IRT-based Aggregation Model of Crowdsourced Pairwise Comparisons for Evaluating Machine Translations

Naoki Otani[1] **Toshiaki Nakazawa**[2] **Daisuke Kawahara**[1] **Sadao Kurohashi**[1]

[1]Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan

[2]Japan Science and Technology Agency, Kawaguchi-shi, Saitama, Japan

otani.naoki.65v@st.kyoto-u.ac.jp nakazawa@pa.jst.jp {dk,kuro}@i.kyoto-u.ac.jp

Abstract

Recent work on machine translation has used crowdsourcing to reduce costs of manual evaluations. However, crowdsourced judgments are often biased and inaccurate. In this paper, we present a statistical model that aggregates many manual pairwise comparisons to robustly measure a machine translation system's performance. Our method applies graded response model from item response theory (IRT), which was originally developed for academic tests. We conducted experiments on a public dataset from the Workshop on Statistical Machine Translation 2013, and found that our approach resulted in highly interpretable estimates and was less affected by noisy judges than previously proposed methods.

1 Introduction

Manual evaluation is a primary means of interpreting the performance of machine translation (MT) systems and evaluating the accuracy of automatic evaluation metrics. It is also essential for natural language processing tasks such as summarization and dialogue systems, where (1) the number of correct outputs is unlimited, and (2) naïve text matching cannot judge the correctness, that is, an evaluator must consider syntactic and semantic information.

Recent work has used crowdsourcing to reduce costs of manual evaluations. However, the judgments of crowd workers are often noisy and unreliable because they are not experts.

To maintain quality, evaluation tasks implemented using crowdsourcing should be simple. Thus, many previous studies focused on pairwise comparisons instead of absolute evaluations. The same task is given to multiple workers, and their responses are aggregated to obtain a reliable answer.

We must, therefore, develop methods that robustly estimate the MT performance based on many pairwise comparisons.

Some aggregation methods have been proposed for MT competitions hosted by the Workshop on Statistical Machine Translation (WMT) (Bojar et al., 2013; Hopkins and May, 2013; Sakaguchi et al., 2014), where a ranking of the submitted systems is produced by aggregating many manual judgments of pairwise comparisons of system outputs.

However, existing methods do not consider the following important issues.

Interpretability of the estimates: For the purpose of evaluation, their results must be interpretable so that we could use the results to improve MT systems and the next MT evaluation campaigns. Existing methods, however, only yield system-level scores.

Judge sensitivity: Some judges can examine the quality of translations with consistent standards, but others cannot (Graham et al., 2015). Sensitivities to the translation quality and judges' own standards are important factors.

Evaluation of a newly submitted system: Previous approaches considered all pairwise combinations of systems and must compare a newly submitted system with all the submitted systems. This made it difficult to allow participants to submit their systems after starting the evaluation step.

To address these issues, we use a model from

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 511–520,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

item response theory (IRT). This theory was originally developed for psychometrics, and has applications to academic tests. IRT models are highly interpretable and are supported by theoretical and empirical studies. For example, we can estimate the informativeness of a question in a test based on the responses of examinees.

We focused on aggregating many pairwise comparisons with a baseline translation so that we could use the analogy of standard academic tests. Figure 1 shows our problem setting. Each system of interest yields translations, and the translations are compared with a baseline translation by multiple human judges. Each judge produces a preference judgment.

The pairwise comparisons correspond to questions in academic tests, a judge's sensitivity to the translation quality is mapped to discrimination of questions, and the relative difficulty of winning the pairwise comparison is mapped to the difficulty of questions. MT systems correspond to students that take academic tests, and IRT models can be naturally applied to estimate the latent performance (ability) of MT systems (students).

Additionally, our approach, fixing baseline translations, can easily evaluate a newly submitted system. We only need to compare the new system with the baseline instead of testing all pairwise combinations of the submitted systems.

Our contributions are summarized as follows.[1]

1. We propose an IRT-based aggregation model of pairwise comparisons with highly interpretable parameters.
2. We simulated noisy judges on the WMT13 dataset and demonstrated that our model is less affected by the noisy judges than previously proposed methods.

2 Related Work

The WMT shared tasks have collected many manual judgments of segment-level pairwise comparisons and used them to produce system-level rankings for MT tasks. Various methods has been proposed to aggregate the judgments to produce reliable rankings.

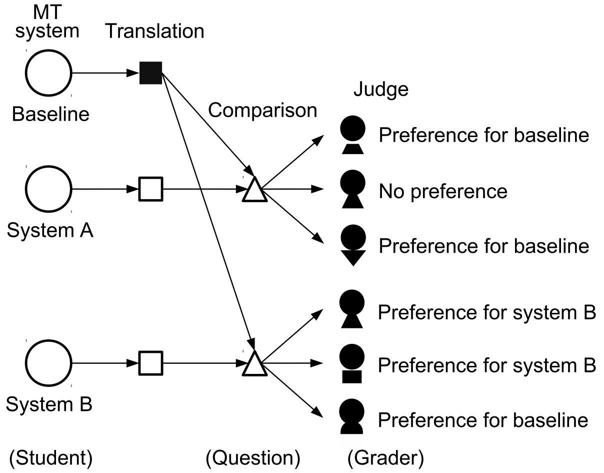


Figure 1: **Illustration of manual pairwise comparison.** Each system yields translations. Judges compare them with a baseline translation and report their preferences. Our goal is to aggregate the judgments to determine the performance of each system.

Frequency based approaches were used to produce the WMT13 official rankings (Bojar et al., 2013), considering statistical significance of the results (Koehn, 2012).

Hopkins and May (2013) noted that we should consider the relative matchup difficulty, and proposed a statistical aggregation model. Their model assumes that the quality of each system can be represented by a Gaussian distribution.

Sakaguchi et al. (2014) applied TrueSkill (Herbrich et al., 2006) to reduce the number of comparisons to reach the final estimate based on an active learning strategy. The same model was recently used for grammatical error correction (Grundkiewicz et al., 2015; Napoles et al., 2015).

These methods acquire the final system-level scores, whereas our model also estimates segment specific and judge specific parameters.

The Bradley–Terry (BT) model was the result of a seminal study on aggregating pairwise comparisons (Bradley and Terry, 1952; Chen et al., 2013; Dras, 2015). Recently, Chen et al. (2013) explicitly incorporated the quality of judges into the BT model, and applied it to quality control in crowdsourcing.

The previously mentioned methods focused on pairwise comparisons of all combination of the MT systems, and thus, the number of comparisons increases rapidly as the number of systems increases.

[1] We also show that our method accurately replicated the WMT13 official system scores using a few comparisons. However, this is not the main focus of this paper.

Our approach, however, only uses comparisons with a fixed baseline. This approach enables to apply IRT models for academic tests and makes it easy to evaluate a newly submitted system.

The work most relevant to our model is the IRT-based crowdsourcing model proposed by Baba and Kashima (2013). Their goal was to estimate the true quality of artifacts such as design works based on ratings assigned by reviewers. They also applied a graded response model to incorporate the authors' latent abilities and the reviewers' biases.

Yet their setting differs from ours in that they focused on the quality of the artifacts, whereas we are interested in the authors. Additionally, their model maps task difficulty and review bias to a difficulty parameter in IRT. However, we naturally extended the model so that standard analysis approaches can be applied to maintain interpretability.

Some studies have focused on absolute evaluations (Goto et al., 2014; Graham et al., 2015). Graham et al. (2015) gathered continuous scale evaluations in terms of adequacy and fluency for many segments, and filtered out noisy judgments based on their consistency. The proposed pipeline results in very accurate evaluations, but 40-50% of all the judgments were filtered out due to inconsistencies. This explains the difficulties of developing absolute evaluation methods in crowdsourcing.

3 Problem Setting

We first describe the problem setting, as shown in Figure 1.

Assume that there are a group of systems $\mathcal{I}$ indexed by i, a set of segments $\mathcal{J}$ indexed by j, and a set of judges $\mathcal{K}$ indexed by k.

Before a manual evaluation, we fix an arbitrary baseline system and use it to translate the segments $\mathcal{J}$. Then, each system $i \in \mathcal{I}$ produces a translation on segment $j \in \mathcal{J}$. One of the judges $k \in \mathcal{K}$ compares it with the baseline translation. The judge produces a preference judgment.

Let $u_{i,j,k}$ be the observed judgment that judge k assigns to a translation by system i on segment j, that is,

$$u_{i,j,k} = \begin{cases} 1 & \text{(preference for baseline)} \\ 2 & \text{(no preference)} \\ 3 & \text{(preference for system } i\text{)} \end{cases},$$

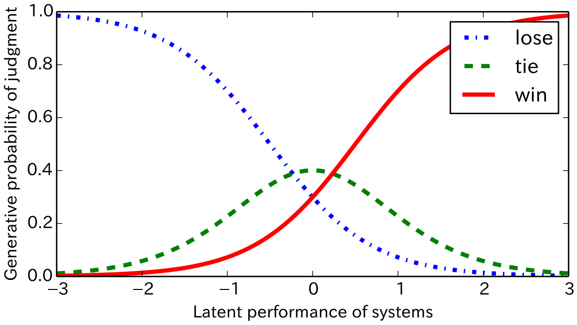


Figure 2: **ICC of graded response model for** $(b_1, b_2) = (-0.5, 0.5)$ **and** $a = 1.7$

and let $c \in \{1, 2, 3\}$ be the judgment label.

Each system i has its own latent performance $\theta_i \in \mathbb{R}$. Our goal is to estimate θ by using the observed judgments $U = \{u_{i,j,k}\}_{i\in\mathcal{I},j\in\mathcal{J},k\in\mathcal{K}}$.

4 Generative Judgment Model

We describe a statistical model for pairwise comparisons based on an IRT model.

4.1 Modified Graded Response Model

Based on the graded response model (GRM) proposed by Samejima (1968), we define a generative model of judgments. GRM deals with responses on ordered categories including ratings such as A+, A, B+ and B, and partial credits in tests. In our problem setting, judgments can be seen as partial credits. When a system beats a baseline translation, the system receives $c = 3$ credit. In the case of a tie, the system receives $c = 2$ credit. The system receives $c = 1$ credit when it lose to the baseline.

Let $\mathrm{P}^*_{jkc}(\theta_i)$ be the probability that judge k assigns judgment $\pi > c$ to a comparison on segment j between system i and a baseline.

$$\mathrm{P}^*_{jkc}(\theta_i) = \frac{1}{1 + \exp(-a_k(\theta_i - b_{jc}))},$$

where $\mathrm{P}^*_{jk0}(\theta_i) = 1, \mathrm{P}^*_{jk3}(\theta_i) = 0$. Parameters a and b are called *discrimination* and *difficulty* parameters, respectively. a represents the discriminablity or sensitivity of the judge, and b represents a segment-specific difficulty parameter. The discrimination parameter (a) is positive, and the difficulty parameter (b) satisfies $b_1 < b_2$, where b_1 corresponds to the difficulty of not losing to the baseline $(c > 1)$, and b_2

corresponds to the difficulty of beating the baseline ($c > 2$).

The generative probability of judgment $u_{i,j,k}$ is defined as the difference in the probabilities defined above, that is,

$$\begin{aligned}\mathrm{P}_{jkc}(\theta_i) &= \mathrm{P}(u_{i,j,k} = c|\theta_i, b_j, a_k)\\ &= \mathrm{P}^*_{jkc-1}(\theta_i) - \mathrm{P}^*_{jkc}(\theta_i).\end{aligned}$$

This function is called *item characteristic curve* (ICC). Figure 2 illustrates the ICC in the GRM. The horizontal axis represents the latent performance of systems, and the vertical axis represents the generative probability of the judgments. This figure shows, for example, that the probability of the system with $\theta = 0$ beating the baseline is 0.3, whereas the system with $\theta = 1.0$ is much more likely to win. The discrimination parameter controls slope of the curves. If a is small, the probability drops a little when θ decreased.

The model described above is different from the original GRM, which assumed that the values of a are independent from question to question, and that each a belongs to exactly one question. However, in our problem setting, the judges evaluate multiple segments, and discrimination parameter a is independent from segment j. This modification means that the GRM can capture the judge's sensitivity.

4.2 Priors

We assign prior distributions to the parameters to obtain estimates stably. We assume Gaussian distributions on θ and b, that is, $\theta \sim \mathcal{N}(0, \tau^2)$ and $b_c \sim \mathcal{N}(\mu_{bc}, \sigma^2_{bc})$ $(c = 1, 2)$. The discrimination parameter is positive, so we assume a log Gaussian distribution on a, i.e., $\log(a) \sim \mathcal{N}(\mu_a, \sigma^2_a)$. Note that τ, μ, and σ are hyper parameters.

5 Parameter Estimation

We find the values of the parameters to maximize the log likelihood based on obtained judgments U:

$$\mathcal{L}(\theta, \xi) = \log\mathrm{P}(U, \theta, \xi).$$

We denote the parameters $a = \{a_k\}_{k \in \mathcal{K}}$ and $b = \{b_{j1}, b_{j2}\}_{j \in \mathcal{J}}$ to be ξ in this section.

5.1 Marginal Likelihood Maximization of Judge Sensitivity and Matchup Difficulty

Estimates are known to be inaccurate when all the parameters are optimized at once, so we first estimate the parameters ξ to maximize the marginal log likelihood w.r.t. the system performance θ.

$$\begin{aligned}m\mathcal{L}(\xi) &= \log\mathrm{P}(U, \xi)\\ &= \sum_{i \in \mathcal{I}} \log \int_{-\infty}^{\infty} \mathrm{P}(\theta)\mathrm{P}(U_i|\theta, \xi) d\theta + \log\mathrm{P}(\xi),\end{aligned}$$

where U_i is the set of judgments given to system i

The equation above can be approximated using Gauss-Hermite quadrature, i.e.,

$$\begin{aligned}m\mathcal{L}(\xi) &\approx \sum_{i \in \mathcal{I}} \log \sum_{t=1}^{T} \frac{1}{\sqrt{\pi}} w_t \mathrm{P}(U_i|\tau x_t, \xi) + \log\mathrm{P}(\xi)\\ w_t &= \frac{2^{T-1} T! \sqrt{\pi}}{T^2 \left(H(x_t)\right)^2}\\ H(x_t) &= \left(2x_t - \frac{d}{dx_t}\right)^{T-1} \cdot 1,\end{aligned}$$

where a practically good approximation is obtained by taking $T \approx 20$.[2]

We solve the optimization problem using the gradient descent methods to maximize the approximated marginal likelihood. The inequality constraints on the parameters are handled by adding log barrier functions to the objective function.

5.2 Maximum A Posteriori (MAP) Estimation of System Performance

Given the estimates of ξ, we estimate the system performance $\theta = \{\theta_i\}_{i \in \mathcal{I}}$ by using MAP estimation.

We maximize the objective function,

$$\begin{aligned}\mathcal{L}(\theta) &= \log\mathrm{P}(U, \theta; \xi)\\ &= \sum_{i \in \mathcal{I}} \log\mathrm{P}(\theta_i) + \sum_{i \in \mathcal{I}} \log\mathrm{P}(U_i|\theta_i; \xi).\end{aligned}$$

The estimates of θ are obtained using the gradient descent method.

5.3 Discussion

So far we have assumed that the estimate is based on batch learning. However, it is known that active

[2] In this study, we set $T = 21$ to include $x = 0$.

learning can reduce the costs (i.e., the total number of comparisons) (Sakaguchi et al., 2014).

To extend our model to the active learning framework, one approach is to optimize the objective function online and actively select the next system to be compared based on criteria such as the uncertainty of the system's performance. We can apply stochastic gradient descent to the online optimization, which updates the estimates of the parameters using the gradients calculated based on a single comparison. This modification was left for future work.

6 Experiments

We conducted experiments on the WMT13 manual evaluation dataset for 10 language pairs.[3] For details of the evaluation data, see the overview of WMT13 (Bojar et al., 2013).

6.1 Setup

Models: Our method **(GRM)** was initialized using $a = 1.7, b = (-0.5, 0.5)$, and a θ value derived by summing up the judgments for each system and scaling θ to fit the prior distribution. For the hyper parameters, we set $\tau = \sqrt{2}, \mu_a = \log(1.7), \sigma_a = 1.0, \mu_b = (-0.5, 0.5), \sigma_b = 2.0$.

To compare with our method, we trained ExpectedWins **(EW)** (Bojar et al., 2013), the model by Hopkins and May (2013), **(HM)** and the two-stage crowdsourcing model proposed by Baba and Kashima (2013) **(TSt)**. We also trained TrueSkill **(TS)** (Sakaguchi et al., 2014), which was used to produce the gold score on this experiment.

We followed Sakaguchi et al. (2014), who also used the WMT13 datasets in their experiments, and initialized the HM and TS parameters. For TSt, we followed Baba and Kashima (2013).

Pairwise comparisons: The WMT dataset contains five-way partial rankings, so we converted the five-way partial rankings into pairwise comparisons. For example, given a five-way partial ranking A > B > C > D > E, we obtain ten pairwise comparisons A > B, A > C, A > D, $\cdots$, and D > E. We randomly sampled 800, 1,600, 3,200 and 6,400 pairwise comparisons from the whole dataset.

The training data differs between the models. For GRM and TSt, we first sampled five-way rankings that contained a baseline translation for each baseline system and obtained pairwise comparisons. For EW and HM, we first converted five-way rankings into pairwise comparisons and selected them at random.[4] TS first receives all the pairwise comparisons and selects the training data based on the active learning strategy, whereas we sampled the comparisons before running the other methods.

Gold scores: We followed the official evaluation procedure of the WMT14-15 (Bojar et al., 2014; Bojar et al., 2015) and made gold scores with TS. We produced 1,000 bootstrap-resampled datasets over all of the available comparisons. We then ran TS and collected the system scores. The gold score is the mean of the scores.

Evaluation metrics: We evaluated the models using the Pearson correlation coefficient and the normalized discounted cumulative gain (nDCG), comparing the estimated scores and gold scores. We used nDCG because we are often interested in ranks and scores, especially in MT competitions such as the WMT translation task.[5] These metrics were also used for experiments in Baba and Kashima (2013).

6.2 Results

Figure 3 shows the correlation and nDCG between the estimated system performance and the gold scores for the WMT13 Spanish–English task. For the GRM and TSt, the baselines used in the evaluation are shown in parentheses in the labels. The other language pairs showed similar tendencies. The complete results for all language pairs can be found in the supplementary data files.

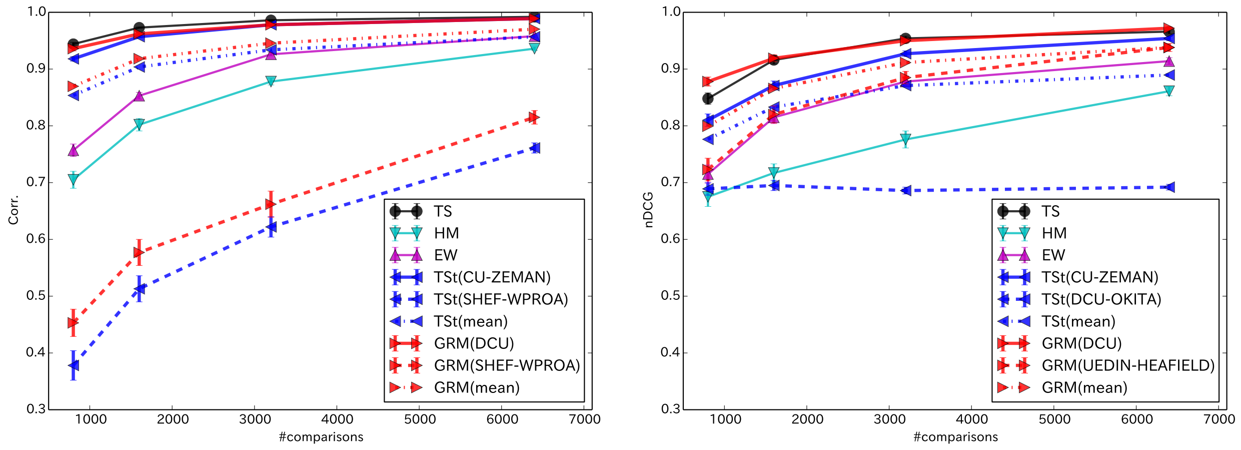
Note that the main contribution of our method is not to perform better than other methods in terms of correlation and nDCG to the gold scores, but to result in highly interpretable and robust estimates discussed later.

TS resulted in the highest correlation and nDCG. It is reasonable because the gold scores themselves were produced by TS, and because it estimates the

[3] `http://statmt.org/wmt13/results.html`

[4] We also applied the sampling procedure of GRM and TSt to EW and HM, but it made their estimation inaccurate.

[5] We did not use Spearman's rank correlation coefficient because it does not consider a margin between ranks.



(a) Correlation

(b) nDCG

Figure 3: **Correlation and nDCG comparing the estimated system performance and gold scores with the number of comparisons for the WMT13 Spanish–English task.** The baseline system is shown in parenthesis for TSt and GRM.

parameters using active learning, unlike the other models.

The GRM with the best baseline system (DCU) achieved almost the same scores as the TS, in terms of correlation and nDCG. Although the TSt with the best baseline resulted in accurate estimates in terms of correlation, it did not in terms of nDCG. With the worst baselines, the GRM and TSt both failed to replicate the gold scores, but the GRM was surprisingly accurate in terms of nDCG (even in the worst case). This implies that the GRM can effectively predict the top ranked systems.

6.3 Baseline Selection

It is likely that single pairwise comparisons do not work well if the baseline is very strong or weak. As shown in Figure 3, the baseline system influences the final result. When we used SHEF-WPROA as baseline, the estimated system performance was not accurate. This is because SHEF-WPROA loses 69.4% of the pairwise comparisons and fails to discriminate between the other systems. In contrast, DCU loses 34.5% and win 34.8% of the comparisons and discriminate the other systems successfully. Thus, when we used DCU as baseline, the best correlation and nDCG were achieved. Therefore, we must determine the appropriate baseline system before the comparisons.

One possible solution is to consider the system-level scores yielded by automatic evaluation metrics such as BLEU and METEOR. Figure 4 shows that we obtained relatively good results when we used a system whose system-level BLEU score and METEOR score[6] were close to the mean of all the systems. [7]

Noise(%)	0	10	20	30	40	50
Correlation						
GRM	.929	.917	.900	.879	.849	.807
HM	+.002	-.005	-.009	-.015	-.025	-.038
EW	-.025	-.028	-.035	-.038	-.040	-.046
nDCG						
GRM	.883	.867	.847	.822	.793	.752
HM	-.024	-.130	-.137	-.144	-.152	-.168
EW	-.035	-.054	-.064	-.060	-.060	-.069

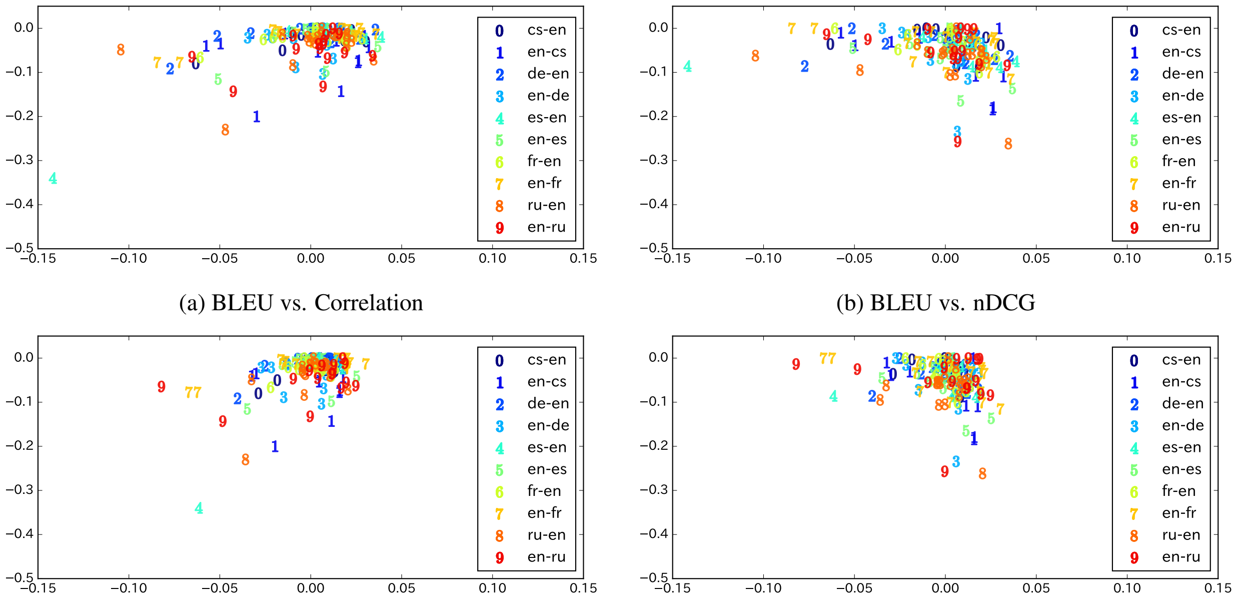
Table 1: **Correlation and nDCG between the estimated system performance and gold scores for the WMT13 Spanish–English task, based on noisy judges.** The values were averaged over all the datasets. The GRM scores were averaged over all baselines. The differences from the GRM are reported for the HM and EW.

6.4 Analysis of Judge Sensitivity

To investigate the robustness of the GRM, we simulated "noisy" judges. We selected a subset of

[6] BLEU and METEOR scores were given by the WMT13 organizers.

[7] The system-level scores can be found in the WMT13 Metrics Task dataset.



(a) BLEU vs. Correlation

(b) BLEU vs. nDCG

(c) METEOR vs. Correlation

(d) METEOR vs. nDCG

Figure 4: **Relationship between system-level BLEU/METEOR scores (horizontal) and correlation/nDCG scores (vertical).** The mean BLEU/METEOR was set to zero, and the best score was set to zero for each language pair.

judges and randomly changed their decisions based on a uniform distribution. The percentage of noisy judges varied between 10% and 50% (in increments of 10%).

We trained HM and EW on the simulated datasets. We excluded TS because it assumes that we can actively request more comparisons from judges when their decisions are ambiguous.

As shown in Table 1, the accuracy of the GRM was less affected by the noisy judges than HM and EW. This is because our model estimates judge-specific sensitivities and automatically reduces the influence of the noisy judges.

6.5 Analysis of the Interpretability of the Estimated Matchup Difficulty

Our model is a natural extension of the GRM Samejima (1968), so we can apply standard analyses for IRT models. Item information is one of the standard analysis methods and corresponds to sensitivity to a latent parameter of interest. Based on the item information, we can find which segment was difficult to be translated better than a baseline translation.

The item information is calculated using the estimated parameters ξ (Samejima, 1968), that is,

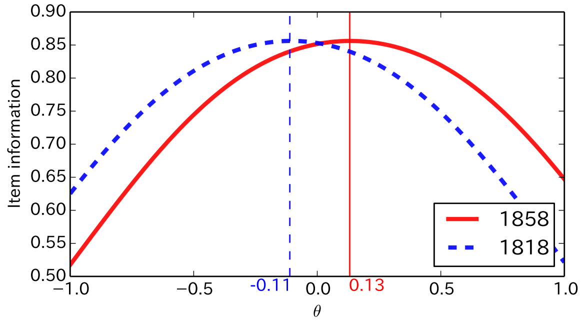


Figure 5: **Item information for the WMT13 Spanish–English task.** The DCU was used as a baseline. We used the averaged estimates of b on 100 sampled datasets with 6,400 comparisons to calculate the item information for all segments.

$$
\begin{aligned}
I_j(\theta) &= -E\left[\frac{\partial^2 \mathcal{L}(\theta;\xi)}{\partial\theta^2}\right] \\
&= \sum_{c=1}^{3}\left[-\frac{\partial^2 \log \mathrm{P}_{jkc}(\theta)}{\partial\theta^2}\right]\mathrm{P}_{jkc} \\
&= \sum_{c=1}^{3}\frac{[\mathrm{P}^{*'}_{jkc-1}(\theta)-\mathrm{P}^{*'}_{jkc}(\theta)]^2}{\mathrm{P}^{*}_{jkc-1}(\theta)-\mathrm{P}^{*}_{jkc}(\theta)},
\end{aligned}
$$

where $\mathrm{P}^{*'} = \partial\mathrm{P}^{*}/\partial\theta$.

Because the item information is only determined

Segment 1858: Difficult to beat the baseline translation.		
Source		Hasta 2007 los dos telescopios Keck situados en el volcán hawaiano de Mauna Kea eran considerados los más grandes del mundo.
Reference		Until 2007, the two Keck telescopes at the Hawaiian volcano, Mauna Kea, were the largest in the world.
DCU[baseline]		Until 2007, the two Keck telescopes located on the Hawaiian volcano Mauna of KEA were considered the largest in the world.
ONLINE-B	($\theta =$) 0.24	Until 2007 the two Keck telescopes located on the Hawaiian volcano Mauna Kea were considered the largest in the world.
UEDIN	0.12	Until 2007, the two Keck telescopes located on the Hawaiian volcano of Mauna Kea were considered the largest in the world.
LIMSI-NCODE-SOUL	0.10	Until 2007 the two Keck telescopes in the Hawaiian Mauna Kea volcano were considered the largest in the world.
CU-ZEMAN	-0.10	Until 2007, the two Keck telescope located in the volcano Mauna Kea hawaiano of were regarded as the world's largest.
JHU	-0.12	Until 2007, the two Telescope Keck located in the Kea volcano hawaiano of Mauna were considered the world's largest.
SHEF-WPROA	-0.92	Until 2007 the two telescope Keck located volcano hawaiano of Mauna KEA were regarded larger of world.

Segment 1818: Easy to beat the baseline translation.		
Source		Dependiendo de las tonalidades, algunas imágenes de galaxias espirales se convierten en verdaderas obras de arte.
Reference		Depending on the colouring, photographs of spiral galaxies can become genuine works of art.
DCU[baseline]		Depending on the drink, some images of galaxias galaxies become true works of art.
ONLINE-B	0.24	Depending on the shades, some images of spiral galaxies become true works of art.
UEDIN	0.12	(Same as ONLINE-B)
LIMSI-NCODE-SOUL	0.10	Depending on the color, some images of galaxies spirals become real works of art.
CU-ZEMAN	-0.10	Depending on the tonalidades, some images of spirals galaxies become true works of art.
JHU	-0.12	Depending on the tonalidades, some images of galaxies spirals become true works of art.
SHEF-WPROA	-0.92	Depending on the tonalidades, some images of galaxies spirals become real artwork.

Table 2: **Translation examples for the WMT13 Spanish–English task.** The reference is a correct translation given by the WMT organizers and was shown to human judges. Estimates of θ (averaged over 100 sampled datasets with 6,400 comparisons) are also reported in the table.

by segments and is independent of the judges, we set $a_k = 1$ ($k \in \mathcal{K}$).

Figure 5 gives two examples of the item information. The horizontal axis corresponds to the system performance θ, and the vertical axis represents the informativeness of a segment. This figure indicates that segment 1858 (red line) can effectively discriminate systems with $\theta \approx 0.13$, whereas segment 1818 (blue dashed line) is sensitive to those with $\theta \approx -0.11$. This means that systems with low θ tend to lose to a baseline translation on segment 1858, and the segment does not tell meaningful information on performance of the systems. However, they sometimes beat a baseline translation on segment 1818, and the segment can measure their performance accurately.

Table 2 shows translations for segments 1858 and 1818. We found that the baseline translation on segment 1818 was relatively good, whereas the baseline translation on segment 1858 contained wrong words such as "drink" and "galaxias". Consequently, systems with low θ tended to lose to the baseline on segment 1858 due to their wrong translation (see the translation of "hawaiano de Mauna Kea"). In contrast, some of the low-ranked systems beat the baseline on segment 1818, and the segment contributed to discriminate them.

The item information is used to design academic tests that can effectively capture students' abilities. It could analogously be used to preselect segments to be translated based on the item information in the MT evaluation.

7 Conclusion

We have addressed the task of manual judgment aggregation for MT evaluations. Our motivation was three folded: (1) to incorporate a judge's sensitivity to robustly measure a system's performance, (2) to maintain highly interpretable estimates, and (3) to handle with a newly submitted system.

To tackle these problems, we focused on pairwise comparisons with a fixed baseline translation so that we could apply the GRM model in IRT by using the analogy of standard academic tests. Unlike testing all pairwise combinations of systems, fixing baseline translations makes it easy to evaluate a newly submitted system. We demonstrated that our model gave robust and highly interpretable estimates on the WMT13 datasets.

In the future work, we will incorporate active learning to the proposed method so that we could reduce the total number of comparisons to obtain final results. Although we evaluated the correlation between the estimated system performance scores and the WMT official scores, other evaluation procedures might also be considered. For example, Hopkins and May (2013) considered model perplexity and Sakaguchi et al. (2014) compared accuracy. However, we cannot directly compare other methods to our method in terms of perplexity or accuracy because our method focuses on comparisons with a baseline translation, whereas they do not. It will be required to investigate correlation between the estimates and expert decisions.

Acknowledgments

We would like to thank Yukino Baba and Hisashi Kashima for providing an implementation of their method. We are also thankful for the useful comments from the anonymous reviewers.

References

Yukino Baba and Hisashi Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 554–562, New York, USA, August. ACM Press.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pages 12–58, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 193–202, New York, New York, USA, February. ACM Press.

Mark Dras. 2015. Evaluating human pairwise preference judgments. *Computational Linguistics*, 41(2):337–345.

Shinsuke Goto, Donghui Lin, and Toru Ishida. 2014. Crowdsourcing for evaluating machine translation quality. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May. European Language Resources Association.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2015. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 461–470, Lisbon, Portugal, June. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill$^{\text{TM}}$: A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 569–576, Vancouver, British Columbia, Canada, Demeber. MIT Press.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1416–1424, Sofia, Bulgaria, August. Association for Computational Linguistics.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, pages 179–184, Hongkong, China, December. International Speech Communication Association.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 588–593, Beijing, China, July. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, pages 1–11, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Fumiko Samejima. 1968. Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, 1968(1):i–169, June.