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 (OpenMT1).

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-internal 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 rank-

1http://www.nist.gov/itl/iad/mig/openmt.cfm
tation) because this reflects the data from the WMT evaluation campaigns (see Figure 1).

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 (= \( 5 \times 2 \)) 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 \( q_{i,a} \) 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 err.

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

| Language Pair          | Win Ratio (J_E) | Win Ratio (J_E) |
|------------------------|----------------|----------------|
| Spanish-English        | 0.6            | 0.5            |
| English-Spanish        | 0.5            | 0.4            |
| German-English         | 0.4            | 0.3            |
| English-German         | 0.3            | 0.2            |
| French-English         | 0.2            | 0.1            |
| English-French         | 0.1            | 0.0            |
| Czech-English          | 0.0            | 0.0            |
| English-Czech          | 0.0            | 0.0            |

Figure 1: Win ratios of the systems in the WMT12 evaluation campaign. Except for the occasional outlier at the low end, the systems follow roughly a uniform distribution. For details on the computation of the win ratios see Section 4.3, our experiments show that uniformly distributed average system qualities lead to uniformly distributed win ratios.
Finding the optimal ranking

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

Systems were ranked by this number. This ranking
method was used for the official ranking of WMT 2012. We
referred 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

score(S_j, S_k) = \max(0, win(S_j, S_k) - loss(S_j, S_k))  \quad (2)

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

score(R) = \sum_{j, k | R(S_j) < R(S_k)} score(S_j, S_k)  \quad (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:

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)}  \quad (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, 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 situ-
ation in the WMT evaluation campaign, with n = 15 sys-
tems and variance \sigma^2 = 10. We randomly draw 10,000 experi-
ments, sample human judgments for each and rank the
systems based on the methods discussed in this section (BO-
JAR, LOPEZ, EXPECTED). We evaluate the rankings R_m for-
tained by each method m against the gold standard ranking
R by computing the ratio of system pairs where the worst
### 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** \([r_j', r_j'']\) 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 \( \text{start}(S_j) \) and \( \text{end}(S_j) \), we seek the largest set of clusters \( \{C_c\} \) that satisfies:

\[
\forall S_j \exists C_j : S_j \in C_j \\
S_1 \in C_j, S_j \in C_k \rightarrow C_j = C_k \\
C_j \neq C_k \rightarrow \forall S_j \in C_j, S_k \in C_k : \\
\text{start}(S_j) > \text{end}(S_k) \text{ or } \text{start}(S_k) > \text{end}(S_j)
\]

#### 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 a 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

---

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 | Pairwise Method | Bootstrap Method |
|-----------|-----------------|------------------|
| \(|J_E|\)  | 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%       |

Figure 3: Errors of the different ranking methods discussed in Section 4: Ratio of system pairs where the worst system is ranked better.

The 9th International Workshop on Spoken Language Translation
Hong Kong, December 6th-7th, 2012
Table 2: Application of our methods to the WMT12 English–Czech evaluation: The 13 systems are split into 6 clusters. About 22,000 judgments were collected.

| Rank | Range | Score | System       |
|------|-------|-------|--------------|
| 1    | 1–3   | 0.626 | LIMSI        |
| 2    | 1–4   | 0.610 | KIT          |
| 3    | 1–5   | 0.592 | ONLINE-A     |
| 4    | 2–6   | 0.571 | CMU          |
| 5    | 3–7   | 0.567 | ONLINE-B     |
| 6    | 4–8   | 0.538 | UEDIN        |
| 7    | 5–8   | 0.522 | LIUM         |
| 8    | 6–9   | 0.510 | RWTH         |
| 9    | 8–12  | 0.463 | RBMT-1       |
| 10   | 9–13  | 0.458 | RBMT-3       |
| 11   | 9–14  | 0.444 | SFU          |
| 12   | 9–14  | 0.441 | UK           |
| 13   | 10–14 | 0.430 | RBMT-4       |
| 14   | 12–14 | 0.409 | JHU          |
| 15   | 15    | 0.319 | ONLINE-C     |

Table 2: Compare to Table 2: In this example, only the last system was split off from the main cluster. Only about 13,000 judgments were collected. Our findings suggest that collecting 30,000 judgments would allow us to break up the systems into about 4 clusters, with each system ranging over only 3 ranks.

Table 3: Guidance on how many pairwise judgments must be collected to obtain a certain ratio of statistically significant (p-level 0.05) distinctions for pairs of systems. In the WMT12 campaign 10,000–20,000 judgments were collected.

| n   | \( \sigma^2 \) | Ratio of significant pairs |
|-----|---------------|-----------------------------|
|     |               | 50% | 70% | 80% | 90% |
| 6   | 8             | 1k  | 4k  | 8k  | 30k |
| 6   | 10            | 2k  | 5k  | 10k | 45k |
| 6   | 12            | 2k  | 7k  | 20k | 60k |
| 8   | 8             | 2k  | 6k  | 14k | 60k |
| 8   | 10            | 3k  | 8k  | 20k | 90k |
| 8   | 12            | 4k  | 14k | 35k | 140k|
| 10  | 8             | 4k  | 10k | 25k | 100k|
| 10  | 10            | 5k  | 16k | 40k | 150k|
| 10  | 12            | 6k  | 20k | 50k | 200k|
| 12  | 8             | 5k  | 15k | 35k | 140k|
| 12  | 10            | 7k  | 25k | 60k | 250k|
| 12  | 12            | 9k  | 35k | 80k | 350k|
| 15  | 8             | 8k  | 25k | 50k | 200k|
| 15  | 10            | 12k | 40k | 80k | 350k|
| 15  | 12            | 15k | 50k | 120k| 500k|

Table 4: Guidance on how many pairwise judgments must be collected to obtain a certain ratio of statistically significant (p-level 0.05) distinctions for pairs of systems. In the WMT12 campaign 10,000–20,000 judgments were collected.
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

[1] C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia, “Findings of the 2012 workshop on statistical machine translation,” in Proceedings of the Seventh Workshop on Statistical Machine Translation, Montreal, Canada: Association for Computational Linguistics, June 2012, pp. 10–48. [Online]. Available: http://www.aclweb.org/anthology/W12-3102

[2] M. Paul, M. Federico, and S. Stücker, “Overview of the IWSLT 2010 Evaluation Campaign,” in Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT), M. Federico, I. Lane, M. Paul, and F. Yvon, Eds., 2010, pp. 3–27.

[3] O. Bojar, M. Ercegovčević, M. Popel, and O. Zaidan, “A grain of salt for the wmt manual evaluation,” in Proceedings of the Sixth Workshop on Statistical Machine Translation. Edinburgh, Scotland: Association for Computational Linguistics, July 2011, pp. 1–11. [Online]. Available: http://www.aclweb.org/anthology/W11-2101

[4] A. Lopez, “Putting human assessments of machine translation systems in order,” in Proceedings of the Seventh Workshop on Statistical Machine Translation. Montreal, Canada: Association for Computational Linguistics, June 2012, pp. 1–9. [Online]. Available: http://www.aclweb.org/anthology/W12-3101

[5] M. Dreyer and D. Marcu, “Hyter: Meaning-equivalent semantics for translation evaluation,” in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 162–171. [Online]. Available: http://www.aclweb.org/anthology/N12-1017

[6] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. Chapman and Hall, 1993.