Models of Translation Competitions

Mark Hopkins and Jonathan May
SDL Research
6060 Center Drive, Suite 150
Los Angeles, CA 90045
{mhopkins,jmay}@sdl.com

Abstract

What do we want to learn from a translation competition and how do we learn it with confidence? We argue that a disproportionate focus on ranking competition participants has led to lots of different rankings, but little insight about which rankings we should trust. In response, we provide the first framework that allows an *empirical* comparison of different analyses of competition results. We then use this framework to compare several analytical models on data from the Workshop on Machine Translation (WMT).

1 The WMT Translation Competition

Every year, the Workshop on Machine Translation (WMT) conducts a competition between machine translation systems. The WMT organizers invite research groups to submit translation systems in eight different tracks: Czech to/from English, French to/from English, German to/from English, and Spanish to/from English.

For each track, the organizers also assemble a panel of judges, typically machine translation specialists.[1] The role of a judge is to repeatedly rank five different translations of the same source text. Ties are permitted. In Table 1, we show an example[2] where a judge (we'll call him "jdoe") has ranked five translations of the French sentence "Il ne va pas."

Each such elicitation encodes ten pairwise comparisons, as shown in Table 2. For each competition track, WMT typically elicits between 5000 and 20000 comparisons. Once the elicitation process is complete, WMT faces a large database of comparisons and a question that must be answered: whose system is the best?

rank	system	translation
1	bbn	"He does not go."
2 (tie)	uedin	"He goes not."
2 (tie)	jhu	"He did not go."
4	cmu	"He go not."
5	kit	"He not go."

Table 1: WMT elicits preferences by asking judges to simultaneously rank five translations, with ties permitted. In this (fictional) example, the source sentence is the French "Il ne va pas."

source text	sys1	sys2	judge	preference
"Il ne va pas."	bbn	cmu	jdoe	1
"Il ne va pas."	bbn	jhu	jdoe	1
"Il ne va pas."	bbn	kit	jdoe	1
"Il ne va pas."	bbn	uedin	jdoe	1
"Il ne va pas."	cmu	jhu	jdoe	2
"Il ne va pas."	cmu	kit	jdoe	1
"Il ne va pas."	cmu	uedin	jdoe	2
"Il ne va pas."	jhu	kit	jdoe	1
"Il ne va pas."	jhu	uedin	jdoe	0
"Il ne va pas."	kit	uedin	jdoe	2

Table 2: Pairwise comparisons encoded by Table 1. A preference of 0 means neither translation was preferred. Otherwise the preference specifies the preferred system.

2 A Ranking Problem

For several years, WMT used the following heuristic for ranking the translation systems:

$$\textsc{OrigWMT}(s) = \frac{\mathsf{win}(s) + \mathsf{tie}(s)}{\mathsf{win}(s) + \mathsf{tie}(s) + \mathsf{loss}(s)}$$

For system s, $\mathsf{win}(s)$ is the number of pairwise comparisons in which s was preferred, $\mathsf{loss}(s)$ is the number of comparisons in which s was dispreferred, and $\mathsf{tie}(s)$ is the number of comparisons in which s participated but neither system was preferred.

Recently, (Bojar et al., 2011) questioned the adequacy of this heuristic through the following ar-

[1] Although in recent competitions, some of the judging has also been crowdsourced (Callison-Burch et al., 2010).

[2] The example does not use actual system output.

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1416–1424,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

gument. Consider a competition with systems A and B. Suppose that the systems are different but equally good, such that one third of the time A is judged better than B, one third of the time B is judged better than A, and one third of the time they are judged to be equal. The expected values of $\textsc{OrigWMT}(A)$ and $\textsc{OrigWMT}(B)$ are both 2/3, so the heuristic accurately judges the systems to be equivalently good. Suppose however that we had duplicated B and had submitted it to the competition a second time as system C. Since B and C produce identical translations, they should always tie with one another. The expected value of $\textsc{OrigWMT}(A)$ would not change, but the expected value of $\textsc{OrigWMT}(B)$ would increase to 5/6, buoyed by its ties with system C.

This vulnerability prompted (Bojar et al., 2011) to offer the following revision:

$$\textsc{Bojar}(s) = \frac{\mathsf{win}(s)}{\mathsf{win}(s) + \mathsf{loss}(s)}$$

The following year, it was BOJAR's turn to be criticized, this time by (Lopez, 2012):

> Superficially, this appears to be an improvement....couldn't a system still be penalized simply by being compared to [good systems] more frequently than its competitors? On the other hand, couldn't a system be rewarded simply by being compared against a bad system more frequently than its competitors?

Lopez's concern, while reasonable, is less obviously damning than (Bojar et al., 2011)'s criticism of ORIGWMT. It depends on whether the collected set of comparisons is small enough or biased enough to make the variance in competition significant. While this hypothesis is plausible, Lopez makes no attempt to verify it. Instead, he offers a ranking heuristic of his own, based on a Minimum Feedback Arc solver.

The proliferation of ranking heuristics continued from there. The WMT 2012 organizers (Callison-Burch et al., 2012) took Lopez's ranking scheme and provided a variant called Most Probable Ranking. Then, noting some potential pitfalls with that, they created two more, called Monte Carlo Playoffs and Expected Wins. While one could raise philosophical objections about each of these, where would it end? Ultimately, the WMT 2012 findings presented five different rankings for the English-German competition track, with no guidance about which ranking we should pay attention to. How can we know whether one ranking is better than other? Or is this even the right question to ask?

3 A Problem with Rankings

Suppose four systems participate in a translation competition. Three of these systems are extremely close in quality. We'll call these close1, close2, and close3. Nevertheless, close1 is very slightly better[3] than close2, and close2 is very slightly better than close3. The fourth system, called terrific, is a really terrific system that far exceeds the other three.

Now which is the better ranking?

$$\mathsf{terrific, close3, close1, close2} \tag{1}$$
$$\mathsf{close1, terrific, close2, close3} \tag{2}$$

Spearman's rho[4] would favor the second ranking, since it is a less disruptive permutation of the gold ranking. But intuition favors the first. While its mistakes are minor, the second ranking makes the hard-to-forgive mistake of placing close1 ahead of the terrific system.

The problem is not with Spearman's rho. The problem is the disconnnect between the knowledge that we want a ranking to reflect and the knowledge that a ranking actually contains. Without this additional knowledge, we cannot determine whether one ranking is better than another, even if we know the gold ranking. We need to determine what information they lack, and define more rigorously what we hope to learn from a translation competition.

4 From Rankings to Relative Ability

Ostensibly the purpose of a translation competition is to determine the relative ability of a set of translation systems. Let $\mathcal{S}$ be the space of all translation systems. Hereafter, we will refer to $\mathcal{S}$ as the space of *students*. We choose this term to evoke the metaphor of a translation competition as a standardized test, which shares the same goal: to assess the relative abilities of a set of participants.

But what exactly do we mean by "ability"? Before formally defining this term, first recognize that it means little without context, namely:

[3] What does "better" mean? We'll return to this question.

[4] Or Pearson's correlation coefficient.

1. **What kind of source text do we want the systems to translate well?** Say system A is great at translating travel-related documents, but terrible at translating newswire. Meanwhile, system B is pretty good at both. The question "which system is better?" requires us to state how much we care about travel versus newswire documents – otherwise the question is underspecified.

2. **Who are we trying to impress?** While it's tempting to think that translation quality is a universal notion, the 50-60% interannotator agreement in WMT evaluations (Callison-Burch et al., 2012) suggests otherwise. It's also easy to imagine reasons why one group of judges might have different priorities than another. Think a Fortune 500 company versus web forum users. Lawyers versus laymen. Non-native versus native speakers. Posteditors versus Google Translate users. Different groups have different uses for translation, and therefore different definitions of what "better" means.

With this in mind, let's define some additional elements of a translation competition. Let $\mathcal{X}$ be the space of all possible segments of source text, $\mathcal{J}$ be the space of all possible judges, and $\Pi = \{0, 1, 2\}$ be the space of pairwise preferences.[5] We assume all spaces are countable. Unless stated otherwise, variables s_1 and s_2 represent students from $\mathcal{S}$, variable x represents a segment from $\mathcal{X}$, variable j represents a judge from $\mathcal{J}$, and variable π represents a preference from Π. Moreover, define the *negation* $\hat{\pi}$ of preference π such that $\hat{\pi} = 2$ (if $\pi = 1$), $\hat{\pi} = 1$ (if $\pi = 2$), and $\hat{\pi} = 0$ (if $\pi = 0$).

Now assume a joint distribution $P(s_1, s_2, x, j, \pi)$ specifying the probability that we ask judge j to evaluate students s_1 and s_2's respective translations of source text x, and that judge j's preference is π. We will further assume that the choice of student pair, source text, and judge are marginally independent of one another. In other words:

$$\begin{aligned} & P(s_1, s_2, x, j, \pi) \\ = & \begin{array}{l} P(\pi|s_1, s_2, x, j) \cdot P(x|s_1, s_2, j) \\ \qquad \cdot P(j|s_1, s_2) \cdot P(s_1, s_2) \end{array} \\ = & \ P(\pi|s_1, s_2, x, j) \cdot P(x) \cdot P(j) \cdot P(s_1, s_2) \\ = & \ P_{\mathcal{X}}(x) \cdot P_{\mathcal{J}}(j) \cdot P(s_1, s_2) \cdot P(\pi|s_1, s_2, x, j) \end{aligned}$$

It will be useful to reserve notation $P_{\mathcal{X}}$ and $P_{\mathcal{J}}$ for the marginal distributions over source text and judges. We can marginalize over the source segments and judges to obtain a useful quantity:

$$\begin{aligned} & P(\pi|s_1, s_2) \\ = & \sum_{x \in \mathcal{X}} \sum_{j \in \mathcal{J}} P_{\mathcal{X}}(x) \cdot P_{\mathcal{J}}(j) \cdot P(\pi|s_1, s_2, x, j) \end{aligned}$$

We refer to this as the $\langle P_{\mathcal{X}}, P_{\mathcal{J}} \rangle$*-relative ability* of students s_1 and s_2. By using different marginal distributions $P_{\mathcal{X}}$, we can specify what kinds of source text interest us (for instance, $P_{\mathcal{X}}$ could focus most of its probability mass on German tweets). Similarly, by using different marginal distributions $P_{\mathcal{J}}$, we can specify what judges we want to impress (for instance, $P_{\mathcal{J}}$ could focus all of its mass on one important corporate customer or evenly among all fluent bilingual speakers of a language pair).

With this machinery, we can express the purpose of a translation competition more clearly: to estimate the $\langle P_{\mathcal{X}}, P_{\mathcal{J}} \rangle$*-relative ability* of a set of students. In the case of WMT, $P_{\mathcal{J}}$ presumably[6] defines a space of competent source-to-target bilingual speakers, while $P_{\mathcal{X}}$ defines a space of newswire documents.

We'll refer to an estimate of $P(\pi|s_1, s_2)$ as a *preference model.* In other words, a preference model is a distribution $Q(\pi|s_1, s_2)$. Given a set of pairwise comparisons (e.g., Table 2), the challenge is to estimate a preference model $Q(\pi|s_1, s_2)$ such that Q is "close" to P. For measuring distributional proximity, a natural choice is KL-divergence (Kullback and Leibler, 1951), but we cannot use it here because P is unknown.

Fortunately, if we have i.i.d. data drawn from P, then we can do the next best thing and compute the perplexity of preference model Q on this heldout test data. Let $\mathcal{D}$ be a sequence of triples $\langle s_1, s_2, \pi \rangle$ where the preferences π are i.i.d. samples from $P(\pi|s_1, s_2)$. The perplexity of preference model Q on test data $\mathcal{D}$ is:

$$\mathsf{perplexity}(Q|\mathcal{D}) = 2^{-\sum_{\langle s_1, s_2, \pi \rangle \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \log_2 Q(\pi|s_1, s_2)}$$

How do we obtain such a test set from competition data? Recall that a WMT competition produces pairwise comparisons like those in Table 2.

[5] As a reminder, 0 indicates no preference.

[6] One could argue that it specifies a space of machine translation specialists, but likely these individuals are thought to be a representative sample of a broader community.

Let $\mathcal{C}$ be the set of *comparisons* $\langle s_1, s_2, x, j, \pi \rangle$ obtained from a translation competition. Competition data $\mathcal{C}$ is not necessarily[7] sampled i.i.d. from $P(s_1, s_2, x, j, \pi)$ because we may intentionally[8] bias data collection towards certain students, judges or source text. Also, because WMT elicits its data in batches (see Table 1), every segment x of source text appears in at least ten comparisons.

To create an appropriately-sized test set that closely resembles i.i.d. data, we isolate the subset $\mathcal{C}'$ of comparisons whose source text appears in at most k comparisons, where k is the smallest positive integer such that $|\mathcal{C}'| >= 2000$. We then create the test set $\mathcal{D}$ from $\mathcal{C}'$:

$$\mathcal{D} = \{\langle s_1, s_2, \pi \rangle | \langle s_1, s_2, x, j, \pi \rangle \in \mathcal{C}'\}$$

We reserve the remaining comparisons for training preference models. Table 3 shows the resulting dataset sizes for each competition track.

Unlike with raw rankings, the claim that one preference model is better than another has testable implications. Given two competing models, we can train them on the same comparisons, and compare their perplexities on the test set. This gives us a quantitative[9] answer to the question of which is the better model. We can then publish a system ranking based on the most trustworthy preference model.

5 Baselines

Let's begin then, and create some simple preference models to serve as baselines.

5.1 Uniform

The simplest preference model is a uniform distribution over preferences, for any choice of students s_1, s_2:

$$Q(\pi|s_1, s_2) = \frac{1}{3} \quad \forall \pi \in \Pi$$

This will be our only model that does not require training data, and its perplexity on any test set will be 3 (i.e. equal to number of possible preferences).

5.2 Adjusted Uniform

Now suppose we have a set $\mathcal{C}$ of comparisons available for training. Let $\mathcal{C}_\pi \subseteq \mathcal{C}$ denote the subset of comparisons with preference π, and let $\mathcal{C}(s_1, s_2)$ denote the subset comparing students s_1 and s_2.

Perhaps the simplest thing we can do with the training data is to estimate the probability of ties (i.e. preference 0). We can then distribute the remaining probability mass uniformly among the other two preferences:

$$Q(\pi|s_1, s_2) = \begin{cases} \dfrac{|\mathcal{C}_0|}{|\mathcal{C}|} & \text{if } \pi = 0 \\ \dfrac{1 - \frac{|\mathcal{C}_0|}{|\mathcal{C}|}}{2} & \text{otherwise} \end{cases}$$

6 Simple Bayesian Models

6.1 Independent Pairs

Another simple model is the direct estimation of each relative ability $P(\pi|s_1, s_2)$ independently. In other words, for each pair of students s_1 and s_2, we estimate a separate preference distribution. The maximum likelihood estimate of each distribution would be:

$$Q(\pi|s_1, s_2) = \frac{|\mathcal{C}_\pi(s_1, s_2)| + |\mathcal{C}_{\hat{\pi}}(s_2, s_1)|}{|\mathcal{C}(s_1, s_2)| + |\mathcal{C}(s_2, s_1)|}$$

However the maximum likelihood estimate would test poorly, since any zero probability estimates for test set preferences would result in infinite perplexity. To make this model practical, we assume a symmetric Dirichlet prior with strength α for each preference distribution. This gives us the following Bayesian estimate:

$$Q(\pi|s_1, s_2) = \frac{\alpha + |\mathcal{C}_\pi(s_1, s_2)| + |\mathcal{C}_{\hat{\pi}}(s_2, s_1)|}{3\alpha + |\mathcal{C}(s_1, s_2)| + |\mathcal{C}(s_2, s_1)|}$$

We call this the Independent Pairs preference model.

6.2 Independent Students

The Independent Pairs model makes a strong independence assumption. It assumes that even if we know that student A is much better than student B, and that student B is much better than student C, we can infer nothing about how student A will fare versus student C. Instead of directly estimating the relative ability $P(\pi|s_1, s_2)$ of students s_1 and s_2, we could instead try to estimate the *universal ability* $P(\pi|s_1) = \sum_{s_2 \in \mathcal{S}} P(\pi|s_1, s_2) \cdot P(s_2|s_1)$ of each individual student s_1 and then try to reconstruct the relative abilities from these estimates.

For the same reasons as before, we assume a symmetric Dirichlet prior with strength α for each

[7] In WMT, it certainly is not.

[8] To collect judge agreement statistics, for instance.

[9] As opposed to philosophical.

preference distribution, which gives us the following Bayesian estimate:

$$Q(\pi|s_1) = \frac{\alpha + \sum_{s_2 \in \mathcal{S}} |\mathcal{C}_\pi(s_1, s_2)| + |\mathcal{C}_{\hat{\pi}}(s_2, s_1)|}{3\alpha + \sum_{s_2 \in \mathcal{S}} |\mathcal{C}(s_1, s_2)| + |\mathcal{C}(s_2, s_1)|}$$

The estimates $Q(\pi|s_1)$ do not yet constitute a preference model. A downside of this approach is that there is no principled way to reconstruct a preference model from the universal ability estimates. We experiment with three ad-hoc reconstructions. The *asymmetric* reconstruction simply ignores any information we have about student s_2:

$$Q(\pi|s_1, s_2) = Q(\pi|s_1)$$

The *arithmetic* and *geometric* reconstructions compute an arithmetic/geometric average of the two universal abilities:

$$Q(\pi|s_1, s_2) = \frac{Q(\pi|s_1) + Q(\hat{\pi}|s_2)}{2}$$
$$Q(\pi|s_1, s_2) = [Q(\pi|s_1) * Q(\hat{\pi}|s_2)]^{\frac{1}{2}}$$

We respectively call these the (Asymmetric/Arithmetic/Geometric) Independent Students preference models. Notice the similarities between the universal ability estimates $Q(\pi|s_1)$ and the BOJAR ranking heuristic. These three models are our attempt to render the BOJAR heuristic as preference models.

7 Item-Response Theoretic (IRT) Models

Let's revisit (Lopez, 2012)'s objection to the BOJAR ranking heuristic: "...couldn't a system still be penalized simply by being compared to [good systems] more frequently than its competitors?" The official WMT 2012 findings (Callison-Burch et al., 2012) echoes this concern in justifying the exclusion of reference translations from the 2012 competition:

> [W]orkers have a very clear preference for reference translations, so including them unduly penalized systems that, through (un)luck of the draw, were pitted against the references more often.

Presuming the students are paired uniformly at random, this issue diminishes as more comparisons are elicited. But preference elicitation is expensive, so it makes sense to assess the relative ability of the students with as few elicitations as possible. Still, WMT 2012's decision to eliminate references entirely is a bit of a draconian measure, a treatment of the symptom rather than the (perceived) disease. If our models cannot function in the presence of training data variation, then we should change the models, not the data. A model that only works when the students are all about the same level is not one we should rely on.

We experiment with a simple model that relaxes some independence assumptions made by previous models, in order to allow training data variation (e.g. who a student has been paired with) to influence the estimation of the student abilities. Figure 1(left) shows plate notation (Koller and Friedman, 2009) for the model's independence structure. First, each student's *ability* distribution is drawn from a common prior distribution. Then a number of translation *items* are generated. Each *item* is *authored* by a student and has a *quality* drawn from the student's *ability* distribution. Then a number of pairwise *comparisons* are generated. Each *comparison* has two *options*, each a translation *item*. The *quality* of each item is *observed* by a judge (possibly noisily) and then the judge states a *preference* by comparing the two *observations*.

We investigate two parameterizations of this model: Gaussian and categorical. Figure 1(right) shows an example of the Gaussian parameterization. The student ability distributions are Gaussians with a known standard deviation σ_a, drawn from a zero-mean Gaussian prior with known standard deviation σ_0. In the example, we show the ability distributions for students 6 (an above-average student, whose mean is 0.4) and 14 (a poor student, whose mean is -0.6). We also show an item authored by each student. Item 43 has a somewhat low quality of -0.3 (drawn from student 14's ability distribution), while item 205 is not student 6's best work (he produces a mean quality of 0.4), but still has a decent quality at 0.2. Comparison 1 pits these items against one another. A judge draws noise from a zero-mean Gaussian with known standard deviation σ_{obs}, then adds this to the item's actual quality to get an observed quality. For the first option (item 43), the judge draws a noise of -0.12 to observe a quality of -0.42 (worse than it actually is). For the second option (item 205), the judge draws a noise of 0.15 to observe a quality of 0.35 (better than it actually is). Finally, the judge compares the two observed qualities. If the absolute difference is lower than his decision

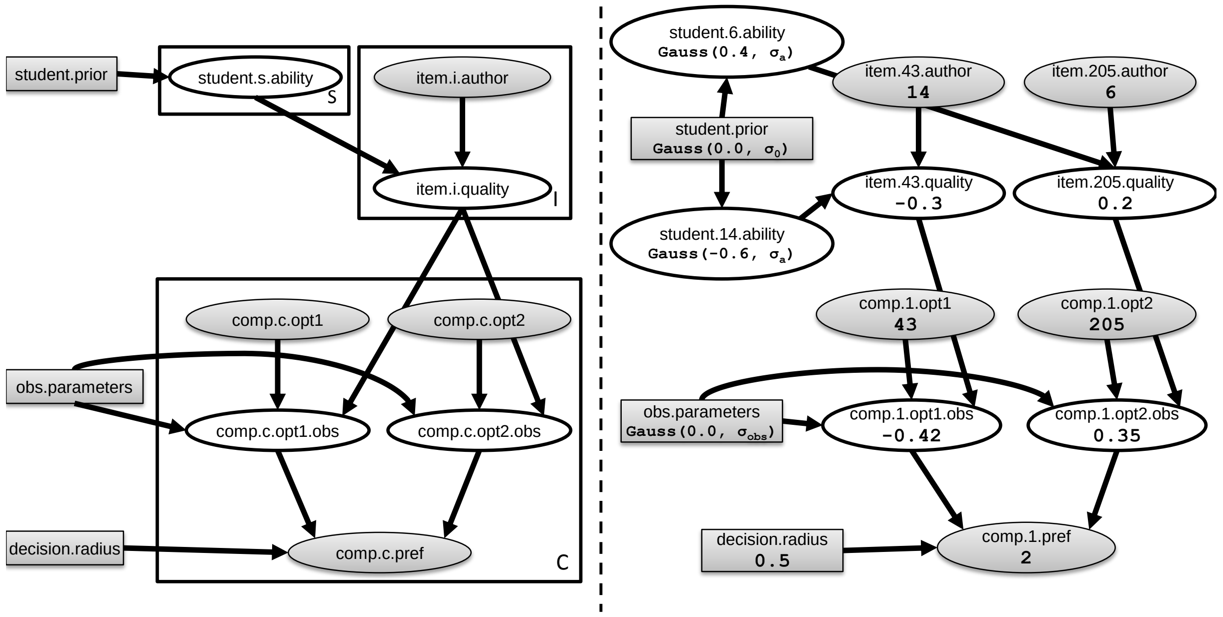

Figure 1: Plate notation (left) showing the independence structure of the IRT Models. Example instantiated subnetwork (right) for the Gaussian parameterization. Shaded rectangles are hyperparameters. Shaded ellipses are variables observable from a set of comparisons.

radius (which here is 0.5), then he states no preference (i.e. a preference of 0). Otherwise he prefers the item with the higher observed quality.

The categorical parameterization is similar to the Gaussian parameterization, with the following differences. Item quality is not continuous, but rather a member of the discrete set $\{1, 2, ..., \Lambda\}$. The student ability distributions are categorical distributions over $\{1, 2, ..., \Lambda\}$, and the student ability prior is a symmetric Dirichlet with strength α_a. Finally, the observed quality is the item quality λ plus an integer-valued noise $\nu \in \{1 - \lambda, ..., \Lambda - \lambda\}$. Noise ν is drawn from a discretized zero-mean Gaussian with standard deviation σ_{obs}. Specifically, $Pr(\nu)$ is proportional to the value of the probability density function of the zero-mean Gaussian $\mathcal{N}(0, \sigma_{obs})$.

We estimated the model parameters with Gibbs sampling (Geman and Geman, 1984). We found that Gibbs sampling converged quickly and consistently[10] for both parameterizations. Given the parameter estimates, we obtain a preference model $Q(\pi|s_1, s_2)$ through the inference query:

$$\begin{aligned} Pr(\mathsf{comp.c'.pref} = \pi \quad | \quad & \mathsf{item.i'.author} = s_1, \\ & \mathsf{item.i''.author} = s_2, \\ & \mathsf{comp.c'.opt1} = \mathsf{i'}, \\ & \mathsf{comp.c'.opt2} = \mathsf{i''}) \end{aligned}$$

where $\mathsf{c'}, \mathsf{i'}, \mathsf{i''}$ are new comparison and item ids that do not appear in the training data.

We call these models Item-Response Theoretic (IRT) models, to acknowledge their roots in the psychometrics (Thurstone, 1927; Bradley and Terry, 1952; Luce, 1959) and item-response theory (Hambleton, 1991; van der Linden and Hambleton, 1996; Baker, 2001) literature. Item-response theory is the basis of modern testing theory and drives adaptive standardized tests like the Graduate Record Exam (GRE). In particular, the Gaussian parameterization of our IRT models strongly resembles[11] the Thurstone (Thurstone, 1927) and Bradley-Terry-Luce (Bradley and Terry, 1952; Luce, 1959) models of paired comparison and the 1PL normal-ogive and Rasch (Rasch, 1960) models of student testing. From the testing perspective, we can view each comparison as two students simultaneously posing a test question to the other: "Give me a translation of the source text which is better than mine." The students can answer the question correctly, incorrectly, or they can provide a translation of analogous quality. An extra dimension of our models is judge noise, not a factor when modeling multiple-choice tests, for which the right answer is not subject to opinion.

[10] We ran 200 iterations with a burn-in of 50.

[11] These models are not traditionally expressed using graphical models, although it is not unprecedented (Mislevy and Almond, 1997; Mislevy et al., 1999).

lp	wmt10		wmt11		wmt12	
	train	test	train	test	train	test
ce	3166	2209	1706	3216	5969	6806
fe	5918	2376	2556	4430	7982	5840
ge	7422	3002	3708	5371	8106	6032
se	8411	2896	1968	3684	3910	7376
ec	10490	3048	8859	9016	13770	9112
ef	5720	2242	3328	5758	7841	7508
eg	10852	2842	5964	7032	10210	7191
es	2962	2212	4768	6362	5664	8928

Table 3: Dataset sizes for each competition track (number of comparisons).

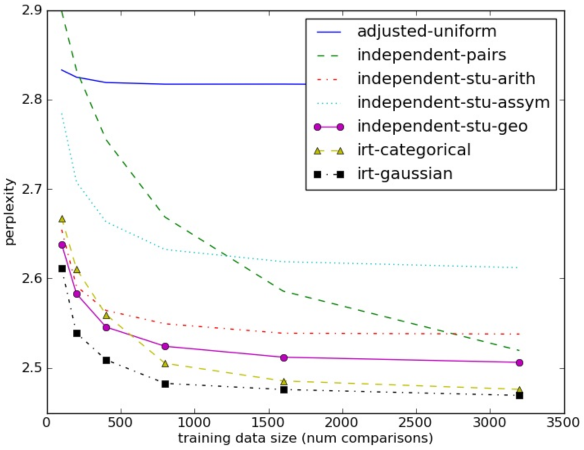

Figure 2: WMT10 model perplexities. The perplexity of the uniform preference model is 3.0 for all training sizes.

8 Experiments

We organized the competition data as described at the end of Section 4. To compare the preference models, we did the following:

- Randomly chose a subset of k comparisons from the training set, for $k \in \{100, 200, 400, 800, 1600, 3200\}$.[12]

- Trained the preference model on these comparisons.

- Evaluated the perplexity of the trained model on the test preferences, as described in Section 4.

For each model and training size, we averaged the perplexities from 5 trials of each competition track. We then plotted average perplexity as a function of training size. These graphs are shown in Figure 2 (WMT10)[13], and Figure 4 (WMT12). For WMT10 and WMT11, the best models were the IRT models, with the Gaussian parameterization converging the most rapidly and reaching the lowest perplexity. For WMT12, in which reference translations were excluded from the competition, four models were nearly indistinguishable: the two IRT models and the two averaged Independent Student models. This somewhat validates the organizers' decision to exclude the references, particularly given WMT's use of the BOJAR ranking heuristic (the nucleus of the Independent Student models) for its official rankings.

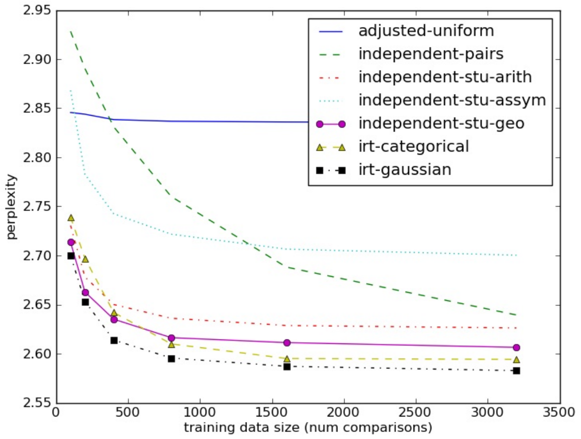

Figure 3: WMT11 model perplexities.

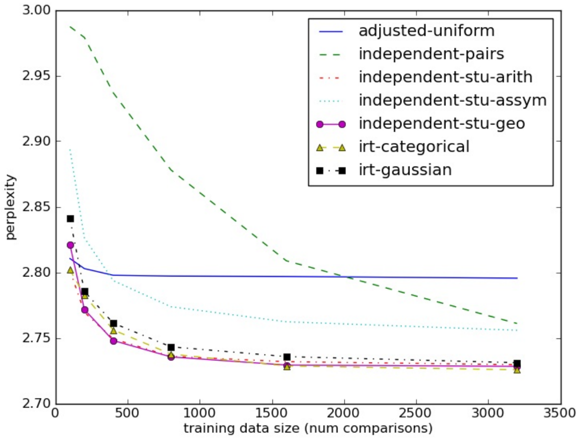

Figure 4: WMT12 model perplexities.

[12] If k was greater than the total number of training comparisons, then we took the entire set.

[13] Results for WMT10 exclude the German-English and English-German tracks, since we used these to tune our model hyperparameters. These were set as follows. The Dirichlet strength for each baseline was 1. For IRT-Gaussian: $\sigma_0 = 1.0, \sigma_{obs} = 1.0, \sigma_a = 0.5$, and the decision radius was 0.4. For IRT-Categorical: $\Lambda = 8, \sigma_{obs} = 1.0, \alpha_a = 0.5$, and the decision radius was 0.

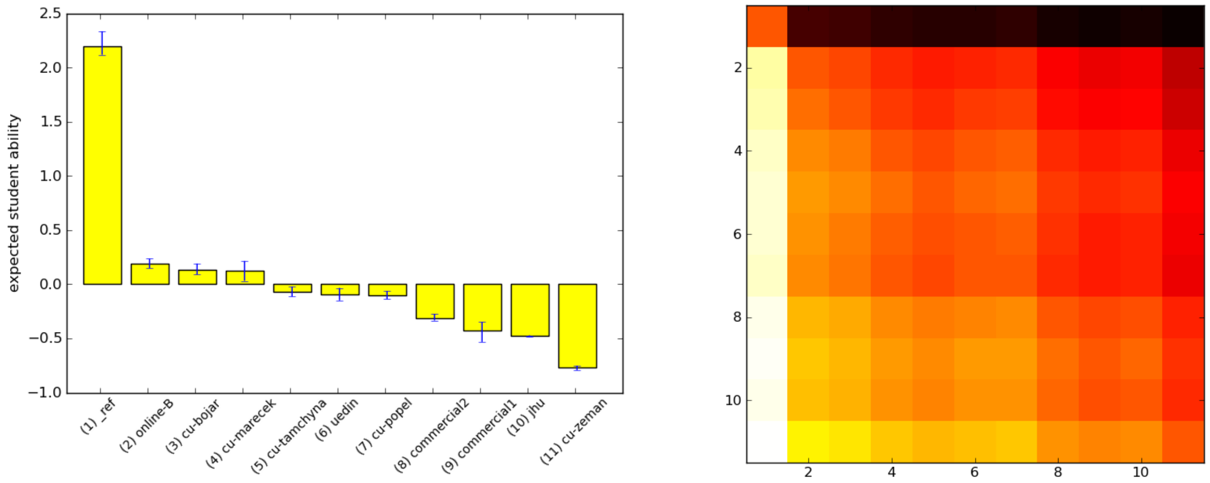

Figure 6: English-Czech WMT11 results (average of 5 trainings on 1600 comparisons). Error bars (left) indicate one stddev of the estimated ability means. In the heatmap (right), cell (s_1, s_2) is darker if preference model $Q(\pi|s_1, s_2)$ skews in favor of student s_1, lighter if it skews in favor of student s_2.

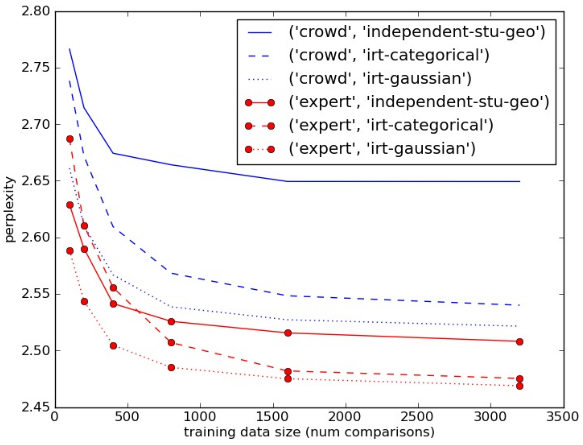

Figure 5: WMT10 model perplexities (crowdsourced versus expert training).

The IRT models proved the most robust at handling judge noise. We repeated the WMT10 experiment using the same test sets, but using the unfiltered crowdsourced comparisons (rather than "expert"[14] comparisons) for training. Figure 5 shows the results. Whereas the crowdsourced noise considerably degraded the Geometric Independent Students model, the IRT models were remarkably robust. IRT-Gaussian in particular came close to replicating the performance of Geometric Independent Students trained on the much cleaner expert data. This is rather impressive, since the crowdsourced judges agree only 46.6% of the time, compared to a 65.8% agreement rate among expert judges (Callison-Burch et al., 2010).

Another nice property of the IRT models is that they explicitly model student ability, so they yield a natural ranking. For training size 1600 of the WMT11 English-Czech track, Figure 6 (left) shows the mean student abilities learned by the IRT-Gaussian model. The error bars show one standard deviation of the ability means (recall that we performed 5 trials, each with a random training subset of size 1600). These results provide further insight into a case analyzed by (Lopez, 2012), which raised concern about the relative ordering of online-B, cu-bojar, and cu-marecek. According to IRT-Gaussian's analysis of the data, these three students are so close in ability that any ordering is essentially arbitrary. Short of a full ranking, the analysis does suggest four strata. Viewing one of IRT-Gaussian's induced preference models as a heatmap[15] (Figure 6, right), four bands are discernable. First, the reference sentences are clearly the darkest (best). Next come students 2-7, followed by the slightly lighter (weaker) students 8-10, followed by the lightest (weakest) student 11.

9 Conclusion

WMT has faced a crisis of confidence lately, with researchers raising (real and conjectured) issues with its analytical methodology. In this paper, we showed how WMT can restore confidence in

[14] I.e., machine translation specialists.

[15] In the heatmap, cell (s_1, s_2) is darker if preference model $Q(\pi|s_1, s_2)$ skews in favor of student s_1, lighter if it skews in favor of student s_2.

its conclusions – by shifting the focus from *rankings* to *relative ability*. Estimates of relative ability (the expected head-to-head performance of system pairs over a probability space of judges and source text) can be empirically compared, granting substance to previously nebulous questions like:

1. **Is my analysis better than your analysis?** Rather than the current anecdotal approach to comparing competition analyses (e.g. presenting example rankings that seem somehow wrong), we can empirically compare the predictive power of the models on test data.

2. **How much of an impact does judge noise have on my conclusions?** We showed that judge noise can have a significant impact on the quality of our conclusions, if we use the wrong models. However, the IRT-Gaussian appears to be quite noise-tolerant, giving similar-quality conclusions on both expert and crowdsourced comparisons.

3. **How many comparisons should I elicit?** Many of our preference models (including IRT-Gaussian and Geometric Independent Students) are close to convergence at around 1000 comparisons. This suggests that we can elicit far fewer comparisons and still derive confident conclusions. This is the first time a concrete answer to this question has been provided.

References

F.B. Baker. 2001. *The basics of item response theory*. ERIC.

Ondej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. 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.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O.F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

S. Geman and D. Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

R.K. Hambleton. 1991. *Fundamentals of item response theory*, volume 2. Sage Publications, Incorporated.

D. Koller and N. Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

S. Kullback and R.A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *Proceedings of WMT*.

R. Ducan Luce. 1959. *Individual Choice Behavior a Theoretical Analysis*. John Wiley and sons.

R.J. Mislevy and R.G. Almond. 1997. Graphical models and computerized adaptive testing. *UCLA CSE Technical Report 434*.

R.J. Mislevy, R.G. Almond, D. Yan, and L.S. Steinberg. 1999. Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*, pages 437–446. Morgan Kaufmann Publishers Inc.

G. Rasch. 1960. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Louis L Thurstone. 1927. A law of comparative judgment. *Psychological review*, 34(4):273–286.

W.J. van der Linden and R.K. Hambleton. 1996. *Handbook of modern item response theory*. Springer.