机器翻译候选对系统融合结果影响分析 *

朱俊国，杨沐昀，张宇，李生，赵铁军

（哈尔滨工业大学，黑龙江省哈尔滨市 150001）

摘要：本文对影响机器翻译系统融合性能的翻译候选进行了实验分析。首先分析了给定翻译方向时参与融合的系统数量对系统融合远景得分的影响，结果显示随着参与融合的系统个数的增加，系统融合的远景得分也随之增加。然后比较了使用来自多个源语言的翻译结果和使用来自单个源语言的翻译结果参与系统融合时对应的系统融合远景得分。数据显示当参与融合的系统个数相同时，使用来自多个源语言的翻译结果所对应的远景得分优于仅使用来自单个源语言的融合结果对应的远景得分。我们检验实际是通融和性能，使用十组不同的单个系统翻译结果组合所对应的融合结果，实验结果表明高质量的候选可以提升系统融合的性能。

关键词：译文候选；系统融合；机器翻译

中图分类号：TP391　　**文献标识码：**A

Re-examination on Translation Candidates for MT System Combination

ZHU Junguo, YANG Muyun, ZHANG Yu, LI Sheng, ZHAO Tiejun

(Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China)

Abstract: This paper presents an empirical investigation into the factors that would impact SMT system combination performance, focusing on the system number. Given a specific source language, the influence of the system number is analyzed at first via the oracle scores corresponding to different number of system involved in the combination. Experimental result shows that the increment of the system involved in the combination brings obvious improvement of oracle performance of system combination. In parallel to this, as for the multiple source languages for the same target language translation, statistical evidences show that the oracle performance of multiple source system combination is better than single source under the same number of involved system. Further, we verify the findings by state-of-the-art system combination technology, experimental results shows that high quality input will contribute to better combination performance.

Key words: Translation Candidates; System Combination; Machine Translation

1 引言

近年来，统计机器翻译（SMT）已经取得显著的发展，提出多种不同的统计机器翻译框架，包括基于短语的统计机器翻译[1]和基于层次短语的统计机器翻译[2]等。不同的翻译方法存在各自的优点和需要改进的方面，系统融合技术作为一种能够利用不同机器翻译系统的输出结果形成质量更高的翻译结果，已经在各种机器翻译评测中作为有效的手段被广泛使用。

类似于自动语音识别中的 ROVER 方法[3]，系统融合对多个机器翻译系统的输出结果使用投票的方给出融合结果。按照融合过程是如何实现的，系统融合存在三种不同的方法：句子级系统融合[4,5,6]、短语级系统融合[7]和词汇级系统融合[8,9]。句子级系统融合

* **收稿日期：** 2013-09-27　　**定稿日期：** 2013-10-13

基金项目：国家自然科学基金(No. 61272384 & 61105072)；国家高技术研究发展计划(863 计划) 重点项目(No. 2011AA01A207)。

作者简介：朱俊国，男，博士，主要研究方向为机器翻译；杨沐昀，男，副教授，主要研究方向为机器翻译以及自然语言处理；张宇，男，硕士，主要研究方向为机器翻译；李生，男，教授，主要研究方向为机器翻译以及自然语言处理。

方法先将参与融合的单个翻译系统 N-best 列表合并，然后对合并后的 N-best 列表进行重排序。短语级系统融合方法先将参与融合的单个系统短语表合并，然后根据新的短语表对源语言句子重新解码。词汇级系统融合方法首先选出骨架翻译，再将其余翻译结果对齐到骨架后根据对齐结果创建混淆网络，最后根据对数线性模型从混淆网络中寻找最优路径作为融合结果。

系统融合方法可以产生质量优于单个系统的融合结果。但是目前系统融合的研究主要集中于具体的对齐方法等改进，对于系统融合性能影响因素并未涉及。本文对系统候选对于机器翻译系统融合结果的影响进行实验分析，主要内容包括：首先，给出参与融合的系统个数对系统融合性能的影响。不同个数的系统参与融合时，提供了不同数量的融合候选词。本文使用系统融合结果的 BLEU 远景得分(即理论上可达到的最佳得分)作为评价指标，给出了不同个数的系统参与融合时所对应的系统融合远景得分，分析参与融合的翻译结果的来源对系统融合性能的影响。多源融合是使用来自多个源语言的翻译结果参与系统融合，与之相对的单源融合仅使用来自单个源语言的翻译结果参与系统融合。将多源融合与单源融合相比较，多源融合有更多的融合系统且来自不同源语言的系统数量和翻译质量均存在较大差异。最后，我们分析了参与融合的单个系统的翻译质量对系统融合的影响。不同的机器翻译系统翻译结果在翻译质量上存在较大的差异，当需要从多个机器翻译系统的翻译结果选择部分参与系统融合时，如何挑选系统组合对性能的影响至关重要。

本文第 2 部分给出系统融合相关工作和计算远景得分的方法，第 3 部分给出实验结果和分析，第 4 部分给出结论和未来工作。

2 机器翻译中的系统融合

机器翻译系统融合的目的是从多个机器翻译系统的输出构建一致的翻译结果，本文回顾句子级系统融合的相关工作并给出计算系统融合远景得分的方法。

2．1 系统融合相关研究

关于系统融合已经有大量的工作，与本文最相关的是句子级系统融合方法和多源系统融合实验。

句子级系统融合方法是从合并后的 N-best 列表中挑选出质量最高的翻译结果。句子级别系统融合是对一个由多个单个机器翻译系统的翻译结果所组成的 N-best 列表进行排序。目前广泛使用的方法包括基于最小贝叶斯风险(MBR)的方法[5,6]和基于通用线性模型的方法[7]。在基于 MBR 的系统融合方法中，需要从组合成的 N-best 列表中选择一个期望损失最小的翻译结果，公式如下：

$$\hat{e} = \underset{e' \in E_h}{argmin} \sum_{\substack{ref \in E_h \\ ref \neq e'}} L(e', ref) * p(ref|f) \tag{1}$$

在这里E_h代表由多个机器翻译系统结果组成的 N-best 列表。$L(e', ref)$代表损失函数，表示相比于标准翻译ref，当前的翻译结果e'的风险，风险函数的值越小，表示选择翻译结果e'的质量越高，目前广泛使用的风险函数是机器翻译自动评价方法的变体，比如 BLEU[10]、TER[11]等。$p(ref|f)$代表从源语言句子f翻译成标准翻译ref的后验概率，由于在实际的融合过程中不存在可用的标准翻译，N-best 列表中的每一个融合候选分别用作这里的ref。系统融合的输入来自多个单个翻译系统的翻译结果，每个翻译系统使用不同的模型和方法，有的翻译系统可以给出后验概率，有的翻译系统则无法给出，比如基于规则的翻译系统，且不同系统给出的后验概率计算的方法不一样，也不具备可比

性，因此在实际使用中 N-best 列表中每个翻译结果的后验概率是相等的。基于通用线性模型方法的想法是对 N-best 列表中的每一个翻译结果赋予一个置信度得分，选择置信度得分最高的翻译结果作为最终的融合结果。通用线性模型假设每个翻译结果置信度得分取对数后可以表示成多个特征的线性组合，在这个模型中使用的特征包括翻译结果在原先的 N-best 列表中的排名，翻译结果中的词数，系统偏置得分和使用 5-gram 语言模型评价而得到的分值。

Matusov 等人[9]给出多源系统融合的实验结果，在实验设置中他们融合两个日文到英文和两个中文到英文的单个系统翻译结果和对应的远景得分。其工作的目的在于给出基于混淆网络的融合方法与远景得分的比较。本文的多源实验目的是为了比较多源和单源在远景得分上的差异。

2．2 系统融合的远景得分

对于系统融合方法的比较，以往的研究一般要通过实验得到融合结果，并以 BLEU 等自动评价测度给出具体数值进行分析。而本文的目标是探索系统融合候选对于系统融合的影响，为了避免具体融合方法对融合结果造成的限制，我们采用了一种 BLEU 远景得分来分析系统融合的性能上限。

假定目前存在一个有 N 个源语言句子的测试集，其中每个源语言句子的翻译候选有 M 个。对于词汇级别系统融合方法来说，最理想的情况是遍历 M^N 个不同的翻译候选组合，从其中挑选出 BLEU 得分最高的一个，但是即便将这种方法应用于一个小规模的测试集上，所需要的计算时间也是非常巨大，因此我们使用贪心算法来获得句子级别的远景得分，考虑到 BLEU 评分的考虑的影响因素，本文使用如下的翻译候选选择规则：

$$\hat{e} = \underset{e' \in E_h}{argmax} \sum_{n=1}^{4} C_n(e', ref) \tag{2}$$

在这里E_h代表合并后的 N-best 列表，$C_n(e', ref)$表示翻译候选e'和标准翻译ref在 N-gram 上的匹配个数。在这个公式使用一元到四元的文法匹配个数的累计值作为翻译候选的质量评价标准，这种评价方法非常的接近于 BLEU 的得分，在于尽可能的增加匹配个数来提高 BLEU 值。

3 实验分析

3．1 实验数据集

本文使用的数据来自 WMT 2011 系统融合评测的数据，使用捷克语(Cz)、德语(De)、法语(Fr)和西班牙(Es)语到英语(En)的双向翻译方向上的训练集和测试集，其中训练集包含 1,003 句，测试集包含 2,000 句，每个语言对方向上均有多个系统提供 1-best 翻译列表。表 1 给出各语言对方向上提供的单个系统翻译结果的个数。

翻译方向	单个系统个数
Cz-En	8
En-Cz	10
De-En	20
En-De	22
Es-En	15
En-Es	15
Fr-En	18
En-Fr	17

表 1 WMT 2011 各语言对提供的单个系统个数

3．2 候选系统数量对于融合结果远景得分影响

本节给出不同个数的系统参与融合的情况下所对应的远景得分，从而可以获得系统数量对远景得分的影响。

在计算不同个数的系统参与融合时所对应融合远景得分时，本文采用的方法是不断递减参与融合的系统个数的方法，这是由于计算全部 C(N, M)个系统组合的代价太大，本章使用贪心法从 N 个系统中选择 M 个系统参与计算，使用这种方法可以节省大量的时间，具体的步骤如下：

1) 在某个语言对上首先使用全部的 N 个翻译系统 1-best 结果参与融合，计算对应的远景得分，起始将 N 个翻译系统作为集合 C，集合 C 中元素的个数为 L；
2) 将集合 C 中每个系统的 1-best 结果依次删除，计算剩余的 L-1 个翻译系统 1-best 结果的远景得分，记录下所有 L 个组合中得分最低和最高的组合；
3) 如果 L-1 等于 2，则停止整个过程，否则选择得分最高的集合作为新的集合 C，返回步骤 2)继续进行实验。

应用上述方法在 WMT 11 年的所有语言对上方向，表 2 给出的是 11 年英语到捷克语方向上参与评测的提交结果中单个翻译系统（共 8 个）以及系统融合的最差和最好的结果，表 3 给出的是该任务在不同个数的系统参与融合的情况下所对应的句子和词汇级别的远景得分。

类别	最差提交结果	最好提交结果
单个系统提交结果	0.0900	0.1985
系统融合提交结果	0.1898	0.2026

表 2 En-Cz 上单个和系统融合的提交结果

系统数	句子远景最低得分	句子远景最高得分
8	0.2490	0.2490
7	0.2317	0.2482
6	0.2303	0.2471
5	0.2291	0.2458
4	0.2268	0.2438
3	0.2234	0.2413
2	0.2186	0.2363

表 3 En-Cz 方向上不同个数的系统参与融合对应的远景得分

将表 3 中的系统融合任务最好的提交结果与翻译任务最好的提交结果相比较，可以发现句子级系统融合显著提高翻译结果的 BLEU 得分，表明系统融合方法可以融合不同翻译结果的优势。再比较表 3 中系统融合的提交结果与表 2 中的远景得分，现有的系统融合提交结果与远景得分有很大的差距，表明系统融合方法仍然有很大的性能提升空间。在仅使用两个单个系统参与融合的情况下，句子级别的远景得分仍然可以达到 0.2186，高于最好的融合提交结果 0.2026。观察表 2 中远景得分随系统个数的变化，可以看到随着系统个数的减少远景得分逐渐降低。

3．3 翻译系统源语言对系统融合性能影响

多源系统融合是使用来自多个源语言的翻译结果参与系统融合，要求产生翻译结果的多个源语言句子互为翻译。WMT 在翻译任务和系统融合任务中所有语言对方向上使用的测试语料均为平行语料。在本节实验中使用捷克语、法语、德语和西班牙语到英语的翻译方向上提供的单个系统翻译 1-best 结果，并且比较多源和单源系统融合在远景得

分上的对比，同时给出多源融合在不同数量的系统参与融合的情况下对应的句子级别的远景得分。

多源融合相比于单源融合有两个不同的影响因素，第一个影响因素是源语言，第二个影响因素是翻译候选个数。已有的研究结果证明不同语言之间翻译的难度是不相同的，从WMT的结果上可以看到捷克语到英语的翻译结果BLEU分较低而西班牙语到英语的翻译结果 BLEU 得分明显偏高，而多源融合将使用这部分信息来提升融合的远景得分。融合候选个数的增加使得存在更好的接近标准译文的翻译结果的可能性增加，这部分数据使得多源融合在远景得分上的提升比简单增加同一种源语言的翻译候选数更大。

实验使用三种不同配置的数据集，三种不同配置的数据集如下：第一种配置使用所有源语言为西班牙语的单个系统翻译结果，第二种配置包含三个数据集，三个数据源中依次包含源语言为捷克语和西班牙语、德语和西班牙语、法语和西班牙语的单个系统翻译结果，第三种配置包括所有源语言为捷克语、德语、西班牙语和法语的单个系统翻译结果。图 1 给出单源和多源系统融合在最高远景得分上的对比。图 2 给出第二种配置和第三种配置在最高远景得分上的对比，图中给出的都是远景得分随系统个数的变化而变化的曲线图，表 4 给出各个数据集中参与融合的系统个数。

翻译方向	系统个数
All	61
Cz-En+Es-En	23
De-En+Es-En	35
Fr-En+Es-En	33
Es-En	15

表 4 第一组实验中单源和多源融合系统个数

从图 1 中可以看到，在相同个数的系统参与融合的情况下，多源系统融合在远景得分上显著优于单源系统融合。从图 2 可以发现在相同个数的系统参与融合的情况下，使用全部数据的多源系统融合在远景得分上优于所有仅使用包含两个源语言的多源融合。因此使用多源融合可以带来远景得分的提升，表示更大可能的系统融合性能提高潜力。

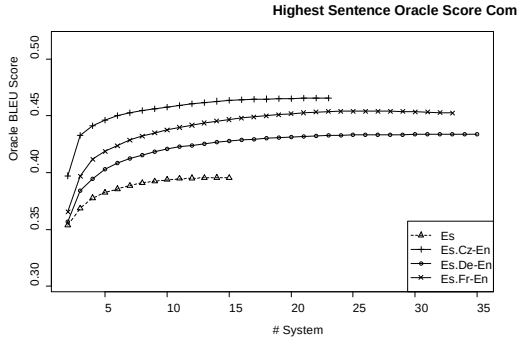

图 1 单源融合和多源融合的对比

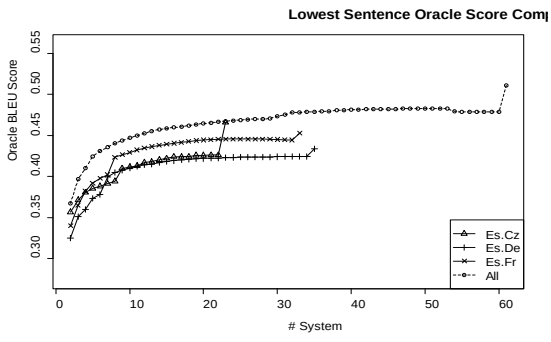

图 2 第二种配置和第三种配置在远景得分上的对比

3．4 多源翻译系统融合的实际性能

Gonzalez-Rubio 等人[6]的文章显示，在所有的融合候选来自同一个源语言的情况，词汇级别的融合效果略优于句子级别的融合，但是在多源数据的情况下，并未有数据支持此种结论，本节给出在特定的数据集下，在使用多源数据时词汇级别融合方法和句子级别融合方法的融合结果比较，使用 BLEU 作为评价方法。

本节的实验使用 WMT2011 数据集上捷克语、德语、法语和西班牙语到英语方向上的单个系统的翻译结果，四种源语言到英语方向上提供翻译结果的系统共有 61 个，可能的多源组合系统非常之多，为了简化实验的流程，对多源数据的组合施加如下的限制：

参与融合的翻译结果必须来自四个源语言，且从每个源语言中仅提供一个系统的提交结果，这样共有 43,200 种可能的组合，这意味着一个句子的所有融合候选仅来自这个组合中使用的单个系统的提交结果。实验按照如下的步骤进行：

1) 计算每种组合的句子级别远景得分，选出远景得分排名前 5 的组合；
2) 从余下的组合中再随机选取 5 个组合；
3) 对这选出的 10 个数据集中的每个数据集，使用词汇级别的融合方法进行系统融合实验，记录对应融合结果的 BLEU 得分。

数据	远景得分	BLEU 得分
远景得分 #1	0.4557	0.3616
远景得分 #2	0.4548	0.3587
远景得分 #3	0.4543	0.3642
远景得分 #4	0.4535	0.3594
远景得分 #5	0.4533	0.363
Random #1	0.3807	0.2925
Random #2	0.3383	0.2863
Random #3	0.3196	0.2792
Random #4	0.3173	0.2637
Random #5	0.3487	0.2852

表 5 词汇级别融合和句子级别融合结果 BLEU 得分

表 5 给出 10 组不同的单个系统组合的情况下，对应的远景得分和融合的 BLEU 得分。比较远景得分排名前 5 的组合（由高到低#1--#5）与随机选择的 5 个组合的融合结果在性能上的差异，可以发现远景得分较高的系统组合在词汇级的融合结果上均优于随机选择的系统组合，这表明远景得分上的较大优势可以直接在现有的框架下被发掘。由于远景得分高表示融合候选与标准译文更接近，因此也表示参与融合的单个系统翻译结果质量更高，从上面数据可以推论高质量的融合候选会产生高质量的融合结果。

4 结论

本文对系统融合性能的影响因素进行了实证分析，影响因素包括参与融合的单个系统的个数、翻译结果的来源和翻译结果的质量。根据实验结果发现，随着参与融合的单个系统个数的增加，系统融合的远景得分也随之增加。相比于单源融合，在相同个数的系统参与融合的情况下，多源融合会获得更高的远景得分。通过比较系统融合远景得分最高的 5 个系统组合的实际融合结果与随机选取的 5 个系统组合的实际融合结果，我们可以发现当参与融合的翻译结果质量较高时所获得的融合结果的质量也较高，因此高质量的单个系统翻译结果可以有效提升融合性能。

对于具有大量单个系统翻译结果的系统融合来说，使用全部的翻译结果未必能获得最好的性能，未来的工作主要是从大量的翻译结果中挑选出高质量的翻译结果，然后进行系统融合。

参考文献

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation[C]//In proceedings of NAACL-HLT, 2003: 48-54.

[2] David Chiang. A Hierarchical Phrase-Based Model For Statistical Machine Translation[C]//In proceedings of the 43rd ACL, 2005: 263-270.

[3] Jonathan G. Fiscus. A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER) [C]// In proceedings of ASRU, 1997: 347-354.

[4] Tadashi Nomoto. Multi-Engine Machine Translation With Voted Language Model[C]//In proceedings of the 42nd ACL. 2004.

[5] Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada and Masaaki Nagata. Generalized Minimum Bayes Risk System Combination[C]//In proceedings of 5th IJCNLP, 2011: 1356-1360.

[6] Jesus Gonzalez-Rubio, Alfons Juan and Francisco Casacuberta. Minimum Bayes-Risk System Combination[C]//In proceedings of the 49th ACL, 2011: 1268-1277.

[7] Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz and Bonnie J. Dorr. Combining Outputs from Multiple Machine Translation Systems[C]//In proceedings of NAACL, 2007.

[8] Shyamsundar Jayaraman and Alon Lavie. Multi-Engine Machine Translation Guided By Explicit Word Matching[C]//In proceedings of the 43rd ACL on Interactive Poster and Demonstration Sessions, 2005: 101-104.

[9] Evgeny Matusov, Nicola Ueffing and Hermann Ney. Computing Consensus Translation From Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment[C]//In proceedings of the 11st EACL, 2006.

[10] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. BLEU: A Method For Automatic Evaluation Of Machine Translation[C]//In proceedings of the 40th ACL, 2002: 311-318.

[11] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, John Makhoul. A Study Of Translation Edit Rate With Targeted Human Annotation[C]//In proceedings of the 7th AMTA, 2006.