Simulating Human Judgment in Machine Translation Evaluation Campaigns

Philipp Koehn

School of Informatics
University of Edinburgh
pkoehn@inf.ed.ac.uk

Abstract
We present a Monte Carlo model to simulate human judgments in machine translation evaluation campaigns, such as WMT or IWSLT. We use the model to compare different ranking methods and to give guidance on the number of judgments that need to be collected to obtain sufficiently significant distinctions between systems.

1. Introduction
An important driver of current machine translation research are annual evaluation campaigns where research labs use the latest prototype of their system to translate a fixed test set, which is then ranked by human judges. Given the nature of the translation problem, where everybody seems to disagree on what the right translation of a sentence is, it comes of no surprise that the methods used to obtain human judgments and rank different systems against each other is also under constant debate.

This paper presents a Monte Carlo simulation that closely follows the current practice in the evaluation campaigns carried out for the Workshop on Statistical Machine Translation (WMT [1]), the International Workshop on Spoken Language Translation (IWSLT [2]), and to a lesser degree, since it mostly relies on automatic metrics, the Open Machine Translation Evaluation organized by NIST (OpenMT[1]).

The main questions we answer are: How many judgments do we need to collect to reach a reasonably definitive statement about the relative quality of submitted systems? Are we ranking systems the right way? How do we obtain proper confidence bounds for the rankings?

2. Related Work
While manual evaluation of machine translation systems has a rich history, most recent evaluation campaigns and lab-intern manual evaluations restrict themselves to a ranking task. A human judge is asked, if, for a given input sentence, she prefers output from system A over output from system B.

While this is a straight-forward procedure, the question how to convert these pairwise rankings into an overall ranking of several machine translation systems has recently received attention. Bojar et al. [3] critiqued the ongoing practice in the WMT evaluation campaigns, which was subsequently changed. Lopez [4] proposed an alternative method to rank systems. We will discuss these methods in more detail below.

An intriguing new development in human involvement in the evaluation of machine translation output is HyTER [5]. Automatic metrics suffer from the fact that a handful of human reference translations cannot expected to be matched by other human or machine translators, even if the latter are perfectly fine translations. The idea behind HyTER is to list all possible correct translations in the compact format of a recursive transition network (RTN). These networks are constructed by a human annotator who has access to the source sentence. Machine translation output is then matched against this network using string edit distance, and the number of edits is used as a metric.

Construction of the networks takes about 1–2 hours per sentence. This cost is currently too expensive for evaluations such as WMT with its annually renewed test set and eight language pairs. But we are hopeful that technical innovations, for instance in automatic paraphrasing, will bring down this cost to make it a more viable option in machine translation evaluation campaigns.

3. Model
We now define a model which consists of machine translation systems that produce translations of randomly distributed quality. We will make design decisions and set the only free parameter (the standard deviation of the systems’ quality distributions) to match statistics from the actual data of the WMT evaluation campaign.

In an evaluation, n systems S = {S1,...,Sn} participate. Each system produces translations with the average quality µn. When simulating an evaluation experiment, the quality µn of each system is chosen from a uniform distribution over the interval [0;10]. So, an experiment is defined by a list of average system qualities E = (µ1,...,µn).

Note: The range of the interval is chosen arbitrarily — the actual quality scores do not matter, only the relative scores of different systems. We use the uniform distribution to chose system qualities (opposed to, say, normal distribu-

1 http://www.nist.gov/itl/iad/mig/openmt.cfm
In each evaluation experiment $E$, a sample of human judgments $J_E$ is drawn. We follow here the procedure of the WMT evaluation campaign: We randomly select sets of 5 different systems $F_{E,i} = \{s_a, s_b, s_c, s_d, s_e\}$ with $1 \leq a, b, c, d, e \leq n$. Each system $j \in F_{E,i}$ produces a translation for the same input sentence, with a translation quality $q_{E,i,j}$ that is chosen from a normal distribution: $N(\mu_j, \sigma^2)$. Based on this set of translations, we extract a set of 10 ($= \frac{5 \times 4}{2}$) pairwise rankings $\{(j_1, j_2) | q_{E,i,j_1} > q_{E,i,j_2}\}$ and add them to the sample of human judgments $J_E$.

Note:
- The variance $\sigma^2$ is the same for all systems. We discuss at the end of this section how the value of the variance is set.
- This procedure may appear unnecessarily complex. We could have just picked two systems, draw translation qualities $q_{E,i,j}$ for each, compare them, and add a pairwise ranking to the judgment sample $J_E$. However, the WMT evaluation campaign follows the described procedure, because comparing a set of 5 systems at once yields 10 pairwise rankings faster than comparing 2 systems at a time, repeated 10 times. It is an open question, if the procedure adds distortions, so we match it in our model.
- The WMT evaluation campaign allows for ties. We ignore this in our model, since it adds an additional parameters (ratio of ties) that we would have to set. It is worth investigating, if allowing for ties changes any of our findings.
- Since it is not possible to tease apart the quality of the system and the perceived quality of a system by a human judge, we do not model the noise introduced by human judgment.

We still have to set the variance $\sigma^2$ which is used to draw translation quality scores $q$ for a translation systems $S_j$ with the average quality of $\mu_j$. We base this number on the ratio of system pairs that we can separate with statistically significance testing, as follows:

Given the sample of human judgments in form of pairwise system rankings $J_E = ((a_1, b_1), (a_2, b_2), \ldots)$ with $1 \leq a_i, b_i \leq n$, $a_i \neq b_i$, we can count how many times a system $S_j$ wins over another system $S_k$ in pairwise rankings: $\text{win}(S_j, S_k) = |\{(a_i, b_i) \in J_E | a_i = j, b_i = k\}|$ — and how many times it loses: $\text{loss}(S_j, S_k) = 1 - \text{win}(S_k, S_j)$. Given these two numbers, we can use the sign test to determine if system $S_j$ is statistically significantly better (or worse) than system $S_k$ at a desired $p$-level (we use $p$-level=0.05).

The more human judgments we have, the more systems we can separate. Figure 2 plots the ratio of system pairs (out of $\frac{n(n-1)}{2}$) that are different according to the sign test against the number of pairwise judgments for all 8 language pairs of the WMT12 evaluation campaign. The variance for our model, chosen to match these curves, ranges from 7 to 12.

### 4. Ranking Methods

There are several ways to use the (actual or simulated) pairwise judgment data $J_E$ to obtain assessments about the relative quality of the systems participating in a given evaluation campaign. We already encountered one such assessment: the statistically significantly better quality of one system over another another at a certain $p$-level according to the sign test. These assessments are reported in large tables in the WMT12 overview paper, but are somewhat unsatisfying because many system pairs are reported as not statistically significantly different.

Instead, we would like to report rankings of the systems. In this section, we will review two ranking methods proposed for this task, introduce a third one, and use our model to assess how often these ranking methods error.

#### 4.1. Bojar

In the recent 2012 WMT evaluation campaign, systems were ranked by the ratio of how often they were ranked better or equal to any of the other systems. Following the argument...
of Bojar et al. [3], this ignores ties and uses the definition of wins and loss as defined above, to compute a ranking score:

\[
\text{score}(S_j) = \frac{\sum_{k,k\neq j} \text{win}(S_j,S_k)}{\sum_{k,k\neq j} \text{win}(S_j,S_k) + \text{loss}(S_j,S_k)}
\] (1)

Systems were ranked by this number. This ranking method was used for the official ranking of WMT 2012. We refer to it here as BOJAR.

4.2. Lopez

Lopez [4] argues against using aggregate statistics over a set of very diverse judgments. Instead, a ranking that has the least number of pairwise ranking violations is said to be preferred. He defines a count function for pairwise order violations

\[
\text{score}(S_j, S_k) = \max(0, \text{win}(S_j, S_k) - \text{loss}(S_j, S_k))
\] (2)

Given a bijective ranking function \(R(j) \rightarrow j'\) with \(j, j' \in \{1, ..., n\}\) the total number of pairwise ranking violations is defined as

\[
\text{score}(R) = \sum_{j,k|R(S_j) < R(S_k)} \text{score}(S_j, S_k)
\] (3)

Finding the optimal ranking \(R\) that minimizes this score is not trivial, but given the number of systems involved in this evaluation campaign, it is manageable.

4.3. Expected Win

In BOJAR, systems are put at an disadvantage, if they are compared more frequently against good systems than against bad systems. We can overcome this by first computing the win ratios between each system pair and then averaging the ratios:

\[
\text{score}(S_j) = \frac{1}{n} \sum_{k,k\neq j} \frac{\text{win}(S_j, S_k)}{\text{win}(S_j, S_k) + \text{loss}(S_j, S_k)}
\] (4)

This score can also be understood as the expectation of a win against a randomly chosen opponent system.

4.4. Evaluation

The three methods above have been justified with an appeal to intuition. But now, with the model that we introduced in Section 3, we are able to run simulations that start with a gold standard ranking based on the systems’ average translation scores \(\mu_i\), generate judgment data, apply the ranking methods, and then check the obtained rankings according to the methods against the gold standard ranking.

We chose an experimental setup that reflects a typical situation in the WMT evaluation campaign, with \(n = 15\) systems and variance \(\sigma^2 = 10\). We randomly draw 10,000 experiments, sample human judgments for each and rank the systems based on the methods discussed in this section (BOJAR, LOPEZ, EXPECTED). We evaluate the rankings \(R_m\) obtained by each method \(m\) against the gold standard ranking \(R\) by computing the ratio of system pairs where the worst
Table 1: Quality of the confidence bounds obtained with the pairwise and bootstrap methods (see Section 5.1. The methods allow us to group the systems into clusters of comparable performance and indicate a range for the rank number in the rankings. Experiment with 15 systems, \( \sigma^2 = 10 \), and p-level 0.05, averaged over 400 runs.

| Judgments \(| J_E \)| | Pairwise Method | | Bootstrap Method |
|---|---|---|---|---|---|---|---|---|
| range size | violations | clusters | violations | range size | violations | clusters | violations |
| 10,000 | 8.1 | 0.8\% | 1.0 | 0\% | 4.6 | 3.4\% | 1.8 | 0.5\% |
| 20,000 | 6.3 | 0.8\% | 1.1 | 0\% | 3.7 | 2.4\% | 3.0 | 0.5\% |
| 30,000 | 5.4 | 0.7\% | 1.4 | 0\% | 3.3 | 2.3\% | 3.9 | 0.4\% |
| 40,000 | 4.9 | 0.9\% | 1.7 | 0.1\% | 3.0 | 2.0\% | 4.7 | 0.4\% |
| 50,000 | 4.5 | 0.9\% | 2.0 | 0.1\% | 2.9 | 2.1\% | 5.3 | 0.7\% |

5. Confidence Bounds

Reporting a definitive ranking hides the uncertainty about it. It is useful to also report, how confident we are that a particular system \( S_j \) is placed on rank \( r_j \). In this section, we aim to give this information in two forms:

- by determining the rank range \([r_j', \ldots, r_j'']\) into which the true rank of the system \( S_j \) falls with a given level of statistical significance, say, p-level 0.05.
- by grouping systems into clusters, to which each system belongs with a given level of statistical significance.

5.1. Methods

We now present two methods to produce this information, discuss how they can be evaluated, and report on experiments.

The first idea is to rely on the pairwise statistically significant distinctions that we can obtain by the sign test from the data. To give an example, if system \( S_j \) is significantly better than \( b = 9 \) systems, worse than \( w = 2 \) systems and indistinguishable from \( e = 3 \) systems, then its rank range is 3–6 (from \( w + 1 \) to \( w + 1 + e \)).

The second idea is to apply bootstrap resampling [6]. Given a fixed set of judgments \( J_E \), we sample pairwise rankings from this set (allowing for multiple drawings of the same ranking). We then compute a ranking with the expected win method based on this resampling. We repeat this process a 1000 times, record each time the rank of a system \( S_j \). We then sort the obtained 1000 ranks, chop off the top 25 and bottom 25 ranks and report the minimum interval containing the remaining ranks as rank range.

Clusters are obtained by grouping systems with overlapping rank ranges. Formally, given ranges defined by start\((S_j)\) and end\((S_j)\), we seek the largest set of clusters \( \{C_c\} \) that satisfies:

\[
\forall S_j \exists C_j : S_j \in C_j
\]

\[
S_j \in C_j, S_j \in C_k \rightarrow C_j = C_k
\]

\[
C_j \neq C_k \rightarrow \forall S_j \in C_j, S_k \in C_k : start(S_j) > end(S_k) \lor start(S_k) > end(S_j)
\]

5.2. Evaluation

We can measure the performance of the confidence bound estimation methods by the tightness of the rank ranges, the number of clusters, and the number of violations for each — a violation happens when the true rank of a system falls outside the rank range or if a system is placed in a cluster with a truly higher ranked system placed into a lower cluster or vice versa.

See Table 1 for results of an experiment with the same settings as above (variance \( \sigma^2 = 10 \), number of systems \( n = 15 \)). The bootstrap resampling method yields smaller rank range sizes (about half) and a larger number of clusters (2–3 times as many). This does come at the cost of increased error, but note that the measured error is well below the statistical significance p-level of 0.05 used to run the bootstrap. If lower error is desired, smaller p-levels may be used.

Table 2 and 3 show the application of the method to two language pairs of the WMT12 evaluation campaign. In the
first example (English–Czech, $\sigma^2 = 9$, $n = 13$, 22,000 judgments) we see a nice separation into 6 clusters, while in the second example (French–English, $\sigma^2 = 10$, $n = 15$, 13,000 judgments) almost all systems are in the same cluster. Our findings in Table 1 suggest that collecting 30,000 judgments would allowed us to separate the systems into about 4 clusters, with each system ranging over only 3 ranks.

### 6. How Many Judgements?

A very practical question that we are trying to answer in this paper is: When we run a manual evaluation, how many judgments do we need to collect?

The answer to this questions depends on how many systems participate in the evaluation and the desired level of certainty — the first number is readily available and the second can be chosen at will. But the answer also depends on the variance $\sigma^2$ of the systems. This is a number that will become only clearer once a large number of judgments have been collected. The findings from the WMT12 evaluation campaign gives some guidance about the value of $\sigma^2$ — numbers between 8 and 12 seem to cover most cases.

Armed with these specifics, Table 4 gives an estimate about the minimum number of judgments required. For instance, for the WMT12 French–English pair ($n = 15$, $\sigma^2 = 10$), the organizers collected 13,000 judgments. This was sufficient to tell about 70% of pairs apart. To raise that number to 80%, about 40,000 judgments are required.

Note that we computed the number in the table with a grid search over the number of judgments, so all numbers are approximate.

### 7. Conclusions

We introduced a Monte Carlo model for the simulation of the methodology underlying current machine translation evalu-
ation campaigns. We used the model to compare different ranking methods, introduced methods to obtain confidence bounds and give guidance on the number of judgment to be collected to obtain satisfying results. The findings show that recent WMT evaluation campaigns do not collect sufficient judgments and that the number of judgments should be doubled or increased three-fold.

8. Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 287658 (EU BRIDGE) and agreement 288487 (MosesCore).

9. References


