Ten Years of WMT Evaluation Campaigns: Lessons Learnt

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, Lucia Specia
Charles University in Prague, Microsoft Research, University of Edinburgh, University of Edinburgh/ JHU, JHU, University of Sheffield
bojar@ufal.mff.cuni.cz, chrife@microsoft.com, bhaddow@inf.ed.ac.uk, phi@jhu.edu
post@cs.jhu.edu, l.specia@sheffield.ac.uk

Abstract

The WMT evaluation campaign (`http://www.statmt.org/wmt16`) has been run annually since 2006. It is a collection of shared tasks related to machine translation, in which researchers compare their techniques against those of others in the field. The longest running task in the campaign is the translation task, where participants translate a common test set with their MT systems. In addition to the translation task, we have also included shared tasks on evaluation: both on automatic metrics (since 2008), which compare the reference to the MT system output, and on quality estimation (since 2012), where system output is evaluated without a reference. An important component of WMT has always been the manual evaluation, wherein human annotators are used to produce the official ranking of the systems in each translation task. This reflects the belief of the WMT organizers that human judgement should be the ultimate arbiter of MT quality. Over the years, we have experimented with different methods of improving the reliability, efficiency and discriminatory power of these judgements. In this paper we report on our experiences in running this evaluation campaign, the current state of the art in MT evaluation (both human and automatic), and our plans for future editions of WMT.

Keywords: Machine Translation, Evaluation, Shared Tasks

1. Introduction

The First Workshop in Statistical Machine Translation was held in 2006, and it has been held annually since then, becoming the First WMT Conference in Machine Translation (WMT 2016) this year. In the first year of WMT there was a shared translation task which attracted 12 task description papers. In 2015 there were 5 different tasks and 46 task description papers, whilst in 2016 there will be 10 different tasks, covering translation of text and images, handling of pronouns in translation, MT evaluation, system tuning, automatic post-editing and document alignment.

The core component of WMT has been the main translation task (which in most years is the only translation task). The first translation task used Europarl (Koehn, 2005) for the test set; since then, we have constructed the test set from news text, with the complex structure and broad topic coverage providing a significant challenge to MT systems. Since 2009 the news test sets have been created specifically for the shared task, by crawling news articles in various languages and translating to the other task languages, providing the MT research community with valuable resources for future research. We have also varied the language pairs from year to year to present different challenges to researchers, although there has always been an emphasis on European languages. The language pairs included in each year's evaluation are shown in Table 1.

A central theme in the WMT shared tasks has been the evaluation of MT. We have explored this extensively, focusing on both human and automatic evaluation. The main translation task has always employed large-scale human evaluation to determine the quality and ranking of the systems; how precisely this is done has varied over the years (Section 2.). The human ranking has enabled the development of automatic metrics by providing a gold standard against which metrics can be compared. Since 2008, the metrics task has asked participants to develop tools to evaluate MT output against one or more references (Section 3.). In 2012, we introduced the quality estimation task, which takes metrics a step further, attempting to evaluate the quality of MT output without use of a reference (Section 4.).

2. Manual Evaluation

Since the very beginning, WMT organizers have taken the position that machine translation performance should be evaluated from time to time against human opinion:

> *While automatic measures are an invaluable tool for the day-to-day development of machine translation systems, they are only a imperfect substitute for human assessment of translation quality* ... (Koehn and Monz, 2006)

This is not to disparage automatic metrics, which have played a crucial role in the progress of the field and the improvement of MT quality over time. It is only to say that they are at best a proxy for what we really care about, and must be regularly anchored to human opinion. The WMT therefore produces an annual *human ranking of systems* for each task, from best to worst. In addition to helping direct researchers to the systems whose features they might wish to copy, this gold-standard system ranking is used to evaluate automatic metrics (a metric metric).

Of course, the question of which system is the best or worst is a fraught one. There are any number of answers, and subsequent questions. The first is best *for what purpose*? For a person trying to understand a foreign-language news article, an MT system that can convey the gist of an article is necessary, but quality might need to be sacrificed for speed. On the other hand, a student trying to learn how to translate an article may require a system that can also correctly generate grammatical and natural-sounding sentences. Evaluations are often broken down along these concepts of *adequacy* and *fluency*.

In fact, in the first two editions of the WMT shared translation task we used adequacy/fluency judgements on a 5-point scale as our main evaluation measure. Not satisfied with the results though, we started experimenting with

Language Pair	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Czech ↔ English		•	•	•	•	•	•	•	•	•	•
Finnish ↔ English										•	•
French ↔ English	•	•	•	•	•	•	•	•	•	•	
German ↔ English	•	•	•	•	•	•	•	•	•	•	•
German ↔ Spanish			•								
Haitian Creole → English						•					
Hindi ↔ English									•		
Hungarian ↔ English			•	•							
Romanian ↔ English											•
Russian ↔ English								•	•	•	•
Spanish ↔ English	•	•	•	•	•	•	•	•			
Turkish ↔ English											•

Table 1: Language pairs in the main translation task.

Metric	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Adequacy / Fluency	•	•									
Sentence Ranking		•	•	•	•	•	•	•	•	•	•
Constituent Ranking		•	•								
Constituent Judgement (Y/N)			•								
Sentence Comprehension				•	•			○			
Direct Assessment											•
Used MTurk					•		•	•			•

Table 2: Metrics used in the human evaluation over the years for all languages pair (•) or only English → Czech (○).

other methods and over the years, WMT has tried several different ones, encoded in different evaluations, summarized in Table 2. Brief explanations of the approaches follow:

- *Fluency / Adequacy*. Annotators were presented with a sentence, and were asked to rank it separately for both fluency and adequacy, on five-point scales.
- *Sentence Ranking*. Annotators are presented with the outputs of multiple systems, along with the source and reference sentence, and asked to rank them, from best to worst.
- *Constituent Ranking*. Annotators were asked to rank the quality of the translations of automatically-identified constituents, instead of the complete sentences.
- *Constituent Judgement (Y/N)*. Annotators were asked to provide a binary judgement on the suitability of the translation of a constituent.
- *Sentence Comprehension*. Annotators were asked to edit MT output for fluency (without providing the reference), and then (separately) to determine via binary judgement whether those edits resulted in good translations.
- *Direct Assessment (DA)*. Annotators are asked to provide a direct assessment of the quality of a single MT output compared to a single reference, using an analog scale.

The adequacy/fluency judgements were abandoned as the 5-point measurements proved to be quite inconsistent and hard to normalize, and they were not popular with the annotators. Viewing the distributions of scores provided by individual annotators showed them to be very different in shape, often skewed in different directions, so there was no clear way to combine judgements from multiple annotators. There was also complaints from annotators about the extreme difficulty in annotating long sentences of, frequently scrambled, MT output.

Two early measures of quality focused only on noun phrase constituents that were automatically identified in the reference and then extracted from system outputs via projections across automatic alignments. Constituent ranking (2007–2008) asked annotators to compare and rank these constituents, while binary constituent judgements (2008) asked them only whether a constituent (provided in context and approximately highlighted) were "acceptable" compared to the reference. An advantage of these binary judgements was very high annotator agreement rates; this is likely due in part to their relatively short length.

Another means of directly assessing output quality (and thereby inferring a system ranking) is Sentence Comprehension, used in 2009 and 2010. In this task, one set of judges was asked to edit a sentence's fluency (without access to the source or reference); these edited sentences were then later evaluated to see whether they "represent[ed] fully fluent and meaning-equivalent alternatives to the reference sentence". This mode of evaluation did not correlate well with relative ranking, however, and was abandoned in 2011 in order to focus annotators' efforts on that method.

In an effort to find a better evaluation method, we introduced Sentence Ranking in 2007. One big advantage of Sentence Ranking is that it is conceptually very simple: of-

fer the annotator two samples of MT output (and a reference) and ask them which they prefer. In practice, in order to gather judgements more efficiently, we present the annotator with 5 different MT outputs at a time, which then yields ten pairwise comparisons. We have experimented with presenting more or fewer sentences at a time, but 5 seems to be a good compromise between efficiency and reliability. We have also experimented with collecting judgements on Amazon's Mechanical Turk (2012 and 2013), in an effort to reduce the effort required from researchers. While relatively effective, the effort required to ensure that the work was completed faithfully, and the even lower annotator agreement rates, caused us to abandon it.

Since 2011, Sentence Ranking has been the only method of human evaluation we have used, but during that time the details have evolved in response to criticism. In particular, Bojar et al. (2011) pointed out various problems with the way the comparisons were collected and interpreted which led to changes in the procedure. A particular problem with Sentence Ranking is that the method involves collecting *relative* judgements of MT performance, but attempts to combine these to give an *absolute* measure of translation performance. Unless a sufficient number of carefully chosen comparisons are made, then systems can be treated unfairly by being compared too often to a very bad, or very good system (or the reference, which may be in there for control). Furthermore, systems were getting credit for ties, so systems which were very similar to others were doing better than they should. Finally, Bojar et al. (2011) showed that the agreement on the Sentence Ranking task falls off rapidly as sentence length increased.

Further analysis of the Sentence Ranking approach was provided by Lopez (2012) who pointed out the difficulties in obtaining a reliable total ordering of systems from the pairwise judgements. Further work (Koehn, 2012) suggested that we really needed to collect more judgements in order to display significant differences between the systems, and also established a means of clustering systems into equivalence classes of mutually indistinguishable systems, based on bootstrap resampling. Thus, since 2013, the system rankings have been presented as a partial ordering over systems, instead of a total ordering, where systems in the same group are considered to be tied. (However, the total ordering is still used when evaluating metrics).

One important point has not been addressed. Over the years, WMT has experimented with many different means of producing a system ranking. These rankings are then used as a gold standard for metrics tasks, and are also published as an official ranking, which researchers make use of in determining which system description papers to plumb for ideas to improve their own systems. Each year, different methods have been evaluated and then kept or discarded according to a number of criteria, such as annotator agreement numbers, or time spent. However, how can we really know which of these is the best? This point was raised by Hopkins and May (2013), who then provided a Bayesian model formulation of the human ranking problem, which allowed them to use perplexity to compare different system rankings. Influenced by this idea, in 2014, we compared the ability of three different models trained on a large set of pairwise rankings, using accuracy on held-out comparisons instead of perplexity. The method that won was a new approach that based on the TrueSkill algorithm (Sakaguchi et al., 2014). This has been in use since.

To conclude, the WMT manual evaluation has engaged in a deep and extensive experimentation over the years. The Sentence Ranking task has formed the core of our evaluation approach, and has seen many variations from year to year. We have made progress on many of the problems with evaluation. However, many problems remain: the relatively low annotator agreement rates, the immense amount of annotator time required, and the difficulty of scaling the sentence ranking task to many systems. In 2016, we plan to run a pilot investigation based on Direct Assessment of machine translation quality, which we hope will further alleviate some of these issues.

3. Automatic Evaluation

Since the second year of the WMT campaigns, targeted effort was also devoted to evaluation of automatic metrics[1] of MT quality, or **metrics task** for short. This meta-evaluation is an important complement to the shared translation task, because automatic metrics are used throughout the development of MT systems and also in automatic system optimization (Neubig and Watanabe, 2016). The utility of some of the metrics in system optimization has been tested in the sister **tuning task** in 2011 and 2015 and also planned for 2016.

Metrics of MT quality are evaluated at two levels:

System-level evaluation tests, how well a metric can replicate the human judgement about the overall quality of MT systems on the given complete set of test set sentences.

Segment-level evaluation tests how well a metric can predict the human judgement for each input sentence.

In both cases, participants of the metrics task are given input sentences, outputs of MT systems and one reference translation. Note that the reliance on a single reference is not ideal. It is well known that the reliability of automatic MT evaluation methods is limited if only one reference is available (see the WMT 2013 overview paper for an empirical evaluation of BLEU with up to 12 references for translation into Czech). The quality estimation task (Section 4.) focuses on the setup where no reference is available at all.

Table 3 summarizes the participation and methods used to evaluate the system-level and segment-level parts of the task. The task had always received a good number of participating teams. The number of evaluated metrics varies considerably across the years, because in some years, multiple variations of some metrics were evaluated.

Starting from 2013, we distinguish "baseline metrics". These metrics are run by the organizer in addition to the submitted ones. Baseline metrics include the `mteval` scoring script and all the metrics available in Moses. We report the exact configuration flags for them, so they should be reliably reproducible.

Throughout the years, the metrics task has always relied on the manual evaluation (Section 2.), so the gold standard

[1] Despite the term "metrics", none of the measures or methods is a metric in the mathematical sense.

	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16
Participating Teams	-	6	8	14	9	8	12	12	11	
Evaluated Metrics	11	16	38	26	21	12	16	23	46	
Baseline Metrics							5	6	7	
System-level evaluation methods										
Spearman Rank Correlation	•	•	•	•	•	•	•	○		
Pearson Correlation Coefficient							○	•	•	•
Segment-level evaluation methods										
Ratio of Concordant Pairs		•	•							
Kendall's τ				•	•	•	$*$	$\star$	$\star$	$\star$
Tuning Task					•				•	•

• main and ○ secondary score reported for the system-level evaluation.
•, $*$ and $\star$ are slightly different variants regarding ties.

Table 3: Summary of metrics tasks over the years.

human judgements do come from different styles of evaluation. A major move from Sentence Ranking to Direct Assessment is considered in 2016, which would particularly affect the segment-level metric evaluation. In Direct Assessment, the judgements have to be sampled differently from the system-level and segment-level evaluation, and there is a concern whether we will be able to find enough distinct speakers for each of the language pairs. Preliminary experiments are now under way.

3.1. How Metrics are Evaluated

As indicated in Table 3, the metrics task has seen a few changes of the exact evaluation method.

Evaluating System-Level Evaluation System-level methods were first evaluated using Spearman rank correlation, comparing the list of systems for a particular language pair as ordered by the metric (given the test set of sentences are reference translations) and as ordered by humans (on the sample of sentences from the test set that actually receive some human judgements). Spearman rank correlation was selected in the first year, because it is applicable also to the ordinal scales of adequacy and fluency which were used in 2006 and 2007. Since 2007, Pearson correlation coefficient could have been also used (as the system scores were on continuous scales), but the switch happened only in 2013. The benefit of Pearson over Spearman is that it considers the distances between the systems, so it should be more stable for systems of similar quality.

Evaluating Segment-Level Evaluation Segment-level evaluation has so far relied on pairwise judgements of translation quality. Given two candidate translations of an input sentence, the segment-level metric gets a credit if it agrees with the human judgement, i.e. the two pairwise judgements are "concordant". The exact calculation of the final score changed throughout the years: in 2008 and 2009, a simple ratio ranging from 0 to 1 was used: the number of concordant pairs out of the total number of pairs evaluated. Starting from 2010, the score was modified to penalize discordant pairs, falling under the general definition of Kendall rank correlation coefficient, or Kendall's τ for short, with $[-1, 1]$ as the range of possible values:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|} \quad (1)$$

There has always been a question of how to handle tied comparisons, either the humans or the metric (or both) assigning the same rank/score to the two candidates. Each type of tied pairs can be included in the denominator and if it is, it may be also included in the numerator (bonified or penalized). After the discussion available in Macháček and Bojar (2013) and Macháček and Bojar (2014), the current method:

- ignores pairs where humans tied altogether,
- does not give any credit or bonus to pairs where the metric predicted a tie,
- but includes these metric-tied pairs in the denominator.

Moving to the Direct Assessment or some other absolute scale in the human evaluation would allow use to use Pearson correlation coefficient instead of Kendall's τ.

Significance From the beginning, it was not quite clear how to establish significance of the observed differences in metric evaluation, especially at the system level where the number of participating systems is less than 20, providing a low sample size.

Starting from 2013, system-level scores for each given language pair were reported with empirical confidence bounds constructed by resampling the "golden truth": given the complete set of human judgements, 1000 variations are constructed by resampling with repetition, leading to 1000 different scorings of the systems.[2] Each participating metric provides a single scoring of the systems and this scoring is correlated with the 1000 golden truths, giving us 1000 results reflecting the variance due to the set of sentences and annotators included in the golden truth.

As noticed by Graham and Liu (2016), confidence intervals obtained from this sampling cannot be used to infer whether one metric significantly outperforms another one, because the number of "significant" pairs would be overestimated. Instead, Graham and Liu (2016) proposes a novel method, artificially generating a large number of MT systems (by

[2]Many of these scorings share the same order of the systems. Unlike Spearman rank correlation, the Pearson correlation coefficient used since 2013 however appreciates also differences in the scores.

mixing the outputs of the real MT systems participating in the translation task) and asking metrics task participants to score e.g. not 5 but 10000 MT systems on the given test set. We will try to adopt this approach in 2016, testing in practice, how many metrics task participants can cope with these enlarged sets of MT systems.

3.2. Observations in Metrics Task

While metrics tasks across the years cannot be directly compared because a whole range of conditions keeps changing, the overall setting remains stable and some general observations can be made:

- BLEU has been surpassed by far by many diverse metrics. On the other hand, we acknowledge that it remains the most widely used and also scores on average well among the baseline metrics, with CDER (Leusch and Ney, 2008) being a competitor.
- The level of 0.9 of system-level correlation into English was reached by the best metrics in 2009, rising up to 0.98 in 2011. These levels were achieved by **aggregate** or **combination metrics** that include many features and standard metrics; sometimes the combination is **trained** on a past dataset. IQmt-ULCh, SVMrank (2010) and MTeRater-Plus (2011) are the early examples, followed by a row of other combination metrics in recent years (e.g. BEER, DPMFcomb, RATATOUILLE in 2014 or 2015). MTeRater is an interesting outlier in that its main component is based on many features from automatic essay scoring (preposition choice, collocations typical for native use, inflection errors, article errors).
- Benefits were confirmed many times from **including paraphrases or synonyms** incl. Wordnet (e.g. Meteor, Tesla in 2010 and 2011), refining the metric to consider the coverage of individual **parts of speech** (e.g. PosBLEU 2008, SemPOS 2009, 2012), focusing on **content words** (Tesla, SemPOS), **dependency relations** (already 2008) or **semantic roles** (already 2007), evaluating at the level of **character sequences** (i-letter-BLEU 2010, chrF 2015, BEER).
- In 2012, we saw a drop in into-English evaluation mainly due to a different set of participating metrics. Such a "**loss of wisdom**" is unfortunate and the baseline metrics run by the organizers are one of possible means to avoiding it. In an ideal world, the authors of the top performing metrics every year would incorporate their metrics to Moses, to ensure that the metric gets evaluated in the coming years. Achieving this state is obviously complicated by the reliance of some of the metrics on diverse language-dependent resources which are not always publicly available. Meteor remains the only such maintained metric throughout the years. Hopefully, some of the trivial but well-performing metrics based on characters (chrF, i-letter-BLEU) will join the baselines soon.

4. Quality Estimation

Quality Estimation (QE) offers an alternative way of assessing translation quality. QE metrics are fully automated and, unlike common evaluation metrics (Section 3.), do not rely on comparisons against human translations. QE metrics aim to provide predictions on translation quality for MT systems in use, for any number of unseen translations. They are trained metrics, built using supervised machine learning algorithms with examples of translations labelled for quality (ideally, by humans). Predictions can be provided at different granularity levels: word, phrase, sentence, paragraph or document. Different levels require different features, label types and algorithms to build prediction models.

While work on QE started back in the early 2000's (Blatz et al., 2004), the use of MT was substantially less widespread back then, and thus the need for this type of metric was less evident. A new surge of interest appeared later (Specia et al., 2009; Soricut and Echihabi, 2010), particularly motivated by the popularisation of MT in commercial settings.

QE was first organised as a shared task (and a track at WMT) in 2012 (Callison-Burch et al., 2012). The main goals were to provide a baseline approach, devise evaluation metrics, benchmark existing approaches (features and algorithms), and establish the state-of-the-art performance in the area. The task focused on quality prediction at sentence level. Only one dataset was provided, for a single language pair (English-Spanish), on the News domain, translated by one MT system. For training and evaluation, translations were manually annotated by professional translators for quality in terms of "perceived" post-editing effort (1-5 scores) . A system to extract baseline QE features and resources to extract additional features were also provided. The baseline system used a Support Vector Machine regression algorithm trained on the features provided. This was found to be a strong baseline (both features and algorithm) and has been used in all subsequent editions of the task.

As we continued running the task in subsequent years (Bojar et al., 2013; Bojar et al., 2014; Bojar et al., 2015), our main goals have been to provide, each year, new subtasks (while keeping the popular ones), additional language pairs, and larger and more reliably labelled datasets. For most subtasks, the evaluation metrics have also been redefined over the years. Table 4 summarises the main components of the shared task over the years.

More specifically, we introduced variants of post-editing effort prediction – edit distance (a.k.a. HTER) and post-editing time – for sentence level (2013), and other subtasks at new granularity levels: (i) a system selection subtask to learn how to rank alternative MTs for the same source sentence, precisely the same goal as the metrics task (Section 3.), but without reference translations (2013); (ii) a word-level subtask concerned with predicting a binary (good/bad) or 3-way (keep, delete, replace) tag for each word in a target sentence (2013), as well as more fine-grained error categories annotated by humans (omission, word order, word form, etc., in 2014); (iii) a paragraph-level subtask to predict a Meteor score for an entire paragraph (2015); (iv) a document-level subtask to predict a task-based human-targeted score for the entire document (2016); and (v) a phrase-level subtask, where binary labels (good/bad) are to be predicted for entire "phrases", as segmented by the MT system (2016). Baseline systems and resources were provided for all these subtasks.

The main language pair has remained English-Spanish

(en→es), the only constant language over all editions for the sentence and word-level subtasks. This was mostly due to the availability of (labelled) data for this pair. However, other language pairs have been explored over the years for most subtasks. English-German (en→de) was used on various occasions, including all subtasks in 2014 and the paragraph-level subtask in 2015. German-English (de→en) was also used in the latter subtask, in all subtask in 2014, and in the MT system selection task in 2013.

The sizes of the datasets varies over the years. A good indicator is the sentence-level subtask. The figures in the last row of Table 4 refer to the largest number of sentences for any score prediction subtask in a given year.

The number of participating teams has remained considerably stable over the years (10–14), but teams tend to submit systems for various subtasks, as well as for the same subtask when multiple languages are available. The submission figures in Table 4 include only submissions for different subtasks and language pairs.

The evaluation of participating systems varies across subtasks. For sentence, paragraph and document levels, systems can be submitted for two variants of each task: scoring (for various labels, e.g. 1-5, 1-3, HTER, time, Meteor) and ranking, where only a relative ranking of test instances is required. Scoring is evaluated using standard error metrics (e.g. Mean Absolute Error) against the true scores and, since 2015, using Pearson's correlation. Ranking is evaluated using Spearman's correlation, as well as a ranking metric proposed for the task in 2012: DeltaAvg, which compares the ranking of instances given by the system against the human ranking for different quality quantiles of the test set. For the word and phrase-level tasks, per-class precision, recall and F-measure metrics are computed, with F-measure for the "bad" class used as main metric in the binary variant.

Overall, the shared tasks have led to many findings and highlighted various open problems in the field of QE. Here we summarise the most important ones:

- **Training data**: The size of the training data is important for all prediction levels, but is even more critical for word and phrase levels. For sentence level, it does not seem to be the case that having more than 2K sentences makes a significant difference in performance. The quality of the data has proved a more important concern. The dataset used for the sentence and word level subtasks in 2015, for example, although large, was of questionable quality (spurious or missing post-editings) and had a very skewed label distribution, which made model learning harder.

- **Algorithms**: There is no consensus on the best algorithm for each subtask. Various popular regression algorithms have ranked best for sentence (and paragraph) level in different years, including SVM, Multilayer Perceptron, and Gaussian Process. For word (and phrase) level, sequence labelling algorithms such as Conditional Random Fields perform best.

- **Tuning**: Feature selection and hyperparameter optimisation proved essential. The winning submissions in most years performed careful (or even exhaustive) search for both features and hyperparameter values.

- **Features**: While a range of features has been used over the years, shallow, often language-independent features, tend to contribute the most. The majority of submissions built on the set of baseline features provided. Recently, word embeddings and other neural inspired features have been sucessfully explored. While features for sentence and word/phrase-level prediction are clearly very distinct from one another, for paragraph level, most systems used virtually sentence level features. We hope that more interesting discourse features will be exploited in 2016 given the much longer documents provided as instances. A critically important feature for all levels is the *pseudo-reference* score, i.e., comparisons between the MT system output and a translation produced by another MT system for the same input sentence.

- **Labels**: Prediction of objective scores, such as post-editing distance and time, has led to better models (in terms of improvements over the baseline system and correlation with human scores) than prediction of subjective scores such as 1-5 labels. Post-editing time seems to be the most effective label. However, given the natural variance across post-editors, this is only the case when data is collected by and a model is built for a single post-editor.

- **Granularity**: The word-level subtask has proved much more challenging than the sentence-level one, often obtaining very marginal improvements over naive baselines. In the tasks we have run so far, this could have been due to: little training data, limited number of examples of words with errors (class unbalance), and potentially noisy automatic word labelling. We attempted to solve some of these limitations by providing data annotated manually for errors (2014), but for cost reasons the largest dataset we could collect has just over 2K segments. A larger dataset (14K segments) was collected based on post-editions in 2015, but the post-editing, and hence the labelling generated from it, are of questionable quality. In 2016, we are providing an even larger dataset (15K segments) post-edited by professional translators. The new phrase-level subtask in 2016 should also help overcome some of the limitations of the word-level one, by providing more natural ways in which to segment the text for errors. The paragraph-level subtask in 2015 did not attract much attention, perhaps due to the use of an automatic metric as quality label (Meteor). In 2016 we provide actual (much longer) documents labelled by humans.

- **Progress over time**: As with any other shared task, measuring progress over time is a challenge since we have new datasets (and often new training sets) every year. Progress in the QE task can however be speculated in relative terms, more specifically, with respect to the improvement of submitted systems over

	'12	'13	'14	'15	'16
Participating Teams	11	14	10	10	-
Evaluated QE Systems	20	55	57	34	-
Subtasks	1	4			
Sentence Level	•	•	•	•	•
Word Level		•	•	•	•
Paragraph Level				•	
Document Level					•
Phrase Level					•
Language Pairs	en→es	en→es, de→en	en↔de, en↔es	en→es, en↔de	en→es
Largest Dataset (snt)	2,254	2,754	4,416	14,088	15,000

Table 4: Details on different editions of the QE task over the years.

the baseline system. This is possible for the sentence-level subtask, since the language pair and baseline system have remained constant over the years. We have observed, year after year, that more systems are able to beat the baseline, and by a larger margin.

5. Plans for Future Editions

In recent years, we have used Sentence Ranking as the sole method of automatic evaluation (refining it according to certain criticisms (Bojar et al., 2011; Lopez, 2012; Koehn, 2012)), but ongoing problems with reliability, interpretability and poor scalability with increasing numbers of systems have driven the search for alternatives. In 2016, we will pilot a new technique for manual evaluation of MT output. This is based on recent work demonstrating an effective means for collecting adequacy and fluency judgements using crowd-sourcing (Graham et al., 2016). This *Direct Assessment* of machine translation quality is similar to our early attempts to judge quality with adequacy and fluency judgements (Koehn and Monz, 2006; Callison-Burch et al., 2007), but improves upon it in critical ways. Crucially, an analog scale is presented to the user in the form of a slider bar, which underneath maps to a 100-point scale, instead of the 5-point Lickert scale we used in the past, which gave us inconsistent results that were difficult to interpret. Annotators are required to do large batches of assessments in a single sitting, which allows their scores to be normalized more reliably. By embedding deformed outputs and comparing their scores to those of their uncorrupted counterpart, inconsistent, unreliable, and untrustworthy annotators can be identified, and their outputs discarded.

The potential advantages of Direct Assessment are:

- It offers good reliability, as measured by inter-annotator agreement;
- the cost of assessment scales linearly in the number of systems assessed (instead of quadratically, as with Sentence Ranking);
- it provides absolute measures which can be compared year-over-year; and
- the concepts of adequacy and fluency are readily interpretable, in a way that the scores derived from Sentence Ranking are not.

Sentence Ranking will remain our primary evaluation for this year, but the results of this evaluation will be compared to those of the DA evaluation in order to help is assess its suitability for future evaluations.

One of the big issues we face in MT evaluation is the question of *for what purpose?* In other words, the way we evaluate our MT system may depend quite strongly on what we want to use it for, whether for gisting, post-editing, direct publication, language learning, automated information extraction, or something else. The Sentence Ranking method is particularly weak in this regard, since we do not give the raters any guidance as to how they should judge the translations. In some sense, we have punted on the difficult question of purpose, allowing each annotator to be guided by his or her own intuitions. This likely explains some of the low annotator agreement rates. Using adequacy and fluency separately is an improvement as the terms have meaningful interpretation, although they are still intrinsic rather than extrinsic measures. In the end, we believe that the work of the WMT manual evaluation has improved our knowledge for how to assess human quality of MT, providing a rich well from which to draw for those wishing to focus on more targeted and specific applications.

For QE, after the 2016 edition we will have covered all possible granularity levels. The plan is to keep the most popular and the most challenging ones, with a particular emphasis on word and phrase-level prediction. Instead of more language pairs, we will prioritise larger and better datasets for fewer language pairs. Another direction we aim to pursue is better integration with other WMT evaluation tasks, e.g. using the test sets and system translations from the translation task, and reusing the manual evaluations as training data. In the past this has proved difficult logistically because of the tasks' timeframe or unsuccessful because the manual evaluations (esp. rankings) were not adequate for QE. The planned changes in the manual evaluation procedure should make this integration possible.

6. Conclusions

The WMT shared tasks have given us a platform to explore all forms of Machine Translation (MT) evaluation; human evaluation, automatic evaluation with a reference, and quality estimation. Not only that, but WMT has helped to drive research in MT evaluation, firstly by having high profile shared tasks to engage the community; and secondly by the extensive data sets that we provide. Each year, we prepare new translation test sets, and annotated data sets for quality estimation. During the tasks, we collect and release all

translation system submissions, all the human judgements, all the submissions to metrics, and all the quality estimation data. These are made available from the WMT website (for this year it is `www.statmt.org/wmt16`) and are used frequently in subsequent research.

MT evaluation is a hard problem, and is capable of generating significant controversy in the MT community, as we have observed when evaluation results were presented. This difficulty is indicated by the number of changes, experiments, and refinements we have introduced over the years. This year, with the piloting of Direct Assessment, we return to a direct measure of the quality of a system output that we abandoned a number of years ago, and are hopeful that the reformulation of the problem will make DA more successful than our earlier experiments. If so, one option for the QE task in subsequent years is for it to model the prediction of DA scores.

Acknowledgements

This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645442 (QT21) and 645357 (Cracker).

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence Estimation for Machine Translation. In *Proc. of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. (2011). A Grain of Salt for the WMT Manual Evaluation. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proc. of the Tenth Workshop on Statistical Machine Translation*, pages 1–46.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada.

Graham, Y. and Liu, Q. (2016). Achieving Accurate Conclusions in Evaluation of Automatic Machine Translation Metrics. In *Proc. of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (to appear)*, San Diego, CA.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.

Hopkins, M. and May, J. (2013). Models of Translation Competitions. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria.

Koehn, P. and Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of MT Summit*.

Koehn, P. (2012). Simulating Human Judgment in Machine Translation Evaluation Campaigns. In *Proc. of IWSLT*, pages 179–184.

Leusch, G. and Ney, H. (2008). BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation Challenge*, Waikiki, Honolulu, Hawaii, October.

Lopez, A. (2012). Putting Human Assessments of Machine Translation Systems in Order. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada.

Macháček, M. and Bojar, O. (2014). Results of the WMT14 Metrics Shared Task. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, MD, USA.

Macháček, M. and Bojar, O. (2013). Results of the WMT13 Metrics Shared Task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria.

Neubig, G. and Watanabe, T. (2016). Optimization for Statistical Machine Translation: A Survey. *Computational Linguistics*, To appear.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for human evaluation of machine translation. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA.

Soricut, R. and Echihabi, A. (2010). TrustRank: Inducing Trust in Automatic Translations via Ranking. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden.

Specia, L., Turchi, M., Cancedda, N., Dymetman, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proc. of the 13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona, Spain.