中国科学技术信息研究所 CWMT'2011 技术报告

何彦青 石崇德 于薇 张均胜 王惠临
中国科学技术信息研究所 信息技术支持中心 北京 100038
E-mail: {heyq,shicd, yuwei,wanghl}@istic.ac.cn

摘 要：本文介绍了中国科学技术信息研究所（ISTIC）参加第七届全国机器翻译研讨会机器翻译评测的情况。本单位参加了英汉科技领域的机器翻译评测项目。本文阐述了本单位机器翻译系统的实现框架以及实施细节，并分析了它们在评测数据上的性能表现。

关键字：机器翻译、自然语言处理、系统融合

ISTIC Evaluation Technical Report for CWMT'2011

Yanqing He, Chongde Shi, Wei Yu, Junsheng Zhang and Huilin Wang
Information Technology Support Center
Institute of Scientific and Technical Information of China, Beijing 100038
E-mail: {heyq, shicd, yuwei, zhangjs, wanghl}@istic.ac.cn

Abstract: *This is an overview of ISTIC evaluation technical report for the 7th China workshop on machine translation. ISTIC participated in the English-to-Chinese machine translation task in scientific and technical domain. This paper describes the implement framework of our machine translation system. We also give the key techniques and analyze the experimental results over the evaluation data.*

Keywords: *machine translation, natural language processing, system combination*

1 引言

中国科学技术信息研究所（Institute of Scientific and Technical Information of China, ISTIC）参加了 2011 年第七届全国机器翻译研讨会（CWMT'2011）组织机器翻译评测活动。在所有的评测项目中，ISTIC 参加了英汉科技领域的机器翻译项目。在该项目的评测中，ISTIC 采用了规则和统计两类机器翻译多引擎相结合进行系统融合的策略，提交了多机器翻译融合的结果。

本文的结构安排如下：第二节给出 ISTIC 机器翻译系统的总体框架和系统融合策略；第三节介绍数据的使用和处理；实验结果及相关分析在第四节；最后在第五节给出结论。

2 系统描述

ISTIC 提交的英汉科技领域的翻译结果为两类机器翻译多引擎的翻译输出基础上的系统融合。两类机器翻译多引擎包括基于统计的机器翻译多引擎(SMT-ME)和基于规则的机器翻译多引擎(RBMT-ME)。这两类多引擎各包含 2 个单引擎，每一个单引擎使用不同的参数来生成 1-Best 组成 1-Best List，进而采用了基于词和短语的系统融合方法[何，2010]进行系统融合。系统的总体框架见图 1。

2.1 两类机器翻译多引擎

基于统计的机器翻译多引擎包括基于短语的统计机器翻译单引擎（Phrase Based statistical machine translation，PBSMT）和基于层次短语的统计机器翻译单引擎(Hierarchical Phrase Based statistical machine translation, HPBSMT)。PBSMT [Koehn et al., 2003; Och et al.,

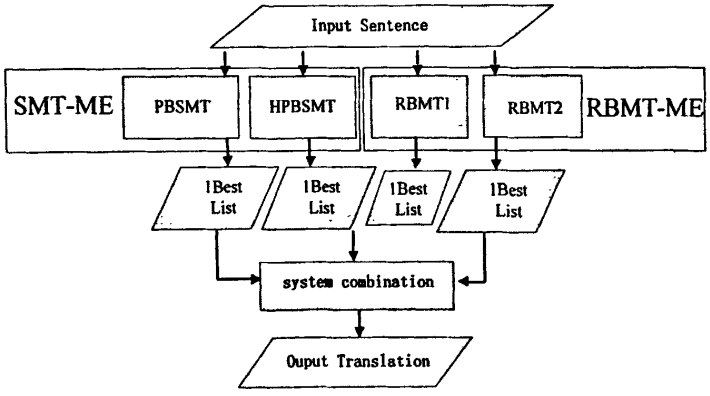

图 1. ISTIC 机器翻译系统框架

2004]使用连续的短语对为翻译单元，能够有效地捕捉训练语料中的连续的翻译对信息。HPBSMT[Chiang, 2005]为基于层次短语的统计机器翻译引擎，采用非句法知识的同步上下文无关文法，自底向上地完成目标翻译的生成。与 PBSMT 不同的是，该模型采用了非连续的短语，能够更为有效地对目标翻译进行排序。这两个引擎都采用了对数线性模型来进行翻译结果的遴选，采用的特征同 Moses。PBSMT 与 HPBSMT 这两个统计翻译引擎相比较，前者是个传统的鲁棒性强的统计机器翻译单引擎，翻译原理比较简单，速度快，对大规模数据的兼容性强。后者翻译原理较为复杂，翻译速度略慢，对大规模数据的兼容性要差，但翻译效果要好。总体而言，作为统计的机器翻译，这两个单引擎的翻译效果对数据的依赖性都比较强，都没有使用任何句法知识，因此翻译结果的可读性要差。

基于规则的机器翻译多引擎包含两个单引擎：规则引擎 1 和 2（Rule Based machine translation， RBMT1 和 RBMT2）。这两个引擎都是采用规则和模板相结合的技术，在传统的基于规则的机器翻译中融入了模板技术、统计技术，属于基于转换的机器翻译引擎。规则引擎对于翻译通用词汇比较擅长，其适应面较宽，翻译可读性好。但是由于科技领域语料中专业词汇较多，这两个引擎的规则覆盖专业词汇的能力有限，因此其翻译效果要差一些。

2.2 基于词和短语的系统融合

针对两类机器翻译多引擎的特点，需要将两类翻译结果有效地进行系统融合，以期在通用词汇翻译的基础上赢得专业词汇知识的补充。

ISTIC 将机器翻译的多个单引擎集成为一个统一的翻译平台，建立了基于词和短语的机器翻译融合系统。常用的多机器翻译系统融合可以从句子、短语和词三个级别上独立进行[Fiscus, 1997; Rosti et al., 2007a; 2007b]。ISTIC 采用词级的系统融合技术来构建混淆网络，将该混淆网络转换为短语表。然后使用该短语表利用短语级的系统融合技术中的重解码技术来进行解码，生成最后的融合结果。这样既保证了融合系统所构建的混淆网络的最大可能性，又可以使用更多的特征进行混淆网络解码。整个系统融合框架如图 2 所示。

对于每一个基于统计的单引擎，采用了不同的语言模型来生成 1-best，组成 1-best List。合并每一个单引擎的 1-Best List 为 1-Best Lists 来进行系统融合。经过在开发集上对比测试，采用了翻译效果最好的单引擎的翻译结果作为骨架翻译。在将每个翻译假设与骨架翻译进行词对齐时选用了 GIZA++ [Och and Ney, 2003]工具生成骨架-假设和假设-骨架双向的词对齐。GIZA++词对齐需要将骨架翻译和其余的每一个翻译假设组成平行句对，需要注意的是，这里的平行句对并不是双语的，而是单语的。由于 GIZA++的词对齐质量受测试集的大小的限制，为了解决这个问题，我们将所有的翻译假设中的单个词和它自身也组合成平行句对，两种平行句对合并在一起后使用 GIZA++工具包进行训练，这样可以保证在词对齐的时候，

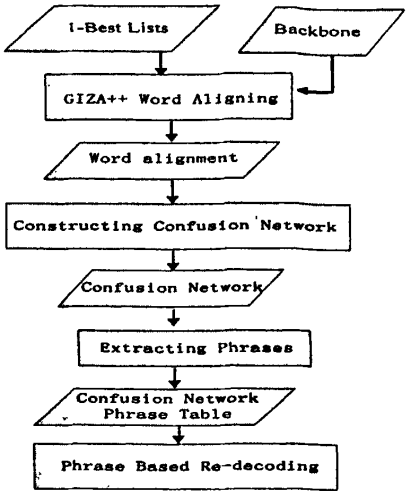

图 2. ISTIC 机器翻译系统融合框架

相同的词肯定可以对齐。同样也采用 Grow-Diag-Final 式启发函数进行词对齐扩展。

利用词对齐构建混淆网络时，为了使搜索路径中的节点尽量减少不可靠的词汇，没有使用空词来扩展混淆网络。直接将对齐参考的每个词作为对照点，利用其余翻译假设与它的词对齐信息，收集在每一个翻译假设中与该词对齐的词汇作为这个词的候选翻译。这样骨架翻译的每个词会有 0 个或者多个候选词汇，重复的词要记录重复次数，这样就形成一个词包。将每个词包放在混淆网络的每条弧上，之后通过投票来计算每个弧上词汇的后验概率，混淆网络就构建完成。在将混淆网络转化为混淆网络短语表时，参考翻译的每个词号看作是一个短语对的源短语，它的目标短语为该词号所对应的混淆网络上每条弧上的词，短语对的概率为该目标词的后验概率。

利用混淆网络短语表和一个基于短语的统计机器翻译系统来生成最后的目标翻译，这个过程类似于短语级的系统融合的再解码过程。但是这里的短语表不是源语言句子和翻译假设重新进行词对齐生成的，而是利用词级系统融合的混淆网络生成的。给定源语言句子 f，融合的过程就是搜索具有最大概率的目标翻译 e：

$$e^* = \arg\max_{e} \sum_{m=1}^{M} \lambda_m h_m(e, f)$$

其中，$h_m(e,f)$ 为特征函数，λ_m 为特征权重。基于短语的重解码过程中，仍然采用对数线性模型来完成最终的翻译融合，使用的特征函数有：

1：短语后验概率；
2：语言模型特征；
3：基于距离的重排序特征；
4：词惩罚；

短语后验概率为该翻译的每个词的后验概率的对数求和，每个词的后验概率来源于生成混淆网络时该词所在词包的后验概率。语言模型特征为该翻译的每个词的语言模型概率 $p(w_n \mid w_{n-1} \ldots w_1)$ 的对数求和。重排序模型为一个简单的基于距离的重排序模型：

$$P_{d(a_k - b_{k-1})} = | a_k - b_{k-1} - 1 |$$

其中，a_k 为第 k 个目标短语的源短语的开始位置，b_{k-1} 为第 $k-1$ 个目标短语的源短语的结

束位置。整句的基于距离的重排序特征为该翻译的每个短语的重排序概率对数求和。词惩罚特征为该翻译的目标词数。解码的搜索策略为柱状搜索算法，最后生成的 1-Best 作为融合结果输出。每个特征函数的权重由最小错误训练算法训练得出。

3 数据的使用和处理

训练语料采用了英汉科技领域项目发布的所有训练语料，语言模型训练数据采用了训练语料的中文部分和搜狗全网新闻语料库（SogouCA）[1]，所有的基于统计的机器翻译的单引擎的参数在该项目发布的开发集上训练。

对中文数据进行的处理有：中文的分词和全角变半角；对英文数据进行的处理为：大小写转换和标点符号的分离处理。采用的 Stanford 的中文分词[2]工具和 Moses[3]的英文 Tokenization 工具。

词对齐工具采用了 GIZA++[4]（全部使用默认的参数）并对该对齐结果进行扩展对齐（grow-diag-final）[Koehn et al., 2004]。

语言模型工具采用了 Srilm[5] [Stolcke, 2002]工具包来获取 5 元文法概率信息。

4 实验

我们分别在英汉科技领域机器翻译项目发布的开发集和测试集上来验证翻译效果的优劣。在开发集上的打分使用了评测组织方发布的打分工具。在测试集上的打分使用了评测组织方开放的在线评测平台。

我们使用 100 个词来限制训练语料的最大长度，获取了 896151 个平行句对作为每个统计机器翻译引擎的训练语料。表 1 列出了使用的所有语料的详细统计量。

表 1. 实验语料的统计量

数据集	语言	句子个数	词汇表	平均句长
训练集	中文	896151	189110	26
	英文	896154	180756	26
开发集	中文	4464	4658	21
	英文	1116	3428	22
测试集	英文	2497	6591	40

4.1 单引擎的翻译结果

表 2 列出了参与系统融合的单个翻译结果在开发集上的打分。其中基于短语的统计机器翻译单引擎使用了两个语言模型来进行翻译，PBSMT-1 只使用了训练语料的中文部分来训练 5 元的语言模型，PBSMT-2 在此基础上又增加了全部的 SogouCA 数据来训练 5 元的语言模型。基于层次短语的统计机器翻译单引擎使用了三个语言模型：HPBSMT-1 只使用了训练

[1] http://www.sogou.com/labs/dl/ca.html
[2] http://www-nlp.stanford.edu/downloads/segmenter.shtml
[3] http://www.statmt.org./moses/
[4] http://giza-pp.googlecode.com/
[5] http://www.speech.sri.com/projects/srilm/download.html

语料的中文部分来训练 5 元的语言模型；HPBSMT-2 在此基础上增加了部分的 SogouCA 数据（前大约 200 万句）；HPBSMT-3 在此基础上增加了全部的 SogouCA 数据。从表 2 可以看出，在英汉科技领域，基于统计的翻译多引擎的翻译表现要优于基于规则的翻译多引擎，基于层次短语的翻译单引擎的表现要优于基于短语的翻译单引擎。不同的语言模型也给出了不同的翻译结果，语言模型数据用的越多，翻译结果越好，但是翻译效果增长的幅度并不大。

表 2：单个翻译结果在开发集上的比较

引擎	BLEU-SBP	BLEU	NIST
RBMT1	0.4298	0.4329	9.6936
RBMT2	0.3554	0.3590	8.7499
PBSMT-1	0.5000	0.5104	10.5639
PBSMT-2	0.5013	0.5104	10.6179
HPBSMT-1	0.5061	0.5152	10.6626
HPBSMT-2	0.5074	0.5177	10.6547
HPBSMT-3	0.5122	0.5216	10.7158

4.2 系统融合结果

除了基于词和短语的系统融合（WPSC）方法，我们也尝试了其他系统融合方法，包括 TER[Sim et al., 2007; Rosti et al., 2007a]、WER[Bangalore et al., 2001]、INHMM[He et al., 2008]。WPSC 选用了 HPBSMT-3 的翻译结果作为骨架翻译。表 3 列出了系统融合的打分。其中 RULE 表示有规则结果参与系统融合，即使用了表 2 中的 7 个翻译结果。反之，NO-RULE 表示没有规则结果参与系统融合，即只使用了表 2 中的从第 4 行到第 8 行的 5 个翻译结果。从表 2 的结果来看，在英汉科技文献翻译中，基于规则的机器翻译多引擎效果不如基于统计的机器翻译多引擎，但从表 3 的结果来看，使用规则引擎的翻译结果和不使用规则引擎的翻译结果来比较，前者融合效果普遍比后者稍好，可见基于规则的机器翻译多引擎能在一定程度上弥补基于统计机器翻译多引擎的一些不足，提高系统融合的翻译结果的质量。在表 3 中，我们使用的 WPSC 与其他融合方法的对比来看，WPSC 的效果比 TER、WER 融合方法略好，与 INHMM 方法比较，使用规则系统融合的结果稍差，但不使用规则系统融合的结果则略好。由此可见，WPSC 融合方法是一种既直观，同时也有效的融合方法。

我们也在此次机器翻译评测的英汉科技领域的测试集上进行了实验。表 4 列出了单引擎和系统融合在测试集上的结果。分析实验结果说明，基于词和短语的系统融合方法（WPSC）能够达到基本的融合要求，能保证融合结果略优于最好的翻译引擎的翻译结果，但是融合的效果没有超过 INHMM，这个结果与在开发集上的实验是一致的。因此，我们在评测中提交了 INHMM 的融合作为 primary 结果，WPSC 的结果作为 contrast 结果。在最后的公布的评测结果中，这两个融合结果都取得了不错的成绩。

WPSC 是个比较保守的系统融合方法。其基本思想在于试图在最好的翻译结果的基础上

使用其他翻译假设的词汇来进行补充，以达到比参与融合的最好的单个翻译结果更好。

表 3：系统融合在开发集上的翻译结果比较

融合方法		BLEU-SBP	BLEU	NIST
TER	RULE	0.5127	0.5240	10.6997
	NO-RULE	0.5070	0.5181	10.6403
WER	RULE	0.5113	0.5205	10.6412
	NO-RULE	0.5047	0.5183	10.5810
INHMM	RULE	0.5268	0.5438	10.7900
	NO-RULE	0.5084	0.5250	10.6018
WPSC	RULE	0.5156	0.5248	10.7829
	NO-RULE	0.5131	0.5221	10.7436

表 3：系统融合在测试集上的翻译结果比较

Results	BLEU5-SBP	BLEU5	BLEU6	NIST6	NIST7	GTM	mWER	mPER	ICT
RBMT1	0.3079	0.3150	0.2551	9.0899	9.1040	0.8099	0.6151	0.3651	0.2819
RBMT2	0.2408	0.2476	0.1930	8.1135	8.1223	0.7588	0.6739	0.4144	0.2364
PBSMT-1	0.3860	0.3989	0.3321	10.4192	10.4415	0.8388	0.6147	0.3085	0.3852
PBSMT-2	0.3881	0.4020	0.3353	10.4731	10.4947	0.8342	0.6140	0.3038	0.3927
HPBSMT-1	0.3853	0.3970	0.3308	10.4071	10.4307	0.8380	0.6064	0.3080	0.3808
HPBSMT-2	0.3986	0.4120	0.3463	10.5815	10.6075	0.8395	0.6060	0.3036	0.3936
HPBSMT-3	0.3965	0.4102	0.3438	10.5567	10.5823	0.8395	0.6132	0.3047	0.3917
INHMM-RULE	0.4083	0.4302	0.3656	10.3527	10.3794	0.8219	0.5687	0.3206	0.4264
WPSC-RULE	0.4009	0.4142	0.348	10.6314	10.6573	0.8435	0.6041	0.3004	0.3979

5 结论

ISTIC 参加了 CWMT'2011 英汉科技领域评测任务的机器翻译项目，并取得了较好的成绩。经过上述实验分析以及整个参评过程，我们积累了下述经验：

1）目前的机器翻译系统模型众多，但是真正一枝独秀的模型很少。机器翻译研究如果想获得大规模应用性发展，系统融合研究是必不可少的。系统融合不但可以博采众家机器翻译模型在理论研究的长处，即使对于同一翻译模型内部，不同翻译参数输出的翻译结果也具备一定的融合空间，更甚之，对于不同的数据集生成的翻译结果也可以一并融合之。

2）囿于翻译的效率问题，目前，我们所采用的融合系统的技术还比较简单，仍有大量的空间可以进一步加强研究。我们的参考翻译策略比较简单，对于所有的句子都采用的单一系统的结果做参考，我们可以采用最小贝叶斯风险解码技术或者其他评价机制来动态地选择参考翻译。我们的词对齐策略也比较简单，只使用了 GIZA++工具来生成，可以采用更多的启发式措施来克服未登录词的问题。

6 致谢

本文受中国科学技术信息研究所学科建设“自然语言处理”课题（XK2011-6）、中国科学技术信息研究所重点工作“多语言信息获取关键技术研究与应用示范”课题（ZD2011-3-3）和中国科学技术信息研究所科研项目预研资金（YY-201122）支持。

参考文献

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proc. ASRU*, pages 351–354.

David Chiang. 2005. A Hierarchical Phrase-based model for Statistical Machine Translation. In *Proceedings of ACL 2005*, pages 263–270.

J.G. Fiscus. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore, Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In *Proceedings of EMNLP 2008.*

何彦青、张均胜、王惠临，基于词和短语的多机器翻译系统融合方法研究，情报学报，2011，已收录。

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology conference / North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2003)*, pages 127-133.

Franz Josef Och, Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Antti-Veikko I.Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, and Richard Schwartz, Bonnie J.Dorr. 2007a. Combining Outputs from Multiple Machine Translation Systems. In *Proceedings of NAACL HLT*, pages 228-235, Rochester,NY, April 2007.

Antti-Veikko I.Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2007b. Improved Word-level System Combination for Machine Translation. In *Proceedings of ACL 2007.*

K.C. Sim, W. Byrne, M. Gales, H. Sahbi and P. Woodland. Consensus Network Decoding For Statistical Machine Translation System [A]. In: *ICASSP*. 2007.

Andreas Stolcke, 2002. SRILM-An extensible language modeling toolkit. In *Proceedings of International Conference on spoken language processing*, volumn 2, pages 901-904.