文章编号：1003-0077(2010)04-0074-11

机器翻译系统融合技术综述

李茂西，宗成庆

（中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190）

摘 要：该文对机器翻译研究中的系统融合方法进行了全面综述和分析。根据在多系统输出结果的基础上进行融合的层次差异，我们将系统融合方法分为三类：句子级系统融合、短语级系统融合和词汇级系统融合。然后，针对这三种融合方法，该文分别介绍了它们各自具有代表性的研究工作，包括实现方法、置信度估计和解码算法等，并着重阐述了近年来使用广泛的词汇级系统融合方法中用于构造混淆网络的词对齐技术。最后，该文对这三类系统融合方法进行了比较、总结和展望。

关键词：人工智能；机器翻译；系统融合；最小贝叶斯风险解码；混淆网络解码；词对齐

中图分类号：TP391　　**文献标识码**：A

A Survey of System Combination for Machine Translation

LI Maoxi, ZONG Chengqing

(National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper presents a survey of system combination for machine translation (MT). According to the different levels of combining the outputs from different machine translation systems, we classify the approaches to system combination into three types: sentence-level combination, phrase-level combination, and word-level combination. The representative work for each type is discussed in this paper, including the methods exploited, confidences estimated, and decoding algorithms, as well as the monolingual sentence alignment approaches which used to build the confusion network in the word-level system combination method. Finally, we discuss the three combination approaches and compare them with each other. The future development prospects of MT system combination are also discussed.

Key words: artificial intelligence; machine translation; system combination; minimum Bayes-risk decoding; confusion network decoding; word alignment

1 引言

在自然语言处理中，几个相似的系统执行同一个任务时，可能有多个输出结果，系统融合就是将这些结果进行融合，抽取其有用信息、归纳得到任务的最终输出结果。系统融合技术已经成功地应用于语音识别、语义角色标注、双语文本的词对齐和词义消岐等任务中。近几年来，随着越来越多机器翻译方法的不断涌现[1-2]，系统融合技术逐渐地应用于机器翻译领域中，并在各种评测活动中取得了较好的成绩。

最早将系统融合技术应用到机器翻译领域中的是 R. Frederking 和 S. Nirenburg[3]，1994 年他们将三个不同的翻译系统（包括基于知识的机器翻译系统、基于实例的机器翻译系统和词转换机器翻译

收稿日期：2009-06-21　**定稿日期**：2009-09-25

基金项目：国家自然科学基金资助项目(60975053,90820303,60736014)；国家支撑计划资助项目(2006BAH03B02)；国家863 计划资助项目(206AA010108-4)；中国新加坡数字媒体研究院资助项目(CSIDM-200804)

作者简介：李茂西(1977—)，男，博士生，主要研究方向为机器翻译；宗成庆(1963—)，男，研究员，博导，主要研究方向为机器翻译、口语信息处理和文本分类。

系统）的输出结果采用图表遍历算法（Chart Walk Algorithm）进行融合，然后对融合结果进行后编辑处理得到最终的系统译文。但是由于当时缺乏有效的译文质量自动评价工具，系统融合后的性能与参与融合的系统性能无法进行定量的可信度比较。2001 年 S. Bangalore，F. Bordel，和 G. Riccardi 将语音识别融合方法中的投票策略（ROVER）[4] 引入到机器翻译系统中[5]，利用负对数投票特征和语言模型特征联合计算最终的一致翻译结果。在融合实验中，他们对五个翻译系统的翻译结果采用多字符串对齐算法（Multiple String Alignment）构造词格网络，实验结果表明，融合后的译文质量不低于最好的单个翻译系统。这引起了机器翻译领域对系统融合技术的关注。随后越来越多的机器翻译方法的涌现和译文质量自动评价方法的发展，促使机器翻译领域中出现了较多的关于系统融合方法的研究。

在机器翻译中进行系统融合可以有多种不同的方法，根据融合过程中操作的目标语言句子层次的不同，本文将其分为三类：

（1）句子级系统融合：针对同一个源语言句子，利用最小贝叶斯风险解码或重打分方法进行比较多个系统的翻译结果，将比较后最优的翻译结果作为最终的一致翻译结果（consensus translation）输出。句子级系统融合方法不会产生新的翻译假设，它只是在已有的翻译假设里挑选出最好的一个，因此该方法不同于下面将要介绍的两种融合方法。句子级系统融合方法也常用于词汇级系统融合方法中选择构建混淆网络的对齐参考假设（或称为对齐骨架）。

（2）短语级系统融合：它利用多系统的输出结果，重新抽取与翻译测试集相关度较高的短语表，并采用加权的方法对翻译概率和词汇化概率进行估计，利用新的短语表对测试集进行解码。短语级系统融合方法的核心思想是重解码（re-decoding）。

（3）词汇级系统融合：借鉴语音识别中混淆网络解码的思想，词汇级系统融合方法首先将多系统输出的翻译假设利用单语句对的词对齐方法构建混淆网络（或称为词转换网络），对混淆网络中每一个位置的候选词进行置信度估计，然后进行混淆网络解码。在解码时通常使用的特征包括：词的置信度得分、语言模型得分、长度惩罚和插入惩罚。

本文 2、3、4 节将分别详细介绍这三种层次的系统融合方法。此外，由于词汇级系统融合方法中构建混淆网络的翻译假设对齐方法是近年来系统融合的研究热点，并且这方面的相关研究工作也比较多，本文将这部分独立出来，在第 5 节进行详细介绍。第 6 节给出近年来国内外对系统融合项目的测评。最后对各种系统融合方法进行了比较、总结和展望。

2　句子级系统融合技术

对于一个源语言句子，经过多个翻译系统翻译后产生多个翻译假设（即一个翻译假设的列表，N-best list），句子级系统融合方法就是从这个翻译假设的列表中，利用贝叶斯风险解码或重打分方法，从中选择一个最优的翻译假设作为最后的一致翻译假设。句子级系统融合的主要技术有两种，分别为：最小贝叶斯风险解码（Minimum Bayes-Risk decoding，MBR）[6] 和通用线性模型（Generalized Linear Model，GLM）[7]。下面分别予以介绍。

2.1　最小贝叶斯风险解码

给定一个源语言句子，最小贝叶斯风险解码是从多个翻译系统产生的翻译假设列表中选出贝叶斯期望风险最低的一个翻译假设作为最终译文。

$$E_{mbr} = \arg\min_{E'} \sum_{E} P(E \mid F) L(E, E') \qquad (1)$$

（1）式中 $P(E|F)$ 是源语言句子 F 翻译成目标语言句子 E 的条件概率，当给定由多个翻译系统产生的翻译假设列表时，$P(E|F)$ 可以近似地由下式计算得到：

$$P(E \mid F) = \frac{P(E, F)}{\sum_{E'} P(E', F)} \qquad (2)$$

（2）式中 $P(E,F)$ 是源语言句子 F 和翻译假设 E 的联合概率分布，当参与融合的翻译系统都是统计机器翻译系统时，它可以根据翻译系统对翻译假设的总打分近似获得。当 $P(E, F)$ 不可获取时，可以假设条件概率 $P(E|F)$ 服从平均分布。

（1）式中的 $L(E,E')$ 是损失函数，当使用译文质量自动评价指标 BLEU 得分[8] 计算最小贝叶斯风险时，它可以表示为：

$$L_{BLEU}(E, E') = 1 - BLEU(E, E') \qquad (3)$$

（3）式中 $BLEU(E,E')$ 是句子级的 $BLEU$ 得分，与语料库级的 BLEU 得分的主要区别在于，为了防止对数运算时，n 元语法为 0 导致数据溢出，它在计算 n 元语法时需要进行加 1 或折半平滑。其他通常使用的损失函数包括基于词错误率（Word Error Rate，WER）或翻译编辑率（Translation Edit

Rate, TER)[9]。

2.2 通用线性模型

通用线性模型融合方法利用重打分策略,对参与融合的每一个翻译假设进行句子置信度估计,将句子置信度的对数和高阶的语言模型及句子长度惩罚进行线性加权联合求取最终译文。计算公式如下:

$$L_j = \log P_j + \nu L_j^{5gr} + \mu W_j \quad (4)$$

(4)式中 P_j 是句子置信度,它可以根据相关翻译假设的排名信息和相关翻译系统给出的得分进行估计。ν、μ 分别是五元语言模型 L_j^{5gr} 和句子长度惩罚 W_j 的权重,它们的值可以在开发集上进行优化调整。

在通用线性模型方法中,由于对翻译假设的句子置信度 P_j 的估计非常复杂,引入可调的参数较多,公式的主观性太强,且融合效果不如最小贝叶斯风险解码,近几年来没有太大的发展。

3 短语级系统融合技术

短语级系统融合方法首先合并参与融合的所有系统的短语表,从中抽取一个新的源语言到目标语言的短语表,然后使用新的短语表和语言模型去重新解码源语言句子。当无法获取参与融合的系统的短语表时,可以通过收集测试集或开发集的源语言句子和每个系统翻译后提供的相应 N-best 列表,产生源语言到目标语言的双语句对,最后使用 GIZA++[10] 工具包生成新的短语表。

3.1 短语表的合并

给定一个测试集,当参与融合的每个系统的短语表都可以获取时,一般可以使用 Moses 解码器[11] 自带的工具包对短语表进行过滤,得到针对特定测试集的过滤后的新短语表。这样产生的小短语表只有原来短语表的 10%到 30%。在收集每个系统过滤后的短语表之后,使用公式(5)对短语的翻译概率进行线性加权以更新短语表:

$$p(e \mid f) = \sum_{i=1}^{Ns} \lambda_i p_i(e \mid f) \quad (5)$$

式(5)中 Ns 表示参与融合的系统个数,λ_i 是第 i 个系统对应的权重 ($1 \leqslant i \leqslant Ns$),$p_i(e \mid f)$ 是第 i 个系统的翻译概率。同样,短语的反向翻译概率和两个词汇化权重的计算方法可以依此类推。

当参与融合的系统的短语表不能直接获取时,需要重新计算该系统的短语表,一般的做法是:将每一个源语言句子和相应的翻译系统生成的 N-best 列表组成新的 N 个双语文本句对,收集测试集的所有源语言句子的 N 个双语句对,形成一个针对特定测试集的语料库,然后使用这个语料库进行 GIZA++ 词对齐,即可得到该融合系统的短语表。使用式(5)的方法可以合并多个系统的短语表得到更新后的短语表。有时为了使排名靠前的翻译假设比排名靠后的翻译假设在短语表的构造时获得更大的权重,可以在语料库构建时,复制多个该翻译假设和源语言句子的双语句对,以增大该翻译假设所产生的短语词条的权重。通常的做法是:将 1-best 复制 $N+1$ 次,2-best 复制 N 次,...,N-best 出现 1 次。

文献[12]测试了短语级系统融合方法对翻译性能提高的上限,通过在短语表中剪除测试集的参考译文中未出现的短语词条,融合后的译文质量比最好的单个系统提高了接近 10 个 BLEU 点。这表明短语级系统融合方法在改善翻译质量上具有很大的潜力。

3.2 一种变形的短语级系统融合

B. Mellebeek 等于 2006 年提出了一种采用迭代算法进行句子分解的方法来实现短语级系统融合[13]。该方法首先对源语言句子进行句法分析,将源语言句子逐步分解成几个语法功能独立的块,然后找出每一块的中心词,最后使用几个翻译系统进行翻译,翻译完成后即进行融合。每个翻译系统每次翻译的单位是句子中独立的块,系统融合就在这些块的多个输出翻译假设上进行。这种方法在选择源语言短语块的最终译文时,依次使用了以下三个启发式特征:

(1) 投票特征:通过少数服从多数的方式选出源短语块的翻译。

(2) 语言模型特征:如果投票特征不能决出优胜的短语翻译,那就选择在得票数最多的几个翻译假设中使语言模型得分最高的那个翻译假设。

(3) 如果经过以上两个步骤都不能选择最终的短语块译文,那就选择置信度最高的系统输出的短语翻译假设作为最终翻译。

4 词汇级系统融合技术

词汇级系统融合技术利用翻译假设中词频信息进行系统融合。词汇级系统融合首先从参与融合的翻译假设中选择一个对齐参考,将其他的翻译假设

对齐到该对齐参考上，通过翻译假设间的单语句对的词对齐信息建立混淆网络(Confusion network)，然后对混淆网络中每两个节点间弧线上的候选词进行置信度估计，最后将候选词的置信度结合语言模型、长度惩罚、插入惩罚等特征进行混淆网络解码，选择通过最优路径的翻译假设作为融合后的译文输出。

4.1　构建混淆网络

在构建混淆网络时，首先需要选择一个翻译假设作为对齐参考假设(alignment reference，有些文献中称它为对齐骨架，skeleton，backbone)。对齐参考假设的选择非常重要，因为它决定了融合后产生译文的词序。通常我们使用 2.1 节中介绍的最小贝叶斯风险解码方法选择对齐参考假设。选择好对齐参考假设后，需要将其他参与融合的翻译假设对齐到该对齐参考假设上。不同于双语文本的词对齐，在词汇级系统融合中进行词对齐时，参与融合的翻译假设都是使用同一种语言，并且翻译假设中还可能存在语法错误，语序不一致，出现大量同义词和同源词等等现象，这使得在翻译假设之间建立词对齐并不容易，这也是目前词汇级系统融合方法中备受关注的问题，我们将在本文第 5 节单独论述这方面的问题。在建立翻译假设词对齐后，词对齐关系中可能存在对空(null)的情况，这在混淆网络中用 ε 符号表示。举例如下，当给定以下三个翻译假设时：

please show me on this map.
please on the map for me.
show me on the map, please.

假定选择第一个翻译假设作为对齐参考，并使用基于词调序的单语句对的词对齐方法[14]进行翻译假设的对齐。对齐后，翻译假设之间的词对齐关系为：

null	please	show	me	on	this	map	.
null	please	for	me	on	the	map	.
,	please	show	me	on	the	map	.

最终形成的混淆网络，见图 1。

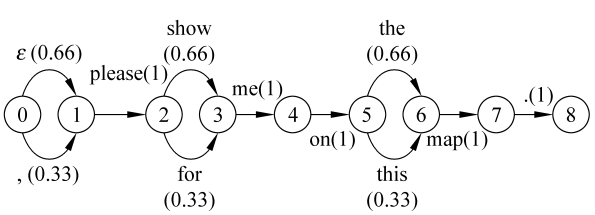

图 1　混淆网络实例

在混淆网络中，每两个节点之间的弧线上的词表示它们是最后融合结果中在相应位置的候选词。词的置信度(词对应的括号中的分值)是在相应位置的候选词中经合并后归一化的分值，例如在 0-1 节点间的弧线上出现了两个“null”(混淆网络中用 ε 符号表示)和一个“,”，则在该位置的候选词“null”和“,”对应的置信度分别为 2/3，1/3，取近似值则为 0.66 和 0.33。

混淆网络解码通常是搜索一条从起始节点到终结节点之间的最优路径，然后把通过最优路径上的候选词连接起来组合成最终的融合译文。当只使用词的置信度特征选择融合结果时，通过图 1 的混淆网络的最优译文是“please show me on the map.”。

在混淆网络解码时，参考对齐的选择影响到最终融合后输出译文的词序，因此十分重要。但是，选用贝叶斯风险最小的翻译假设作为对齐参考假设时，并没有考虑到同一个源语言句子可以翻译成多个合理的不同词序的目标语言句子，并且先验概率较大的翻译假设比较小的翻译假设的词序合理的可能性大，为了解决这个问题，Rosti 等提出了一种多混淆网络[15]方法，它轮流将每一个参与融合的系统的 1-best 作为对齐参考假设，并构建相应的混淆网络，将这些单个混淆网络连接在一起时，它们就形成了一个多混淆网络，图 2 给出了一个带先验概率的多混淆网络[7]。每个混淆网络起点都连接到一个空词(null，图中表示为 ε)所对的弧，空词后的概率是相应的混淆网络的对齐参考假设所在系统的先验概率，终点也连接到一个空词所对的弧，空词后括号的分值是 1，1 取对数后为 0，所以该弧线只起连接作用。在多混淆网络解码时，一般把起始弧线空词后所对应的分值同后面的特征值相乘，以保证先验概率大的翻译假设的词序有更大的概率成为融合后译文的词序。

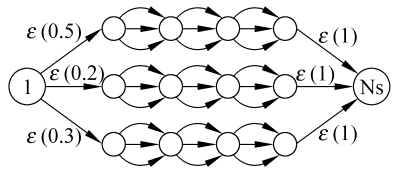

图 2　带先验概率的多混淆网络解码

4.2　解码时常用的特征和特征权重的优化调整

单纯使用词的置信度进行混淆网络解码时，在融合后的译文中容易插入一些冗余单词。这些冗余

的单词破坏了原来翻译假设中短语的连续性，打破了原来翻译假设的词序，从而导致融合后最终输出的译文不符合语法规则。为了解决这个问题，文献[15-19]通过引入空词插入惩罚因子和语言模型等方法来规范融合后产生的新的翻译假设，同时为了平衡计算语言模型得分容易导致最终的译文较短，所以，又引入了句子长度惩罚特征。在混淆网络解码中引入语言模型得分、插入惩罚因子和长度惩罚因子后，可以建立类似于机器翻译中的对数线性模型。假设给定一个源语言的句子 F，混淆网络解码就是求满足下面式(6)中的目标语言句子 E^*：

$$E^* = \arg\max_E(\alpha\log P_{AL} + \beta N_{nulls}(E) + \chi\log P_{LM} + \delta N_{words}(E)) \quad (6)$$

其中 α、β、γ、δ 分别对应融合过程中产生翻译假设的词的置信度 P_{AL}、插入惩罚 $N_{nulls}(E)$、语言模型得分 P_{LM}、长度惩罚 $N_{words}(E)$ 的权重。

对于混淆网络节点 i 和 $i+1$ 弧线上的候选词中第 j 个候选词的置信得分，由(7)式给出：

$$w_{i,j} = \mu\sum_{u=1}^{Ns}\sum_{v=1}^{N}\lambda_u\lambda_v c_w \quad (7)$$

(7)式给出了在有 N_s 个系统，每个系统提供 N 个翻译假设参与融合时，词的置信度计算公式。其中 λ_u 是系统 u 对应的先验概率，λ_v 是词所在翻译假设的权重，一般采用均匀权重，但是有时为了给排名靠前的翻译假设中的词赋以更高的权重，也可以采用基于排名的权重(rank-based)，即出自第 v 个翻译假设中的每一个词的概率都要乘上 $1/(1+v)$，c_w 是第 u 个系统第 v 个翻译假设中的词，如果在混淆网络节点 i 和 $i+1$ 之间的弧线上出现候选词 $w_{i,j}$，则该值取 1，否则取 0。μ 为归一化因子，它保证在节点 i 和 $i+1$ 之间出现的所有候选词的总置信度为 1。

在上面的混淆网络解码中有 N_s 个系统先验概率，4 个特征权重需要调整，一般采用改进的 Powell 参数调整算法[20]进行调整。该算法把需要调整的每个特征的权重看成是 N 维向量空间中的向量，在每一轮迭代中，使用一个基于网格(grid-based)的线性最小化算法优化每一维向量，并产生新的向量来加速优化过程。同样的算法也可以应用到机器翻译中对数线性模型的特征权重的调整(即最小错误率训练)[21]，但是在混淆网络解码时，需要同时调整特征的权重和系统的先验概率，所以它同最小错误率训练算法并不完全相同。

图 3 给出了多混淆网络解码的流程图，多混淆网络解码时参数的调整是在给定的开发集上进行的，在参数调整的每一轮循环中，都要执行图 3 的流程，直到每一个权重和先验概率的变化小于规定的阈值。

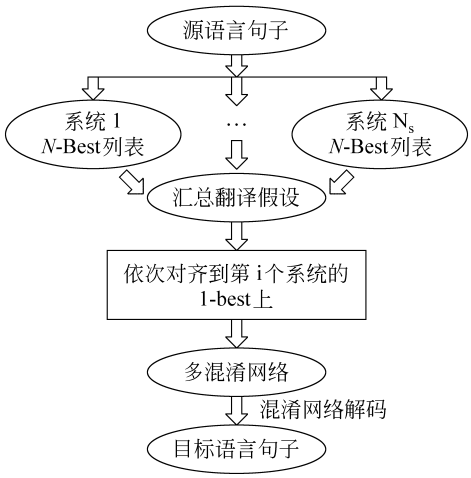

图 3 多混淆网络解码流程

4.3 一种变形的词汇级系统融合方法

在 4.1 节中提到，词汇级系统融合后输出的译文中较易插入一些冗余词，破坏了短语的连续性。K. C. Sim 等 2007 年提出了一种变形的词汇级系统融合方法[22]，他将这种方法称为一致网络最小贝叶斯风险解码(Consensus Network MBR，ConMBR)，该方法不同于上文介绍的通过引入语言模型、插入惩罚等特征来解决这个问题，ConMBR 方法把参与融合的每个系统的 1-best 翻译假设同词汇级系统融合后输出的译文进行比较，选取其中与融合产生的译文的贝叶斯风险最小的 1-best，并用这个翻译假设作为最终的输出译文。ConMBR 方法在混淆网络解码时并没有使用语言模型、插入惩罚、长度惩罚等特征，它只使用了词的置信度特征。这种词汇级系统融合方法并没有产生新的翻译假设，它只是从原来参与融合的多个系统的 1-best 中选出一个最优的翻译假设。ConMBR 方法用数学公式表示为：

$$E_{ConMBR} = \arg\min_{E'} L(E', E_{Con}) \quad (8)$$

5 构建混淆网络的词对齐技术

在机器翻译领域中，利用混淆网络解码进行系统融合的思想来源于语音识别领域。在语音识别

中，多个系统对口语句子的识别结果通过词错误率准则产生词对齐，利用词对齐信息构建混淆网络，解码后输出一致的语音识别文本[4]。不同于语音识别领域中识别文本之间的词对齐，机器翻译的系统融合在进行翻译假设的对齐时，不同的翻译假设之间存在着词序不一致、同义词、同根词、同源词等等难以处理的情况。而且，它也不同于统计机器翻译中在大量训练语料上的双语词对齐，系统融合中在翻译假设之间进行词对齐时，缺乏足够的语料。因此，机器翻译的系统融合中，翻译假设之间的单语句对的词对齐是目前词汇级系统融合研究的一个难点，也是目前研究的一个热点。

本文根据词对齐工作方式的不同，将它们分为基于编辑距离的词对齐、基于语料库的词对齐和基于语言学知识的词对齐。

5.1　基于编辑距离的单语句对的词对齐

基于编辑距离的单语句对的词对齐是计算将一个字符串(句子)转换成另一个字符串(另一个句子)所需的最少编辑次数时，附加产生的一种单语句对的词对齐。在字符串转换时，编辑的单元是单词。

基于词错误率准则(Word Error Rate，WER)的词对齐：字符串转换时允许的编辑操作包括单词的插入(*Ins*)、删除(*Del*)、替换(*Sub*)。词错误率的计算公式：

$$WER(E,Er)=\frac{Ins+Del+Sub}{Nr}\times 100\% \quad (9)$$

(9)式中 E 是需对齐的字符串，Er 是目标字符串，Nr 是目标字符串中所含的单词数，*Ins*、*Del* 和 *Sub* 分别是插入、删除和替换操作的次数。

基于翻译编辑率准则(Translation Edit Rate，TER)[9]的词对齐：字符串转换时允许的编辑操作包括单词的插入(*Ins*)、删除(*Del*)、替换(*Sub*)和语块的移位(*shift*)。翻译编辑率的计算公式如下：

$$TER(E,Er)=\frac{Ins+Del+Sub+Shift}{Nr}\times 100\% \quad (10)$$

与(9)式相比，(10)中分子多了一个移位次数 *shift*。在计算翻译编辑率的脚本程序 *Tercom* 中①，一般采用动态规划算法计算单词的插入、删除、替换次数，而采用贪婪算法进行语块的移位操作：通过反复试探，最终选择一个需要最少的插入、删除、替换编辑操作数的移位组合。因此，它不是全局最优搜索算法。针对翻译编辑率准则产生的词对齐所存在的问题，Li 等提出了一种直接调序的单语句对的词对齐方法[14]。基于词调序的词对齐方法(Word Reordering Alignment，WRA)首先找出待对齐的翻译假设和参考对齐之间的所有公共的连续短语块，然后对它们进行局部对齐，在局部对齐关系中寻找交叉的短语块对齐，最后利用启发式方法进行短语块之间的调序。

举例如下，给定以下两个翻译假设，当第二个翻译假设选为对齐参考时：

this color do you think suits me
do you think that color suits me

基于 WER 的词对齐、基于 TER 的词对齐和 WRA 词对齐如表 1，表 2 和表 3 所示。

表 1　基于 WER 的词对齐

This	color	do	you	think	null	null	suits	me
null	null	do	you	think	that	color	suits	me

表 2　基于 TER 的词对齐

this	do	you	think	null	color	suits	me
null	do	you	think	that	color	suits	me

表 3　WRA 词对齐

do	you	think	this	color	suits	me
do	you	think	that	color	suits	me

5.2　基于语料库单语句对的词对齐

给定一个源语言句子，将参与融合的每个翻译系统的翻译结果组合起来，生成一个翻译假设列表。基于语料库的单语句对的词对齐方法利用这些输出的翻译假设列表构建语料库，然后在这种小型的语料库上训练单语句对的词对齐关系。

E. Matusov 等 2006 年提出了直接使用统计机器翻译中双语文本词对齐工具包GIZA++ 进行单语句对的词对齐训练方法[23]。他的理论建模过程如下：

条件概率 $Pr(En|Em)$ 是给定翻译假设 Em 的情况下得到翻译假设 En 的概率，它可以通过引入一个隐含的词对齐关系 A 来计算：

$$Pr(En\mid Em)=\sum_{A} Pr(En,A\mid Em) \quad (11)$$

将(11)式等号右边的概率进行分解得到：

① http://www.cs.umd.edu/~snover/tercom/

$$Pr(En, A \mid Em) = Pr(A \mid Em)\Pr(En \mid A, Em) \quad (12)$$

把(11)和(12)式中 Em 看成是 IBM 模型中的源语言句子 F,即可套用 IBM 模型使用 EM 算法来进行词对齐训练。

在实际的词对齐训练中,单语语料库的构建方式如下:给定一个包含 M 个源语言句子的测试集,N_s 个参与融合的翻译系统对每一个源语言句子提供 N 个翻译假设,对应于测试集中的每一个源语言句子,将收集的 $N_s \times N$ 个翻译假设按任意排列两两组合得到 $N_s \times N \times (N_s \times N\text{-}1)$ 个对齐的单语句对,汇总后得到的单语语料库总共包含 $N_s \times N \times (N_s \times N - 1) \times M$ 对对齐句对。使用这种方式构建的语料库由于 N_s 和 N 的值太小,容易导致数据稀疏,一般需要将开发集的数据也添加进训练语料库。

微软的 X. He 等 2008 年针对单语文本的词对齐与双语文本的词对齐的不同之处,提出了一种利用间接隐马模型(Indirect HMM)获取翻译假设之间对齐的方法[18]。该方法把对齐骨架中的词看成是隐马模型的状态,翻译假设中的词看成是隐马模型的观察序列,对齐骨架和翻译假设之间的词对齐关系当作隐藏变量,使用一阶隐马模型来估计给定对齐骨架时生成翻译假设的条件概率:

$$p(e'^{J}_{1} \mid e^{I}_{1}) = \sum_{a^{J}_{1}} \prod_{j=1}^{J} [p(a_j \mid a_{j-1}, I) p(e'_j \mid e_{a_j})] \quad (13)$$

在式(13)中,发射概率 $p(e'_j \mid e_{a_j})$ 利用对齐骨架中的词和翻译假设中的词之间的相似度进行建模,又称为相似模型(similarity model);而转移概率 $p(a_j \mid a_{j-1}, I)$ 对翻译假设和对齐骨架的词序重排序进行建模,又称为位变模型(distortion model)。在计算时,相似概率是语义相似(semantic similarity)和词形相似(surface similarity)的线性插值。在双语文本词对齐时,源语言单词和目标语言单词只需考虑语义上的相似概率 $p_{sem}(e_i f_j)$;而单语文本词对齐时,语义相似可以处理同义词问题,而词形相似则可以很好地处理同根词、动词时态、形容词比较级等等使用GIZA++ 进行词对齐训练时很难处理的困难。位变概率计算主要取决于对齐的词之间的跳转距离,文章中把它们分成几个经验值来计算。在得到翻译假设之间的对齐关系后,该方法采用一种启发式对齐归一化规则来处理对齐过程中产生的一对多和对空等不利于转换成混淆网络的特殊词对齐情况。

杜金华等于 2008 年提出了一种融合语料库和编辑距离的单语文本的词对齐方法 GIZA-TER[17]。它将翻译假设按照上述 E. Matusov 等使用的 GIZA++ 方法,采用 Grow-Diag-Final 扩展规则[10]训练短语的词对齐。然后采用穷举法搜索最小化词错误率的一种短语移位组合。这种方法减少了短语被拆分的可能性,融合后的译文对句子的局部连贯性破坏较小。

5.3 基于语言学知识的单语句对的词对齐

基于编辑距离的单语句对的词对齐方法在计算时仅仅依靠词形的信息来获取翻译假设中词之间的对齐关系,而对于同义词、同源词的对齐它仅仅依靠位置关系来判断;基于语料库的单语句对的词对齐方法借鉴了双语文本的词对齐建模方法,通过建立相似模型来处理词义相似的单词之间的对齐关系。这两种方法在翻译假设对齐时没有或很少考虑到使用语言学知识来进行翻译假设的对齐。

N. F. Ayan 等在 2008 年提出了一种单语句对的词对齐方法。这种方法使用 WordNet 同义词典来处理词义相似的单词:包括同义词和不同词性的同根词。通过查词典(WordNet)对参与对齐的两个翻译假设中出现的单词词条进行相互求交处理,来判断它们是否为同义词。值得注意的是,WordNet 中只收录了具有实体意义(open-class)的单词,对于限定词、小品词等等它并没有收录。对于这个问题,N. F. Ayan 等对这些词分别创建了一个词性等价类,词性等价类中的词可以认为是词义相似的词。

使用同义词典的翻译假设对齐步骤描述如下:

(1) 使用 WordNet 同义词典抽取同义词;

(2) 利用同义词信息对对齐参考假设进行扩展;

(3) 修改 Tercom 脚本程序来处理同义词匹配。

值得注意的是,N. F. Ayan 等在这篇文章中还提到过一种两步法(two-pass)来构建混淆网络的对齐策略,它和 A.-V. I. Rosti 等在同年提出的一种递增的假设对齐(Incremental Hypothesis Alignment)方法[24]相似,两种方法都是解决翻译假设对齐时产生的同一个问题。下面对两步法进行简要的介绍。

通常我们在利用翻译假设之间的词对齐构建混淆网络时,多个翻译假设和对齐参考假设之间的对齐是独立的,它们分别对齐到参考对齐上,这种情况

导致翻译假设中对空的词之间不能很好地建立对齐关系。举例如下，给定下面三个翻译假设：

I like balloons
I like big blue balloons
I like blue kites

当选择第一个假设为对齐参考假设时，它们产生的两两对齐如下：

I	like	null	null	balloons	null
I	like	big	blue	balloons	null

I	like	null	null	balloons	null
I	like	null	null	blue	kites

将“I like blue kites”对齐到参考对齐“I like balloons”时，它并没有联系到“I like big blue balloons”和“I like balloons”对齐中的“big blue”这两个对空的词，这使得“I like blue kites”中的“blue kites”这两个词错误地对齐到对齐参考假设中的词“balloons null”。两步法在翻译假设词对齐时，首先将所有的翻译假设对齐到对齐参考上，构建一个混淆网络，然后使用这个混淆网络创建一个新的对齐骨架（也可称为对齐参考，主要是为了区分起见），在对齐骨架中每一个位置上的词都是通过投票从该位置的候选词中选出，再次将所有的翻译假设对齐到更新后的对齐骨架上形成最终的混淆网络。

另一种基于语言学知识的单语句对的词对齐方法是使用基于句法知识：反向转录文法（Inversion Transduction Grammar，ITG）[25]时产生的词对齐[26]。这种翻译假设对齐方法是计算 invWER 翻译质量评价尺度[27]时产生的一种单语句对的词对齐。invWER 评价尺度是将一个字符串转化成另一个字符串时最小的编辑次数，同翻译质量评价尺度 WER 和 TER 的不同之处在于，这些编辑操作是反向转录文法容许的在句法树节点上插入、删除、替换和语块的移动操作。基于 invWER 的翻译假设对齐方法的计算复杂度比 WER 和 TER 高，但是，融合后输出译文的句法结构比使用翻译编辑率产生的译文合理。

5.4　单语句对的词对齐质量对融合性能的影响

在统计机器翻译中，双语文本的词对齐精度的少许提高并不能保证翻译质量的提高[28]。在系统融合中，针对翻译假设之间单语句对的词对齐目前并不存在有效的评价指标，这导致单语句对的词对齐质量和系统融合的性能之间缺乏定量关联的尺度。用某种翻译假设对齐方法进行系统融合，融合后译文的质量优于使用另一种翻译假设对齐方法，也只是存在于特定的测试集或开发集上。目前看来，判断一种翻译假设对齐方法绝对优于另一种方法还缺乏理论证据和经验数据，这也是这几种翻译假设对齐方法共存的原因。

表 4　三种单语句对的词对齐方法对系统融合性能的影响

	IWSLT07CE	SSMT07CE
Primary	**31.52**	**29.81**
WER	31.38	**30.51**
TER	32.32	29.82
WRA	**32.40**	29.77

表 4 给出了使用三种不同的基于编辑距离的翻译假设对齐方法对 2007 年国际口语翻译评测（IWSLT'07）的汉英测试集和 2007 年全国统计机器翻译研讨会（SSMT'07）汉英测试集的几个系统翻译结果进行融合的结果。Primary 是最好的单个系统的 BLEU 得分。从融合结果上看，WRA 方法在 IWSLT'07 汉英测试集（IWSLT07CE）上获得了最好的得分，但是在 SSMT'07 汉英测试集（SSMT07CE）上融合的得分却最低，并低于参与融合的最好单个系统的性能。基于 WER 的翻译假设对齐方法则恰恰相反，它在 SSMT07CE 任务上取得了最好的成绩，却在 IWSLT07CE 上取得了最差的成绩。这可能是由于 WRA 方法对于短句（IWSLT07CE 测试集为口语领域）有较好的调序能力，而对于长句（SSMT07CE 测试集为新闻领域），过多的调序反而破坏了原来翻译假设的连续性，从而导致了融合性能的降低。

6　关于系统融合方法的评测

近几年来，机器翻译领域涌现出了越来越多基于不同方法的机器翻译模型，如基于句法的统计机器翻译模型、基于层次短语的统计机器翻译模型等等。这些多样化翻译模型的出现使得我们可以容易地获取多个翻译系统的输出译文，这大大推进了机器翻译系统融合的发展。针对系统融合的评测项目也逐渐出现在各种机器翻译的评测活动[29]中。

我国第四届全国机器翻译研讨会(CWMT'08)[①]是最早开展系统融合评测项目的会议。它是在"机器翻译"项目评测结果提交后,将所有参评单位的 *N*-best 结果发给"系统融合"参评单位;各系统融合参评单位在上述的多家机器翻译系统输出结果基础上进行系统融合。这次系统融合评测采用的开发集是 SSMT'07 提供的语料。共有 6 家单位参与了系统融合评测项目,他们的 BLEU 值和 mWER 得分如表 5 所示。

表 5 CWMT'08 系统融合评测结果

	BLEU4	mWER
Primary	**28.09**	**68.24**
Unit 1	**29.44**	**67.61**
Unit 2	**29.06**	**67.85**
Unit 3	**28.73**	69.52
Unit 4	27.21	69.95
Unit 5	26.79	69.86
Unit 6	25.09	71.54

其中,Primary 是最好的单个系统,Unit 1-6 是参与系统融合项目的单位编号(数据来源于文献[30])。

表 6 中 Sampling 列表示短语表训练时随机抽取的双语语料占总语料的比例。Primary 是参与融合的最好的单个系统, Sentence-level, Phrase-level, Word-Level 分别为句子级,短语级,词汇级的融合系统性能(数据来源于文献[28])。

表 6 三种系统融合方案的性能比较

Sampling /%	Primary BLEU	Sentence-level BLEU	Phrase-level BLEU	Word-Level BLEU
5	27.82	29.51	29.93	30.25
10	29.70	31.42	31.75	31.99
20	31.37	32.56	32.76	33.17
40	32.66	33.52	33.88	33.98
80	33.67	**34.17**	34.35	**34.38**
100	**33.90**	34.03	**34.08**	34.02

如表 5 所示,参评的 6 家单位中,只有 3 家在 BLEU 得分上比最好的单个系统有提高,2 家参评单位在 mWER 得分上比最好的单个系统有所提高。这一方面是由于参加"机器翻译"项目评测的单位提交的翻译结果质量参差不齐,最好的系统(BLEU:28.09, mWER:68.24)比排名第二的系统(BLEU:24.12, mWER:70.58)高出近 4 个 BLEU 点。另一方面也说明系统融合的性能缺乏稳定性,还有很多可做的研究工作。

另一个开展系统融合项目评测的是 NIST'09 机器翻译评测[②],这也是 NIST 评测第一次将系统融合作为一个单独的项目进行评测。NIST'09 系统融合项目是在各机器翻译参评单位提交翻译结果后进行的,它分为两个任务:阿拉伯语—英语和乌尔都语—英语。对于每一个系统融合任务,它将机器翻译的测试集分成两部分,接近 30%机器翻译的测试集数据用来做系统融合的开发集,系统融合的开发集对每一个源语言句子提供 4 个参考译文用于系统融合的参数调整,接近 70%机器翻译的测试集数据作为系统融合的测试集,以比较各系统融合参评单位的融合性能。

7 比较、总结和展望

7.1 三种融合方法的比较

在机器翻译系统融合中,一般情况下,最优的输出译文不同于原始输入译文中的任何一个。

根据前面的介绍,句子级系统融合方法利用参与融合的翻译假设的句子级别的知识,通过对翻译假设进行互相比较,或者利用一些反映翻译性能的本质特征对翻译假设进行重打分,从中选择一个最优的翻译假设。由于该方法并没有生成新的翻译假设,所以它能有效地保护原来翻译假设中短语的连续性和句子的词序。但是,它融合后输出的译文并没有吸收借鉴其他翻译假设中词或短语层次的知识,它只是从句子层面对翻译假设进行横向比较,因此它对融合性能的提高不如其他两种融合方法高。词汇级系统融合方法将翻译假设进行对齐,把参与融合的所有翻译假设的信息转化成词汇层面的知识,然后通过混淆网络解码将零散的词汇重新组织成完整的输出译文。这种融合方法从词的层次重组了输出译文,因此它能充分利用各个翻译假设的词汇级别的知识,取长补短。但是混淆网络解码在生成新的翻译假设时,并不能保证新生成的翻译假设和参与融合的翻译假设的词序的一致性以及短语连

① http://www.nlpr.ia.ac.cn/cwmt-2008/

② http://www.nist.gov/speech/tests/mt/2009/

贯性，因此，可能出现尽管最终的融合输出译文的自动打分较高，但是不符合语法的情况。短语级系统融合方法借鉴其他翻译系统的短语表知识，利用传统的基于短语的翻译引擎来重新解码源语言的句子。它能有效地保持短语的连续性和译文的局部词序。但是目前来看它不能很好地利用非连续短语和句法结构知识来克服译文的远距离调序问题。因此，短语级系统融合方法的性能介于前两者之间。

在实际融合性能上，W. Macherey 等 2007 年对这三种融合方法进行了一个经验性的比较[31]，他们通过对训练数据进行不同比例的抽样来观察参与融合的翻译系统的输出结果的相关度和最终融合译文质量的关系。在实验中，抽样尺寸分别为 5%，10%，20%，40%，80%，100%，抽样尺寸越小的翻译系统之间的相关度越小，每一种抽样尺寸抽出 10 组样本，用这 10 组样本单独进行词对齐训练，衍生出 10 个翻译系统，将这 10 个翻译系统的输出结果进行融合。融合结果如表 6。实验结果显示，相关度较小的翻译系统之间进行融合，三种融合方法的性能：词汇级系统融合＞短语级系统融合＞句子级系统融合，而当参与融合的翻译系统之间相关性较强时，三种融合方法的性能相当。该文给出的建议是，在进行系统融合时，尽量选用相关度较小的几个翻译系统进行融合，这样融合后的译文能获得较大的性能提升。

7.2　总结

本文对机器翻译系统融合方法进行了全面的综述和分析，介绍了三个层次的系统融合方法：句子级系统融合方法、短语级系统融合方法和词汇级系统融合方法，阐述了这三种融合方法各自的代表性研究工作，并比较了它们的优缺点和性能。对于当今主流的词汇级系统融合方法，本文分析了它的关键技术：单语句对的词对齐方法，并将它们分为三类，介绍了它们之中典型的八种方法。本文同时也介绍了当前开展机器翻译系统融合项目的评测活动，包括 NIST'09 机器翻译评测活动。

在对这三种系统融合方法的分析比较中我们可以看出，融合后的译文质量与参与融合的翻译系统之间的相关性有关。影响翻译系统的相关性的因素有很多，包括使用的模型差异，参数训练方法的互异等等。为了获得更好的翻译性能，我们应该将几个相关性较小的翻译系统利用词汇级系统融合方法进行融合。

在介绍词汇级系统融合的关键技术：单语句对的词对齐方法时，本文将三种基于编辑距离的单语句对的词对齐技术对系统融合的性能影响进行了比较。实验数据表明，这三种词对齐方法在不同的测试集上，有不同的表现，但是没有一种方法明显优于另外一种方法。这可能是由于基于编辑距离的词对齐仅仅考虑词形完全一致时的情形，并没有考虑同义词、同根词和同源词的对齐。基于语料库的词对齐方法为词形相似和词义相似的词建模，较好地解决了这个问题。而基于语言学知识的词对齐引入了同义词典或句法分析器来解决词对齐问题。它们分别用不同的方式试图获取质量更高的单语句对的词对齐。

目前，尽管机器翻译中的系统融合方法已经在某种程度上证明了，它能有效地改善翻译译文的质量，但是对系统融合性能持怀疑态度的研究者依然很多。这主要是由于当前主流的词级系统融合方法容易打破短语的连续性，插入一些对译文可读性破坏较大的词或者引入一些较严重的语法错误，而自动评价译文生成质量的 BLEU 值并不能很好的捕捉这些情况。BLEU 值的少许提高并不真正意味着系统融合对机器翻译质量的提高。

另一方面，系统融合方法的多样化导致了融合质量的参差不齐，而且各种方法在所有语料上的性能并不一致。例如，词汇级系统融合中各种单语句对的词对齐方法就存在八种以上，另外，还存在各种分配系统先验权重的方法、词的置信度估计方法等等，对这些方法组合对比，工程量很大。因此，目前缺乏对系统融合中的各种方法做深入的研究和比较工作。

7.3　展望

机器翻译模型的金字塔框架[32]把翻译的发展过程分为基于词、短语、句法、语义等几个阶段。套用这个发展模式，系统融合的发展目前还处于词和短语阶段：利用词或短语在各翻译假设中出现的频度信息来进行词或短语的置信度估计。我们认为，通过源语言或目标语言的句法或语义知识来深层次的指导融合，将能较好地克服系统融合中目前所困扰的译文短语不连续或译文不符合语法结构、融合性能不稳定等等难题，最终达到多种翻译方法的水乳交融。

参考文献

[1] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008.

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

[3] R. Frederking, S. Nirenburg. Three heads are better than one[C]//Proceedings of the fourth Conference on Applied Natural Language Processing. 1994：95-100.

[4] J. G. Fiscus. A post-processing system to yield reduced word error rates：Recognizer output voting error reduction (ROVER)[C]//IEEE Workshop on Automatic Speech Recognition and Understanding. 1997：347-354.

[5] S. Bangalore, F. Bordel, G. Riccardi. Computing consensus translation from multiple machine translation systems[C]//IEEE Workshop on Automatic Speech Recognition and Understanding. ASRU'01, 2001：351-354.

[6] S. Kumar, W. Byrne. Minimum bayes-risk decoding for statistical machine translation[C]//Proc. HLT-NAACL. Boston, MA, USA, 2004：196-176.

[7] A.-V. I. Rosti, N. F. Ayan, B. Xiang, et al. Combining outputs from multiple machine translation systems[C]//Proceedings of NAACL HLT. Rochester, NY, 2007：228-235.

[8] K. Papineni, S. Roukos, T. Ward, et al. BLEU：a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). Philadelphia, PA, 2002：311-318.

[9] M. Snover, B. Dorr, R. Schwartz, et al. A study of translation edit rate with targeted human annotation [C]//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. Cambridge, 2006：223-231.

[10] F. J. Och, H. Ney. A systematic comparison of various statistical alignment models[J]. Computational Linguistics. 2003, 29(1)：19-51.

[11] P. Koehn, H. Hoang, A. Birch, et al. Moses：Open Source Toolkit for Statistical Machine Translation[C]//Proceedings of the ACL 2007 Demo and Poster Sessions. Prague, 2007：177-180.

[12] F. Huang, K. Papineni. Hierarchical system combination for machine translation[C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, 2007：277-286.

[13] B. Mellebeek, K. Owczarzak, J. V. Genabith, et al. Multi-engine machine translation by recursive sentence decomposition[C]//Proceedings of the 7th Conference of the Association for Machine Translation in the Americas. Cambridge, 2006：110-118.

[14] M. Li, C. Zong. Word reordering alignment for combination of statistical machine translation systems [C]//International Symposium on Chinese Spoken Language Processing (ISCSLP). Kunming, China, 2008：273-276..

[15] A.-V. I. Rosti, S. Matsoukas, R. Schwartz. Improved Word-Level System Combination for Machine Translation[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, 2007：312-319.

[16] B. Chen, M. Zhang, A. Aw, et al. Regenerating hypotheses for statistical machine translation[C]// Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, 2008：105-112.

[17] 杜金华，魏玮，徐波. 基于混淆网络解码的机器翻译多系统融合[J]. 中文信息学报，2008,22(4)：48-54.

[18] X. He, M. Yang, J. Gao, et al. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems[C]//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, 2008：98-107.

[19] N. F. Ayan, J. Zheng, W. Wang. Improving alignments for better confusion networks for combining machine translation systems[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, 2008：33-40.

[20] R. P. Brent. Algorithms for minimization without derivatives[M]. Prentice-Hall, 1973.

[21] F. J. Och. Minimum error rate training in statistical machine translation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. Sapporo, Japan, 2003.

[22] K. C. Sim, W. J. Byrne, M. J. F. Gales, et al. Consensus Network Decoding for Statistical Machine Translation System Combination[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). 2007：105-108.

[23] E. Matusov, N. Ueffing, H. Ney. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment[C]// The 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006). Trento, Italy, 2006：33-40.

(下转第118页)