Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation

Maoxi Li, Aiwen Jiang, Mingwen Wang
School of Computer Information Engineering, Jiangxi Normal University,
Nanchang, China, 330022
moseli@yeah.net, aiwen.jiang@ia.ac.cn, mwwang@jxnu.edu.cn

Abstract

The listwise approach to learning to rank has been applied successfully to information retrieval. However, it has not drawn much attention in research on the automatic evaluation of machine translation. In this paper, we present the listwise approach to learning to rank for the automatic evaluation of machine translation. Unlike previous automatic metrics that give absolute scores to translation outputs, our approach directly ranks the translation outputs relative to each other using features extracted from the translation outputs. Two representative listwise approaches, ListNet and ListMLE, are applied to automatic evaluation of machine translation. When evaluated using the dataset of the WMT 2012 Metrics task, the proposed approach achieves higher segment-level correlation with human judgments than the pairwise approach, RankNet, and with all the other metrics that were evaluated during the workshop, and it achieves honorably a comparable system-level correlation with the performance of most competitors.

1 Introduction

Human assessment of fluency and adequacy is one of the earliest methods widely used by researchers to manually evaluate the quality of machine translation. The fluency score indicates how the translation output sounds to a native speaker of the target language; while the adequacy score indicates how much of the meaning expressed in the source text is also expressed in the translation output. A five-point scale is often used to score fluency and adequacy (Callison-Burch et al., 2007; Paul, 2008). Translation outputs are seldom evaluated by skilled human judges due to high evaluation costs and the time consumed. The machine translation evaluation campaign employs human assessment to rank translation systems relative to each other. The human judgment collection released by the evaluation campaign is often defined as the gold standard for validating automatic evaluation metrics.

In recent years, many state-of-the-art automatic evaluation metrics, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), MAXSIM (Chan and Ng, 2008), TESLA (Liu et al., 2010), PORT (Chen et al., 2012) and others have been proposed as automated understudies to human assessment; they quickly calculate similar absolute scores between translation outputs and human references. They are heuristic metrics and do not utilise supervised learning to model human judgments directly (Song and Cohn, 2011).

Let us take human judgment of the five-point fluency and adequacy scores as the class label of translation quality. Then, the issue of automatic evaluation of machine translation can be converted to the question of classification or regression. Corston-Oliver et al. propose an approach that constructs decision trees to distinguish machine translation from human translation using linguistic features (Corston-Oliver et al., 2001), Kulesza and Shieber employ support vector machines to improve segment-level machine translation evaluation (Kulesza and Shieber, 2004). While Albrecht and Hwa introduce a regression approach that directly optimises the metrics to predict the translation quality of translation output according to human adequacy and fluency judgments (Albrecht and Hwa, 2008), Specia and Gimenez combine confidence estimation and

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 51–59.

reference-based metrics together in a regression framework to measure the segment-level machine translation quality (Specia and Giménez, 2010).

In addition to human assessment of fluency and adequacy, another approach is to manually rank translation outputs from multiple machine translation systems according to their translation quality. The typical instructions for human ranking are as follows:

> *You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).*
> (Callison-Burch et al., 2012)

Compared with the human assessment of fluency and adequacy, human evaluation of ranking is more intuitive and consistent (Callison-Burch et al., 2008). Therefore, the approach of human assessment of fluency and adequacy has been abandoned by the WMT evaluation campaign since 2008. If human ranking judgment is adopted as the gold standard for automatic evaluation metrics, the task is more like a "learning to rank" problem than classification or regression.

Duh argues for the ranking approach for the automatic evaluation of machine translation (Duh, 2008) and employs a pairwise approach to learning to rank, Rank SVM (Joachims et al., 2009), to directly rank the translation outputs using *n*-gram precision features, such as BLEU metrics. Song and Cohn use a pointwise approach to learning to rank, SVM$^{\text{light}}$ (Joachims, 1999), to train a new metric, ROSE, with different kernel functions using simple features (Song and Cohn, 2011).

In contrast to previous work that applied a pointwise or pairwise approach to learning to rank for the automatic evaluation of machine translation, we utilise a "listwise" approach to learning to rank that directly rank translation outputs. It is shown that the listwise approach performs better than the pointwise and pairwise approaches in information retrieval, for which the central problem is ranking. Two representative listwise approaches, ListNet (Cao et al., 2007) and ListMLE (Xia et al., 2008), are introduced and applied to the automatic evaluation of machine translation. We train the ranking model on the dataset of the WMT 2011 Metrics task using the n-gram precision, language model, and bidirectional translation probability features, and we test the segment-level and system-level correlations with human judgments on the dataset of the WMT 2012 Metrics task. The experimental results show that the new approach is promising.

2 Formulation

Learning to rank is a type of supervised machine learning problem in which the goal is to automatically construct a ranking model from training data, and the ranking model is used to rank, i.e., produce a permutation of items in the test set. Learning to rank has been successfully applied to information retrieval, natural language processing, and data mining.

The state-of-the-art learning to rank methods can be classified into three categories based on their input representation and loss function (Liu, 2009; Li, 2011). The pointwise approach attempts to solve the problem of ranking using existing learning methods, such as classification, regression, etc. Therefore, the group structure of ranking is ignored. The pairwise approach addresses the ranking problem by pairwise comparison, and many pairwise ranking algorithms have been proposed, such as RankNet (Burges et al., 2005) and Rank SVM. The listwise approach solves the ranking problem straightforwardly by taking the total ranking lists as instances in both training and testing. Thus, the group structure of ranking is maintained. However, some of the listwise approaches, SVM$^{\text{map}}$ (Yue et al., 2007) and SoftRank (Taylor et al., 2008) and so on, are designed to directly optimise the evaluation measures of information retrieval, including MAP, and NDCG et al.; this hampers their application in other research areas. In this work, we restrict the listwise approaches to those whose optimisation aim is not explicitly related to the evaluation measures used in information retrieval.

2.1 The Listwise approach

In this subsection, we provide a formal description of the listwise approach to learning to rank for the automatic evaluation of machine translation, as well as two representative listwise approaches, ListNet (Cao et al., 2007) and ListMLE (Xia et al., 2008), explored in our experiments in the following subsection.

Suppose that the test set for the machine translation task consists of *m* source segments and their corresponding references. Each source segment to be translated, $t^{(i)}$, and its human reference, $R^{(i)}$, can be represented by $T^{(i)} = (t^{(i)}, R^{(i)})$ ($i = 1, \dots, m$). If there are $n^{(i)}$ machine translation systems translating the source segment $t^{(i)}$, a set of translation hypotheses, $H^{(i)} = (h_1^{(i)}, \dots, h_{n^{(i)}}^{(i)})$, can be generated. Here, $h_j^{(i)}$ denotes to the j-th translation hypothesis associated with the source segment $t^{(i)}$ ($j = 1, \dots, n^{(\mathrm{i})}$). Assume that the translation hypotheses have been ranked relative to each other according to their translation quality by human annotators, and the human ranking list is $y^{(\mathrm{i})} = (y_1^{(\mathrm{i})}, \dots, y_{n^{(i)}}^{(i)})$; ties are allowed in the human ranking. The human ranking list is taken as the gold standard for automatic metrics. We denote this set as $S = \{(T^{(i)}, H^{(i)}), y^{(i)}\}_{i=1}^{m}$, and it can be used as the training set for a listwise approach to learning to rank.

A feature vector $x_j^{(i)}$ is created with which we hope to estimate the translation quality of the translation hypotheses $h_j^{(i)}$ given the source segment $t^{(i)}$, ($i = 1, \dots, m$; $j = 1, \dots, n^{(\mathrm{i})}$), such as n-gram precision between a translation hypothesis and human references. For a given source segment $t^{(i)}$, if we denote all its translation hypotheses $H^{(i)} = (h_1^{(i)}, \dots, h_{n^{(i)}}^{(i)})$ to feature vectors, a list of feature vectors is formed, $x^{(i)} = (x_1^{(i)}, \dots, x_{n^{(i)}}^{(i)})$. We can take the list of feature vectors $x^{(i)}$ and the corresponding human ranking list $y^{(i)}$ as an instance for a listwise approach to learning to rank. Therefore, the original training set S is further represented by $S' = \{x^{(i)}, y^{(i)}\}_{i=1}^{m}$.

Our aim is to train a ranking function f that can assign a score $f(x_j^{(i)})$ to a feature vector $x_j^{(i)}$ (or the triple of source segment $t^{(i)}$, translation references $R^{(i)}$, and translation hypothesis $h_j^{(i)}$), while for the list of feature vectors $x^{(i)}$, ranking function f outputs a sequence of values $z^{(i)} = (f(x_1^{(i)}), \cdots, f(x_{n^{(i)}}^{(i)}))$. The loss between the predicated ranking list $z^{(i)}$ and the gold standard ranking list $y^{(i)}$ (the human ranking list) is represented by $Loss(z^{(\mathrm{i})}, y^{(\mathrm{i})})$. The objective of training for the listwise approach to learning to rank is to minimise the sum of losses $\sum_{i=1}^{m} Loss(y^{(i)}, z^{(i)})$ over the training data.

Given the ranking scores of the translation hypotheses by the ranking function f, the Plackett-Luce model (Guiver and Snelson, 2009) defines a permutation probability for each possible permutation of the translation hypotheses. Let π denote a permutation (ranking list) of the translation hypotheses, $\pi^{-1}(i)$ denote the translation hypothesis in the i-th rank in π, and $s = \{s_1, s_2, \dots, s_n\}$ denote the ranking scores of the translation hypotheses. The probability of permutation π based on scores s is defined as follows:

$$P_s(\pi) = \prod_{j=1}^{n} \frac{s_{\pi^{-1}(j)}}{\sum_{k=j}^{n} s_{\pi^{-1}(k)}} \qquad (1)$$

Suppose that the ranking function f is a Neural Network model with parameter ω; it can assign a score $f_\omega(x_j^{(i)})$ to a feature vector $x_j^{(i)}$. Given a list of feature vectors $x^{(i)}$, the ranking function f_ω outputs a sequence of values $z^{(i)}(f_\omega) = (f_w(x_1^{(i)}), \cdots, f_\omega(x_{n^{(i)}}^{(i)}))$. Thus, the permutation probability of the translation hypotheses is calculated as:

$$P_{z^{(i)}(f_w)}(x^{(i)}) = \prod_{j=1}^{n^{(i)}} \frac{\exp(f_w(x_{z(j)}^{(i)}))}{\sum_{k=j}^{n^{(i)}} \exp(f_w(x_{z(k)}^{(i)}))} \qquad (2)$$

where $z(j)$, $z(k)$ denote to the j-th and k-th translation hypotheses ranked according to their translation quality.

2.2 The ListNet approach

ListNet defines the loss function using the KL divergence between the probability distribution for the ranking model and that for the human ranking list (Cao et al., 2007).

$$Loss(y^{(i)}, z^{(i)}(f_w)) = -\sum_{j=1}^{n^{(i)}} P_{y^{(i)}}(x_j^{(i)}) \log(P_{z^{(i)}(f_w)}(x_j^{(i)})) \qquad (3)$$

The Gradient Descent algorithm is utilised to tune the parameter ω of the Neural Network.

$$\begin{aligned} \Delta w = {} & -\sum_{j=1}^{n^{(i)}} P_{y^{(i)}}(x_j^{(i)}) \frac{\partial f_w(x_j^{(i)})}{\partial w} + \\ & + \frac{1}{\sum_{j=1}^{n^{(i)}} \exp(f_w(x_j^{(i)}))} \sum_{j=1}^{n^{(i)}} \exp(f_w(x_j^{(i)})) \frac{\partial f_w(x_j^{(i)})}{\partial w} \end{aligned} \qquad (4)$$

2.3 The ListMLE approach

ListMLE is a variation of ListNet that employs the negative log likelihood of the permutation probability of human ranking as the loss function (Xia et al., 2008).

$$L(y, z(f_w)) = -\sum_{i=1}^{m} \log \prod_{j=1}^{n^{(i)}} \frac{\exp(f_w(x_{y(j)}^{(i)}))}{\sum_{k=j}^{n^{(i)}} \exp(f_w(x_{y(k)}^{(i)}))} \qquad (5)$$

One can deduce that the loss function of ListMLE has some nice statistical properties, including soundness and convexity.

When training, we minimise the sum of the losses with respect to all the training instances;

the Gradient Descent algorithm is employed to tune the parameter ω of the Neural Network.

$$\Delta w = \sum_{j=1}^{n^{(i)}} \left(\frac{\sum_{k=j}^{n^{(i)}} x^{(i)}_{y(k)} \exp(f_w(x^{(i)}_{y(k)}))}{\sum_{k=j}^{n^{(i)}} \exp(f_w(x^{(i)}_{y(k)}))} - x^{(i)}_{y(k)} \right) \quad (6)$$

3 Segment-level Scoring and System-level Scoring

Once we have the Neural Network parameter *ω*, we can establish a ranking model that can output a score *f*(*x*) for a feature vector *x*. The feature vector of the translation hypothesis in the test set is extracted from the triple of source segment, translation references, and itself.

Because the ranking function can only output segment scores, system-level scores cannot be calculated directly. We assign scores to the systems based on the percentages of their translation hypotheses that are better than or equal to the translation hypotheses of any other machine translation systems by pairwise comparison. Note that this algorithm is similar to the approach used to compute the published final system rankings by the WMT metrics task (Callison-Burch et al., 2012); the difference lies in human references being excluded in the pairwise comparison in our algorithm.

4 Feature Set

The selection of features used in the listwise approach to learning to rank for the automatic evaluation of machine translation is motivated by BLEU metrics and the phrased-based statistical translation model. For each translation output, we automatically extract nine features to form a feature vector; the complete feature set is shown in Table 1. The features are classified into three categories, including *n*-gram matching precision between the translation output and human references, language model probability of the translation output, and approximate bidirectional translation probabilities.

4.1 *N*-gram matching precision

The *n*-gram matching precision between the translation output and multiple human references is exploited by the automatic metrics BLEU and NIST, which can measure the quality of the translation output to a certain extent. We use the open source script "mteval-v13a.pl"[1] to calculate the *n*-gram matching precision of the translation output. To avoid overflowing when calculating the logarithmic segment-level *n*-gram matching precision, the halves-smoothed algorithm is used. In addition, a brevity penalty is also introduced to penalise short translation outputs such as in the BLEU metrics.

ID	Description
1-4	*n*-gram precision, *n*=1..4
5	brevity penalty
6	language model probability *P*(*e*)
7	length penalty
8	the approximate conditional probability *P* (*e* | *f*)
9	the approximate conditional probability *P* (*f* | *e*)

Table 1: Feature set. Features 1-5 can be combined to form the smoothing segment-level BLEU score. Features 6-7 are the language model probability of the translation output, Features 8-9 are approximate bidirectional translation probabilities.

4.2 Language model probability

The statistical language model probability of the translation output quantitatively analyses the likelihood that the translation output is generated from the monolingual training data, which can measure the fluency of the translation output. We combine the target language side of the training corpus for statistical machine translation with human references to form the monolingual training data, train a 4-gram language model on the data, and compute the language model probability of the translation output of the model. In addition, we introduce length features of the translation output to normalise the language model probability.

4.3 Approximate bidirectional translation probabilities

To indicate how much of the meaning expressed in the source segment is also expressed in the translation output, namely the translation adequacy, we use the following formula to approximately calculate the conditional probability of translation output *e* given source sentence *f* in the absence of word alignments between them:

$$P(e \mid f) \approx \prod_{i=1}^{n} \sum_{j=1}^{m} p(e_i \mid f_j) \quad (7)$$

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13a.pl

where $p(e_i \mid f_j)$ denotes the lexical translation probability of target word e_i in the translation output given the source word f_j in the source segment. The lexical translation probability can be estimated using the IBM model (Brown et al., 1993) with the bilingual training corpus for statistical machine translation.

To further measure the translation adequacy given translation output e, we introduce the inverted conditional probability $P(f \mid e)$ whose source sentence was f.

5 Experimental Results

To test the performance of the listwise approach to learning to rank for the automatic evaluation of machine translation, we conducted experiments on datasets released by the Metrics tasks of WMT 2011 and WMT 2012. The dataset of WMT 2011 was used as a training set to optimise the ranking model of the listwise approach to learning to rank, while the WMT 2012 dataset was used to test the correlation with human judgments.

The datasets of WMT 2011 and WMT 2012 metrics tasks both included evaluation of the system outputs of 8 translation tasks, namely Czech-English (CZ-EN), German-English (DE-EN), Spanish-English (ES-EN), and French-English (FR-EN), and the opposite translation directions. The WMT 2011 Metrics task, in addition to containing the translation outputs of individual systems, also contained the translation outputs from the combination systems (Callison-Burch et al., 2011). For simplicity, we only used the dataset of individual systems to optimise the ranking model. Due to space limitations, we do not present the segment-level and system-level correlation with human judgments for the Metrics task of WMT 2011. The segment-level correlation and system-level correlation between human judgment and automatic metrics are calculated with the scripts[2] officially released by the WMT evaluation campaign.

In addition to comparing the listwise approaches with the relevant metrics, we also compared the listwise approaches with the pairwise approach, RankNet (Burges et al., 2005), using the same features and the same training set as the listwise approach.

5.1 Segment-level correlation

The Kendall's tau rank correlation coefficient is used to compute the segment-level correlation between human judgments and automatic metrics. We calculate segment-level correlation as follows:

$$\tau = \frac{\textit{num concordant pairs} - \textit{num disconcordant pairs}}{\textit{total pairs}}$$

Ties are excluded in the pairwise comparison. The possible values for τ range between 1 and -1; the higher the value for τ, the more closely the automatic metrics correlated with human judgments.

	FR-EN (11594 PAIRS)	DE-EN (11934 PAIRS)	ES-EN (9796 PAIRS)	CZ-EN (11021 PAIRS)	AVERAGE
Segment-level correlation for translations into English					
ListNet	**0.26**	0.25	**0.30**	0.21	**0.26**
ListMLE	0.20	**0.28**	**0.30**	**0.25**	**0.26**
spede07_pP	**0.26**	**0.28**	0.26	0.21	0.25
Meteor	0.25	0.27	0.25	0.21	0.25
RankNet	0.20	0.24	0.24	0.22	0.23
AMBER	0.24	0.25	0.23	0.19	0.23
TerrorCat	0.18	0.19	0.18	0.19	0.19
SIMPBLEU	0.19	0.17	0.19	0.13	0.17
XEnErrCats	0.17	0.18	0.18	0.13	0.17
posF	0.16	0.18	0.15	0.12	0.15
WordBlockEC	0.15	0.16	0.17	0.13	0.15
BlockErrCats	0.07	0.08	0.08	0.06	0.07
SAGAN_STS	n/a	n/a	0.21	0.20	n/a

Table 2: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for metrics scoring of translations into English on WMT 2012, ordered by average absolute value.

Table 2 and Table 3 summarise the segment-level correlation between the listwise approaches to learning to rank and human judgments of the WMT 2012 Metrics task, along with the segment-level correlation of the participated metrics

[2] http://www.statmt.org/wmt12/results.html

released by the official report (Callison-Burch et al., 2012). Here, the names of the proposed metrics are abbreviated and italicised.

	EN-FR (11562 PAIRS)	EN-DE (14553 PAIRS)	EN-ES (11834 PAIRS)	EN-CZ (18805 PAIRS)	AVERAGE
Segment-level correlation for translations out of English					
ListNet	0.25	**0.19**	**0.26**	**0.19**	**0.22**
ListMLE	0.25	0.18	0.24	0.15	0.21
RankNet	0.25	0.16	0.21	0.16	0.20
Meteor	**0.26**	0.18	0.21	0.16	0.20
AMBER	0.23	0.17	0.22	0.15	0.19
TerrorCat	0.18	**0.19**	0.18	0.18	0.18
SIMPBLEU	0.20	0.13	0.18	0.1	0.15
EnXErrCats	0.20	0.11	0.17	0.09	0.14
posF	0.15	0.13	0.15	0.13	0.14
WordBlockEC	0.19	0.1	0.17	0.1	0.14
BlockErrCats	0.13	0.04	0.12	0.01	0.08

Table 3: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for metrics scoring translations out-of-English on WMT 2012, ordered by average absolute correlation value.

As shown in Table 2, we compared the "*ListNet*" and "*ListMLE*" approaches with the "*RankNet*" approach and the associated metrics for scoring translations into English. The "*ListNet*" and "*ListMLE*" approach achieved the best average correlations with human judgments, and outperformed the best associated metrics, "spede07_pP (Wang and Manning, 2012)", 1%. The pairwise approach to learning to rank, "*RankNet*", is 3% lower than the "*ListNet*" and "*ListMLE*" approaches, but still outperformed most of the associated metrics. The "*ListNet*" approach also achieved the best average segment-level correlation with human judgments on out-of-English tasks as shown in Table 3, and outperformed the best associated metrics, "Meteor (Denkowski and Lavie, 2011)", by 2%. Note that the "*ListNet*" approach had the best average performance both on into-English and on out-of-English tasks. However, the gap of segment-level correlation with human judgments on different translation tasks between the two listwise approaches is very small, approximately 6%. For the pairwise approach, "*RankNet*" also achieved a good segment-level correlation with human judgments, but the correlation was still lower than the listwise approaches. It was confirmed that the features we used were efficient to measure the translation quality.

5.2 System-level correlation

	CZ-EN - 6 SYSTEMS	DE-EN - 16 SYSTEMS	ES-EN - 12 SYSTEMS	FR-EN - 15 SYSTEMS	AVERAGE
System-level correlation for translations into English					
SEMPOS	**0.94**	0.92	**0.94**	0.80	**0.90**
ListMLE	**-0.94**	**-0.94**	-0.85	-0.78	0.88
AMBER	0.83	0.79	0.97	0.85	0.86
ListNet	-0.83	-0.86	-0.87	-0.82	0.84
Meteor	0.66	0.89	0.95	0.84	0.83
TerrorCat	0.71	0.76	0.97	**0.88**	0.83
RankNet	-0.89	-0.66	-0.92	-0.81	0.82
SIMPBLEU	0.89	0.70	0.89	0.82	0.82
TER	-0.89	-0.62	-0.92	-0.82	0.81
BLEU	0.89	0.67	0.87	0.81	0.81
posF	0.66	0.66	0.87	0.83	0.75
BlockErrCats	-0.64	-0.75	-0.88	-0.74	0.75
WordBlockEC	-0.66	-0.67	-0.85	-0.77	0.74
XEnErrCats	-0.66	-0.64	-0.87	-0.77	0.74
SAGAN_STS	0.66	n/a	0.91	n/a	n/a

Table 4: System-level Spearman's rank correlation of the automatic evaluation metrics with human judgments for metrics scoring translations into English on WMT 2012, ordered by average absolute value.

We measured the correlation between the automatic metrics and human judgments at the system-level using Spearman's rank correlation coefficient. The system-level correlations with human judgments of the listwise approaches, the pairwise approach, and the associated metrics on

WMT 2012 metrics task are summarised in Table 4 and Table 5.

	EN-CZ - 13 SYSTEMS	EN-DE - 15 SYSTEMS	EN-ES - 11 SYSTEMS	EN-FR - 15 SYSTEMS	AVERAGE
System-level correlation for translations out of English					
ListMLE	-0.78	**-0.81**	-0.36	-0.68	**0.66**
SIMPBLEU	**0.83**	0.46	0.42	**0.94**	**0.66**
BlockErrCats	-0.65	-0.53	-0.47	-0.93	0.64
EnXErrCats	-0.74	-0.38	-0.47	-0.93	0.63
posF	0.80	0.54	0.37	0.69	0.60
WordBlockEC	-0.71	-0.37	-0.47	-0.81	0.59
TerrorCat	0.65	0.48	**0.58**	0.53	0.56
ListNet	-0.73	-0.53	-0.32	-0.64	0.55
AMBER	0.71	0.25	0.50	0.75	0.55
TER	-0.69	-0.41	-0.45	-0.66	0.55
Meteor	0.73	0.18	0.45	0.82	0.54
BLEU	0.80	0.22	0.40	0.71	0.53
RankNet	-0.38	-0.57	-0.17	-0.64	0.44
SEMPOS	0.52	n/a	n/a	n/a	n/a

Table 5: System-level Spearman's rank correlation of the automatic evaluation metrics with human judgments for metrics scoring translations out of English on WMT 2012, ordered by average absolute value.

Because a lower rank value indicates better translation quality on human rank judgments, we trained the ranking model of the listwise approaches and the pairwise approach using the human ranking scores on the dataset of the WMT 2011 Metrics task. We achieved a lower score, signifying a better translation system. Thus, for Spearman's system-level rank correlation, the values of the listwise approaches and the pairwise approach are negative, such as the system-level correlation of TER metrics. To facilitate comparing the proposed approaches with the metrics we used, we converted the average negative value to absolute value.

As shown in Table 4 and Table 5, for scoring the into-English translation task, the "*ListMLE*" approach had the best average system-level correlation with human judgments among the learning to rank approaches. However, it was lower than the best associated metrics, "SEMPOS", of 2%, while for scoring the out-of-English translation task, the "*ListMLE*" approach and its associated metrics, "SIMPBLEU", tied for first place. Although the "*ListMLE*" approach achieved high average correlation with human judgments for the out-of-English task, the "*ListNet*" and "*RankNet*" approaches had lower system-level correlation with human judgments than most of the associated metrics. When we looked carefully, we found that the associated metrics "Meteor" and "AMBER" are in the same situation. That is, they had higher segment-level and system-level correlations for scoring translation into English but still had lower system-level correlations for scoring out-of-English translation. This phenomenon may be induced by the fluctuation of system-level correlation with human judgments on different translation directions for this task.

6 Discussion and Conclusion

In this paper, we introduce the listwise approach to learning to rank for the automatic evaluation of machine translation and explore two representative listwise approaches, ListNet and ListMLE. The experimental results suggested that the proposed approaches achieve the best segment-level correlation with human judgments and have a comparable system-level correlation with the associated metrics. It is confirmed that the listwise approach to learning to rank is promising for the automatic evaluation of machine translation.

There are several advantages of the listwise approach:

- Because the listwise approaches take the whole human ranking list as instances for training, they can maintain the group structure of the whole ranking list and predict the rank of translation outputs more precisely than the pairwise or pointwise approaches.
- The listwise approach can effectively help integrate many features into the automatic evaluation of machine translation. In this work, we only utilise *n*-gram matching precision between the translation output and human references, language model probability of the translation output, and approximate bidirectional translation probabilities. Future work includes inte-

grating syntactic and semantic linguistic features to further improve correlation with human judgments.

- The scoring values of the listwise approach are interpretable. Given a source segment, the values of segment-level scores of translation outputs indicate the translation quality relative to each other.

However, we should recognise that the listwise approach to learning to rank for automatic evaluation of machine translation cannot score a single machine translation system. It can only rank multiple translation systems relative to each other.

Acknowledgments

This research has been funded by the Natural Science Foundation of China under Grant No. 6120 3313, 6127 2212, and 6116 3006, and supported by the Natural Science Foundation of Jiangxi Provincial Department of Science and Technology of China under Grant No 20132BAB201038 and 20132BAB201030, and also supported by the Natural Science Foundation of Jiangxi Educational Committee of China under Grant No. GJJ12212.

References

Joshua S. Albrecht and Rebecca Hwa, 2008. *Regression for Machine Translation Evaluation at the Sentence Level*. Machine Translation, 22 (1-2). pages 1-27.

Satanjeev Banerjee and Alon Lavie, 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65-72, Ann Arbor.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer, 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, 19 (2). pages 263-311.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton and Greg Hullender, 2005. Learning to Rank Using Gradient Descent. *Proceedings of the 22nd international conference on Machine learning*, pages 89-96.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder, 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136-158, Prague, Czech Republic.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder, 2008. Further meta-evaluation of machine translation. *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70-106, Columbus, Ohio.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia, 2012. Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10-51, Montreal, Canada.

Chris Callison-Burch, Philipp Koehn, Christof Monz and Omar Zaidan, 2011. Findings of the 2011 Workshop on Statistical Machine Translation. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22-64, Edinburgh, Scotland.

Z. Cao, T. Qin, T.Y. Liu, M.F. Tsai and H. Li, 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. *Proceedings of the 24th international conference on Machine learning*, pages 129-136.

Yee Seng Chan and Hwee Tou Ng, 2008. MAXSIM: A Maximum Similarity Metric for Machine Translation Evaluation. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, pages 55-62, Columbus, Ohio.

Boxing Chen, Roland Kuhn and Samuel Larkin, 2012. PORT: a Precision-Order-Recall MT Evaluation Metric for Tuning. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 930-939, Jeju Island, Korea.

Michael Denkowski and Alon Lavie, 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

S. Corston-Oliver, M. Gamon and C. Brockett, 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 148-155.

George Doddington, 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrence Statistics. *Proceedings of the second international conference on Human Language Technology Research (HLT'02)*, pages 138-145, San Diego, California, CA, USA.