# Re-assessing the WMT2013 Human Evaluation with Professional Translators Trainees

**Mihaela Vela**
Saarland University
m.vela@mx.uni-saarland.de

**Josef van Genabith**
German Research Center for Artificial Intelligence
Josef.Van_Genabith@dfki.de

## Abstract

This paper presents experiments on the human ranking task performed during WMT2013. The goal of these experiments is to re-run the human evaluation task with translation studies students and to compare the results with the human rankings performed by the WMT development teams during WMT2013. More specifically, we test whether we can reproduce, and if yes to what extent, the WMT2013 ranking task and whether specialised knowledge from translation studies influences the results in terms of intra- and inter-annotator agreement as well as in terms of system ranking. We present two experiments on the English-German WMT2013 machine translation output. Analysis of the data follows the methods described in the official WMT2013 report. The results indicate a higher inter- and intra-annotator agreement, less ties and slight differences in ranking for the translation studies students as compared to the WMT development teams.

## 1 Introduction

Machine translation evaluation is an important element in the process of building MT systems. The Workshop for Statistical Machine Translation (WMT) compares new techniques for MT through human and automatic MT evaluation and provides also tracks for evaluation metrics, quality estimation of MT as well as post-editing of MT.

To date, the most popular MT evaluation metrics essentially measure lexical overlap between reference and hypothesis translation such as IBM BLEU (Papineni et al., 2002), NIST (Doddington, 2002), Meteor (Denkowski and Lavie, 2014), WER (Levenshtein, 1966), position-independent error rate metric PER (Tillmann et al., 1997) and the translation edit rate metric TER (Snover et al., 2006) and TERp (Snover et al., 2009). Gonzàlez et al. (2014) as well as Comelles and Atserias (2014) introduce their fully automatic approaches to machine translation evaluation using lexical, syntactic and semantic information when comparing the machine translation output with reference translations.

Human machine translation evaluation can be performed with different methods. Lo and Wu (2011) propose HMEANT, a metric based on MEANT (Lo et al., 2012) that measures meaning preservation between hypothesis and reference translation on the basis of verb frames and their role fillers. Another method is HTER (Snover et al., 2006) which produces targeted reference translations by post-editing MT output. Another method is HTER (Snover et al., 2006) which produces targeted reference translations by post-editing MT output. Human evaluation can also be performed by measuring post-editing time, or by asking evaluators to assess the fluency and adequacy of a hypothesis translation on a Likert scale. Another popular human evaluation method is ranking: ordering a set of translation hypotheses according to their quality. This is also the method applied during the recent WMTs, where humans are asked to rank machine translation output by using APPRAISE (Federmann, 2012), a software tool that integrates facilities for such a ranking task. In WMT, human MT evaluation is carried out by the MT development teams, usually computer scientists or computational linguists, sometimes involving crowd-sourcing based on Amazon's Mechanical Turk.

Being aware of the two communities, machine translation and translation studies, we took the available online data from the WMT2013[1] and tried to reproduce the ranking task with translation studies students for the English to German translations. The three questions we want to answer are:

- Can we reproduce at all the WMT2013 results for the language pair English-German?

- Are translation studies students (future translators) evaluating different from the WMT development teams, or in other words does specialised knowledge from translation studies influence the outcome of the ranking task?

- Are translation studies students more consistent as a group and with themselves in terms of intra- and inter-agreement?

We concentrate on English-German data since the majority of our evaluators were native speakers of German and since, from a translation studies point of view, professional translation should be performed only into the mother tongue.

## 2 The WMT2013 English-German Data

Before presenting the experimental setting and outcomes, we present the WMT data. We are aware of the fact that the main objective of the WMT is to evaluate the state-of-the-art in machine translation. In this context evaluation plays an important role, since a robust and reliable evaluation method makes it easier to perform a more in-depth differentiation between different machine translation outputs.

In 2013 during the WMT human evaluation campaign, the evaluation was performed both by the WMT development teams (further named researchers) and by turkers. The researcher group comprised all the participants in the WMT machine translation task. The turkers group was composed of non-experts on Amazon's Mechanical Turk (MTurk). Both groups were asked to rank randomly selected machine translation outputs, organised as quintuples of 5 outputs produced by different MT systems. The researchers were asked to rank quintuples for 300 source sentences whereas the turkers were paid per MTurk unit. Such a unit is called a human intelligence Task (HIT) and consisted of three source sentences and the corresponding quintuples. For each HIT turkers were paid $0.25.

In our experiments we focus on the language pair English-German, we compare our results with those obtained in the English-German human evaluation task. We concentrate on the evaluation performed by researchers, assuming that translation studies students will be at least as consistent as researchers and having in mind that intra- and inter-annotator agreement for the turkers' group was lower than for the researchers' group. Researchers are a well defined group, or at least a better defined group, than the turkers about whom we had no information.

From the WMT2013 English-German data, which we took as reference for our experiments, we observed that there were in total 38 researchers taking part in the English-German manual evaluation task. The range of the evaluated source sentences and their quintuples is from 3 to 1059. From the 38 evaluators 12 evaluated the same sentences more than once, the range in this case being from 3 to 240 repeated sentences. From here we can conclude that for the English-German task just 12 researchers can be considered for the intra-annotator agreement. The sentence overlap between researchers (relevant for the inter-annotator agreement) has also a wide range: from sentences evaluated in common with 2 researchers to sentences evaluated in common with 36 researchers. In total the researchers in WMT2013 produced 39582 ranking pairs, without counting ties, based on which the final agreement scores and the system ranking was computed.

Another observation from the WMT2013 data is related to the systems researchers had to rank. The data shows that researchers ranked only 14 out of the 21 participating systems. The anonymised commercial and online systems were excluded from the human evaluation task.

The main criticism towards this kind of evaluation of MT output is that the evaluation does not provide evidence of the absolute quality of the MT output, but evidence of the quality of a machine translation system compared to other MT systems. If the evaluators had to decide on the ranking of 5 bad MT outputs, it might happen that even the MT system ranked first, scores bad in terms of adequacy and fluency. On the other hand, in such ranking tasks the specific skills, required for example in translation studies, are not necessary activated, since the ranking task is in fact a comparison task. Therefore, we assume that researchers and

[1] http://www.statmt.org/wmt13/

translations studies students will achieve at least comparable scores since no task-specific knowledge is required and the two groups, different from the turkers' group, can be considered homogeneous groups.

## 3 Experimental Design

We conducted the experiments as similar as possible to the manual ranking task in WMT2013. Like in WMT2013, evaluators were presented with a source sentence, a reference translation and five outputs produced by five anonymised and randomised machine translations systems. The instructions for the evaluators remained the same as in WMT2013:

> *You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed)*

For performing the ranking task we implemented the Java-based ranking tool depicted in Figure 1.[2] Similar to APPRAISE (Federmann, 2012) the ranking can be performed on a scale from 1 to 5, with 1 being the best translation and 5 being the worst translation.

For a given source sentence, each ranking of the five MT outputs has the potential to produce 10 ranking pairs. Before applying the corresponding formulas on the data, the ranking pairs from all evaluators and for all systems are collected in a matrix like the one in Table 1. The matrix records the number of times system $\mathrm{S_i}$ was ranked better than $\mathrm{S_j}$ and vice-versa.

For example, if we look at the two systems $\mathrm{S_1}$ and $\mathrm{S_3}$ in the matrix, we can see that $\mathrm{S_3}$ was ranked 2 times higher (from the left triangle) and 4 times lower (from the right triangle) than system $\mathrm{S_1}$.

From the matrix, the final score for each system - as defined by Koehn (2012) and applied in WMT2013 - can be computed. From the matrix in Table 1 the score for system $\mathrm{S_1}$ is computed by counting for each pair of systems ($\mathrm{S_1}$, $\mathrm{S_2}$), ($\mathrm{S_1}$, $\mathrm{S_3}$), ($\mathrm{S_1}$, $\mathrm{S_4}$), ($\mathrm{S_1}$, $\mathrm{S_5}$) the number of times $\mathrm{S_1}$ was ranked higher than the other system divided by the total number of rankings for each pair. The results for each pair of systems including $\mathrm{S_1}$ are then summed and divided by the number of systems, this being the final score for $\mathrm{S_1}$.

|  | $\mathrm{S_1}$ | $\mathrm{S_2}$ | $\mathrm{S_3}$ | $\mathrm{S_4}$ | $\mathrm{S_5}$ |
|---|---|---|---|---|---|
| $\mathrm{S_1}$ | 0 | 3 | 4 | 2 | 2 |
| $\mathrm{S_2}$ | 0 | 0 | 1 | 0 | 1 |
| $\mathrm{S_3}$ | 2 | 2 | 0 | 2 | 2 |
| $\mathrm{S_4}$ | 4 | 3 | 4 | 0 | 5 |
| $\mathrm{S_5}$ | 1 | 2 | 1 | 1 | 0 |

Table 1: Representation of the ranking pairs as a matrix

Considering having a system $\mathrm{S_i}$ from a set of systems S of size k and a set of rankings for each system pair ($\mathrm{S_i}$, $\mathrm{S_j}$), where $j = 1 \dots k$, $S_{\mathrm{j}} \in S$ and $i \neq j$ the score for $\mathrm{S_i}$ is defined as follows:

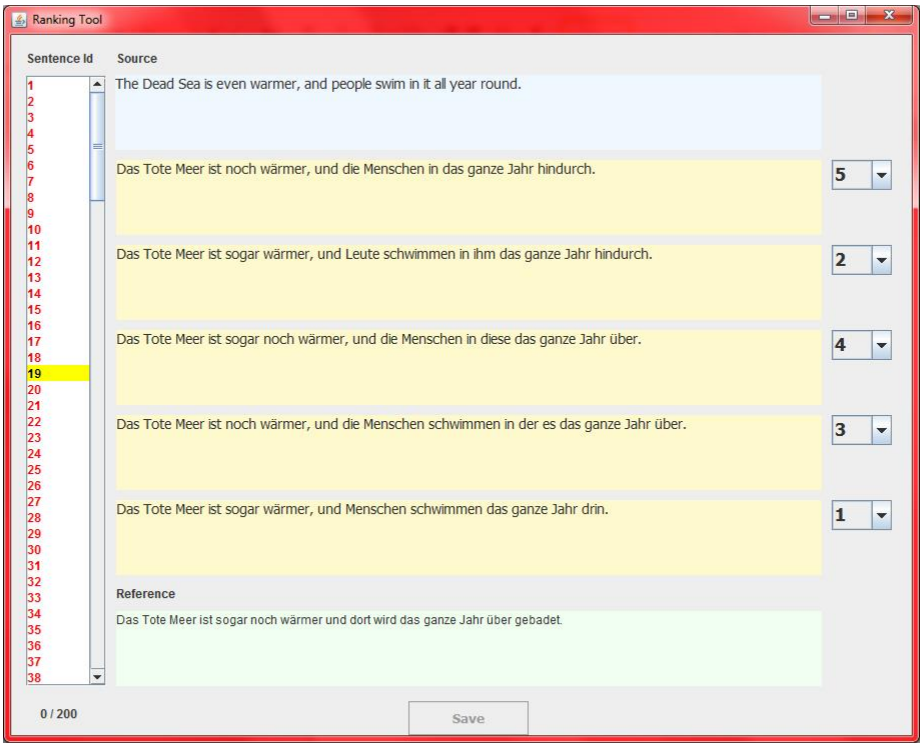
$$score(S_i) = \frac{1}{k} \sum_{i,j \neq i}^{k} \frac{\mid S_{\mathrm{i}} > S_{\mathrm{j}} \mid}{\mid S_{\mathrm{i}} > S_{\mathrm{j}} \mid + \mid S_{\mathrm{i}} < S_{\mathrm{j}} \mid}$$

Based on Koehn's (2012) formula each system gets a score and a ranking among the set of systems. After performing the ranking the systems are clustered by using bootstrap resampling, thus returning the final score and the cluster for each system.

Different from WMT2013 we run two evaluation rounds for the ranking task. The first round was a pilot study on which all evaluators had to evaluate the same set of randomised and anonymised sentences selected from the published WMT2013 ranking task data set. The set contained 200 source sentences and five anonymised and randomised MT outputs for each source sentence. In the pilot study we selected, as in WMT2013, only the above mentioned 14 machine translation systems for evaluation, disregarding the remaining anonymised commercial and online systems.

Regarding the sampling of the data, the second evaluation round followed the ranking task performed in WMT2013: each evaluator ranked a different randomised and anonymised sample consisting of 200 source sentences and five anonymised and randomised MT outputs for each source sentence. The individual samples were built out of all 21 machine translations outputs of the 3000 source sentences provided for the translation task.

[2] The implementation of a new tool was motivated by the accessibility of a server for the evaluators. This way each evaluator had his own evaluation set containing both the tool and the data set.



Source

The Dead Sea is even warmer, and people swim in it all year round.

Das Tote Meer ist noch wärmer, und die Menschen in das ganze Jahr hindurch. 5

Das Tote Meer ist sogar wärmer, und Leute schwimmen in ihm das ganze Jahr hindurch. 2

Das Tote Meer ist sogar noch wärmer, und die Menschen in diese das ganze Jahr über. 4

Das Tote Meer ist noch wärmer, und die Menschen schwimmen in der es das ganze Jahr über. 3

Das Tote Meer ist sogar wärmer, und Menschen schwimmen das ganze Jahr drin. 1

Reference

Das Tote Meer ist sogar noch wärmer und dort wird das ganze Jahr über gebadet.

Figure 1: The Java-based ranking tool.

### 3.1 The Pilot Study

During the pilot study, the translation studies students had to manually rank 200 source sentences and their corresponding randomised and anonymised 5 translations. The specifics of the pilot was that each evaluator received the same data set for evaluation. In fact we randomly retrieved 180 sentences and their 5 corresponding machine translation outputs from the WMT2013 manual evaluation data set, from the rankings performed by the researchers. Out of the 180 sentences we randomly selected 20 sentences which were repeated in the data set. Based on the 200 source sentences, out of which 10% were repeated, we could compute both the inter-annotator agreement and the intra-annotator agreement. For the inter-annotator agreement we took all 200 sentences into consideration, whereas for the intra-annotator agreement we considered the preselected 20 sentences which were repeated in the data set.

During the pilot study 25 translation students and a translation lecturer took part in the experiment. Except for three students, the remaining 23 evaluators were native speakers of German with at least a B2 level[3] for English. The three non-native speakers of English had at least a C1 knowledge level of German and B2 for English. Out of the 26 evaluators 14 completed the task by ranking the quintuples for all 200 source sentences, the remaining group evaluated between 2 and 26 source sentences. In total we collected 25780 ranking pairs in the pilot study.

Based on the collected rankings the intra-annotator agreement could be computed just for 17 evaluators, the ones who evaluated sentences more than once. On the other hand, the inter-agreement was computed pairwise between all evaluators, the fact that all evaluators received the same set of sentences made this possible.

Both types of agreement (intra and inter) were measured by computing Cohen's kappa coefficient (Cohen, 1960), as it was defined by Bojar et al. (2013)

$$\kappa = \frac{P_{\text{agree}}(S_{\text{i}}, S_{\text{j}}) - P_{\text{chance}}(S_{\text{i}}, S_{\text{j}})}{1 - P_{\text{chance}}(S_{\text{i}}, S_{\text{j}})} \qquad (1)$$

where $\mathrm{P}_{\text{agree}}(\mathrm{S}_{\text{i}},\mathrm{S}_{\text{j}})$ is the proportion of times that evaluators agree on the ranking of the systems $\mathrm{S}_{\text{i}}$ and $\mathrm{S}_{\text{j}}$ ($\mathrm{S}_{\text{i}} < \mathrm{S}_{\text{j}}$ or $\mathrm{S}_{\text{i}} = \mathrm{S}_{\text{j}}$ or $\mathrm{S}_{\text{i}} > \mathrm{S}_{\text{j}}$) and

[3] http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages#Common_reference_levels

$\mathrm{P_{chance}(S_i,S_j)}$ is the number of times they agree by chance. $\mathrm{P_{chance}(S_i,S_j)}$ itself is defined as

$$P_{\text{chance}}(S_{\text{i}}, S_{\text{j}}) =$$
$$P(S_{\text{i}} > S_{\text{j}})^2 + P(S_{\text{i}} = S_{\text{j}})^2 + P(S_{\text{i}} < S_{\text{j}})^2 \tag{2}$$

Table 2 list the values for $\mathrm{P_{agree}}$, $\mathrm{P_{chance}}$ and $\kappa$. The final $\kappa$ is then the arithmetic mean of the fourth column, resulting in an overall intra-annotator agreement of 0.745 as compared to 0.649 during WMT2013.

| User | $\mathrm{P_{agree}}$ | $\mathrm{P_{chance}}$ | $\kappa$ |
|---|---|---|---|
| uds1 | 1.000 | 0.431 | 1.000 |
| uds2 | 0.915 | 0.387 | 0.861 |
| uds3 | 0.674 | 0.157 | 0.613 |
| uds4 | 0.661 | 0.148 | 0.602 |
| uds5 | 1.000 | 0.360 | 1.000 |
| uds6 | 0.746 | 0.271 | 0.651 |
| uds7 | 0.710 | 0.199 | 0.637 |
| uds8 | 0.638 | 0.142 | 0.578 |
| uds9 | 1.000 | 0.467 | 1.000 |
| uds10 | 0.520 | 0.095 | 0.469 |
| uds11 | 0.974 | 0.392 | 0.957 |
| uds12 | 0.884 | 0.373 | 0.815 |
| uds13 | 0.792 | 0.302 | 0.702 |
| uds14 | 0.710 | 0.172 | 0.649 |
| uds15 | 0.792 | 0.302 | 0.702 |
| uds19 | 0.900 | 0.352 | 0.845 |
| uds25 | 0.666 | 0.190 | 0.579 |

Table 2: Intra-annotator agreement for the pilot study.

For the inter-annotator agreement $\kappa$ is computed by comparing each evaluator with other evaluators with whom she/he shared sentences in the ranking task. Each evaluator has been compared with the other 25 evaluators, the pairwise comparison of the 26 evaluators resulting in 325 evaluators pairs. For each of these pairs we calculated Cohen's $\kappa$, the overall inter-annotator agreement being the arithmetic mean from the inter-annotator agreement of the evaluator pairs. In the pilot study the inter-annotator agreement achieved a value of 0.494 as compared to 0.454 during WMT2013.

The system scores were calculated according to Koehn (2012). The results are listed in Table 3. In this stage we performed no clustering, since the experiments with bootstrap resampling have shown, that the cluster varied a lot depending on the sample size. Since we had no information about the sample size during bootstrap resampling performed during WMT2013 and because we collected less rankings (25780 vs. 39582 during WMT2013), we stopped here with the computation of system rankings.

| Rank | Score | System |
|---|---|---|
| 1 | 0.647 | PROMT |
| 2 | 0.572 | UEDIN-SYNTAX |
| 3 | 0.546 | ONLINE-B |
| 4 | 0.516 | LIMSI-SOUL |
| 5 | 0.505 | STANFORD |
| 6 | 0.504 | UEDIN |
| 7 | 0.490 | KIT |
| 8 | 0.462 | CU-ZEMAN |
| 9 | 0.456 | TUBITAK |
| 10 | 0.453 | MES-REORDER |
| 11 | 0.404 | JHU |
| 12 | 0.331 | SHEF-WPROA |
| 13 | 0.314 | RWTH-JANE |
| 14 | 0.294 | UU |

Table 3: System ranking in the pilot study without bootstrap resampling

The pilot study proved that performing the re-ranking of the English to German MT output from WMT2013 is a feasible task. Moreover, the $\kappa$ scores indicate that translation studies students are more consistent when ranking MT output.

### 3.2 Main Study

In the main phase of our re-ranking experiment each evaluator received a different sample consisting of 200 source sentences, the reference translation for each source sentence and five anonymised and randomised machine translation outputs. Because we sampled the data from the 3000 source sentences and the 21 available system outputs, during the main study we collected information about all systems and ignored the fact, that in WMT2013 evaluators were shown only preselected systems. The software as well as the requirements for performing the ranking task remained the same as in the pilot study.

Similar to the pilot study, in each sample consisting of the 200 source sentences and the corresponding 5 machine translation outputs 10% of the data was repeated, in order to compute the intra-annotator agreement. For inter-annotator agreement we selected 20 source sentences and their corresponding reference translation as well as the corresponding 5 machine translation outputs which were common to each sample. In this phase we had 37 evaluators, all of them being 2nd or 3rd BA translation studies students. With the exception of 3 students, all of the students were native speakers of German with at least a B2 level of English. The three non-native speaker of German had a C1 level of English. From the 37 students, 19 ranked all 200 sentences completing the task. The other 18 students ranked between between 20 and 60 sentences. From all the rankings performed by the evaluators in the main study we collected 37318 ranking pairs[4], a comparable number to the 39582 ranking pairs collected during WMT2013.

From the collected data we computed Cohen's $\kappa$ for the intra-annotator agreement based on the rankings collected from 22 evaluators. We obtain a $\kappa$ of 0.772 for the intra-annotator agreement. From all possible pairs of evaluators, here 666, only 536 pairs had ranked sentences in common and had therefore an inter-annotator $\kappa$ greater than 0. The arithmetic mean of these pairs gave us the overall inter-annotator agreement resulting in $\kappa$ of 0.510.

Since in the second run of the experiment we collected almost the same number of ranking pairs as during WMT2013, we performed the ranking of the systems with and without bootstrap resampling. Table 4 lists the ranking scores without bootstrap resampling.

For bootstrap resampling we sampled from the set of pairwise rankings ($S_i$, $S_j$) collected from all evaluators and computed the score for each system with the formula in equation 3. By iterating this procedure a 1000 times, we determined the range of ranks into which a system falls in 95% of the cases[5], corresponding to a p-level of $p \leq 0.05$. The systems with overlapping ranges we clustered by taking into account that Bojar et al. (2013) recommend to build the largest set of clusters. Actually we performed the bootstrap resampling twice, once by picking 100 rankings pairs from each evaluator[6], and once by selecting 200 ranking pairs for each evaluator. The results show that the difference between 100 and 200 ranking pairs had no impact on the final ranking of the systems, and a minimal one on the way how systems were grouped to clusters. On the right side of Table 5 we present the ranking and clustering results based on samples build of 100 randomly picked rankings pairs per evaluator.

| Rank | Score | System |
|---|---|---|
| 1 | 0.593 | ONLINE-B |
| 3 | 0.573 | UEDIN-SYNTAX |
| 4 | 0.552 | PROMT |
| 5 | 0.541 | UEDIN |
| 6 | 0.511 | KIT |
| 7 | 0.480 | MES-REORDER |
| 8 | 0.478 | LIMSI-SOUL |
| 9 | 0.465 | CU-ZEMAN |
| 10 | 0.463 | STANFORD |
| 11 | 0.426 | TUBITAK |
| 12 | 0.422 | JHU |
| 13 | 0.352 | UU |
| 14 | 0.345 | SHEF-WPROA |
| 15 | 0.311 | RWTH-JANE |

Table 4: System ranking in the main study without bootstrap resampling

## 4 Discussion on Results

The motivation for running the experiments presented in the previous sections was guided by the main question whether future translators, in our case translations studies students, would rank MT output differently than the WMT2013 development teams. Being aware that translation studies students are language and translation experts, we expected them to be more consistent and more discriminative in their decisions as the WMT development teams.

With this in mind, we conducted two experiments, a pilot study and a main study, for the language pair English-German investigating whether translation studies students would evaluate MT output very differently from the WMT development teams and if yes, to what extent and how could we quantify these differences. During the pilot study we observed that the results are similar to those from WMT2013, achieving an intra-annotator agreement of 0.745 and an inter-annotator agreement of 0.494 as compared to 0.649 and 0.457 during WMT2013, we run the main study described in Section 3.2. The results from the main experiment show that translation

[4] For the 14 systems evaluated by researchers during WMT2013 we collected 24202 ranking pairs

[5] This means that the best and worst 2.25% scores for a system are not taken into consideration

[6] Repetitions were allowed.

| WMT2013 | | | Main Study | | |
|---|---|---|---|---|---|
| Rank | Score | System | Rank | Score | System |
| 1 | 0.637 | ONLINE-B | 1 | 0.594 | ONLINE-B |
| | 0.636 | PROMT | 2 | 0.572 | UEDIN-SYNTAX |
| 3 | 0.614 | UEDIN-SYNTAX | | 0.556 | PROMT |
| | 0.571 | UEDIN | | 0.540 | UEDIN |
| | 0.571 | KIT | 6 | 0.510 | KIT |
| 7 | 0.523 | STANFORD | 7 | 0.482 | MES-REORDER |
| 8 | 0.507 | LIMSI-SOUL | | 0.480 | LIMSI-SOUL |
| 9 | 0.477 | MES-REORDER | | 0.460 | STANFORD |
| | 0.476 | JHU | | 0.459 | CU-ZEMAN |
| | 0.460 | CU-ZEMAN | 11 | 0.427 | TUBITAK |
| | 0.453 | TUBITAK | | 0.426 | JHU |
| 13 | 0.361 | UU | 13 | 0.351 | UU |
| 14 | 0.329 | SHEF-WPROA | | 0.344 | SHEF-WPROA |
| | 0.323 | RWTH-JANE | 15 | 0.308 | RWTH |

Table 5: System ranking with bootstrap resampling in WMT2013 and in the main study

studies students achieve an intra-annotator agreement of 0.772 and an inter-annotator agreement of 0.510. The values are slightly higher than the ones of the researchers during WMT2013, but the differences are not really that pronounced. One interpretation of these results is that this task did not require specialised knowledge neither from the researchers nor from the translation studies students. Although researchers are probably not so familiar with translation studies theories and translation students are not specialists in machine translation, from the results, we notice an overlap in decision taking/making between the two groups. This overlap can be, as mentioned before, due to the nature of the evaluation task, since evaluators from both groups had to rank the machine translation output given the source text and the reference translation and the knowledge about the source and target language was enough.

The higher agreement values for the students' group can be an indicator that students ranked the machine translation output more thoroughly, a fact that was confirmed also by the non-formal feedback we got from the evaluators. Most of them them complained that it was very difficult to rank machine translation output of roughly similar overall quality. They reported that they had first to rank for themselves the errors they saw in the machine translation output before ranking the sentences.

Another aspect which probably influenced the results is the number of evaluators (for intra-annotator agreement) and evaluator pairs (for the inter-annotator agreement) considered in the computation of $\kappa$. The lower the number of evaluators and evaluator pairs the higher the influence of each evaluator and pair on the final $\kappa$.

Concerning the system rankings presented by Bojar et al. (2013) and computed based on the expected wins described by Koehn (2012), we can remark a shifting of ranks between the systems listed in the WMT2013 report and the rankings obtained by the translation studies students. Still, this rank shifting is more preeminent in the middle part of the table, than at the bottom, proving that systems with similar quality of MT output are harder to rank than MT output which is very different. Table 5 gives an overview of the WMT2013 system rankings as well as of the system rankings in our main experiment. ONLINE-B was ranked by both groups as the best system, UEDIN-SYNTAX and UEDIN kept their ranks as well as KIT, UU, SHEF-WPROA and RWTH. Although the other systems changed their rankings by moving up or down, there is no real striking position change in the ranking list. From Table 5 we can also notice that the scores for the systems have suffered a slight decrease in our main experiment as compared to the WMT2013 results. This is due to the fact that students made a clearer distinction between good and bad translations by trying to avoid ties, this being reflected into the final systems scores.

| | WMT2013 | Pilot | Main Study |
|---|---|---|---|
| Total number of evaluators | 38 | 26 | 37 |
| Total number of rankings pairs | 39582 | 25780 | 37318 |
| Evaluators considered for intra-annotator agreement | 12 | 16 | 22 |
| $\kappa$ (Intra-annotator agreement) | 0.649 | 0.745 | 0.772 |
| Evaluators pairs considered for inter-annotator agreement | 372 | 325 | 536 |
| $\kappa$ (Inter-annotator agreement) | 0.457 | 0.494 | 0.510 |

Table 6: Overview over collected data and Cohen's $\kappa$ for the language pair English-German

## 5 Conclusion

From our pilot study as well as from our main experiment on evaluating machine translation by ranking sentence level machine translation output we found that the MT development teams in WMT2013 are not so different from the translation studies students we had as evaluators in our experiments. Turning back to the questions we asked in Section 1, we can say that our experiments overall reproduced the WMT2013 ranking task with some differences in the results. Indeed, we observed that the group of students achieved higher agreement score $\kappa$ meaning that they were more consistent individually and as a group. On the other hand, from the computation of the system rankings the students confirmed at least the first and last places in the WMT2013 system ranking, although the scores achieved by all systems were slightly lower. The slight decrease of ranking scores is due to the fact that translation studies students were more discriminative and produced less ties. Based on the results presented in the previous sections we consider that the human ranking task does not required any specialised knowledge. Moreover, we argue that a homogeneous group and a good command of the source and target language are enough to replicate the results of the ranking task in the WMT2013.

## References

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Hervé Saint-Amand, Radu Soricut, and Lucia Specia, editors. 2013. *Proceedings of the 8th Workshop on SMT*. ACL.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Comelles, Elisabet and Jordi Atserias. 2014. Verta participation in the wmt14 metrics task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 368–375, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Denkowski, Michael and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on HLT*, pages 138–145.

Federmann, Christian. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *PBML*, 98:25–35, 9.

Gonzàlez, Meritxell, Alberto Barrón-Cedeño, and Lluís Màrquez. 2014. Ipa and stout: Leveraging linguistic and source-based features for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 394–401, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Koehn, Philipp. 2012. Simulating human judgment in machine translation evaluation campaigns. In *IWSLT*, pages 179–184.

Levenshtein, Vladimir Iosifovich. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lo, Chi-Kiu and Dekai Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 220–229.

Lo, Chi-kiu, Anand Karthik Tumuluru, and Dekai Wu. 2012. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252, Montréal, Canada, June. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on SMT*, pages 259–268.

Tillmann, Christoph, Stephan Vogel, Hermann Ney, Alexander Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *Proceedings of the EUROSPEECH*, pages 2667–2670.