开发集选用NIST-02汉英测试集(878句)，并且选用NIST-05(1082句)和NIST-08(1357句)汉英两组测试集进行对比实验；使用SRILM在Gigaxinhua语料上训练出一个4-gram英语语言模型。对于翻译模型的训练，首先利用GIZA++[5]从汉英、英汉两个方向进行训练，获得词语对齐，并采用grow-diag-final[9]方法优化对齐，然后进行短语抽取[9]得到短语翻译概率表。使用最小错误率训练方法(minimum error rate training, MERT)[10]来训练对数线性模型中的特征参数，而翻译质量的评价则使用mteval-v11b.pl工具计算大小写不敏感条件下的BLEU-4[11]和NIST[12]值。

对比实验设置如下：以Moses在未进行改进的对齐文件上进行解码作为“Baseline”，添加改进对齐效果模块的系统作为新系统，记为“Baseline + 词语对齐改进”。

在NIST-05和NIST-08两组测试集上的实验结果参见表1：

表1 在测试集上的系统对比结果

测试集	系统	BLEU	NIST
NIST-05	Baseline	0.2188	7.1628
	+ 词语对齐改进	0.2135	7.1709
NIST-08	Baseline	0.1756	6.2556
	+ 词语对齐改进	0.1734	6.3761

分析实验结果，我们可以看出，以BLEU为评价指标时，在NIST-05和NIST-08上，我们提出的改进方案略差于Baseline，而以NIST为评价指标时，在NIST-05和NIST-08上，改进方案均略好于Baseline。根据Zhang Ying等人[13]的分析，词语对齐的改进方案能在一定程度上帮助Moses在解码时选出更好的词语，但是没能加强词语调序的能力。根据这个思路，返回去查看有关的n-gram正确率，我们确实发现，在改进的系统中，1-gram的正确率要好于Baseline，而2-gram至4-gram要略差于Baseline。

虽然可以在实施针对GIZA++对齐结果的改进方案后的对齐结果中看到比Baseline更好的对齐结果，然而改进方案在Moses这个解码器中并不十分有效。以下原因导致了这一现象的发生：

(1) GIZA++是一个无监督的对齐模型，而我们并不十分了解GIZA++的行为。仅仅从命名实体的角度出发，去考虑修正双语命名实体之间的对齐可能是不够的。

(2) 系统在“翻译替换”中使用了翻译规则，来实现双语命名实体的识别这个目标。然而，语言应用的灵活性以及汉语端分词错误等因素，在很大程度上限制了标记替换系统的能力。

5 总结及未来工作

本文提出一种尝试改进词语对齐效果的方案，它通过约束双语命名实体之间的对齐，以期改善词语对齐效果，并提高解码效果。目前的实验结果表明，该方案对 Moses 这个基于短语的机器翻译系统并不十分有效，但是该方法还有改进空间。

我们还有一些工作需要做，它们包括：(1) 在有人工标注过词对齐的句对齐语料中检查该方案的 AER 值是否提高。(2) 改进“基于命名实体翻译的替换”模块的性能，增强其健壮性，以更好地应对语言应用灵活性这一特点。(3) 可以考虑在双语命名实体识别后，生成双语命名实体词典；而在后面的“GIZA++重新对齐”一步中，考虑利用该词典来改善词语对齐效果。

参 考 文 献

[1] Franz Josef Och, and Hermann Ney. A comparison of alignment models for statistical machine translation. In COLING '00: The 18th International Conference on Computational Linguistics, pages 1086–1090. (2000)

[2] Franz Josef Och, and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, 2003, 29(1):19-51

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. Moses: Open Source Toolkit for Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

[4] Peter.F.Brown, Stephen A. Della Pietra, Vincent J.Della Pietra, and Robert L.Mercer. The mathematics of statistical machine translation: parameter estimation, computational linguistics, 1993, 19(2):263-311.

[5] Franz Josef Och, and Hermann Ney. Improved statistical alignment models. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pages 440–447. (2000)

[6] 赵军. 命名实体识别、排歧和跨语言关联. 中文信息学报. 2009 年 2 期

[7] 翟飞飞, 夏睿, 周玉, 宗成庆. 汉英双向时间和数字命名实体的识别与翻译系统. 第五届全国机器翻译研讨会（CWMT）论文集，南京. 第 172－179 页. (2009)

[8] Stolcke, Andreas. 2002. Srilm – an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken language Processing, volume 2, pages 901–904.

[9] Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proceedings of HLT-NAACL 2003, pages 127–133.

[10] Franz Josef Och, Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167. (2003)

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318. (2002)

[12] George Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of ARPA Workshop on Human Language Technology (http://www.nist.gov/speech/tests/mt/)

[13] Ying Zhang, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), pages 2051-2054.