文章编号：1003-0077(2018)09-11-09

基于融合策略的机器翻译自动评价方法

马青松[1,2,3]，张金超[1,2,3]，刘 群[1,4]

(1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院大学，北京 100049；3. 腾讯科技(北京)有限公司，北京 100080；4. 都柏林城市大学，都柏林 爱尔兰)

摘 要：机器翻译自动评价发展至今，各种自动评价方法不断涌现。不同的自动评价方法从不同的角度评价机器译文的质量。该文提出了基于融合策略的自动评价方法，该方法可以融合多个自动评价方法，多角度地综合评价机器译文质量。该文主要在以下几个方面探索进行：(1)对比分别使用相对排序(RR)和直接评估(DA)两种人工评价方法指导训练融合自动评价方法，实验表明使用可靠性高的 DA 形成的融合自动评价方法(Blend)性能更好；(2)对比 Blend 分别使用支持向量机(SVM)和全连接神经网络(FFNN)机器学习算法，实验表明在当前数据集上，使用 SVM 效果更好；(3)进而在 SVM 基础上，探索使用不同的评价方法对 Blend 的影响，为 Blend 寻找在性能和效率上的平衡；(4)把 Blend 推广应用到其他语言对上，说明它的稳定性及通用性。在 WMT16 评测数据上的实验，以及参加 WMT17 评测的结果均表明，Blend 与人工评价的一致性达到领先水平。

关键词：机器翻译自动评价；融合；直接评估

中图分类号：TP391 **文献标识码**：A

A Novel MT Metric Based on the Hybrid Strategy

MA Qingsong[1,2,3], ZHANG Jinchao[1,2,3], LIU Qun[1,4]

(1. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Tencent Technology (Beijing) CO. Ltd., Beijing 100080, China;
4. Dublin City University, Dublin, Ireland)

Abstract: With the development of machine translation (MT) evaluation, various MT metrics have been proposed. Different metrics evaluate the quality of MT hypotheses from different perspectives. This paper proposes a novel MT metric that combines the merits of a range of metrics. Our investigation includes several aspects: (1) Comparing the performance of combined metrics that using Direct Assessment manual evaluation (DA) or Relative Ranking human evaluation (RR) to guide the training process. Experiments show that reliable DA human evaluation benefits the combined metric, Blend. (2) Comparing the performance of Blend using SVM or FFNN as the training algorithm. (3) Examining the contribution of metrics incorporated in Blend tentatively, in order to find a trade-off between performance and efficiency. (4) Applying Blend to other language pairs if incorporated metrics support the specific language pair. Experiments on WMT16 and WMT17 Metrics tasks show that Blend achieves the start-of-the-art performance.

Key words: machine translation metric; combined; direct assessment

0 引言

机器翻译自动评价旨在为机器翻译系统提供快速、可靠的质量评估。近些年来，随着机器翻译技术的发展，自动评价也受到越来越广泛的关注。机器翻译自动评价方法通常通过计算机器译文和参考译文的相似度来衡量机器译文质量，不同的自动评价

收稿日期：2017-11-07 **定稿日期**：2018-02-01
基金项目：国家自然科学基金(61379086)

方法从不同的角度计算二者之间的相似度。比如，基于词汇的自动评价方法中，BLEU[1]和NIST[2]统计机器译文和参考译文的共现N元组，Meteor[3]和GTM[4]捕捉机器译文和参考译文之间的词对齐信息，WER[5]、PER[6]和TER[7]计算从机器译文到参考译文的编辑距离。基于句法的自动评价方法主要比较机器译文和参考译文在浅层语法结构[8]、依存句法结构[9]或成分句法结构[10]上的相似度。

虽然各个评价方法都不尽完美，但它们都各自从不同的角度衡量机器译文和参考译文的相似度，反映机器译文在不同评价角度上的质量。那么，多角度的评价将会更全面地反映机器译文的真实质量。一个直接又有效的方法，就是利用各个评价方法的评分，把它们融合成一个新的评价方法。各评价方法的评分代表对机器译文在不同角度上的评价，融合后新的评价方法是对机器译文的多角度综合评价。

文献[11]提出寻找最优组合的方法，各个评价方法按照与人工评价的相关度降序排列，依次尝试加到最优集合里，如果能提高最优集合的性能则加入；否则不加入。这是一种无参数的组合方法。另外，也可以采用有参数的组合方法，最直观的就是线性组合，基本形式如式(1)所示。

$$\text{score} = \sum_i w_i x_i + b \tag{1}$$

其中，w_i表示第i个评价方法x_i的权重。

文献[11]中的无参数组合方式是一种贪心算法，可能会得到局部最优的组合。为了避免这种情况的发生，我们提出有参数的融合自动评价方法，采用机器学习算法进行训练，并进行多方面的实验探索，主要包括以下几个方面。

(1) 根据人工评价方法的不同，我们提出两种融合自动评价方法，分别是DPMFcomb和Blend，实验表明Blend性能更好；

(2) 在Blend上，对比使用支持向量机(SVM)[12]和全连接神经网络(FFNN)两种机器学习算法的性能，实验发现在当前数据集上，使用SVM效果更好。

(3) 进而在SVM基础上，探索融合不同的评价方法对Blend的影响，为Blend寻找在性能和效率上的平衡。

(4) 把Blend推广应用到其他语言对上，验证了它的稳定性及通用性。

后续组织结构如下：第一节介绍模型方法，第二节介绍实验，第三节介绍Blend参加WMT17评测的结果，第四节进行总结。

1 基于融合策略的自动评价方法

我们首先介绍两种人工评价方法，相对排序(relative ranking，RR)和直接评估(direct assessment，DA)；然后介绍分别使用RR和DA指导训练的两种融合自动评价方法：DPMFcomb和Blend。

1.1 两种人工评价方法

在WMT评测任务的发展过程中，先后使用两种人工评价方法，分别是相对排序(RR)和直接评估(DA)。本节中我们将分别介绍这两种人工评价方法。

相对排序的人工评价方法，让评价者对同一个源端句子的五个不同机器译文进行1～5排名，从1到5表示机器译文质量依次下降，并且允许并列排名。表1是RR评价结果的一个示例，它表示对编号为103的句子，评价者给五个机器译文(MTsys1-5)的排名结果。

表1 相对排序(RR)结果的示例

segId	MTsys	Rank
103	MTsys1	3
103	MTsys2	2
103	MTsys3	1
103	MTsys4	3
103	MTsys5	4

直接评估(DA)[13]给出对机器译文绝对的评分，在给定一个机器译文和一个相应的参考译文情况下，评价者通过衡量机器译文在多大程度上充分表达了参考译文的含义，拖动表征机器译文质量的取值范围为1～100的滑动条给出评分。每个评价者的评分都要通过严格的质量控制，并做归一化处理。最后，每个机器译文的评分Score是多个评价者评分(归一化后的评分)的平均值。表2表示评价者使用DA方法对不同编号句子的机器译文的评分。

表2 直接评估(DA)结果的示例

segId	MTsys	Score
100	MTsys1	0.34
101	MTsys2	0.78
102	MTsys3	0.21
…	…	…

相对排序从 2008 年 WMT 自动评价任务开始时使用，一直到 2016 年，积累了多年的数据。相对排序能在一定程度上反映机器译文的质量，但它有两个明显的缺点。首先，相对排序只提供五个给定机器译文的相对排名，这只反映它们之间的相对质量高低，不能反映它们各自的整体质量。其次，相对排序存在人工评价者间的一致性较低的问题[14]，这降低了相对排序的可靠性。与相对排序相比，直接评估能给出机器译文的绝对评分，且设计一系列措施保证其可靠性。因此在 WMT17 评测任务中，直接评估已经取代相对排序，成为唯一的人工评价方法。

1.2　DPMFcomb：相对排序(RR)指导训练的融合自动评价方法

DPMFcomb 使用 RR 人工评价数据，以各个评价方法的评分为特征，使用 SVM 进行训练，是一个与人工评价一致性很高的自动评价方法。DPMFcomb 融合的评价方法，包含 Asiya[15]① 工具中目标端为英语的默认评价方法，包括 55 个基于词汇、句法和语义的自动评价方法(如 BLEU，NIST 等)，以及另外三个自动评价方法，分别是 ENTF[16]，REDp[17]② 和 DPMF[18]。

若把 RR 给出的 1 到 5 的排名看作五个不同的类别，那么 DPMFcomb 的训练过程就可以看作是多分类问题，因此可以用 SVM[12] 进行训练。SVM 是 Vapnik 等人于 1995 年提出的一种学习器，可以用于分类和回归分析。以线性分类问题为例，SVM 可以从训练数据中学习找到一个最优超平面(图 1 的中间一条直线)，实现线性分类。对于线性不可分问题，SVM 通过引入核函数对当前空间进行非线性变换，在高维空间实现线性分类。

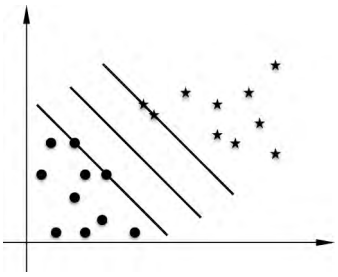

图 1　SVM 寻找最优超平面

具体的，DPMFcomb 使用 SVMrank，训练数据如表 3 所示，第一列是目标类别，即 RR 排名；第二列表示句子编号；从第三列开始，每列代表一个特征，即为融入的各个评价方法的评分。

表 3　DPMFcomb 的训练数据格式

3	qid:103	1:0.32	2:0.37	…	57:0.42
2	qid:103	1:0.54	2:0.48	…	57:0.66
1	qid:103	1:0.56	2:0.57	…	57:0.76
3	qid:103	1:0.35	2:0.39	…	57:0.44
4	qid:103	1:0.22	2:0.30	…	57:0.29
2	qid:104	1:0.54	2:0.50	…	57:0.64
3	qid:104	1:0.29	2:0.39	…	57:0.49
…					

在排序任务中，在测试阶段 SVM 生成的预测值可以转化为对测试集的排序；而在机器翻译评价任务中，自动评价方法通常给出机器译文的质量分数，所以此预测值不必再转化，可直接表示为 DPMFcomb 对机器译文的评分，如式(2)所示。

$$\mathrm{DPMFcomb}=\sum_i w_i\phi(x_i)+b \tag{2}$$

其中，w 和 b 是模型参数，ϕ 表示使用的核函数，x_i 表示融入的第 i 个评价方法的评分。

DPMFcomb 参加了 WMT15-16 评测的自动评价任务，连续两年获得目标端为英语的语言对中与人工评价的平均一致性最高的成绩，其设置及结果可以参考文献[19-20]。

1.3　Blend：直接评估(DA)指导训练的融合自动评价方法

我们提出 DA 指导训练的融合自动评价方法，命名为 Blend，它可以利用任意的自动评价方法的优点，形成一个新的基于融合策略的自动评价方法③。

Blend 与 DPMFcomb 的基本思想一致，但二者在训练数据及训练方法上并不相同。Blend 分别使用回归支持向量机(SVM regression)和全连接神经网络(FFNN)训练，找到使其性能最优的训练方式。

(1) 使用 libsvm[21] 中的 SVM regression 训练时，训练数据如表 4 所示。

① http://asiya.lsi.upc.edu/

② DPMFcomb 在 WMT15 评测中融入 REDp，在 WMT16 评测中没有融入 REDp。下文实验使用 DPMFcomb 在 WMT16 评测中的配置。

③ https://github.com/qingsongma/Blend

表4 Blend的训练数据格式

0.34	1:0.36	2:0.43	3:0.52	…	57:0.43
0.78	1:0.66	2:0.72	3:0.43	…	57:0.57
0.21	1:0.28	2:0.30	3:0.19	…	57:0.20
…					

其中，第一列表示目标值，即为DA评分；之后每列代表一个特征，即融入的各个评价方法的评分。最终Blend评分如式(3)所示。

$$\text{Blend} = \sum_i w_i \phi(x_i) + b \tag{3}$$

(2) FFNN是由输入层、隐含层(一层或多层)和输出层构成的神经网络模型，其隐含层和输出层的每一个神经元与相邻层的所有神经元连接(即全连接)，如图2所示。

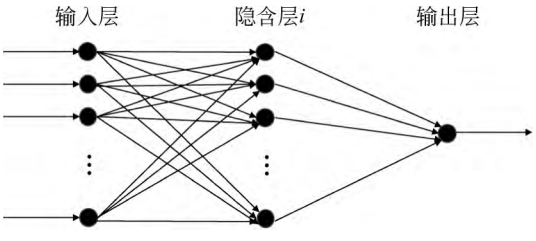

图2 全连接神经网络(FFNN)模型图

当Blend采用FFNN的训练时，输入层的每个输入表示各个评价方法的评分，输出层的输出为Blend对机器译文的评分。设输入层有M个输入节点，隐含层有N个节点，输出层是一个节点，则有：

$$y_j = \sum_{i=1}^{M} w_{ij}\, x_i + b_j \tag{4}$$

$$O_j = f(y_j) \tag{5}$$

$$\text{Score} = \sum_{j=1}^{N} O_j\, w_j + b \tag{6}$$

其中，x_i表示i个输入节点的输入值，即第i个评价方法的评分；w_{ij}表示第i个输入节点到第j个隐含层节点的权重；$f(\cdot)$表示激励函数；w_j表示第j个隐含层到输出层的权重；b_j和b表示偏置值；Score是输出层的输出，代表Blend对机器译文的评分。

2 实验

我们进行了四组实验：(1)探索基于相对排序数据的DPMFcomb和基于直接评估数据的Blend在目标端为英语的语言对上的表现，对比两种模型的性能；(2)分别实现基于SVM和FFNN的Blend训练方法，对比二者性能；(3)实验了融合不同种类的自动评价方法，为Blend寻找在性能和效率上的平衡；(4)在其他语言对上验证Blend的有效性。模型评价指标是模型输出与标准人工评价分数的皮尔逊(Pearson)一致性系数。

2.1 实验设置

我们在WMT16评测任务中目标端为英语的各语言对上和英语—俄语语言对上测试。DA评价数据从WMT15-16评测任务中获得，数据量情况如表5所示。因为目前只有少数有限的DA评价数据，当我们测试WMT16中每一个目标端为英语的语言对(560句)时，使用WMT16的其他目标端为英语的语言对和WMT15的所有目标端为英语的语言对数据进行训练(共4 800句)。对于英语—俄语语言对，我们使用WMT15的英语—俄语数据(500句)训练，在WMT16的英语—俄语(560句)上测试。

表5 WMT15-16评测任务DA评价数据量

	捷克语—英语	德语—英语	芬兰语—英语	罗马尼亚语—英语	俄语—英语	土耳其语—英语	英语—俄语
WMT15	500	500	500	—	500	—	500
WMT16	560	560	560	560	560	560	560

使用SVM regression训练时，训练数据和测试数据的特征都归一化到[−1,1]区间。我们使用epsilon-SVR，选择RBF核函数，epsilon设置为0.1。使用FFNN训练时，训练集与测试集保持与使用SVM regression时一致，并从训练集中随机抽取500句作为开发集，其他设置在下文中详细介绍。

2.2 Blend与DPMFcomb的对比实验

在WMT16评测中，DPMFcomb融合57个自动评价方法，使用SVMrank，从WMT12-WMT14评测任务的所有目标端为英语的语言对中，根据RR评价结果，抽取约445 000的训练数据。为了对比，Blend融合同样的57个自动评价方法，使用

SVM regression，从 WMT15-WMT16 的目标端为英语的语言对上，抽取 4 800 句训练数据进行训练，训练得到的模型称为 Blend. all。

表 6 和表 7 分别列出了系统级和句子级的 Pearson 一致性系数。表 6 显示 Blend. all 在 WMT16 的目标端为英语的语言对中，在系统级上与人工评价的平均一致性(0. 951)达到最高，超过了当年评测中表现最好的两个自动评价方法，MPEDA(0. 941)和 BEER (0. 920)。表 7 列出 WMT16 评测的目标端为英语的语言对中，Blend. all 和另外两个表现最好的自动评价方法 DPMFcomb 和 EMTRICS-F 在句子级上的 Pearson 系数。DPMFcomb 在 WMT16 评测的目标端为英语的语言对上表现最好，说明融合评价方法的有效性。表 7 显示 Blend. all 在所有目标端为英语的语言对的平均 Pearson 系数最高。值得一提的是，虽然 Blend. all 的训练集远远少于 DPMFcomb 的训练集，Blend. all 的平均 Pearson 系数(0. 641)却高于 DPMFcomb(0. 633)。

所以，以上结果说明在 WMT16 评测的目标端为英语的语言对中，DA 指导训练的 Blend，在性能上优于 RR 指导训练的 DPMFcomb。这在一定程度上是由于 DA 数据比 RR 数据可靠：RR 数据只反映机器译文间的相对质量，且存在评价者间一致性较低的问题；而 DA 数据给出机器译文的绝对评分，并且设计一系列措施保证其可靠性。因此，我们后面的实验在 Blend 上进行。

表 6　在 WMT16 评测数据上各自动评价方法的 10K 系统级的 Pearson 系数

	捷克语—英语	德语—英语	芬兰语—英语	罗马尼亚语—英语	俄语—英语	土耳其语—英语	平均
Blend. all	**0. 991**	**0. 954**	0. 969	0. 879	**0. 942**	0. 972	**0. 951**
MPEDA	0. 988	0. 923	**0. 971**	**0. 905**	0. 923	**0. 975**	0. 948
BEER	0. 985	0. 871	0. 964	0. 828	0. 894	**0. 975**	0. 920

表 7　在 WMT16 评测数据上各自动评价方法的句子级 Pearson 系数

	捷克语—英语	德语—英语	芬兰语—英语	罗马尼亚语—英语	俄语—英语	土耳其语—英语	平均
Blend. NN	0. 710	0. 603	**0. 605**	0. 631	**0. 631**	0. 651	0. 639
Blend. all	0. 710	**0. 615**	0. 602	0. 636	0. 622	0. 658	**0. 641**
DPMFcomb	**0. 713**	0. 598	0. 584	0. 627	0. 615	**0. 663**	0. 633
METRICS-F	0. 696	0. 601	0. 557	**0. 662**	0. 618	0. 649	0. 631

2. 3　Blend 分别使用 SVM regression 和 FFNN 的对比实验

Blend 设计分别使用 SVM regression 和 FFNN 训练的对比实验，从中选择一个更优的训练方式。首先，我们在捷克语—英语上尝试多组实验，寻找使得 Blend 在使用 FFNN 训练时的最优实验参数设置。实验结果如表 8 所示。

表 8(a)探索使用不同的数据形式，即原始数据(各个评价方法的评分)、使用 libsvm 中的 svm_scale(表 8 中记为 svm_std)归一化数据，以及 Z 值数据。不同的数据形式，分别与一层或两层全连接神经网络组合，其他设置相同，具体如下：采用 SGD 优化方法，学习率设为 0. 01，使用 sigmoid 激励函数，隐层维度设为 57(与输入向量维度一致，即为融入的评价方法的个数)。由表 8(a)可知，2NN-orgin，即使用原始数据及两层神经网络的实验设置，与 DA 人工评价的 Pearson 一致性系数最高。表 8(b)首先在 2NN-orgin 基础上尝试不同的隐层维度，分别为 64、128、256、512 和 1024。实验发现当隐层维度为 256 时，Pearson 系数相对较高。之后在 2NN-orgin-256 上增加 L1、L2 正则项，其 Pearson 系数有所增加；继而将 sigmoid 分别换为 tanh 和 ReLU 激活函数，发现使用 tanh 时效果有明显提升。表 8(c)在表 8(b)基础上，把三种数据形式与设置为 0. 5 的 dropout 分别组合，发现当使用 svm_scale 与 dropout 组合设置时，Pearson 系数再次显著提升。表 8(d)尝试不同的 dropout 值，发现当其设置为 0. 1 时效果最好；继而尝试更深的网络层数，发现效果稍微下降。

所以，我们采用 2NN-svm__std-256-L-tanh-drop0.1 的实验设置，记作 Blend.NN，并采用此设置在其他到英语端的语言对上实验，其结果与使用 SVM regression 训练得到的模型 Blend.all 比较，结果如表 7 所示。由表 7 可知，在当前数据集上，Blend 使用 SVM 的训练方式(Blend.all，0.641)略优于使用 FFNN(0.639)，由此可以说明 SVM 在小数据集上就有较好的表现，我们下文的实验均在 SVM regression 上进行。

表 8 各模型在 WMT16 的捷克语—英语上的 Pearson 系数

NN-orgin	-0.124
NN-svm_std	0.606
NN-z_std	0.593
2NN-orgin	**0.634**
2NN-svm_std	0.625
2NN-z_std	0.592

(a)

2NN-orgin-256	**0.637**
2NN-orgin-256-L	**0.638**
2NN-orgin-256-L-tanh	**0.663**

(b)

2NN-svm_std-256-L-tanh	0.657
2NN-z_std-256-L-tanh	0.681
2NN-orgin-256-L-tanh-drop0.5	0.692
2NN-svm_std-256-L-tanh-drop0.5	**0.698**
2NN-z_std-256-L-tanh-drop0.5	0.681

(c)

2NN-svm_std-256-L-tanh-drop0.1	**0.710**
3NN-svm_std-256-L-tanh-drop0.1	0.708
4NN-svm_std-256-L-tanh-drop0.1	0.708

(d)

2.4 Blend 在性能和效率上的平衡

原则上，为获得与人工评价数据更高的一致性，Blend 能够融入更多数量的自动评价方法。然而，是否有些评价方法在性能上没有对 Blend 起很大的作用，同时还降低了 Blend 的效率呢？为了探寻这点，我们把 Asiya 工具中适用于目标端为英语的语言对的默认自动评价方法分为三类，分别是基于词汇、基于句法和基于语义的评价方法。下文中 Blend.lex 只融合了默认的基于词汇的自动评价方法，Blend.syn 和 Blend.sem 分别表示只融合了基于句法和基于语义的自动评价方法。Blend.lex 包含 25 种自动评价方法，但实际只有九种自动评价方法①，因为其中有些自动评价方法只是一种自动评价方法的不同变种。Blend.syn 和 Blend.sem 分别包含 17 种和 13 种自动评价方法，但实际各自对应三种不同的自动评价方法(详见文献[15])。

在 WMT16 评测的句子级实验结果如表 9 所示。Blend.all，包含 Asiya 所有默认的评价方法，在五个目标端为英语的语言对(共 6 种)上与人工评价的一致性，以及平均一致性达到最高。然而，值得注意的是：Blend.lex 在句子级上与人工评价的平均一致性与 Blend.all 相比仅差 0.009，而 Blend.syn 和 Blend.sem 的性能远低于 Blend.all，甚至低于 Blend.lex。基于句法和基于语义的自动评价方法通常比较复杂，耗时较长。经测试，基于词汇、句法和语义的评价方法在服务器上的平均用时为 19.3ms/句、85.5ms/句和 181.4ms/句②。Blend.lex 的性能与 Blend.all 相当，所以 Blend 可以只融合 Asiya 工具中基于词汇的评价方法，在达到高性能的同时提高效率。

我们又继续增加了四种其他的自动评价方法到 Blend.lex 中：CharacTer[22]，一种基于字符的自动

表 9 在 WMT16 评测数据上 Blend 融合不同类型的评价方法时的句子级 Pearson 系数

	捷克语—英语	德语—英语	芬兰语—英语	罗马尼亚语—英语	俄语—英语	土耳其语—英语	平均
Blend.all	**0.710**	**0.615**	**0.602**	**0.636**	**0.622**	0.658	**0.641**
Blend.lex	0.704	0.589	0.583	0.625	0.620	**0.674**	0.632
Blend.syn	0.656	0.528	0.494	0.560	0.533	0.610	0.564
Blend.sem	0.610	0.533	0.492	0.507	0.501	0.554	0.533

① 分别是 BLEU，NIST，GTM，METEOR，ROUGE，Ol，WER，TER 和 PER。

② CPU：AMD Opteron(TM)，8 核，8 线程；内存：96GB

评价方法；BEER[23]，一种融入多角度特征的自动评价方法；DPMF 和 ENTF（在 DPMFcomb 的实验中证明了它们的有效性）。新增的四种自动评价方法分别从字符、句法等角度衡量机器译文质量，且都方便使用。表 10 说明 Blend. lex＋4(0. 640)的性能优于 Blend. lex(0. 632)，并且与表 9 中的 Blend. all (0. 641)非常接近，可以作为 Blend 在性能和效率上的一个很好的平衡。

表 10　在 WMT16 评测数据上 Blend. lex 加入 4 个不同类型的评价方法时的句子级 Pearson 系数

	捷克语—英语	德语—英语	芬兰语—英语	罗马尼亚语—英语	俄语—英语	土耳其语—英语	平均
Blend. lex	0. 704	0. 589	0. 583	0. 625	0. 620	0. 674	0. 632
Blend. lex＋CharacTer	0. 707	0. 596	0. 575	0. 628	0. 620	**0. 680**	0. 634
Blend. lex＋BEER	**0. 709**	0. 589	0. 580	0. 627	0. 622	0. 673	0. 634
Blend. lex＋DPMF	0. 706	0. 592	**0. 590**	0. 632	0. 626	0. 670	0. 636
Blend. lex＋ENTF	0. 703	0. 595	0. 588	0. 629	0. 629	0. 676	0. 637
Blend. lex＋4	**0. 709**	**0. 601**	0. 584	**0. 636**	**0. 633**	0. 675	**0. 640**

2. 5　Blend 在其他语言对上的实验

Blend 可以适用于任何语言对，只要融入的评价方法支持这种语言对。因为目前除了目标端为英语的语言对外，只有英语—俄语的 DA 评价数据，所以我们在 WMT16 评测的英语—俄语语言对上实验来说明这一点，其句子级一致性结果如表 11 所示。

表 11　在 WMT16 评测的英语—俄语语言对中各自动评价方法的句子级 Pearson 系数

	英语—俄语
Blend. default	0. 613
Blend. default＋2	0. 675
BEER	0. 666

Blend. default 融合 Asiya 提供的适用于英语—俄语的默认自动评价方法，共 20 个，实质为九种①。模型在 500 句训练集上训练得到。Blend. default＋2 在 Blend. default 基础上，只加入 BEER 和 CharacTer，在句子级的 Pearson 系数上取得很大提升，从 0. 613 上升到 0. 675。BEER 是在 WMT16 评测中英语—俄语的最好的自动评价方法(0. 666)，此实验结果显示，BEER 可以在性能上给 Blend 带来很大提升，同时 Blend 可以进一步提升性能已经很好的自动评价方法，再一次说明融合策略的有效性。

3　Blend 在 WMT17 评测上的结果

Blend 参加了 WMT17 评测的自动评价任务。在目标端为英语的语言对中，提交 Blend. lex＋4，其训练数据包括 WMT15 和 WMT16 所有目标端为英语的语言对的数据，共 5 360 句。在句子级上，Blend 在所有七种目标端为英语的语言对中，均获得了第一名的成绩；在系统级上，在六种目标端为英语的语言对（共七种）中取得了第一名的成绩；在 10K 系统级(10 000 个翻译系统)上，在两种目标端为英语的语言对（共七种）中获得了第一名。

此外，Blend 参加了英语—俄语语言对的自动评价任务，提交 Blend. default＋2，训练数据包括 WMT15 和 WMT16 两年英语—俄语的数据，共 1 060 句。Blend 在英语—俄语语言对中，取得在句子级上第五（与最高的一致性系数相差 0. 058）、系统级第一、10K 系统级上第二的成绩。WMT17 评测结果的详细报告参见文献[24]，Blend 的系统报告参见文献[25]。文献[25]是本文提出的融合评价方法系列探索性工作的一部分，本文相比于文献[25]，有更系统的探索、实验和分析。

4　总结

本文提出基于融合策略的自动评价方法，融合多个自动评价方法，以形成一个新的、与人工评价有更高一致性的自动评价方法。根据人工评价方法的不同，我们提出两种融合自动评价方法，分别是 DPMFcomb 和 Blend，实验结果表明：使用 DA 指导训练的 Blend，即使在较少的训练数据上，其性能

① 与 Blend. lex 一样的 9 种。

也优于 DPMFcomb；在 Blend 上，对比使用 SVM 和 FFNN 两种机器学习算法的性能，发现在当前数据集上使用 SVM 效果略好（此结论仅限于当前数据集）；我们进一步探索了在 SVM 基础上融合不同的评价方法对 Blend 的影响，为 Blend 寻找在性能和效率上的平衡；在多个语言对上进行了实验，证明了 Blend 的稳定性及通用性。该文提出的 Blend 方法参加了 WMT17 评测，取得了多项第一的优异成绩。

参考文献

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

[2] Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurence statistics [C]//Proceedings of the 2nd International Conference on Human Language Technology Research. San Diego, California, 2002: 138-145.

[3] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language [C]//Proceedings of the 9th Workshop on Statistical Machine Translation. Baltimore, Maryland USA, 2014: 376-380.

[4] Melamed I D, Green R, Turian J P. Precision and recall of machine translation [C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003—short papers-Volume 2. Edmonton, Canada, 2003: 61-63.

[5] Nießen S, Och F J, Leusch G, et al. An evaluation tool for machine translation: Fast evaluation for MT research[C]//Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Greece, 2000.

[6] Tillmann C, Vogel S, Ney H, et al. Accelerated DP based search for statistical translation[C]//Proceedings of the 5th European Conference on Speech Communication and Technology. Rhodes, Greece, 1997.

[7] Snover M, Madnani N, Dorr B J, et al. Fluency, adequacy, or HTER?: Exploring different human judgements with a tunable MT metric[C]//Proceedings of the 4th Workshop on Statistical Machine Translation. Athens, Greece, 2009: 259-268.

[8] Chan Y S, Ng H T. MAXSIM: a maximum similarity metric for machine translation evaluation[C]//Proceedings of ACL-08: HLT. Columbus, Ohio, USA, 2008: 55-62.

[9] Owczarzak K, van Genabith J, Way A. Evaluating machine translation with LFG dependencies[J]. Machine Translation, 2007, 21(2): 95-119.

[10] Liu D, Gildea D. Syntactic features for evaluation of machine translation[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 25-32.

[11] Giménez J, Màrquez L. Heterogeneous automatic MT evaluation through non-parametric metric combinations[C]//Proceedings of the 3rd International Joint Conference on Natural Language Processing. Hyderabad, India, 2008.

[12] Vapnik V N. Statistical learning theory(Vol. 1)[M]. New York: Wiley, 1998.

[13] Graham Y, Baldwin T, Moffat A, et al. Continuous measurement scales in human evaluation of machine translation[C]//Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse. Sofia, Bulgaria, 2013: 33-41.

[14] Callison-Burch C, Fordyce C, Koehn P, et al. (Meta-) evaluation of machine translation[C]//Proceedings of the 2nd Workshop on Statistical Machine Translation. Prague, Czech Republic, 2007: 136-158.

[15] Giménez J, Màrquez L. Asiya: An open toolkit for automatic machine translation (Meta-) Evaluation [J]. Prague Bull. Math. Linguistics, 1994(1): 77-86.

[16] Yu H, Weizhi X, Liu Q, et al. ENTF: An Entropy-Based MT evaluation Metric [C]//Proceedings of 13th China Workshop, CWMT. Dalian, China, 2017: 68-77.

[17] Yu H, Wu X, Xie J, et al. RED: A reference dependency rased MT evaluation metric[C]//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 2042-2051.

[18] Yu H, Wu X, Jiang W, et al. An automatic machine translation evaluation metric based on dependency parsing model[J]. arXiv preprint. 2015. arXiv:1508.01996.

[19] UFAL, M. Results of the WMT15 metrics shared task[C]//Proceedings of the 10th Workshop on Statistical Machine Translation. Lisboa, Portugal, 2015: 256-273.

[20] Bojar O, Graham Y, Kamran A, et al. Results of the WMT16 metrics shared task[C]//Proceedings of the 1st Conference on Machine Translation. Berlin, Germany, 2016: 199-231.

[21] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3), 27.

[22] Wang W, Peter J T, Rosendahl H, et al. Character: Translation edit rate on character level[C]//Proceedings of the 1st Conference on Machine Translation. Berlin, Germany, 2016: 505-510.

[23] Stanojevic M, Sima'an K. BEER 1.1: ILLC UvA submission to metrics and tuning task[C]//Proceedings of the 10th Workshop on Statistical Machine Translation. Lisboa, Portugal, 2015: 396-401.

[24] Bojar O, Graham Y, Kamran A. Results of the WMT17 metrics shared task[C]//Proceedings of the Conference on Machine Translation (WMT). Copenhagen, Denmark, 2017: 489-513.

[25] Ma Q, Graham Y, Wang S, et al. Blend: a novel combined MT metric based on direct assessment—CASICT-DCU submission to WMT17 metrics task[C]//Proceedings of the Conference on Machine Translation (WMT). Copenhagen, Denmark, 2017: 598-603.

马青松(1993—)，博士，主要研究领域为自然语言处理、机器翻译、机器翻译评价。
E-mail: maqingsong@ict.ac.cn

张金超(1989—)，博士，主要研究领域为自然语言处理、机器翻译。
E-mail: zhangjinchao@ict.ac.cn

刘群(1966—)，博士，研究员，教授，主要研究领域为自然语言处理、机器翻译。
E-mail: liuqun@ict.ac.cn

中国中文信息学会大数据安全与隐私保护专业委员会成立大会暨第二届网络空间安全学术前沿与学科建设研讨会在西安召开

2018 年 7 月 25 日下午，中国中文信息学会大数据安全与隐私保护专业委员会在西安召开成立大会，中国中文信息学会理事长方滨兴院士、副理事长兼秘书长孙乐研究员，中国科学院信息工程研究所副总工李凤华研究员及中文信息学会大数据安全与隐私保护专业委员会委员等 50 余名专家学者出席大会，会议由中国中文信息学会副理事长兼秘书长孙乐主持。

大数据安全与隐私保护已成为电子政务、电子商务、医疗、教育、金融证券、社会治理、军事国防、智慧城市等国家重大行业发展的瓶颈，是关系到国计民生、经济发展、社会稳定的核心关键技术，学术界、工业界以及政府机构都高度关注该问题。大数据安全与隐私保护涉及网络空间安全、计算机科学与技术等相关学科，成立学会大数据安全与隐私保护专业委员会的目的就是促进大数据安全与隐私保护的学术交流与技术进步、推动大数据安全与隐私保护的技术与产业发展。

经委员投票选举，会议选举中国科学院信息工程研究所李凤华研究员为主任委员，西安电子科技大学李晖教授、浙江大学任奎教授、上海交通大学邱卫东教授、暨南大学翁健教授、百度主任架构师吴烨等五人为副主任委员，西安电子科技大学沈玉龙教授为秘书长，陈金俊、邹德清、殷丽华三人为副秘书长。

随后，李凤华主任委员发言表示专委会将在中国中文信息学会的领导下，重点开展大数据安全与隐私保护方面的科学研究、学术交流、成果转化、科技普及、智库支持等方面工作。副主任委员李晖代表专委会挂靠单位发言表示将努力为专委会做好服务和保障工作。会议还对专委会 2018 年工作计划进行了讨论。

最后，理事长方滨兴院士做了总结讲话，他首先对专委会的成立和第一届领导集体表示祝贺，并期望新成立的大数据安全与隐私保护专业委员会能够通过定期举办大数据安全及隐私保护方面的研讨，着力推动解决云安全、数据安全及应用安全等安全领域前沿问题。