统计机器翻译中非线性参数学习的研究

陈华栋，黄书剑，赵迎功，戴新宇，陈家骏
南京大学计算机软件新技术国家重点实验室，南京 210023
Email: {chenhd, huangsj, zhaoyg, daixy, chenjj}@nlp.nju.edu.cn

摘　要：尽管对数线性模型在统计机器翻译系统中取得了显著的效果，但是该模型仍然存在着缺陷：首先，对数线性模型限制了使用的特征与模型得分之间必须成线性关系。其次，对数线性模型无法深度的刻画和表达特征中蕴含的信息。这使得特征的设计需要大量的人为知识和经验的干预，简单的线性组合无法充分使用特征，致使很多有效的特征被忽视。本文探讨了非线性模型在统计机器翻译中的应用的可能性，并进行了目标函数和采样方法选择的实验。

关键字：对数线性模型；统计机器翻译；非线性模型。

Research on Parameter Learning of Non-linear Model in Statistic Machine Translation

Huadong Chen, Shujian Huang, Yinggong Zhao, Xinyu Dai and Jiajun Chen
State Key Laboratory for Novel Software Technology, Nanjing University, 210023
Email: {chenhd, huangsj, zhaoyg, daixy, chenjj}@nlp.nju.edu.cn

Abstract: *Although the log-linear model achieves great success in SMT, it still suffers from some drawbacks: first, the features which are used in the model must be linear with respect to the model itself; then, it cannot further interpret the features to reach their potential information, which makes feature designation hard and often needs human knowledge and understanding to help. The log-linear model combines features with mere simply dot product operation which cannot make full use of the features and also makes some of efficient features are ignored. This paper focuses on the possibility of using non-linear model in SMT and conducts experiments to verify which object function and sample method are suitable for this kind of situation.*

Keywords: *log-linear model, SMT, non-linear model*

1　引言

自从[Och et al., 2002]提出使用对数线性模型进行统计机器翻译系统的训练以来，大部分的统计机器翻译系统[Yamada et al., 2001, Koehn et al., 2003, Liu et al., 2007, Chiang et al., 2007, Shen et al., 2010]都是基于对数线性模型进行训练的。该模型使得统计机器翻译的翻译效果取得了显著的进步。对数线性模型的优点是使得任意的特征能够简单的加入到当前的统计机器翻译系统中。后续的大量研究工作集中于怎样设计有效的特征用以描述翻译过程，从而使得双语语言对之间复杂的翻译问题转化成了特征的设计与选择的问题。现阶段常见的统计机器翻译系统分为两个主要部分：系统相关的特征描述和基于对数线性模型的特征参数学习。

尽管对数线性模型促进了统计机器翻译的发展，在翻译性能上取得了显著的提高，该模型仍然存在着一些潜在的不足之处：首先，对数线性模型限制了模型所使用特征的对数

值与模型得分之间必须成线性关系，然而潜在的特征是否与模型得分成线性关系往往难以得到保证，这使得特征的设计需要大量的知识和理解，这通常需要经验性的设计和大量的实验验证。其次，对数线性模型通过将一系列的特征进行简单的线性组合而得到最终的模型得分，而特征之间并非都是相互独立的，对于特征之间存在的非线性组合以及耦合现象，例如，词翻译概率特征在累加计算过程中与词计数特征之间的信息是有相互重叠的，并不是简单的线性组合所能够描述的，同样的对于短语翻译概率特征和短语计数特征也是如此。对数线性模型无法进行有效的刻画。

基于上述分析，本文对统计机器翻译中常用的特征之间采用非线性组合的方式进行了实验尝试，研究了非线性模型对于统计机器翻译的特征表达和组合的影响。本文的后续部分中，第二部分介绍了相关工作背景，第三部分介绍了传统的对数线性模型和本文使用的非线性模型，第四部分给出利用非线性模型进行参数学习的若干可能的学习方法和策略，第五部分给出了实验比较和分析，第六部分对全文工作进行了小结。

2 相关工作

[Maskey et al., 2012]通过对原有的翻译概率采用无监督的深度置信网络（Deep Belief Network，简写为 DBN）生成新的特征，而这些特征相比于原来的特征更有区分能力，有利于翻译质量的提高。[Lu et al., 2014]通过半监督的深度自编码网络（Deep Auto-encoder ，简写为 DAE）对短语系统中的一系列特征进行了重新表达，得到了一系列新的特征，并且通过统计分析得到新的特征具有更大的方差分布，即更具区分能力。将新的特征加入翻译系统中，取得了翻译性能的提升。这两项工作，都是将原来的特征采用特征重新表达模型进行重新表达，然后作为新的特征加入翻译系统，从而取得性能的提升，体现了当前的对数线性模型并未能充分的解释和表达现有的一些特征。[Zhao et al., 2014]对于翻译系统中的稀疏化特征之间相互重叠耦合的现象进行了研究，采用深度自编码网络（DAE）对稀疏化特征进行重新组合，对稀疏化特征进行降维解决特征之间的信息冗余和过拟合问题，其后将处理后获得的新特征作为额外特征加入翻译系统，并在实验中验证了非线性的特征组合是编码特征的附加信息的重要因素。

[Duh et al., 2008]对于对数线性模型可能造成的在训练数据上的欠拟合问题进行了研究，在重排序模块利用单个的最小错误率训练的对数线性模型作为一个弱学习器，采用 Boosting 的思想来组合多个弱学习器，提出了 BoostedMERT 模型，在重排序上取得了不错的效果。[Liu et al., 2013]针对对数线性模型的线性假设和深度表达和解释特征的缺陷，在对数线性模型的基础上，提出了附加神经网络（Additive Neural Network，简写为 AdNN）模型，该模型分为两个部分，线性部分用来处理非局部特征，非线性部分用来处理局部特征，采用单层神经网络对局部特征进行重新解释和组合，一定程度上缓解了对数线性模型的缺陷。

上述的工作分别体现了对数线性模型在线性假设和特征的解释表达方面的问题，也对此作出了一些改进，然而这些方法仍然未能脱离对数线性模型的范畴，往往是通过对原有部分特征或引入的新的特征采用非线性模型进行重新表达，将重新表达后的特征作为额外特征加入对数线性模型中，即，对该新特征与原有特征间的组合仍然是采用对数线性模型。本文将对于直接采用非线性模型组合和表达特征的有效性进行研究和验证，并给出相关实验结果。

3 模型

3.1 对数线性模型

[Och et al., 2002]提出了对数线性模型，该模型可以表示为如下的公式：

$$P(e,d|f;W)=\frac{\exp\{W^{\top}\cdot h(f,e,d)\}}{\sum_{e',d'}\exp\{W^{\top}\cdot h(f,e',d')\}} \quad (1)$$

其中f标记了源语言句子，e、e′标记了源语言句子的翻译候选，d、d′标记语言对$<\mathrm{f,e}>$的推导，即层次短语模型中的同步规则集合，或者短语模型中的短语对集合。$\mathrm{h(f,e,d)}$是定义在$<\mathrm{f,e,d}>$上的特征向量，W是h对应的权重向量，即模型参数。对数线性模型不同于生成式模型，它能够方便的加入任意的特征。

对数线性模型的解码获得最佳候选译文ê的过程，即选择对数线性模型下概率最高的候选译文的过程，可以表示为如下公式：

$$<\hat{e},\hat{d}>=\mathrm{argmax}_{e,d}P(e,d|f;W)=\mathrm{argmax}_{e,d}(W^{\top}\cdot h(f,e,d)) \quad (2)$$

3.2 非线性模型

尽管对数线性模型在统计机器翻译的性能上取得了显著的效果，该模型仍然存在着一些缺陷：首先，对数线性模型假设特征与模型得分之间必须成线性关系，限制了特征的使用。其次，对数线性模型将一系列的特征进行简单的线性组合，无法深度的解释和刻画组合特征所表达的信息。

非线性模型无疑可以更加多样的组合现有特征，而神经网络在特征表达方面的应用已经在很多方面取得了显著的效果，所以本文采用神经网络取代当前的对数线性模型，研究其在统计机器翻译中的表现。本文采用简单的单隐层神经网络模型，该模型可以表示为如下的公式：

$$\mathrm{S(f,e,d;W,M,B,B')}=\sigma\left(\mathrm{W}^{\top}\cdot\sigma\left(\mathrm{M}\cdot\mathrm{h(f,e,d)}+\mathrm{B}\right)+\mathrm{B}'\right) \quad (3)$$

其中f标记了源语言句子，e标记了源语言句子的翻译候选，d标记语言对$<\mathrm{f,e}>$的推导，$\mathrm{M}\in\mathbb{R}^{\mathrm{u}\times\mathrm{k}}$是第一层网络的权重矩阵，$\mathrm{B}\in\mathbb{R}^{\mathrm{u}}$是一个向量，为第一层网络的偏置，$\mathrm{W}\in\mathbb{R}^{\mathrm{u}}$是隐层到输出层的权重向量，B′是输出层的权重，σ表示神经网络的激活函数，本文采用sigmoid 激活函数。

利用非线性模型取代对数线性模型，实际改变了得分函数的计算过程，解码的搜索目标变为：

$$<\hat{\mathrm{e}},\hat{\mathrm{d}}>=\mathrm{argmax}_{\mathrm{e,d}}\mathrm{S}\left(\mathrm{f,e,d;W,M,B,B}'\right) \quad (4)$$

4 非线性参数学习

4.1 基本学习框架

在对数线性模型的参数训练中最常使用的[Och, 2003]提出的最小错误率训练算法（Minimum Error Rate Training，简写为 MERT）通过逐方向线性探查的方法搜索最优的参数以优化 BLEU 值，因而该方法只能用于线性组合关系的模型参数学习而难以直接使用到非线性参数学习中。相关文献中使用了一些其他的线性模型训练算法，比如[Watanabe et al., 2007; Chiang et al., 2008]采用的 Margin Infused Relaxed Algorithm(MIRA), Hopkins and

May, 2011 提出的 Pairwise Ranking Optimization(PRO)等等，通过迭代的优化参数，使得输出的 1-best 的损失最小，则可以为非线性参数学习提供一些可借鉴之处。

本文借鉴 PRO 中的思想，将训练问题看作假设对之间的排序（pairwise ranking）问题处理，当在给定的源语言句子f的 N-best 中存在着两个候选译文$< e^*, d^* >$和$< e', d' >$，如果知道e^*的 BLEU 值大于e'，那么我们期望权重向量能够使得e^*的模型得分值大于e'，表示为如下的公式：

$$\mathrm{BLEU}(e^*) > \mathrm{BLEU}(e') \Leftrightarrow \mathrm{S}_w(e^*) > \mathrm{S}_w(e') \tag{5}$$

本文将$< e^*, e' >$这样的候选译文对称为f的偏好对（preference pair）。本文后续内容中的非线性参数学习的输入即为若干这样的偏好对。

由于使用偏好对作为输入，传统的基于 N-best 的 BLEU 得分计算不再可行，我们需要通过设计一定的方法来评估给定模型在偏好对上的表现，以此评估当前系统的性能，我们把这样的评估方法称为目标函数。另一方面，由于翻译过程中解码得到的候选翻译一般表现为一个 N-best 列表，为得到学习所需的偏好对，我们需要对 N-best 候选翻译列表进行采样处理。此外，不同的优化算法也可能对系统性能产生影响。因此，下文中，本文分目标函数、采样方法和优化算法三个部分对我们的非线性参数学习进行说明。

4.2 目标函数

4.2.1 最大间隔

对于对间排序问题（pairwise ranking），已经有很多相关的工作，大部分工作将其视为分类问题，因此本文利用最大间隔的思想，将$< e^*, e' >$这样的偏好对分为正例，而将$< e', e^* >$这样的偏好对分为负例，并且以最大的间隔分类，确保其分类的置信度，因此定义了如下的目标函数：

$$\sum_{f} \sum_{e^*, d^*, e', d'} \delta(f, e^*, d^*, e', d'; \theta)$$

$$\delta(\cdot) = Max\{S(f, e', d'; \theta) - S(f, e^*, d^*; \theta) + 1), 0\} \tag{6}$$

其中θ为模型参数，$\theta = \{\mathrm{W}, \mathrm{M}, \mathrm{B}, \mathrm{B}'\}$，即公式（3）中的$\mathrm{W}, \mathrm{M}, \mathrm{B}, \mathrm{B}'$权重的集合。$\delta$即hinge loss的表现形式。

4.2.2 排序损失

由于将训练参数当作排序问题来处理，那么也可以采用排序相关的目标函数，仍然遵循模型得分高的样本，期望其 BLEU 值高。本文由此采用了如下排序相关的目标函数（Rank）：

$$\sum_{e_i, e_j} (S(f, e_i, d_i; \theta) - S(f, e_j, d_j; \theta))(BLEU(e_j) - BLEU(e_i)) \tag{7}$$

其中$\mathrm{e_i}, \mathrm{e_j}$为源语言语句对应的候选翻译集合中的任意两个候选翻译，$\mathrm{S}(\cdot)$表示候选翻译对应的模型得分。

该模型的直观含义是当$\mathrm{e_i}$的模型得分高于$\mathrm{e_j}$的模型得分时，如果$\mathrm{e_i}$的 BLEU 值小于$\mathrm{e_j}$的 BLEU 值，那么损失函数的将会增加，如果模型得分和 BLEU 值一致，将会得到奖励。该模型与最大间隔模型的区别在于，该模型加入了 BLEU 差值对损失函数的影响，如果 BLEU 值相差较大的偏好对被模型分错了将会导致较大的损失。因此，该模型实际偏向于 BLEU 差值较大的偏好对。

4.2.3 采样方法

在目标函数中，需要获取给定的源语言句子 f的候选译文的偏好对，当 N-best 非常大的时候，选取所有的候选译文对是不可取的，同时选取所有的候选译文对也会导致引入过多的噪音，因此，需要采用一定的采样策略从所有可能的偏好对中选取一部分进行训练。基于部分现有工作，本文总结并设计了如下几种采样策略。

4.3.1 PRO 采样 （PROS）

遵循 PRO 中提到的采样方法（PRO Sampling ，简写为 PROS），对于第 i 个源语言句子，反复的从其 N-best 中同分布的选取一个候选译文对$< e_j, e_k >$，并且以$\alpha_i(\left|\mathrm{BLEU}(e_j) - \mathrm{BLEU}(e_k)\right|)$的概率采样，知道采到规定数目的样本。其中$\alpha_i(\cdot)$如下：

$$\alpha_i(n) = \mathrm{sig}(n, \bar{g}_i) \tag{8}$$

其中$\bar{g}_i$为第 i 个源语言句子 N-best 中所有候选译文对 BLEU 差值的平均，$\mathrm{sig}(n, m)$表示以 m 为中心的 sigmoid 函数在 n 点的取值。

该采样方法根据不同的源语言语句的候选译文的 BLEU 差值平均作为参照点，在采样过程中，期望大于 BLEU 差值大于平均差值的偏好对更容易被采样，倾向于使用差别比较明显的样本进行训练。

4.3.2 基于 Oracle 的采样 （ORS）

根据[Taro Watanabe, 2012]中所提到的，根据 BLEU 值得到 N-best 中的 Oracle，并以 Oracle 和剩余的 N-best 组成偏好对（Oracle-with-Rest Sampling，简写为 ORS）。

该采样策略的出发点是由于，翻译系统最终的输出仅仅是 1-best，那么只需要将最好的与剩下的候选译文区分出来，那么便能够得到最好的翻译结果。

4.3.3 基于 1-best 的采样 （MBRS）

选择模型得分最高的候选译文与剩下的候选译文中 BLEU 得分低于其的候选译文组成偏好对（Model-best-with-Rest Sampling，简写为 MBRS）。

该采样策略在第三种采样策略的基础上进行了弱化，期望通过这样的样本将实际上更为优质却被模型低估的候选译文与当前模型最好的候选译文区分出来。

4.3.4 基于比例的采样 （PTS）

按照 BLEU 值对 N-best 排序，其中最高的 10%与中间的 80%以及最低的 10%之间组成偏好对（Percentage based Sampling，简写为 PTS）。

该采样策略进一步模糊了所有候选译文之间的全序关系，仅仅要求区分出，好的，中等的和坏的。从而减轻了分类难度，期望通过较高的分类精度，得到更好的排序结果。

4.3 优化算法

由于采用非线性模型，拥有比较多的参数，在优化算法方面，一般会采用梯度下降的方法，由于本文定义的最大间隔的目标函数是一个 min-max 类型的目标函数，是无法采用梯度方法的，那么次梯度方法便是一个很好的替代选择，本文采用[Liu et al., 2013]所采用的共轭次梯度算法（conjugate subgradient，简写为 CG），来进行参数优化：

1. 首先解码获得 dev 集合中所有句子的 N-best
2. 对每个源语言句子进行采样，获得偏好对
3. 利用 dev 集合中采到的所有样本，计算损失函数及其梯度，利用 CG 算法迭代训练直至达到最大迭代次数
4. 获得新的权重，重排序实验中利用该权重进行重排序，解码训练中利用该权重重

新解码并重复上述步骤，直至达到最大迭代次数。

5 实验和结果

5.1 实验配置

本文在中英任务上进行实验，训练集为LDC2002E18、LDC2003E14、LDC2004E12、LDC2007T09等LDC提供的中英平行语料（共计约860万句对），开发集为NIST03，测试集为NIST02，NIST04和NIST05。本文的基线系统采用组内开发的基于层次短语的翻译系统，采用如下特征：正向翻译概率，正向词汇化概率，反向翻译概率，反向词汇化概率，词计数，短语计数，规则计数，粘合规则计数，未登录词计数，翻空计数以及语言模型，并采用最小错误率算法进行参数训练。上述特征在非线性模型中作为输入层输入，本文中的非线性模型采用单隐层结构，隐层节点数设置为20。语言模型采用组内开发的训练工具在GigaWord上训练得到5-gram，利用GiZA++在双语语料上获得词对齐并依此获得翻译规则。本文实验的评价标准采用BLEU4 [Papineni et al., 2002]，本文中所用到的句子级的BLEU值都是使用BLEU+1[Liang et al.,2006]。

5.2 实验结果

本文的实验主要分为重排序实验和翻译系统参数训练实验。在验证目标函数和采样策略时采用重排序实验，重排序实验排除了解码过程中其他因素的影响，仅仅以 N-best 列表的形式作为实验的输入，对于验证目标函数和采样策略可以更加快速并且明显地得到结论。本文的主要目的是研究非线性参数学习在机器翻译中的使用，设计翻译系统的参数训练实验，可以通过与对数线性模型进行比较进一步验证非线性参数学习在统计机器翻译中的可行性。下面实验中的 BLEU 值取三次独立训练结果的平均值。

5.2.1 目标函数实验

首先，对于本文提到的两种目标函数，在重排序实验上进行实验验证，采用如下特征：正向翻译概率，正向词汇化概率，反向翻译概率，反向词汇化概率，词计数，短语计数，规则计数以及语言模型：

表 1 目标函数比较

20-best	dev-03	test02	test04	test05	avg
Max-Margin	40.61%	38.97%	38.75%	38.46%	38.73%
Rank	38.72%	38.64%	38.31%	38.32%	38.42%

该实验在 20-best 上进行，采用 PROS 方法进行采样，可以发现采用 Max-Margin 的目标函数要明显优于 Rank 的目标函数，并且在开发集上高了 1.89，在测试集上也高了 0.31。这是由于 Max-Margin 目标函数在考虑 BLEU 值大的候选译文模型得分也大的基础上，兼顾了分类准确性的置信度，正负例之间需以最大间隔分开，而 Rank 目标函数仅仅考虑了模型得分和 BLEU 值之间不一致会使损失增加。下面的实验中的目标函数将采用最大间隔的目标函数。

5.2.2 采样策略实验

对于不同的采样策略进行重排序实验（采用 Max-Margin 目标函数）：

表 2 采样策略比较

20-best	dev-03	test02	test04	test05	avg

PROS	40.61%	38.97%	38.75%	38.46%	38.73%
ORS	39.15%	39.04%	38.75%	37.77%	38.52%
OWS	39.56%	38.74%	39.02%	38.33%	38.70%
MBRS	39.70%	38.63%	38.81%	38.39%	38.61%
PS	39.60%	38.76%	38.87%	38.36%	38.66%

从实验结果看，在本文中的不同的采样方法，仍然是 PROS 的采样方法最好，PROS 的采样较为均匀，并且依据每一个源语言句子的候选译文集合的 BLEU 差值的平均值作为指导可以采到更加具有效果的样本。

5.2.3 正则化项实验

从上述结果可以看出，在开发集上已经可以取得较为明显的效果，然而在测试集上并未能取得性能的提高，考虑是否是发生了过拟合，在原来的目标函数上加入正则化项。

表 3 加入正则化项的实验

20-best	dev-03	test02	test04	test05	avg
Baseline	39.37%	39.08%	39.11%	38.37%	38.85%
PRO + 0.01L2	40.44%	38.67%	38.72%	38.17%	38.52%
PRO + 0.05L2	40.34%	38.84%	38.90%	38.15%	38.63%
PRO + 0.08L2	40.27%	38.63%	38.91%	38.45%	38.66%
PRO + 0.1L2	40.57%	38.97%	38.86%	38.57%	38.80%
PRO + 0.12L2	40.41%	38.69%	38.73%	38.46%	38.63%
PRO + 0.15L2	40.41%	38.98%	38.83%	38.81%	**38.87%**
PRO + 0.2L2	40.60%	38.69%	38.84%	38.46%	38.66%

从实验结果可以看到，选择一个合适的正则化项，可以取得一定的效果。并且在选择和合适的目标函数和采样策略后，可以在开发集上提高 1.04，并且在在测试集 NIST05 上取得 0.44 的提升，在测试集上平均结果与基线系统结果大体相当。

5.2.4 翻译系统参数训练

采用如下特征：正向翻译概率，正向词汇化概率，反向翻译概率，反向词汇化概率，词计数，短语计数，规则计数，粘合规则计数，未登录词计数，翻空计数以及语言模型。

由于当前的非线性模型较为简单，考虑采用较为复杂的网络结构，借鉴自编码网络的思想，在翻译系统参数训练中补充了一组目标函数上加入了重构误差项（reconstruct error，简写为 rec）的设置进行实验，目标函数变为如下格式：

$$\sum_{f}\sum_{e^*,d^*,e',d'}\delta(f,e^*,d^*,e',d';\theta)+\frac{1}{2}(|W'\sigma(\mathrm{M}\cdot\mathrm{h}(f,e^*,d^*)+\mathrm{B})-h(f,e^*,d^*)|^2 + |W'\sigma(\mathrm{M}\cdot\mathrm{h}(f,e',d')+\mathrm{B})-h(f,e',d')|^2)$$

$$\delta(\cdot)=Max\{S(f,e',d';\theta)-S(f,e^*,d^*;\theta)+1),0\} \quad (9)$$

其中W'为隐层到重构层的权重矩阵。

表 4 非线性模型的翻译系统参数训练实验

	dev-03	test02	test04	test05	avg

Baseline	39.25%	39.07%	38.81%	38.01%	38.63%
PROS-CG-linear	39.46%	38.61%	38.54%	38.40%	38.51%
PROS-CG-nonlinear	36.51%	36.45%	36.05%	35.67%	36.06%
PROS-CG-nonlinear-rec	37.54%	38.94%	37.79%	36.74%	37.83%

从实验中可以看出，线性模型下 CG 算法进行训练与基线系统是可比的，这说明了 CG 算法的有效性。在翻译系统的解码实验中采用更为复杂的神经网络模型要优于简单的单隐层神经网络，虽然非线性模型未能取代对数线性模型，但是该实验验证了非线性模型在统计机器翻译系统中仍然具有提升性能的空间。

6 总结

本文实现了一套非线性参数训练的框架，同时实现了 re-rank 的相应框架，在采样方法和目标函数的选择上进行了验证实验，说明了最大间隔目标函数和 PROS 的采样方法更加适合非线性模型。非线性模型部分取得了与 MERT 接近的结果。在部分实验中，在 dev 集合上取得了明显的提升，但并未能在测试集上显著超过基线系统。可能存在的原因：当前的网络较为简单，输入层的稠密特征（dense feature）较少，无法学到具有较强区别能力和泛化能力的特征以取代原有的特征；通过观察训练过程，发现非线性模型的训练波动比较大，可能陷入局部最优。下一阶段将加入稀疏化特征，与当前的 dense 特征一起利用非线性模型进行学习，利用深度的网络来重新表达和组合特征，对非线性模型预训练，使得学习较为稳定。

7 参 考 文 献

Yamada, K. and K. Knight. 2001. A syntax-based statistical translation model. In Proceedings of ACL.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Proc. Of HLT-NAACL. ACL.

Liu, Yang and Liu Qun and Lin Shouxun. 2006. Tree-to-String Alignment Template for Statistical Machine Translation. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics

Chiang, David. 2007. Hierarchical phrase-based translation. Comput. Linguist., 33(2):201–228, June.

Shen, Libin and Xu Jinxi and Weischedel Ralph. 2010. String-to-dependency Statistical Machine Translation. Volume 36, pages 295–302. Comput. Linguist.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 295–302, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maskey, S and Zhou B. 2012. Unsupervised Deep Belief Features for Speech Translation[C] INTERSPEECH.

Lu, Shixiang and Chen, Zhenbiao and Xu, Bo. 2014. Learning New Semi-Supervised Deep Auto-encoder Features for Statistical Machine Translation. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Zhao, B, Tam Y C, Zheng J. 2014. AN AUTOENCODER WITH BILINGUAL SPARSE FEATURES FOR IMPROVED STATISTICAL MACHINE TRANSLATION[J]. ICASSP.

Duh, Kevin and Katrin Kirchhoff. 2008. Beyond loglinear models: Boosted minimum error rate training for

n-best re-ranking. In Proceedings of ACL-08: HLT, Short Papers, pages 37–40, Columbus, Ohio, June. Association for Computational Linguistics.

Liu, Lemao and Taro Watanabe and Eiichiro Sumita and Zhao Tiejun. 2013. Additive Neural Networks for Statistical Machine Translation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

Watanabe, Taro, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLPCoNLL), pages 764–773, Prague, Czech Republic, June. Association for Computational Linguistics.

Chiang, David, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In Proc. of EMNLP. ACL.

Hopkins, Mark and Jonathan May. 2011. Tuning as ranking. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1352–1362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Watanabe, Taro. 2012. Optimized Online Rank Learning for Machine Translation. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Papineni, Kishore, Salim Roukos, ToddWard, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Liang, Percy, Alexandre Bouchard-Cˆot´e, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 761–768, Sydney, Australia, July. Association for Computational Linguistics.