Robustness in Machine Translation Evaluation

Nitika Mathur

ORCID: 0000-0002-7694-5065

School of Computing and Information Systems

The University of Melbourne

This dissertation is submitted for the degree of

Doctor of Philosophy

December 2021

To my dearest sister . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 100,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Nitika Mathur

December 2021

Preface

Large portions of Chapter 3 have appeared in the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Towards efficient machine translation evaluation by modelling annotators. *In Proceedings of the Australasian Language Technology Association Workshop* 2018, pages 77–82, Dunedin, New Zealand, December 2018.

Large portions of Chapter 4 have appeared in the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Sequence effects in crowdsourced annotations. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2860–2865, Copenhagen, Denmark, September 2017.

Large portions of Chapter 5 have appeared in the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* pages 2799–2808, Florence, Italy, July 2019.

Large portions of Chapter 6 have appeared in the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *In*

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online, July 2020.

Acknowledgements

The PhD was a period of intense personal growth, and I am grateful to so many people for being a part of this journey.

First, I'm indebted to Trevor and Tim for being the best supervisors I could ask for. Thank you for all those valuable insights and feedback on my research, and for challenging me to get better as a researcher and writer. I couldn't have made it through without all your support and encouragement in moments of struggle and self-doubt.

I am so grateful to Justin Zobel and Oliver Adams for all the long, introspective conversations over these years. I am a better person because of you, and this thesis is so much stronger for your advice.

I'd like to thank Yvette Graham (who supervised my first research project and introduced me to this field) and Ondrej Bojar for the chance to be coordinate the organization of the metrics shared task at WMT 2020, and implementing some of the ideas in this thesis as a part of the evaluation. Thank you also to the rest of the team: Johnny Wei, Qingsong Ma, Markus Freitag. It was great working with you all, and I learned a lot. Markus, thank you for taking over the reigns for WMT 2021, bringing in more amazing people to the team, and expanding the scope of the task.

I deeply appreciate the thoughtful reviews of the two external reviewers of this thesis. You have given me much to think about. In particular, I value the perspectives of Reviewer 2 that comes from decades of experience in the field.

Thank you, to my dearest sister Nitya for proofreading parts of this thesis, and catching so many of the extraneous commas I tend to use. I bear full responsibility for any that remain.

I'm thankful for the financial support provided by the Melbourne Research Scholarship by the University of Melbourne.

I was fortunate to be part of an amazing research group at the University of Melbourne. Thank you to Long, Bahar, Julian, Doris, Aili, Shiva, Miji, Ping Ping, Mel, Lea, Jey Han, Daniel, AJ, Felix, Yitong, Afshin, Kat, Brian, Ned, Philip, AJ, Shrae, Daniel, and so many more for stimulating conversations and new perspectives. Thank you also to Oscar, Alex, Farah, Mohammad and my other friends in the school of CIS for being good company.

To my mother, sister and brother-in-law, you have my deepest gratitude for your love and unconditional support of all my decisions. And for all the laughter and silliness. I could not have done this without you.

Jagruti, you are the one constant that all the equations of my universe depend on. Neha and Alekhya, talking to you is always a comfort and a delight. Chris, you helped me stay on course so many times when I was ready to give up. To Soumya, Kartheik, Yossf, Rebecca, Sumedha, Shiksha, Shravya, Divya, Supriti, Sumanka, Krithika, Vijay, Aditya, Anusha, Apeksha, Rashmi, Abhishek, Pallavi, Pranita jiji, Rinesh Jijaji, Kerul, and the rest of my wonderful friends and family, I am so lucky to have you all.

Dad, we dreamed a thousand dreams together, listening to your beloved songs, watching the stars at night. This thesis is one of those dreams and I wish you were here to share this moment with me.

Abstract

We need reliable and efficient methods to measure the quality of machine translation (MT) systems. An ideal translation is a fluent sentence in the target language that preserves the meaning of the source sentence. To what degree does an output of an MT system satisfy these criteria? This is typically measured by eliciting human judgements, for example, by asking them to rate the quality of MT system translations. Human evaluation is expensive and laborious; consequently, automatic metrics were introduced to provide immediate feedback to MT system developers. Importantly, automatic metrics are also frequently used as the primary measure for reporting empirical results in the MT literature.

In this thesis, we address three aspects of MT evaluation: (1) improving the efficiency of human evaluation, (2) developing new automatic metrics, and (3) improving the evaluation of automatic metrics to aid in metric selection and analysis of metric outputs.

Human judgements are inherently noisy; to obtain accurate scores for individual translations, multiple judgements are collected and averaged. We design unsupervised Bayesian methods to improve aggregation of human judgements that take annotator reliability into account. With these methods, we can compute more accurate scores with fewer judgements per translation, thus decreasing costs. We also explore sequence effects in the data: we show that annotators' judgement can be affected by the context of the decision, specifically quality of preceding items, and propose a simple fix to mitigate this problem.

We propose new automatic MT metrics that rely on contextual word embeddings to measure similarity: these embeddings are influenced by the sentence context of the words in addition to the word itself, and result in a substantial improvement in correlation with human judgements compared to previous metrics.

Finally, we look at evaluation of automatic metrics. The research community has developed more sophisticated metrics, but MT system developers are slow to move away from BLEU, a metric that computes the surface similarity between the MT output and the human reference. This is due to a generally held belief that using BLEU is valid to compare similar MT systems, along with the fact that BLEU is fast to compute and doesn't require additional resources. While BLEU appears to have a reasonable correlation with human judgements in recent studies, this is possibly an artefact of the MT systems considered in the evaluation. We show that outlier MT systems (those that are much better or worse than other systems that are included in the evaluation) can lead to an over-estimate of the correlation, leading to misplaced trust in the metric. We then look at efficacy of metrics when comparing any two MT systems, which is the most common use for automatic metrics. We show that small improvements in automatic metrics often disagree with human judgements, and empirical MT research must always be supported by human judgements.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Advances in natural language processing in recent years have been driven by empirical evaluation. Accordingly, to make progress in any task, it is essential that our evaluation methods are meaningful and reflect the true capability of our models. This thesis focuses on improving the robustness of evaluation of machine translation (MT) systems.

With machine translation, we are not only attempting to teach a computer to understand language, but also produce words in a new language that preserve the meaning of the original text. How do we know how well we've succeeded in this challenging task? When we make changes to a system, how do we know that this has resulted in an improvement? How do we decide if it's better than other MT systems?

As MT is intended for human consumption, it is natural to ask humans to judge translation quality based on predetermined characteristics such as their adequacy and fluency. Adequacy measures how much meaning is preserved in the translation, and fluency measures the clarity and grammaticality of the translation. Ideally, we'd obtain these judgements from experts such as translators or linguists who are skilled in picking up nuance in translation quality.

Crowdsourcing, i.e., paying a "crowd" of non-experts to complete our annotation tasks, reduces the total cost and time involved, but introduces new complexities into the process as annotator reliability can not simply be assumed. Human opinions are generally noisy and inconsistent, and this is further exacerbated by personal preferences and cognitive biases, which can have a significant impact on the quality of annotations. Exactly how this annotation task is structured has been evolving since the beginning of MT research, and this is still an active research area.

Expert human evaluation can take weeks or even months. While crowdsourcing can be significantly faster, there is a need for immediate feedback during MT system development: to test whether an idea works, or to compare different iterations of an MT system. Automatic metrics have been developed for this purpose. Automatic metrics are based on the assumption that "the closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002). They assign a numerical score to Machine Translation output based on how similar it is to a reference translation by a human expert. There are many ways to correctly translate a sentence, and many more ways to be wrong. Any difference in the MT output compared to the reference could be a valid way of expressing the same idea, could reduce the fidelity or intelligibility of the translation, or even change the meaning of the sentence completely. MT metrics must be flexible enough to allow valid variations, and yet discriminate against invalid translations. While humans can intuitively make this judgement, this task is extremely difficult to automate, and is still unsolved, despite many years of effort by the research community.

BLEU (Papineni et al., 2002), which essentially computes overlap of n-grams in the MT output when compared to multiple human references, was the first metric that reported a high correlation with human judgements. It was widely adopted by the MT community, and triggered research into understanding the metric: it has several flaws, such as a failure to recognise valid synonyms and paraphrases, or to discriminate between the relative importance of the words in the sentence (Callison-Burch et al., 2006; Stent et al., 2005). There have been a variety of approaches to improve on BLEU. Shallow surface-level metrics, such as BLEU and TER (Snover et al., 2006) predominate in practice, due in part to their reasonable correlation to human judgements, and their being parameter free, making them easily portable to new languages. In contrast, trained metrics (Song and Cohn, 2011; Stanojević and Sima'an, 2014; Ma et al., 2017; Shimanaka et al., 2018), which are learned to match human evaluation data, have been shown to result in a large boost in performance.

In addition to their use during system development, automatic metrics often serve as the primary method of evaluation to report the quality of MT systems, serving as a cheaper alternative to human evaluation. For instance, the academic community uses small improvements in automatic metrics to claim improvement in the state of the art. Making decisions based on unreliable metrics could lead to wasted effort in attempting to reproduce spurious improvements or discarding promising ideas (Freitag et al., 2020; Kocmi et al., 2021). It is thus essential to pick the best metrics to report experiment results, and to understand the limitations of these metrics when making conclusions.

This thesis contains several contributions to increase the efficiency and reliability of MT evaluation, covering three different aspects of MT evaluation: *human evaluation*, *automatic evaluation*, and the *evaluation of automatic metrics*.

Throughout this thesis, we use data from the Conference on Machine Translation (WMT; previously Workshop on Machine Translation), which is run annually to provide a forum for MT research (Koehn and Monz, 2006 through to Barrault et al., 2019). The `news translation` shared task is an important feature of the conference, where participants can submit MT systems to translate new test sets in multiple directions. The organisers have a firm belief that "automatic evaluation is an imperfect substitute for human evaluation" (Koehn and Monz, 2006), and accordingly base their primary results on a large scale human evaluation of the participating MT systems. This shared task is a major source of innovation in new MT techniques, and confirms whether ideas from the literature that were validated using automatic metrics hold up to human evaluation and generalise to new, unseen test sets.

The data collected from the human evaluation in WMT is used to evaluate automatic metrics in the metrics shared task, which serves to validate existing automatic metrics and drive the development of new metrics that refine existing methods or introduce novel approaches to estimate MT quality. All data from WMT is publicly available, and we use this data to evaluate our contributions to both human and automatic MT evaluation.

1.1 Research Questions and Contributions

Human Evaluation Our first set of research questions is concerned with collecting and aggregating annotations for human evaluation of machine translation:

- Can we model annotator reliability when aggregating translation ratings from multiple annotators?
- Does the order of annotations introduce bias in the data?

Human annotations are inherently noisy, and annotator reliability can vary, particularly when the data is crowdsourced. To obtain an accurate label or rating for individual items, typically multiple annotations are collected per item, which are first filtered based on quality control items, and then aggregated using the majority vote (for discrete tasks) or average scores (for numeric tasks). We propose a simple probabilistic model for MT human evaluation data that infers annotator precision, and essentially weights each annotator's scores based on their precision when estimating the quality of a translation. This yields more accurate scores that require fewer annotations per translation, compared to the recommended best practice of using the mean score of workers who pass quality control.

When humans are evaluating a set of instances, they would ideally score each instance independently. However, we are all subconsciously influenced by cognitive biases. We show evidence of sequence effects in MT evaluation data, where the score of an instance is affected by scores assigned to previous instances. We suggest a simple method to mitigate the impact of sequence effects when redundant annotations are obtained.

Automatic Metrics We next move on to designing automatic metrics:

- How can we leverage contextual word embeddings to improve Machine Translation evaluation metrics?
- Can a supervised MT metric effectively learn from extremely noisy training inputs?

Contextual word embeddings map words to vector representations that depend on their sentence context. We develop new MT evaluation metrics which rely on contextual word embeddings to encode the translation and reference. Our first metric is a simple yet effective unsupervised metric that approximates the precision, recall and F-score of the information in the reference. Our supervised metrics compute sentence representations from the contextualised word embeddings, and then map these to a similarity score. We train these on human evaluation data, and find that the metric reliability improves when evaluating on a large, but extremely noisy dataset that was previously unexplored by supervised metrics. We show that when there is limited budget for the number of annotations, model training is more efficient with single annotations on more instances compared to accurate scores from aggregating multiple annotations on fewer instances.

Evaluation of Metrics We finally re-evaluate the evaluation of automatic metrics:

- How is the correlation of metrics with human judgements affected by the set of MT systems evaluated?
- How does metric reliability depend on the quality of the MT systems evaluated?
- When comparing two MT systems, how do conclusions based on automatic metrics compare with those based on human evaluation?

Metrics are typically evaluated based on the Pearson correlation with human judgements on a small set of MT systems that are not an unbiased sample, or representative in any meaningful way. We find that outlier MT systems whose quality is much better or much worse than the rest of the systems have a disproportionate influence on the computed correlation. We identify a robust method for identifying outliers, and demonstrate their effect on correlation, which for some metrics can result in radically different conclusions about their utility.

The findings of the WMT 2019 shared task on metric evaluation show that the correlation of metrics decreases dramatically when evaluated only on the systems with the highest human scores (Ma et al., 2019). We suggest that this result can be attributed to the instability of computing correlations at small sample sizes, and find that no empirical evidence that metrics become less reliable as the MT system quality increases.

To determine how much we can trust automatic metrics, we quantify how often metric conclusions agree with human decisions, given the difference between metric scores of two MT systems. The academic community regularly uses small differences in BLEU scores to claim a new state of the art, but we find that when the difference in BLEU scores is small, the metric is not good at predicting the result of human judgements. On the other hand, even large BLEU differences do not always guarantee agreement with human decisions. We show that BLEU is clearly outperformed by other metrics, but ultimately, all metrics are an inadequate substitute for high quality human evaluation.

1.2 Thesis Structure

The field of Machine Translation evaluation has a rich history, which we present in Chapter 2. We first chronicle methods used for large-scale human evaluation of multiple systems, focussing on direct assessment (DA), which is the current method used by the annual Conference on Machine Translation (WMT). Next, we introduce various approaches towards designing automatic metrics. We describe popular metrics as well as the current state of the art metrics that form the baselines and benchmarks to our proposed metrics. Finally, we review the methods used to evaluate and compare these metrics.

Chapters 3 and 4 focus on two aspects of improving human evaluation: better aggregation and mitigating cognitive biases. We begin Chapter 3 with a review of probabilistic models to aggregate data from various sources. We present an analysis of behaviour of MT annotators, and then describe our model for MT quality scores. We present results on two multiply-annotated MT adequacy datasets, and show that these models are more accurate than simply averaging scores, which can be improved further when we remove the least reliable annotators. Finally, we show that we can use heatmaps showing pairwise correlation of annotator scores to determine whether we have collected sufficient annotations to yield accurate scores.

In Chapter 4, we first review cognitive biases, focussing on the biases that people are susceptible to when making a sequence of decisions: the *gambler's fallacy*, *sequential contrast effects* and *assimilation effects*. We then present a simple linear model that can be used to detect sequence effects, where the annotator response of the current item is influenced by their response to the previous item. We provide evidence that sequence effects are present in MT adequacy data, as well as other independent crowdsourced datasets in Natural Language Processing.

Chapter 5 moves on to automatic evaluation. We begin with a review of contextualised word embeddings. We describe our proposed new metrics: simple metrics that compute semantic similarity between the embeddings of the MT output and the reference translation, followed by supervised neural models that learn sentence representations and then predict the translation quality. We also explore alternatives to our trained metrics, where we compare MT outputs directly with the source instead of the reference. We present results on WMT datasets in various settings, then provide an error analysis that shows where our metrics are right and where they are wrong.

Chapter 6 explores evaluation of automatic metrics. The first half of this chapter builds on recent findings on the metrics evaluation task at WMT, which indicate that the correlation of automatic metrics is affected by outlier systems, and the correlation falls dramatically when restricted to the top MT systems. We look into the data to understand the significance of these findings: (a) we propose means to identify outlier MT systems and illustrate the effect of these systems when computing correlation between the metric and human scores, and (b) we analyse whether the metrics are less reliable when comparing high quality MT systems. In the second part of the chapter, we focus on the reliability of metric decisions when comparing two systems. Using human judgements as the ground truth, we quantify the errors made by metrics when: (a) a metric can not detect a difference in quality of two systems but humans can and (b) a metric incorrectly concludes that system X is better than system Y when humans either judge system X to be similar or worse than system Y.

We summarise the contributions of the thesis in Chapter 7. We discuss the limitations of this work and present avenues for future work.

Chapter 2

Background: Machine Translation Evaluation

In this chapter, we present a broad introduction to evaluation of machine translation. The first part of the chapter is about human evaluation, where humans are asked to rate the quality of MT system outputs; we focus on large-scale evaluation campaigns where multiple MT systems are compared. We then move on to automatic metrics which are used as a cheaper and faster alternative to human evaluation. We present approaches to designing automatic metrics, and finally, we review meta-evaluation of automatic metrics which are assessed based on their correlation with human judgements.

2.1 Human Evaluation Methodologies

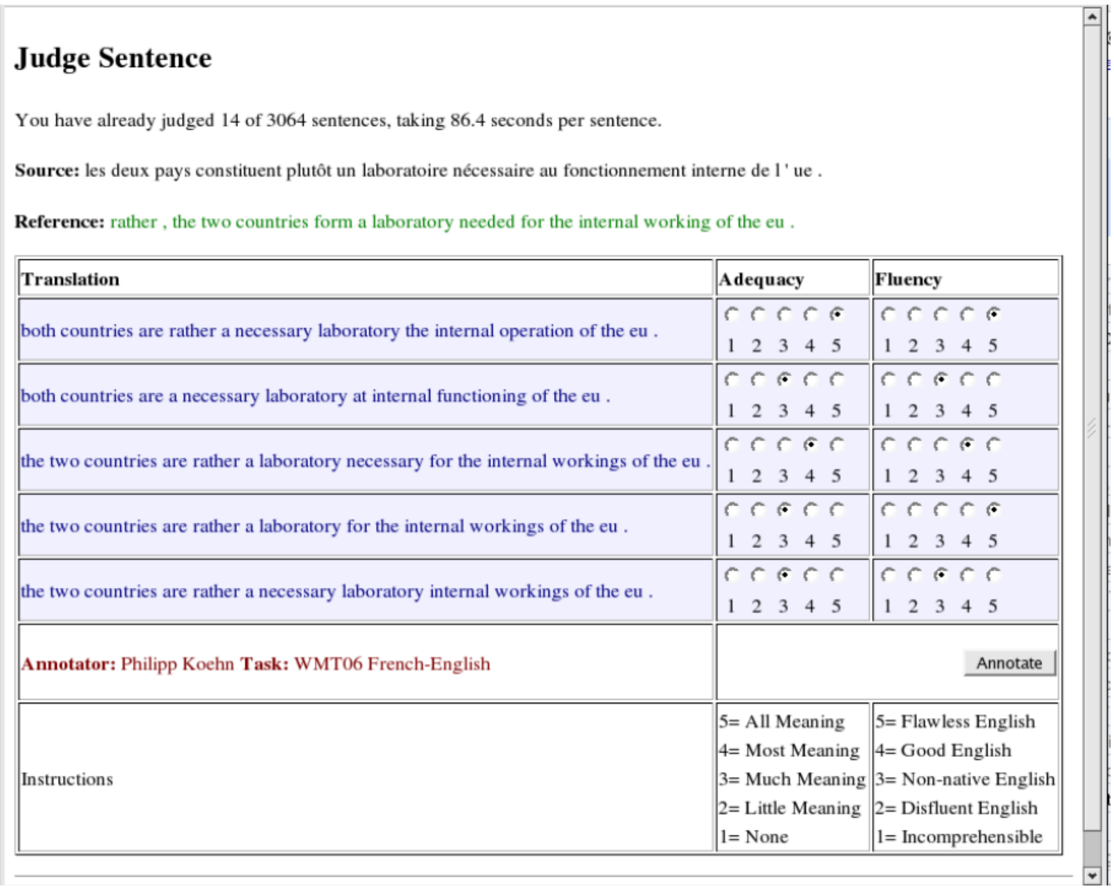
Reliable evaluation is critical when measuring progress of MT research, or acceptability of an MT system for a particular task. MT output is intended for human use, so the best way of evaluating quality is to obtain human judgements.

We use human judgements on data from the Conference on Machine Translation (abbreviated WMT for historical reasons) throughout the thesis, both to demonstrate methods to improve human evaluation, and to evaluate automatic metrics. We begin this section with pre-WMT evaluation campaigns that influenced WMT evaluation, before focussing on the evolution of human evaluation methods at WMT. After a brief detour to review crowdsourcing annotations, we describe direct assessment, the current method at WMT, in more detail.

2.1.1 Early MT Evaluation Campaigns

The Automatic Language Processing Advisory Committee conducted the first large-scale rigorous evaluation of multiple MT systems (Pierce and Carroll, 1966). They asked human experts to rate translations for intelligibility and fidelity. Intelligibility was measured directly on a discrete 9-point scale, ranging from “perfectly clear and intelligible" to “hopelessly unintelligible". To measure fidelity, annotators were presented with the translation first, and then had to rate the informativeness of a reference by an expert translator, when compared to the initial translation. For both tasks, evaluators were provided training, and given guidelines for each available option. In addition, they measured informativeness of an MT system based on scores on reading comprehension tests of the passages translated by MT systems when compared with those by a human expert.

In the early 1990s, the DARPA MT initiative evaluated translations on metrics that were modelled on the metric used by the US government to evaluate work by professional human translators (White et al., 1994; White and O’Connell, 1994). Annotators were asked to assess fluency and adequacy of translations on a discrete 7-point scale. Fluency is intended to capture the grammaticality and idiomatic word choice of the translated sentence, irrespective of the accuracy of the information. On the other hand, adequacy is intended to measure how much information in the source sentence is preserved in the translation, without taking fluency into account. Fluency was evaluated on whole sentences. Adequacy ratings were collected at the level of linguistic components of a sentence of length ranging between 5 to 20 words, which



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	○ ○ ○ ○ ● 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	○ ○ ○ ● ○ 1 2 3 4 5	○ ○ ○ ● ○ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ○ ○ ● 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	○ ○ ● ○ ○ 1 2 3 4 5	○ ○ ● ○ ○ 1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Fig. 2.1 Screenshot of the evaluation interface used to collect adequacy and fluency ratings in WMT 2006 and 2007 (Koehn and Monz, 2006).

were extracted using information from parse trees. The human evaluation of the NIST Open MT challenges in the 2000s was based on the DARPA method (Consortium, 2002; NIST, 2002).

2.1.2 WMT Evaluation

Since 2006, the Conference on Machine Translation (WMT)[1] has organised a large scale human evaluation of the MT systems submitted in the translation task.

[1] Previously Workshop on Statistical Machine Translation (2006-2015). In 2016, the name was changed to Conference on Machine Translation, but it has retained the acronym of WMT.

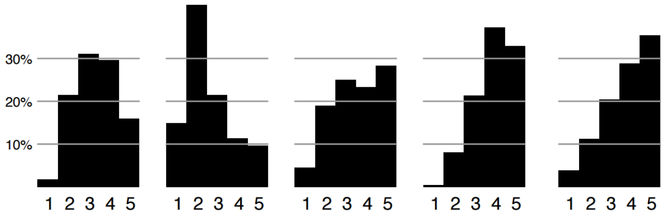


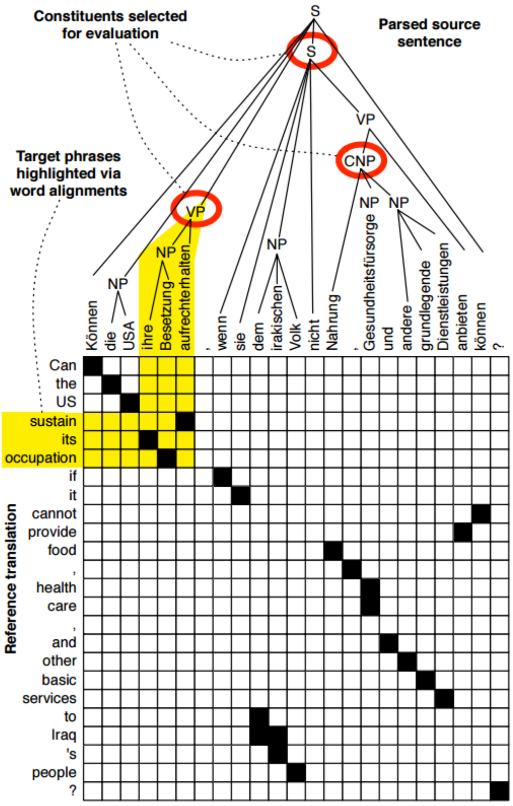
Fig. 2.2 Distribution of adequacy scores given by five different judges in WMT 2006 (Koehn and Monz, 2006)

Adequacy and Fluency

The human evaluation at the first WMT (Koehn and Monz, 2006) was influenced by the evaluation at the NIST worshop. Judgements were collected by the participants of the news translation task, a tradition that has continued till date.

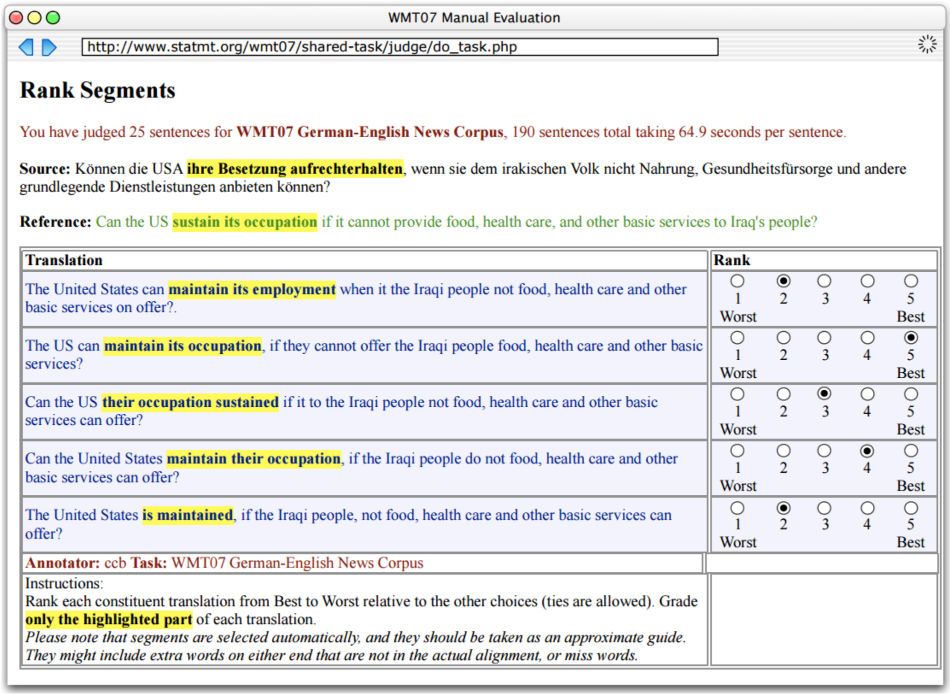
Annotators were asked to rate the adequacy and fluency of a set of five MT system outputs on a five-point scale based on the source and reference translation (Figure 2.1). The evaluators reported that it was difficult to assign scores to long translations riddled with multiple errors. They developed their own rules of thumb to decide between categories. This leads to different distributions of scores for each annotator (Figure 2.2): some are more lenient than others and aggregating these judgements into one final score is not straightforward.

In addition, there was a high correlation between fluency and adequacy scores. Fluency and adequacy are expected to be naturally correlated, for example, a highly disfluent translation is likely to be incomprehensible, and thus highly inadequate. In the WMT evaluation (Koehn and Monz, 2006), this relationship could have been exacerbated due to both fluency and adequacy being presented together, and the presence of the source translation and reference biasing the annotator to consider the meaning of the translation even when evaluating fluency. Furthermore, the presence of multiple translations on the screen enables annotators to consider it a ranking task instead of providing absolute scores. Assessing the fluency and adequacy of each translation independently could mitigate these sources of bias.



(a).

(b.)



Rank Segments

You have judged 25 sentences for **WMT07 German-English News Corpus**, 190 sentences total taking 64.9 seconds per sentence.

Source: Können die USA **ihre Besetzung aufrechterhalten**, wenn sie dem irakischen Volk nicht Nahrung, Gesundheitsfürsorge und andere grundlegende Dienstleistungen anbieten können?

Reference: Can the US **sustain its occupation** if it cannot provide food, health care, and other basic services to Iraq's people?

Translation	**Rank**
The United States can **maintain its employment** when it the Iraqi people not food, health care and other basic services on offer?.	○ 1 Worst ● 2 ○ 3 ○ 4 ○ 5 Best
The US can **maintain its occupation**, if they cannot offer the Iraqi people food, health care and other basic services?	○ 1 Worst ○ 2 ○ 3 ○ 4 ● 5 Best
Can the US **their occupation sustained** if it to the Iraqi people not food, health care and other basic services can offer?	○ 1 Worst ○ 2 ● 3 ○ 4 ○ 5 Best
Can the United States **maintain their occupation**, if the Iraqi people do not food, health care and other basic services can offer?	○ 1 Worst ○ 2 ○ 3 ● 4 ○ 5 Best
The United States **is maintained**, if the Iraqi people, not food, health care and other basic services can offer?	○ 1 Worst ● 2 ○ 3 ○ 4 ○ 5 Best
Annotator: ccb **Task:** WMT07 German-English News Corpus	
Instructions: Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the highlighted part** of each translation. *Please note that segments are selected automatically, and they should be taken as an approximate guide. They might include extra words on either end that are not in the actual alignment, or miss words.*	

Fig. 2.3 Figure illustrating (a) the extraction of syntactic constituents from the source sentence and alignment with the MT output, and (b) screenshot of the Evaluation interface used to collect judgements for syntactic constituents in WMT 2006 (Koehn and Monz, 2006).

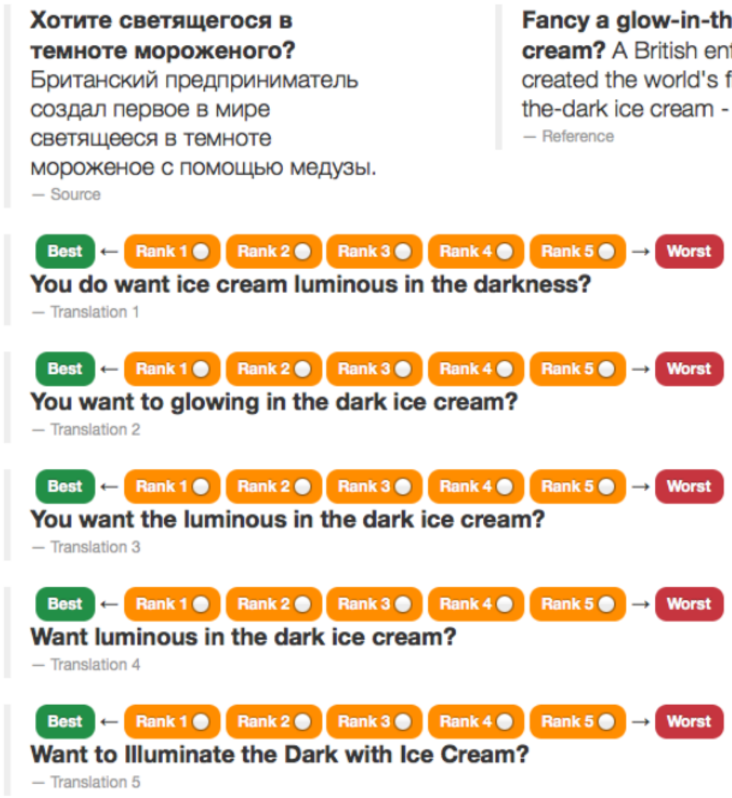
In the next workshop (Callison-Burch et al., 2007), the organisers introduced alternative evaluation methods: sentence ranking, ranking of syntactic constituents and binary acceptability of these constituents.

Syntactic Constituent Ranking and Acceptability Judgements

Syntactic constituents between 3 and 15 words were automatically extracted from the source sentence based on the output of parse trees, and mapped automatically to system translations (Fig. 2.3 a) using a word alignment tool such as GIZA++ (Och and Ney, 2003). Annotators are presented with the source, reference and MT system outputs with the selected constituent highlighted, and are asked to rank them (Figure 2.3b) or to judge binary yes/no acceptability, depending on the task. Only constituents which can be aligned to all five systems are used, and this possibly introduces bias in the evaluation. Annotators for constituent ranking and judgements were the fastest and most consistent, as measured by both inter- and intra-annotator agreement. In addition, binary judgements offer the potential for re-use when evaluating another system on the same test set. Nevertheless, these methods were discontinued after two years, possibly due to difficulty in extracting constituents and aligning them to the translations.

Sentence Ranking

Sentence ranking was used as the official measure of evaluation between 2007 and 2016. Annotators are presented with a block of three consecutive source sentences, and are asked to rank a set of five system translations of each source sentence (Figure 2.4). Although this method is simpler, annotator agreement is still very low; in WMT 2016, inter-annotator agreement was only 0.357, as measured by Cohen's Kappa. Systems can be unfairly advantaged (or disadvantaged) if they are often compared to systems that are very bad (or very good). In addition, information on the degree of difference between translations is lost, and it is not



Хотите светящегося в темноте мороженого? Британский предприниматель создал первое в мире светящееся в темноте мороженое с помощью медузы.
— Source

Fancy a glow-in-the-dark ice cream? A British entrepreneur has created the world's first glow-in-the-dark ice cream - using jellyfish.
— Reference

You do want ice cream luminous in the darkness?
— Translation 1

You want to glowing in the dark ice cream?
— Translation 2

You want the luminous in the dark ice cream?
— Translation 3

Want luminous in the dark ice cream?
— Translation 4

Want to Illuminate the Dark with Ice Cream?
— Translation 5

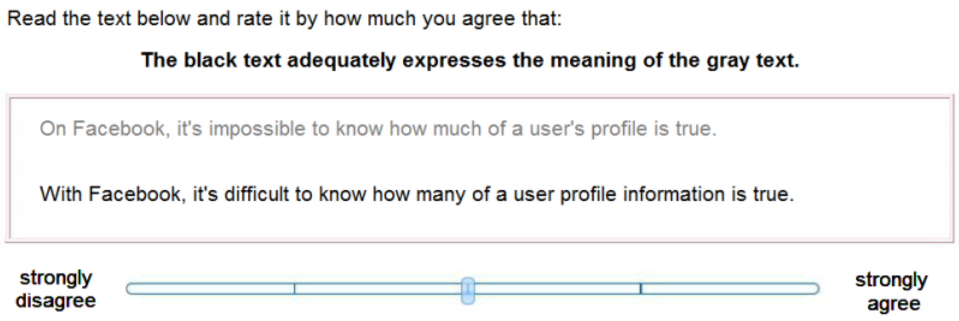
Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst
Best ← Rank 1 ○ Rank 2 ○ Rank 3 ○ Rank 4 ○ Rank 5 ○ → Worst

Fig. 2.4 Screenshot of Appraise, the tool used to collect translation rankings at WMT (Bojar et al., 2014).

straightforward to compute a final ranking of all systems from multiple partial rankings (Lopez, 2012; Hopkins and May, 2013; Sakaguchi et al., 2014).

Direct Assessment

In 2016, a new method to collect absolute judgements (Graham et al., 2013) was trialled, and has been officially adopted since 2017. Annotators score translation quality using a visual analogue scale (Fig. 2.5), which is mapped to continuous scores that range between 0-100. The score of an MT system is the average score of all its translations. This method has a high correlation with sentence ranking, but requires considerably fewer annotations as the number of assessments per system scales linearly instead of quadratically. Finally, annotators are not



Read the text below and rate it by how much you agree that:

The black text adequately expresses the meaning of the gray text.

On Facebook, it's impossible to know how much of a user's profile is true.

With Facebook, it's difficult to know how many of a user profile information is true.

Fig. 2.5 Screenshot of the annotation interface for direct assessment (Graham et al., 2013).

forced to choose between categories, and a wider array of statistical methods can be applied on continuous scores to solve the problem of differing internal scale.

Direct assessment also contains built-in measures for quality control, that allows it to be crowdsourced. This thesis relies heavily on DA: we present methods to improve the collection and aggregation of crowdsourced DA judgements, and primarily use DA scores to evaluate automatic metrics. We describe direct assessment in detail in Sec. 2.1.4, but before that, we make a small detour to crowdsourcing. This next section contains a brief description of crowdsourcing annotations and early experiments on obtaining crowdsourced annotations in the area of natural language processing and more specifically, machine translation evaluation, along with simple attempts to solve the challenges that come with crowdsourcing.

2.1.3 Crowdsourcing Annotations

Howe (2008) defines crowdsourcing as "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call." Platforms such as Amazon Mechanical Turk and Appen allow *requesters* to pay *workers* for completing *Human Intelligence Tasks* (HITS). Workers (also called "Turkers" on Amazon Mechanical Turk) are free to choose HITS that interest them, and are paid on completing the HIT. Requesters have the option to reject payment for poor quality work, or to offer bonuses for excellent work. The task needs to be designed carefully to accommodate untrained workers and filter out low quality annotations from workers who do not complete the task in good faith.

Typically, HITS include quality control items to help identify poor quality work. For example, items with known answers can give an estimate on worker accuracy. And finally, to obtain more reliable labels, multiple annotations are collected for each item, which can then be aggregated. Typically, this means using the majority label for categorical tasks, and the arithmetic mean for continuous labels. However, this ignores the differences in reliability and expertise of annotators, and we can potentially obtain better results when we model annotators.

Crowdsourcing in NLP and MT

Snow et al. (2008) were the first to systematically evaluate crowdsourcing data for natural language processing. They obtained and analysed data for five tasks that require different kinds of annotations: affective text analysis (continuous ratings between 1 and 100 for six emotions), word similarity (ordinal scale between 1-10), textual entailment (binary yes/no judgements), temporal event annotation (binary judgements), and word sense disambiguation (varying number of senses). Simple aggregation of ten annotations per item was sufficient to match the performance of human experts, in a fraction of the time and at considerably less expense. They didn't employ any methods to filter out low-quality annotations, but experimented with methods to model the reliability and bias of workers, which we describe in the next chapter (Sec. 3.2).

In Machine Translation, crowdsourcing has been used to obtain parallel corpora (Ambati et al., 2010), create reference translations (Bloodgood and Callison-Burch, 2010), obtain paraphrases to supplement existing references (Denkowski et al., 2010) as well as evaluate MT systems. Zaidan and Callison-Burch (2011) have used MTurk to crowdsource translations, with additional HITS to post-edit, and then filter out bad translations.

To test the possibility of using crowdsourced judgements for ranking MT systems, Callison-Burch (2009) used Amazon Mechanical Turk to replicate the translation ranking judgements of WMT 2008. This task is to rank a set of five translations of the same source sentence, and they obtained 5 repeat annotations for each set of translations. For each pair of translations in the set, they used the weighted majority of worker judgements to decide which translation is better. They experimented with two criteria for weighting workers: (a) agreement with experts on 10 initial judgements, and (b) agreement with the four other workers. The Spearman correlation of system ranks[2] using these crowdsourced judgements and expert judgements matched the expert-expert correlation (around 0.8). Weighting workers improved agreement between workers, but made no significant difference to correlation of MT system ranks with expert judgements. Crowdsourced judgements were used to supplement expert judgements between WMT 2010 to 2013 (Callison-Burch et al., 2008; Bojar et al., 2013), but were subsequently dropped due to very low agreement rates.

Denkowski and Lavie (2010b) collected adequacy ratings for machine translation output from English to Arabic on an ordinal scale of 1-4. They collected scores from 10 workers per translation, and also collected expert scores for around 10% of the translations. They tried different strategies to improve the quality of the data collected from Mechanical Turk, some of which were inspired by Callison-Burch (2009):

1. Removing Low-Agreement Judges: Remove workers with low inter-annotator agreement with other workers, then compute the average score of remaining workers.

2. Removing Outlying Judgements: For every translation, remove scores whose distance from the average score exceeds a threshold, then recompute the average.

3. Weighted Voting: weight each worker's scores based on their agreement with expert scores.

[2] An MT system's scores were calculated as the average number of times that it was preferred to any other system.

4. Scaling Judgements: Shift the scores of each worker by a constant such that their mean score matches the average of expert scores on the set of overlapping translations.

In the end, they chose a two-stage normalisation scheme that relies on the availability of expert judgements: First, the scores of workers that consistently score above or below the expert scores are scaled. The final score was the average score weighted by agreement with the gold standard, after discarding workers with poor quality annotations. This two-stage method resulted in improving sentence-level correlation [3] with expert judgements to 0.487, which is a substantial improvement from the correlation of the raw data (0.078). While the quality of these crowdsourced judgements is considerably low, they are also a lot cheaper to obtain compared to expert judgements.

When crowdsourcing judgements on a continuous scale, as done with direct assessment, it is difficult to directly use the usual controls established for quality control of discrete or ordinal labels. We can not expect annotators to exactly replicate the score by a human expert, as it is possible that they are more or less lenient than the expert; their scores might disagree with the "gold" scores, but be internally consistent. This necessitates other creative methods to filter out low-quality data. The next section describes DA and its quality control methods.

2.1.4 Direct Assessment

Graham et al. (2013) proposed using continuous scale judgements to evaluate MT. Annotators are asked to rate the MT output using a continuous slider, which maps to an underlying scale of 0-100. The slider has markings that divide the scale into four equal parts to help annotators with internal calibration.

To measure adequacy, annotators are asked to rate the similarity between an MT output and a human reference translation. In a separate task, annotators are asked to rate the fluency of the

[3] The type of correlation is not specified.

sentences. They are not shown the reference during fluency assessment, and this serves as an independent evaluation of the MT outputs, without any potential for reference bias.

To reduce costs, these annotations are crowdsourced using Amazon Mechanical Turk. The HITs are designed to filter out inconsistent workers and obtain enough judgements to reduce worker bias. Each HIT contains 100 items:

- 70 outputs of the MT systems being evaluated
- degraded versions of 10 of these translations,
- 10 reference translations by a human expert, corresponding to 10 system translations, and
- repeats of another 10 translations.

There is a gap of 40 sentences between an MT output and the corresponding quality control item (repeat, reference, or degraded translation), to minimise chances of annotators remembering the score they assigned earlier. The scores on the quality control items are used to filter out workers who either click randomly or on the same score continuously. The repeat sentences are used to measure intra annotator consistency. A conscientious worker would give a near perfect score to reference translations, and give a lower score degraded translations when compared to the corresponding MT system translation. Other indications of low quality work are clicking the same score in sequence, or completing the 100 annotations in a few seconds. After filtering out workers who do not meet the quality control requirement, annotations for the 70 MT outputs and the 10 repeats are used in the evaluation.

In the next chapter, we show that this quality control mechanism often filters out useful data, unnecessarily increasing the cost of the evaluation.

Individual workers may have their own internal scale, and to make the worker score distributions more consistent, every individual worker's scores are standardised by subtracting the worker's mean, and dividing by their standard deviation. The final score of an MT system is the mean score of all its translations. While continuous scores of individual translations may be noisier when compared to interval level ratings, this noise is expected to cancel out when a large number of translations are averaged.

Accurate Scores for Individual Translations

Manual MT evaluation is subjective and difficult, and it is not possible even for a diligent human to be entirely consistent on a continuous scale. Thus, any human annotations are noisy by nature.

To obtain accurate scores of individual translations, multiple judgements are collected and averaged (Graham et al., 2015). This is validated by the law of large numbers, which states that for independently, identically distributed (i.i.d) samples, the sample mean approaches the population mean as the size of the sample increases. This was empirically tested by obtaining two independent sets of judgements: for a sample size of 15, the mean scores have a correlation greater than 0.9 for all language pairs, and sample size of 40 are almost perfectly correlated for Spanish-English translations. We verify this in Sec. 3.5, and propose a better method to aggregate these scores.

However, it is possible that the i.i.d assumption is violated, and all annotators have the same bias. In Chapter 4, we explore one such source of bias.

Direct Assessment at WMT

Direct assessment has been used to collect human judgements at WMT since 2016, and the methodology has been refined over the years. This section presents the changes to DA to improve reliability.

The annotations were entirely crowdsourced in 2016, and it was difficult to source skilled workers to evaluate translations in languages other than English. Since 2017, DA adequacy judgements have been the official evaluation method at WMT (Bojar et al., 2017b). The evaluation of English translations is crowdsourced, and judgements for translations into other languages are collected from participants of the WMT news translation shared task. The correlation of MT system scores sourced through researchers and crowd workers was above 0.98 for three language pairs, confirming the reliability of the crowdsourced evaluation. Fluency evaluations were dropped, probably due to high correlation with adequacy scores and to save costs.

In 2018, WMT organisers investigated bilingual assessment for English → Czech translations, where MT was compared directly with the source input and the reference translations were not displayed to the annotators (Bojar et al., 2018). This helps remove reference bias (Fomicheva and Specia, 2016) and also guards against potentially low-quality reference translations. Finally, since the reference translation is not required to evaluate MT systems, it can instead be included in the evaluation as a benchmark for MT systems. Since 2019, all evaluations in languages other than English have been bilingual (Barrault et al., 2019). Monolingual evaluation is still used for translations into English where the DA judgements are crowdsourced.

When constructing HITs in the original proposal of DA, annotators were presented randomly sampled translations from all MT systems included in the evaluation. Since most MT systems translate MT sentences independently, their translations of a document can be incoherent even if individual sentences appear to be of high quality (Läubli et al., 2018; Toral et al., 2018; Graham et al., 2020). In 2019, the evaluation was updated to included document context: annotators score translated sentences of an MT system in document order.

2.2 Automatic Metrics

While human judgements are more reliable for evaluating machine translation systems, they are expensive and time-consuming. The need for cheap, immediate feedback motivates the development of automatic methods. Automatic metrics evaluate MT system performance by comparing the semantic similarity of an MT system output (also referred to as a hypothesis or candidate translation in the literature) with one or more reference translations provided by professional human translators. Automatically determining the validity of variations in the MT output from the reference is not easy, and this is still an active research area.

In addition to comparing MT systems, automatic metrics play a vital role in tuning the parameters of statistical MT systems (Och, 2003; Watanabe et al., 2007). In recent years, reinforcement learning approaches have been designed to directly optimise neural MT systems to BLEU (Wu et al., 2018).

This thesis heavily focuses on automatic metrics: in Chapters 3 and 4, we focus on human annotations which are used to train and evaluate metrics; in Chapter 5, we propose new automatic metrics; and in Chapter 6, we revisit evaluation of these metrics.

This section provides a detailed background on automatic metrics. We begin with the desiderata for automatic metrics. We provide historical context into the development of automatic metrics and we provide more detail about the challenges for designing metrics to evaluate machine translation. We introduce BLEU, the first metric that attempted to solve these challenges. We then summarise various approaches taken by existing metrics to improve on BLEU, and identify areas of improvement that serve as the basis for our metrics in Chapter 5. We present a brief overview of reference-free automatic metrics. We describe selected metrics in more detail: BLEU, METEOR, TER and chrF, which are our baselines, and YISI-1 and RUSE, which were the state of the art at the time when we developed our metrics. In Chapter 6, when we are re-evaluating metric evaluation, we focus on a subset of these metrics. Finally, we present a detailed overview of methods to evaluate automatic metrics (meta-evaluation), focussing on measures used at the annual WMT metrics shared task which drives research into automatic metrics today.

2.2.1 Criteria

What are the criteria of an effective automatic metric?

The most important requirement is that it reflect the true quality of MT systems. While human evaluation is not perfect, a carefully designed evaluation is the closest we have to the ground truth, and automatic metrics are evaluated based on their fidelity to human scores. This is typically measured by the correlation of metric scores on a set of MT systems against human scores. It is also useful to understand which sentences are translated well or poorly, and we would also like our metrics to correlate with human judgements when scoring individual sentences (see Sec. 2.3 for more details on evaluating metrics).

Other criteria for metrics include:

- High discriminative power: we need metrics to correctly choose the best system among systems with very similar quality.
- Cross-lingual portability: it should be easily portable to new languages.
- Immune to adversarial attacks: when MT systems are directly optimised to metrics, it is important that they do not exploit loopholes in the metric that allow them to obtain high metric scores that do not reflect the true quality.
- Speed and usability: a metric that satisfies all these criteria would not be adopted by the research community unless it is freely available, easy to use, and reasonably fast when scoring a test set.

It would be truly difficult to design a one-size-fits-all perfect metric. Many are developed with English or other European languages in mind, and are directly ported to other languages depending on the availability of the external resources required. A metric that is reliable when evaluating translations in one language may not be reliable over other languages. Languages differ widely in how words are formed, and how much information is contained in a single word. With highly analytic languages like Chinese, a word typically represents a single concept. On the other extreme are polysynthetic languages like Inuktitut, where an entire clause can be represented in a single word (Micher, 2017). In addition, the importance of word order differs across languages, partly due to a difference in morphological complexity (Comrie, 1989).

Furthermore, there might be variance in metric reliability even when evaluating MT systems translating from different languages into the same language. For example, if the two languages are from the same language family, then it might be easy for MT systems to get the word order right. In this case, a metric that doesn't contain a penalty for wrong word order could still be reliable. This would not hold when evaluating translations from a distantly related language where it is more likely that MT systems make critical mistakes with the word order.

And finally, even when translating from the same source language to the same target language, metric reliability depends on the set of systems evaluated, the domain, and possibly even the test set and the references.

2.2.2 Approaches to Designing Automatic Metrics

In this section, we present a history of the development of automatic metrics, reviewing the various approaches used to compute translation quality when comparing the MT output with a reference translation. We then present selected metrics in more detail in Sec. 2.2.4 The first automatic metric used in MT research, the word error rate (WER), was borrowed from the speech recognition community (Olive et al., 2011). This is based on the Levenstein distance (Levenshtein, 1966) of the MT system output with the reference translation, that computes the minimum number of words that need to be substituted, inserted or deleted to change the candidate to the reference. WER is computed as the total number of errors normalised by the length of the reference.

$$WER = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{reference length}} \tag{2.1}$$

However, such a straightforward string comparison to an independently translated reference does not work well, because there are multiple valid ways to translate the same source sentence. When comparing with a reference, any variation in word choice or phrase order of the candidate from the reference could either mean the translation is still valid or that it has errors of varying severity. For example, a valid paraphrase of the sentence would preserve the meaning but possibly have a large WER. On the other hand, a missing negation is a small deviation from the reference with a drastic impact on meaning. Automatically evaluating MT is an active research area, and there are a variety of approaches to solve this problem.

An early approach to allow for different choices in phrase-ordering was position-independent error rate (PER), which was introduced as an alternative to WER in a paper that presented an algorithm to speed up the decoding step of the translation process of statistical MT systems (Nießen et al., 1998). PER ignores word order completely and considers the sentence as a bag of words. It counts the number of substitutions, and either deletions or insertions required such that all the words in the translation are matched with the words in the reference. Where WER harshly penalises any deviation from the reference, PER gives high scores to completely scrambled translations as long as there is lexical overlap. Neither metric recognises valid word variation in word choice.

One way to mitigate this was to introduce multiple references: Alshawi et al. (1998) computed WER of the translation with multiple references, and chose the reference closest to the translation, i.e., the one that gives the minimum error. However, the closest "correct" translation is unlikely to be a part of the available set of references, and this still underestimates the translation quality.

BLEU was developed in 2002 by researchers at IBM specifically to evaluate MT systems (Papineni et al., 2002). BLEU computes the precision of n-grams of the translation compared to the reference, rewarding local ordering of words while not requiring the same global order. BLEU was designed to exploit multiple references more fully than WER: n-grams in the hypothesis are marked correct if they match with *n*-grams from any of the available references. These references were expected to be produced by translators with different styles, with the hope that this would mitigate any biases towards or against any particular MT systems. Even if the metric was not always accurate when evaluating individual sentences, these errors would cancel out when computed across the whole test set. The metric was introduced with some carefully-thought out measures to avoid potential pitfalls. Most importantly, it was supported with empirical data: it had a high correlation with human adequacy and fluency scores over a set of three commercial MT systems and two human translators (one a native speaker and the other not) translating from Chinese to English.

BLEU was very quickly adopted by the MT community. This popularity was further cemented by its use in tuning statistical MT systems: in 2003, methods were proposed to directly optimise to automatic metrics instead of maximising likelihood (Och, 2003). The authors tested several metrics including WER, PER and BLEU, but made no definitive conclusions as to the best metric. However, by 2004, BLEU appears to have become the chosen metric for both system tuning (Shen et al., 2004; Och et al., 2004) and evaluation (Koehn, 2004).

It also sparked research in understanding the usefulness of this metric (Doddington, 2002; Coughlin, 2003), and prompted the development of metrics that identify and address its limitations. These metrics use a wide variety of approaches to compute a similarity score between the MT system output and the reference(s): (a) exact matches: rewarding lexical similarity of words, *n*-grams, or, more recently, characters; (b) synonyms and paraphrases: rewarding matches of words or phrases that have the same meaning and can be used interchangeably; and (c) embeddings: moving from symbolic to distributed representations allows for a continuous measure of similarity between words. Finally, some metrics use additional linguistic information such as part of speech, constituency and dependency trees, semantic roles and discourse roles.

Many metrics combine these components with rules or heuristics to obtain a single score (as with BLEU and WER). Some metrics such as METEOR and TER use a development set to tune the weights of individual components (Snover et al., 2009; Denkowski and Lavie, 2010a; Lo and Wu, 2013). Finally, other metrics are fully supervised, where they use a selection of these individual components as features (Albrecht and Hwa, 2007; Song and Cohn, 2011; Fishel et al., 2012; Stanojević and Sima'an, 2014). Some of these metrics are ensembles of other existing metrics (Joty et al., 2014; Yu et al., 2015; Ma et al., 2017). In recent years, we have seen the rise of neural end-to-end metrics that learn sentence representations from scratch or beginning with pre-trained embeddings (Gupta et al., 2015; Shimanaka et al., 2018). The supervised metrics are trained on human evaluation data; depending on what data is available, they are either framed as a regression task directly to predict scores, or use a learning to rank framework (Li, 2011) to distinguish between different translations of the same source sentence.

In the next section, we describe the key components that existing metrics use when matching the contents of the two sentences, and how they use this information to compute the final metric score.

Exact Matching: Word and Character Level

Exact word matching, as with WER and BLEU, is an obvious criterion for measuring sentence similarity.

Metrics like CDER and TER are based on modifications to edit distance, in an attempt to find the middle ground between WER and PER. When computing WER, if an MT system generates a phrase of n words in a correct, but different position to the reference, WER counts this movement as n additions and n deletions. CDER (Leusch et al., 2006) and TER (Snover et al., 2006) are modifications of the Levenshtein distance that reduce this harsh penalty for movement of phrases.

Other metrics follow BLEU in computing overlap between word n-grams, and use various strategies for combining this information. NIST (Doddington, 2002) is a direct modification of BLEU, which, among other tweaks, weighs n-gram matches based on their frequency in the test set. The ROUGE family of metrics (Lin and Och, 2004) computes n-gram recall instead of precision, and introduces skip-bigrams which allow for gaps between the two words matched.[4] GTM introduced the F-score (the harmonic mean of recall and precision) over matched n-grams, as a way to balance between the two components that measure whether all content in the translation is correct and whether all the content in the reference is available in the translation. Other strategies include computing an alignment between the words in the translation and reference, then computing precision, recall or F-score on the matches in the alignment. For example, METEOR (Banerjee and Lavie, 2005) includes a greedy aligner, and MAXSIM and TESLA (Chan and Ng, 2008; Liu et al., 2010) compute an optimal alignment.

Finally, in any sentence, some words are more important than others towards expressing the general idea of a sentence. For example, when comparing *Arrietty borrowed a sugar cube* and *Arrietty stole sugar cube*, the loss of *a* renders the sentence less fluent, but does not make much difference to adequacy. For this reason, some metrics assign a lower weight to function words (Denkowski and Lavie, 2010c), or use idf-weighting that gives more importance to less frequent words (Lo, 2017). This also helps with reducing the reward for spurious matches of highly frequent words like *the*. However, this can be problematic. For example, *with* can be considered a function word, but when comparing *she ate with a spoon* and *she ate a spoon*, removing the word makes the sentence slightly nonsensical and definitely incorrect.

Languages like Chinese and Japanese do not use white space as a word delimiter and require a pre-processing step for word segmentation. Metrics were designed specifically for these languages which compute the n-gram similarity at the character level instead; skipping the noisy word segmentation step improves correlation of the metric (Li et al., 2011; Liu and Ng, 2012).

[4] ROUGE was originally developed to evaluate automatic text summarisation.

Computing character-level similarity also helps with matching morphological variants of words in all languages, and are particularly useful for languages like Finnish and Russian that have rich morphology. Character-level metrics like CharacTER (Wang et al., 2016) and EED (Stanchev et al., 2019), which compute edit distance over characters, and chrF (Popović, 2015), which computes the F-score of character n-grams, have the simplicity, efficiency and portability of BLEU. And they typically outperform not just BLEU, but also some fairly complex metrics. BEER is a trained metric, whose features included word matching and precision, recall and F-score of character n-grams, as well as syntactic features and permutation trees to model reordering (Stanojević and Sima'an, 2015). Later versions of the metric (Stanojević, 2017) only consisted of lexical features as the high correlation was mostly attributed to the character-level features; any gain in correlation from including the more complex features was not worth the decrease in efficiency.

Flexible Matching: Synonyms and Paraphrases

Requiring exact matches with the references means that metrics assign lower scores to translations with a valid but different word choice from the reference. While this can be mitigated to some extent by using multiple references, it is expensive to generate multiple references, and a small set of references can not capture the entire set of valid variations. This motivates the use of more flexible matching.

In addition to surface-level lexical comparison, some metrics use external resources to match synonyms and paraphrases. Princeton WordNet (Fellbaum, 1998) is a manually constructed lexical database that contains semantic relationships between more than a hundred thousand words. In particular, it groups synonyms, words which refer to the same concept and can largely be used interchangeably, into *synsets*. METEOR introduced the concept of synonym matching to automatic MT evaluation. The METEOR aligner matches words in the translation and reference if they belong in the same WordNet synset. Including WordNet synonyms improved correlation for METEOR, and the strategy was also adopted by metrics like TESLA and TER-PLUS.

This variation in expressing the same concept is not restricted to the word level; phrases that have low surface-level similarity can have the same meaning. Moreover, developing a wordnet requires several years of human effort, and mature wordnets are available only for a limited set of languages. Paraphrase tables can be computed automatically from parallel corpora for any language (Bannard and Callison-Burch, 2005). Many metrics incorporate paraphrase similarity, either before computing the alignment between the two sentences (Denkowski and Lavie, 2010c; Marie and Apidianaki, 2015), or using paraphases as additional references when computing the score (Pang et al., 2003; Barančíková, 2014). Synonym and paraphrase matches are not unlikely to be noise, and the word context plays a crucial role in determining whether a word can be replaced with a synonym or not. Most metrics address this issue by simply assigning a lower weight to these matches, but some attempt to determine the validity of the substitution by training a classifier (Kauchak and Barzilay, 2006) or using existing word sense disambiguation tools (Marie and Apidianaki, 2015).

Flexible Matching with Embeddings

So far, we have treated words (or groups of words) as symbolic units. When comparing them, we look at exact matches, or use synonym or paraphrases to match words that can be used interchangeably in a given context. So the matching is binary: two words or phrases are either matched or not. However, meaning similarity is a gradient: the word *ancient* does not have the same meaning as *past*, *old* or *history*, but it is more similar to these words than *gift*, *amorphous* or *physics*.

Word embeddings map each word to a dense vector in high dimensional space, such that semantically similar words are closer in the space compared to words with very different meanings. These distributed representations can then be used to automatically compare word similarity in a continuous space, for example, by computing the cosine similarity between the embeddings of the two words.

Metrics extend BLEU directly by also rewarding fuzzy *n*-gram matches in addition to exact matches (Wang and Merlo, 2016; Tättar and Fishel, 2017). Servan et al. (2016) and Fomicheva et al. (2015) use embedding similarity in addition to or in place of synonym matching, to reward fuzzy matches if the word embedding similarity is greater than a threshold that needs to be carefully chosen to balance between precision and recall. MEANT 2.0 (Lo, 2017) directly uses the cosine similarity of the word embeddings.

However, classic word embeddings are independent of word context, and context is captured instead using hand-crafted features or heuristics. For example, UPF-COBALT (Fomicheva et al., 2015) uses additional rules to only reward matches where the words are in the same context. More recently, methods for creating contextual word embeddings (Peters et al., 2018; Devlin et al., 2019) have been developed, where the representation of a word is dependent on its context in the sentence. When comparing embeddings of the same word in two sentences, the similarity will be higher if the word context is similar in both sentences. In Chapter 5, we use off-the-shelf contextual word embeddings to design a metric that is highly effective despite, or perhaps because of, its simplicity.

Another approach is to match the sentence embeddings of the reference and translation. The DREEM metric (Chen and Guo, 2015) computes sentence embeddings as a concatenation of three different methods: one-hot representations based on the counts of the words in the sentence, the average of pre-trained word embeddings, and unsupervised sentence embeddings computed from unsupervised recursive auto-encoders. The final score is the cosine similarity of the sentence embeddings. This method only requires pre-trained word embeddings, and is not further trained on human annotations.

Metrics such as UOW-REVAL and RUSE directly learn embeddings of the entire translation and reference sentences, and are trained end-to-end to predict scores. UOW-REVAL (Gupta et al., 2015) learns sentence representations of the MT output and reference translation as a Tree-LSTM (Tai et al., 2015), and then models their interactions using the element-wise difference and the Hadamard product between the two representations. RUSE (Shimanaka et al., 2018) has a similar architecture, but it uses pre-trained sentence representations instead of learning them from scratch on the training dataset

In both these "neural" metrics, the two sentence representations are learned independently, and are then compared. To improve on this, we look to other NLP tasks that have sentence-pairs as inputs such as natural language inference (Bowman et al., 2015), paraphrase detection, and semantic textual similarity (STS) (Cer et al., 2017). These tasks share similarities with the automatic MT evaluation: a good translation entails the reference and vice-versa, and an irrelevant or wrong translation would be neutral or contradictory compared to the reference. The similarity with paraphrase detection and STS is even more direct, except that MT outputs are not always fluent. For these tasks, models that include pairwise word interactions when learning sentence representations have a higher accuracy than systems that process the two sentences independently (Rocktäschel et al., 2016; Chen et al., 2017; Wang et al., 2017). In Chapter 5, we adopt this idea of learning sentence representations conditioned on the other sentence to the task of automatic MT evaluation, which we then use to predict translation quality. Additionally, we build on contextual word embeddings when learning these sentence embeddings.

Since we designed these metrics, there has been other work that uses contextual embeddings. Most notably, BERTSCORE uses similar ideas as our pre-trained metrics, and was developed independently and released during the review period of our paper. In Chapter 4, we talk about this metric in more detail and compare it with our work. YISI-1 was updated in 2019 to use contextual word embeddings instead of word2vec. And more recently, there have been a plethora of supervised metrics that build on contextual word embeddings such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020a).

Other Approaches

Other metrics use additional resources to make more linguistically-informed matching. Instead of lexical overlap, some syntax-based metrics compute overlap over part of speech tags, syntactic constituents, dependency chains or semantic roles (Liu and Gildea, 2005; Giménez and Màrquez, 2007a; Mehay and Brew, 2007; Popović and Ney, 2009; Yu et al., 2014).

Errors from these pre-processing tools can propagate to the metric, particularly when the input is machine translation outputs that are not fluent. Some metrics attempt to solve this by computing the linguistic structures on the reference sentence only (Yu et al., 2014). But a more compelling argument against these methods is that they are computationally expensive and require resources that are available only in a limited set of languages. Ma et al. (2017) show that including these expensive syntax-oriented metrics only results in a small gain in their ensemble metric, so only use lexical metrics.

2.2.3 Reference-free Automatic Evaluation

Automatic metrics require a one-time investment of generating high quality metrics. It would be very useful to evaluate MT system quality by comparing MT system outputs directly with the source sentence.

Reference-based automatic metrics can be biased towards MT system translations that are superficially similar to the reference, so comparing directly with the source can potentially remove reference-bias. However, these can be biased towards MT systems that produce literal translations of the source that are not fluent, but also potentially inaccurate.

This task of automatic reference-free evaluation has a rich history in the application of estimating the quality of translations of a single MT system. Possible use-cases include:

- to determine whether a translation is perfect, needs post-editing, or if the translator is better off starting from scratch.
- to estimate post-editing cost.

- to automatically estimate whether a translation is adequate for gisting purpose

WMT has included a shared task on MT quality estimation (QE) since 2012 (Callison-Burch et al., 2012). Early QE systems were typically supervised models, with features from the source sentence (measuring translation difficulty), system output (measuring translation fluency and comprehensibility), and both (measuring translation adequacy)(Blatz et al., 2004; Specia et al., 2009, 2013). Some features require the corpus that was used to train the MT system, linguistic tools such as parsers or part-of-speech taggers, or even internal information from the MT system which indicate its confidence in the translation. Another set of important features can be obtained by using the outputs of a high-quality MT system as pseudo-references, and using scores of automatic metrics such as METEOR.

End-to-end trained deep neural networks resulted in a significant boost in performance. The predictor-estimator model (Kim and Lee, 2016), a key neural architecture, consists of two stacked modules: (a) an encoder-decoder that is trained on parallel data to predict translation probabilities of the MT output; and (b) an estimator which uses the representations from the predictor module to estimate translation quality. The latest state of the art quality estimation systems use cross-lingual pre-trained contextual embeddings (Specia et al., 2020).

WMT 2019 metrics task introduced a new track inviting reference-free metrics that were evaluated in the same setting as traditional reference-based metrics. The best-performing metrics, UNI (Yankovskaya et al., 2019) and YiSi-2 (Lo, 2019), were both based on cross-lingual embeddings. UNI is a trained neural model that predicts translation quality from sentence representations that are obtained from pre-trained cross-lingual embeddings LASER and BERT. YiSi-2 is a reference-free version of YiSi (which we describe in Sec. 2.2.4) which uses multilingual BERT embeddings.

More recently, the state of the art in reference-free evaluation of MT systems has been advanced by metrics that use the probabilities of a large, multilingual neural machine translation model to score the MT outputs given the source sentence (Thompson and Post, 2020; Agrawal et al., 2021).

2.2.4 Selected Metrics

We describe selected metrics in more detail, along with their strengths and weaknesses based on our criteria.

- Baselines: BLEU, METEOR, TER, CHRF. BLEU is still the most widely used evaluation metric. In addition to BLEU, some research papers also include the results of METEOR and TER, and more recently, CHRF.
- Benchmarks: RUSE and YISI-1 are both embedding–based metrics that were the best metrics at WMT 2017 and 2018.

BLEU

BLEU is based on precision of n-grams of the MT output when compared with one or more reference translations. BLEU allows some flexibility in word order of the translation, by allowing matches of n-grams anywhere in the sentence. BLEU was originally designed to use multiple translations to account for variation in word choice and word order. Since multiple translations are used, it is not possible to use recall directly. In place of recall, BLEU uses a brevity penalty for translations shorter than the (shortest) reference, as shorter MT outputs usually can not contain all the information in the source sentence.

Translations that include multiple instances of words with high-likelihood of being in the reference will have an inflated value of precision. To solve this, the count of each n-gram in the candidate translation is clipped by the maximum count of the n-gram in the reference.

$$\text{count}_{\text{clip}}(x) = \text{min}(\text{count}(x), \text{max\_ref\_count}(x)) \tag{2.2}$$

The values of the precision will necessarily get smaller as the n-gram order increases. For example, there will be fewer exact matches of 4-grams compared to unigrams. Noting that this decrease in n-gram precisions is roughly exponential with n, Papineni et al. (2002) chose to convert them to the logarithmic scale before averaging them, which is equivalent to using the geometric mean.

The final BLEU score is the weighted geometric mean of the clipped n-gram precisions, scaled down by the brevity penalty.

$$\text{BLEU-N} = \text{BP}.\left(\prod_{N=1}^{n} \text{p}_n\right)^{\frac{1}{n}}, \tag{2.3}$$

where

$$\text{p}_n = \frac{\sum_{s \in \text{candidate sentences}} \sum_{n\text{-gram} \in s} \text{Count}_{\text{clip}} n\text{-gram}}{\sum_{s' \in \text{candidate sentences}} \sum_{n\text{-gram}' \in s'} \text{Count}_{\text{clip}} n\text{-gram}'} \tag{2.4}$$

$$\text{Brevity penalty BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases} \tag{2.5}$$

Although BLEU was originally designed to be used with multiple references, it is typically used with a single reference today.

When computing BLEU scores for short documents or single sentences, it is very likely that there are no matches of the higher order n-grams. As BLEU uses a geometric average, a precision of zero over 4-grams or even 3-grams results in a BLEU score of zero for the sentence. This necessitates smoothing of zero counts. A variety of techniques such as the add 1 smoothing (Lin and Och, 2004) have been proposed, and including any of them results in a big improvement over the sentence-level correlation with human scores (Chen and Cherry, 2014).

BLEU scores range between 0 and 1, where a translation closer to the reference gets a higher BLEU score. Often, the scores are multiplied by hundred to increase readability, and these values are referred to as BLEU points. However, there is no intuitive interpretation, and a BLEU score on its own can give little information on how good a system is. It is highly dependent on the test set and reference(s) used (Culy and Riehemann, 2003). So BLEU scores are only useful when comparing two systems on the same test set and references. This also holds for all the other metrics that followed BLEU.

BLEU has a number of parameters. Some, like the maximum length of *n*-grams, is almost always set at the default value of 4. However, differences in pre-processing the text –whether text is lowercased, for example, or what tokenization is used – can have a large influence on the value of BLEU. To solve this, sacreBLEU (Post, 2018) was developed to help researchers use BLEU with consistent tokenization and parameters, and clearly report all parameters of BLEU in their paper.

Despite its flaws, the MT research community continues to rely on BLEU as the default automatic metric; it is well studied, quick to compute and portable to any language. Tools like compare-mt (Neubig et al., 2019) have been developed to analyse BLEU score differences between systems.

METEOR

METEOR (Banerjee and Lavie, 2005) was developed to specifically address some of the deficiencies of BLEU. To allow for variation in word choice, METEOR also rewards matches of stems, paraphrases and synonyms. Instead of relying on higher order *n*-grams to check for fluency, METEOR explicitly computes an alignment between the translation and reference sentences based on the full set of word matches, and introduces a fragmentation penalty that accounts for word-reordering. A translation that is essentially word salad would receive a higher fragmentation penalty compared to a coherent sentence.

The METEOR score is defined as the F-score of the unigrams multiplied by the fragmentation penalty. Then F-score is computed as the harmonic mean of the precision and recall of the words in this alignment, weighted to give more importance recall. The fragmentation penalty depends on the number of chunks in the alignment, where a chunk is a group of contiguous word matches that have the same order in both the MT system output and the reference.

$$\text{Meteor}_{\alpha,\beta,\gamma} = \text{F}_{\alpha}.\text{frag}_{\beta,\gamma} \tag{2.6}$$

, where

$$\text{P} = \frac{\text{\# matches}}{\text{length of translation}}, \text{R} = \frac{\text{\# matches}}{\text{length of reference}}, \text{F}_{\alpha} = \frac{\text{PR}}{\alpha\text{P} + (1 - \alpha\text{R})} \tag{2.7}$$

$$\text{frag}_{\beta,\gamma} = 1 - \gamma(\frac{\text{\# chunks}}{\text{\# matches}})^{\beta} \tag{2.8}$$

METEOR is relatively fast to compute, but adoption for different languages depends on the availability of resources such as stemmers and WordNet (to check for synonyms).

METEOR places a higher importance on recall, as this was found to empirically correlate better with human scores when comparing different MT systems. It is recommended to place equal importance for precision and recall when directly using METEOR to optimise an MT system.

METEOR has been refined over the years. For example, the weights for individual components can be tuned over the gold standard scores on a development set. The latest version of METEOR can use automatically computed paraphrase tables instead of WordNet, allowing it to be more portable into new languages.

TER

The Translation Edit Rate (Snover et al., 2006), like WER, computes the number of edits required to change an MT output to a correct and fluent translation. When a candidate has a different phrase-order to the reference, WER requires the edit operations for each word in the phrase.

To avoid harshly penalising such translations, TER introduces a new edit operation: the movement of a sequence of words.

$$\text{TER} = \frac{\text{\# edits}}{\text{\# reference words}}, \tag{2.9}$$

where an edit could be insertions, deletions, substitutions of either single words, or a “shift” in a block of words from the hypothesis to another location in the sentence.

When multiple references are available, TER chooses the reference that has the smallest distance from the hypothesis. It is normalised by the average length of all reference translations.

Computing the optimum edits that include moving blocks of words is an NP-complete problem; in practice, beam search is used to find an efficient approximate edit distance.

While TER’s penalty for different phrase-orderings is considerably more lenient compared to WER, TER still penalises translations that have a different word-choice from the reference. To mitigate this, TER was later extended to TER-plus (Snover et al., 2009) by adding support for matching words with the same stems, synonyms, paraphrases. In addition, the cost for each type of edit can be tuned based on a human gold standard in a development set.

CHARACTER (Wang et al., 2016) and EED (Stanchev et al., 2019) are both inspired by TER, and they compute the edit distance on the character-level instead of the word level.

HTER: To eliminate reference bias, translations are evaluated against a “targeted reference” that is obtained by post-editing the MT output into a fluent and faithful translation of the source sentence. HTER is defined as the TER score calculated using this targeted reference. HTER had a Pearson correlation of 0.630 with human judgements. Like TER, HTER assigns the same cost for every edit, whether it’s a trivial morphological error, or an edit that changes the meaning completely (a negation, for example).

CHRF

CHRF (Popović, 2015) is the F-score of character n-grams of the MT output and the reference. White space characters are discarded before computing n-grams; experiments showed that explicitly including white space did not yield an improvement in correlation. This makes CHRF tokenisation-independent.

$$\text{CHRF}_\beta = \left(1+\beta^2\right)\frac{\text{CHRP}\cdot\text{CHRR}}{\beta^2\cdot\text{CHRP}+\text{CHRR}} \tag{2.10}$$

CHRP and CHRR are the arithmetic mean of the precision and recall respectively of character n-grams, where n ranges from 1-6. CHRF is the weighted harmonic mean of CHRP and CHRR, where β determines the importance of recall over precision. A value of 1 implies that both precision and recall are equally important

Like BLEU, CHRF also computes clipped counts when computing precision and recall. As matching character n-grams is more likely than matching word n-grams, the scores of chrF tend to be high. Finally, it is rare to have zero counts even of the longest character n-gram, and does not require smoothing.

CHRF has been extended to CHRF++ (Popović, 2017), which includes the F-scores of word unigrams and bigrams in addition to character n-grams when calculating precision and recall. This helps temper the probably overly-optimistic scores of CHRF and yields a small (often inconsistent) improvement in correlation.

CHRF is a lexical metric that doesn't require any additional resources. As such, it is quick to compute and language-independent. It has high correlation with human scores, and is often competitive with the top metrics that require additional resources, particularly on languages other than English.

MEANT, MEANT 2.0 and YiSi-1

MEANT (Lo et al., 2012) is a semantic similarity metric which uses linguistic information from semantic frames and word embeddings to compute similarity between words in matched frames. The metric aims to compute whether the MT outputs preserve the key information of the source input: *who* did *what* to *whom*, for *whom*, *when*, *where*, *why* and *how*? Each frame contains a predicate and its arguments.

MEANT first computes and aligns the shallow semantic frames of the MT output and the reference translation based on the similarity of the predicates. For each pair of aligned frames, the role fillers of the arguments of the MT output and reference translation are aligned based on their semantic similarity. MEANT computes similarity of two words as the cosine similarity of their embeddings, and similarity of two phrases is computed by aggregating word similarity. The final score is the micro-averaged semantic similarity of the phrases in aligned semantic frames of the MT output and the reference translation.

MEANT 2.0 (Lo, 2017) computes n-gram similarity instead of of word similarity, and introduces inverse-document-frequency-weighting for each n-gram. When shallow semantic parsers are not available, MEANT_2.0-NOSRL computes the weighted F-score on the entire sentence instead of on the aligned semantic frames.

The YiSi-1 family of metrics is inspired by MEANT, and optionally incorporates information from semantic roles. YiSi-0 is a degenerate version that does not require any resources beyond the reference translation, and it computes lexical similarity by computing LCS distance. YiSi-1 computes semantic similarity as the cosine similarity of word2vec (Mikolov et al., 2013) embeddings in 2018, and was updated to use contextual embeddings (Devlin et al., 2019) in 2019. YiSi-2 is a reference-free version that uses cross-lingual contextual embeddings to compute the similarity between source and MT outputs.

RUSE

Regressor Using Sentence Embeddings (Shimanaka et al., 2018) is an end-to-end neural metric that is trained on human evaluation scores. As the name suggested, the metric uses pre-trained sentence embeddings to obtain representations of the MT system output and the reference translation.

The version submitted to the WMT metrics task in 2017 used the concatenation of three different pre-trained sentence embeddings — Infersent (Conneau et al., 2017), Quickthought (Logeswaran and Lee, 2018) and USE (Cer et al., 2018) — that are obtained by training models with a diverse set of objectives on different datasets.

The representation of the sentence pair is the concatenation of the representations of the MT system output and the reference translation, and their element-wise product and difference (a heuristic first proposed by Mou et al. (2016) that has proved useful for text classification tasks with sentence-pairs as inputs). This helps the model extract the interactions between the contents of the two sentences. This representation is given as input to a regressor, which is trained against Direct Assessment (Graham et al., 2015) scores. The recommended choice of regressor is a carefully tuned feedforward neural net.[5] The sentence embeddings are kept constant during training.

RUSE has a high correlation with human scores when evaluating MT systems translating to English, and was, at the time of introduction, not outperformed by any other metric. However, RUSE is not very efficient; computing three different sets of pre-trained models is not quick, even when it has been speeded up with a GPU. And finally, RUSE requires pre-trained sentence representations, that could, in theory, be obtained for any new language given a monolingual corpus. But in practice, high quality sentence embeddings are only available for a small set of languages, which makes RUSE difficult to port to new languages.

[5] They optimise for the number of layers, number of hidden units, dropout and batch size.

2.3 Evaluating automatic Metrics

Although human evaluation is not perfect, we assume that it is the ground truth, and we evaluate metrics based on their agreement with human scores. BLEU (Papineni et al., 2002) was the first metric introduced with strong empirical justification: it had high correlation with human fluency and adequacy judgements over 5 “systems” translating from Chinese to English. This included three commercial MT systems and 2 human non-professional translators to test if BLEU can correctly distinguish between high quality translations. The test set contained around 500 sentences that were taken from 40 news articles.

Coughlin (2003) conducted a large scale correlation study of metric and human scores across seven language pairs; they computed the correlation between human and metric score differences on pairs of systems.

Subsequent metrics like CDER (Leusch et al., 2006) and METEOR (Banerjee and Lavie, 2005) were all evaluated based on correlation with human judgements on MT systems.

Since 2007, the metrics shared task at WMT has played a key role in the development and evaluation of automatic metrics. The data collected from the large-scale human evaluation of the news translation task at WMT serves as the ideal test bed for evaluating automatic metrics. The metrics shared task has been an important component of WMT, where participating metrics, along with baselines, are evaluated on two levels:

- system level, which measures how much automatic metrics measure the overall quality of the MT systems submitted for each language pair in the translation task
- segment level, which evaluates metrics on a more fine-grained scale, typically sentences. Through this thesis, we use sentence-level to refer to this, though the official term at WMT is segment-level.

We next describe the evaluation of metrics at the system and sentence level.

2.3.1 System-level Evaluation

From 2007 to 2012, the metrics were evaluated based on their Spearman rank correlation coefficient with human scores on the set of MT systems.

The **Spearman correlation** measures the strength of the monotonic relationship between the metric and human scores. When automatic metric evaluation was first introduced at WMT, human scores were obtained on an ordinal scale, and Spearman correlation was the natural choice of evaluation measure.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{2.11}$$

where

n is the total number of observations (MT systems in our case)

$d_i = \text{rank}(m_i) - \text{rank}(h_i)$ is the difference in ranks of metrics and human judgements for system i

Spearman correlation is 1 when the metric ranks the MT systems in the same order as human evaluation. The correlation decreases if the metrics disagree with humans when ordering any two given systems, irrespective of how close the systems are in quality.

Given that many MT system pairs have very small score differences (and in some cases these differences are not statistically significant), evaluating with the Spearman correlation harshly penalises metrics that have a different ordering for these systems.

This motivates the switch to **Pearson correlation**, which takes the score differences into account. Pearson correlation coefficient measures the strength of the linear relationship between the human and metric scores.

$$r = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(m_i - \overline{m})}{\sqrt{\sum_{i=1}^{n}(h_i - \overline{h})^2(m_i - \overline{m})^2}} \tag{2.12}$$

where

h_i and m_i are the human and metric score respectively of system i

$\overline{h}$ and $\overline{m}$ are the mean scores of humans and metrics respectively.

If a metric errs when ordering two similar systems, the decrease in Pearson correlation is smaller than when the systems are widely different in quality. Finally, the Spearman correlation does not take into account the magnitude of score difference between two systems, as long as the order is right. If human scores of two systems are really close, Pearson correlation penalises metrics if there is a wide margin between the scores of these systems, whether or not the ranking is correct.

In this sense, Pearson correlation more closely reflects our criteria for metric evaluation. But Pearson correlation has its drawbacks. For example, it assumes that the relationship is linear. At WMT 2014, when the organisers of the metrics task switched from Spearman to Pearson correlation, they noted that this assumption is unlikely to be violated since the MT systems evaluated are typically in a small quality range.

Pearson correlation is highly sensitive to the presence of outliers in the dataset. This is because it is highly dependent on the mean, and the presence of any outliers will skew the mean towards the outlier. We elaborate on this in Chapter 5, and present the implications of this on metric evaluation.

Until WMT 2016, the gold standard was the MT system scores obtained through relative ranking (RR) method. In 2016, DA was trialled and has been the official method since 2017. In 2016, for the to-English language pairs, metrics were evaluated using both the RR and DA methods. At the system level, metric rankings with both methods were similar for Czech, Turkish and Finnish, but there were discrepancies with the other three languages: the ranking of the top three metrics were inverted for German-English; the Russian-English evaluation has a single outlier, CHARACTER, which has a high ranking with DA but not RR; and finally, there was only a moderate agreement for Romanian-English. The metrics results paper (Bojar et al., 2016b) did not investigate any of these discrepancies.

2.3.2 Sentence-level Evaluation

When metrics can reliably separate good translations from the bad at the sentence level, this could help with error analysis during system development. Since 2008, the WMT metrics task has also evaluated metrics at the sentence level (Callison-Burch et al., 2008). This evaluation is necessarily dependent on the kind of human annotations available, and the methods of sentence-level meta-evaluation has necessarily evolved along with changes in human evaluation. In addition, these measures have been refined over the years.

We describe the three methods used over the years, based on the correlation method computed over the method of obtaining human scores.

Kendall's Tau over Relative Ranking (RR)

When human evaluation was collected in the form of rankings of 5 MT system translations of the same source sentences, metrics were evaluated based on how well they reproduced the same ranking as humans.

This was initially computed as average accuracy of metrics correctly ordering every pair of MT systems evaluated, then was replaced by a modified version of the Kendall's Tau correlation coefficient.

Kendall's Tau has the intuitive explanation as the difference between the probability of the metrics evaluating pairs of sentences in the right order, and the probability that they are in a different order.

For a pair of translations by systems i and j for the same source sentence, a pair of ranks $(h_i, m_i), (h_j, m_j)$ is said to be:

- concordant, if $(h_i - h_j)(m_i - m_j) > 0$

- discordant, if $(h_i - h_j)(m_i - m_j) < 0$
- neither, if $(h_i - h_j) = 0$

The Kendall's tau correlation is then calculated over all available human pairwise judgements,as:

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total number of pairs}} \tag{2.13}$$

Note that there are many options to deal with ties in human and metric scores (Macháček and Bojar, 2014). This variation, which has been used since 2014, ignores all translation pairs where human scores indicate ties and penalises ties in metric scores if the human judgements indicate that one is better than the other.

Since the WMT annotations are restricted only to ranks of a limited set of translations for the same source sentence, the number of concordant and discordant pairs are similarly restricted, so what is being calculated is not exactly Kendall's Tau. It tests whether a metric correctly ranks different system translations of the same source sentence, but doesn't directly evaluate whether the metric rightly assigns low scores to low quality translations and vice-versa.

Given the low values of inter- and intra-annotator agreement in the human scores, there is a lot of noise in the human judgements we use as "ground truth". Worse, when there are contradictory annotations for a given translation-pair, these are counted as separate annotations, so even a hypothetically "perfect" metric would not receive a perfect correlation of 1. Thus, we believe that this measure does not give a very accurate understanding of metric performance.

Pearson Correlation over Direct Assessment (DA)

In 2016, when WMT trialled Direct Assessment to evaluate MT systems, they also collected additional human scores specifically for more accurate sentence level evaluation of automatic metrics. See Sec. 2.1.4 for details.

To elaborate, the test sets included accurate scores of 560 translations per language pair, which were sampled randomly from all MT system translations available. Metrics were then evaluated based on their Pearson correlation (Eq. (2.12)) with human scores over these translations.

Since the test set typically consists of thousands of sentences, it is rare that the small sample of 560 sentences includes translations of the same source sentence by different systems. Thus, this evaluation tests the reverse of the modified Kendal's Tau over pairwise judgements: it tests whether a metric correctly assigns low scores to low quality translations and vice-versa, but not whether it ranks different system translations of the same source sentence in the correct order. For example, if a given source sentence is hard to translate, then we can expect all systems to produce low quality translations. This method of evaluating with Pearson does not test whether the metric is correctly distinguishing between the different translations of this same sentence.

There is only moderate agreement over the WMT 2016 sentence-level correlations computed using RR and DA, which was not investigated further.

Kendall's Tau over Relative Ranking from Direct Assessment (DARR)

While ideally we would like to have 15-way annotated accurate scores for each language-pair, it is expensive to collect all these additional annotations. When insufficient scores are collected, it can be argued that the singly-annotated data contains too much noise to be suitable for sentence-level evaluation using Pearson Correlation Coefficient. When we have continuous scores for at least two system-translations of the same source sentence (after filtering out annotators that do not pass quality control), these are converted to pairwise judgements. In an attempt to mitigate noise, these translation pairs are included in the gold standard only if the difference in their raw scores is greater than a predefined threshold. The threshold is set at 25 points as the DA analogue scale has markers that divide the scale into four regions: 0-25,25-50,50-75 and 75-100. Once we have these pairwise judgements, we evaluate metrics using the same formulation of Kendall's Tau as with RR (Eq. (2.13)). This method of evaluation is denoted as DARR.

While these measures sound reasonable, there have been no empirical experiments presented that investigate how the noise affects metric correlations computed, and how the metric rankings obtained using this method compare to using the Pearson Correlation over 15-way annotated scores or Kendall's Tau over relative ranking judgements.

2.3.3 Statistical Significance testing

To test whether the difference in correlations of two metrics M_1 and M_2 can be attributed to chance, the following statisitical significance tests are used:

- **Pearson Correlation**:

  The William's test for dependent correlations that share a variable (Williams, 1959) is used to test whether the Pearson correlation between M_1 and human scores is equals the correlation between M_2 and human scores,

$$T(n-3) = \frac{(r_{13}-r_{23})\sqrt{(n-1)(1+r_{12})}}{\sqrt{2K\frac{(n-1)}{(n-3)} + \frac{(r_{13}+r_{23})^3}{4}(1-r_{12})^3}}, \tag{2.14}$$

  where n is the samples size (number of systems or sentences), r_12 is the correlation between M_1 and M_2, r_13 is the correlation between M_1 and human scores, and r_23 is the correlation between M_2 and human scores, and $K = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}$.

  The p-value is calculated from the t-distribution with n-3 degrees of freedom, and a cutoff of $p = 0.05$ is used to determine significance.

  The William's test is recommended for small or moderate samples (Neill and Dunn, 1975). Graham and Baldwin (2014) proposed the test as suitable for comparing automatic metrics. It takes the correlation between the metrics into account, and is more powerful than the equivalent test for independent samples (Fisher z-transformation). The power of the William's test to differentiate between metrics increases with the correlation between the two metrics.

- **Kendall's Tau**:

  The bootstrap test is used to test for statistical significance between Kendall's Tau correlations of metrics. 95% confidence intervals are estimated from 1000 Bootstrap samples of human judgements, and metric correlation differences are deemed as statistically significant if the they have non-overlapping confidence intervals.

2.4 Summary

In this chapter, we presented a review of the three main aspects of machine translation evaluation: human evaluation, automatic evaluation, and evaluation of automatic metrics. We identified potential areas of improvement, and in the next four chapters, we present our research on improving aggregation of human judgements (Chapter 3), detecting bias in human annotations (Chapter 4), leveraging contextual embeddings to develop automatic metrics (Chapter 5), and re-evaluating the evaluation of automatic metrics (Chapter 6).

Chapter 3

A Probabilistic Model for Aggregating Human Judgements

This chapter builds on the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Towards efficient machine translation evaluation by modelling annotators. *In Proceedings of the Australasian Language Technology Association Workshop* 2018, pages 77–82, Dunedin, New Zealand, December 2018.

3.1 Introduction

Human evaluation is a fundamental requirement for reliable assessment of machine translation, despite progress in automatic evaluation methods over the years. Further, human judgements serve as a gold standard to evaluate automatic metrics. Accordingly, accurate human judgements are crucial for progress in automatic evaluation. The process of collecting human annotations is time-consuming and expensive. The data is inevitably noisy due to the subjective nature of the task. This problem is exacerbated when the annotations are crowdsourced from anonymous unskilled workers who are less invested in the task. The question of how to efficiently collect this data has evolved over the years, but there is still scope for improvement.

We described direct assessment ("DA"; Graham et al., 2017) in Sec. 2.1.4 of the previous chapter. DA is currently accepted as the best practice for human evaluation of machine translation, and has been the official method of collecting human annotations at the annual Conference for Machine Translation since 2017 (Bojar et al., 2017b). This method was designed for crowdsourcing judgements: every crowd worker scores a set of 100 translations on a continuous scale. 30 out of the 100 translations are quality control items designed to filter out unreliable workers; they check whether workers give high scores to high quality translations, similar scores to repeat items and comparatively low scores to translations that have been deliberately degraded.

In this chapter, we show that this quality control process is typically effective at identifying the extremely unreliable workers, but has a low recall for good workers: about a third of good quality data is discarded, which increases the cost of the evaluation. On the other hand, despite variation in quality, in the DA method, all the workers who have passed quality control are given equal weights to obtain the final average 'true' score. To solve both these problems, we explore Bayesian methods that explicitly model the reliability of all annotators, and use this information to infer the true quality of the translations. This provides a more cost-effective way to collect more accurate estimates of the true translation quality.

We begin the chapter with a brief description of different probabilistic models that provide principled means to obtain the true labels from multiply-annotated data for a variety of tasks: discrete labels, ordinal labels, pairwise preference judgements and continuous labels, focusing, when possible, on methods designed specifically for text processing and machine translation evaluation. Next, we present the multiply-annotated direct assessment datasets, along with an analysis of worker scores to understand the scope of improvement from modelling annotator precision.

We then present our model to infer the quality of machine translation outputs and the results of our experiments: compared to the current method of averaging scores of workers who pass quality control, our models typically produce more accurate estimates of the quality of translations for a given set of annotations collected. We explore *spammer removal*, where we rerun the model with the top k workers with the highest inferred precision and show that this can potentially improve the accuracy of the model estimate. Finally, we present an analysis of selected individual HITs to look for leads to answer two questions: what is the ideal k, and is there a need to collect more annotations?

3.2 Background: Probabilistic models of aggregating data

The simplest way to aggregate multiple annotations is to use the mean or the majority vote. However, this ignores the differing reliabilities and biases of workers. In Sec. 2.1.3, we described some strategies to mitigate these problems when crowdsourcing data, for example, filtering out scores that have low agreement with experts and weighting worker's scores based on their agreement with experts. These methods require expert annotations, which are not always available. Other methods look at inter-annotator agreement: weight worker scores based on their agreement with the majority or mean of the remaining workers, which, in turn is often biased due to the presence of scores from unreliable workers.

In scenarios where multiple conflicting sources of information are available per instance, there is always a potential for improved aggregation using probabilistic models. These models can be particularly beneficial when aggregating crowdsourced data, given the high variance in annotator reliability and instance difficulty.

In this section, we describe methods that aim to infer the true labels/scores of the instances that use Bayesian models. They were developed to infer the ground truth from multiple conflicting sources, whether it is crowdsourced workers, expert annotators (Dawid and Skene, 1979), or classifier combination (Kim and Ghahramani, 2012). Some of these models are

Model	Input type	Requires features	Models annotators
D&S (Dawid and Skene, 1979)	Discrete	N	Y
MACE (Hovy et al., 2013)	Discrete	N	Y
Raykar (Raykar et al., 2010)	Discrete,continuous	Y	Y
Reviewer Calibration (Flach et al., 2009)	Ordinal	N	Y
H&M (Hopkins and May, 2013)	Relative Ranking	N	N
Trueskill (Sakaguchi et al., 2014)	Relative Ranking	N	N
EASL (Sakaguchi and Van Durme, 2018)	Continuous	N	N
GP (Groot et al., 2011)	Continuous	Y	Y
MTQE (Cohn and Specia, 2013)	Continuous	Y	Y

Table 3.1 Summary of probabilistic models included in Sec. 3.2.

supervised learners that primary designed for predicting the labels of new data based on the features of the instances; these models take annotator reliability into account and jointly learn the ground truth and the parameters of the learner. We also include probabilistic models that aggregate MT evaluation data but do not take annotator reliability into account. These models are summarised in Tab. 3.1.

A variety of methods can be used to infer the ground truth for these models, such as expectation maximisation (Dawid and Skene, 1979), Gibbs sampling (Hopkins and May, 2013; Kim and Ghahramani, 2012), expectation propagation (Flach et al., 2009), and variational inference (Hovy et al., 2013).

Notation Through out this section, we use $r_{i,j}$ to denote the response by worker j for item i, where the ground truth is z_i. For discrete or ordinal data, we denote the total number of classes as K. The models use various probability distributions to model the data, and we denote the normal distribution as $\mathscr{N}$, the gamma distribution as $\mathscr{G}$, the multinomial distribution as $\mathscr{M}$, the beta distribution as $\mathscr{B}$ and the uniform distribution as $\mathscr{U}$

Discrete Labels

In seminal work on modelling annotators to aggregate data, Dawid and Skene (1979) assumed that worker performance depends on the true label of the items. Their model assumes that the true labels are drawn from categorical distribution with probability π, and each worker is parameterised by a KxK confusion matrix where $\theta_{j,r,z}$ is the probability of worker j to have a response r when the true label is z. They designed an expectation maximisation algorithm to jointly learn the true labels as well as the parameters $\theta_{j,r,z}$ and π. These inferred labels were more accurate compared to majority voting. Snow et al. (2008) used this model to obtain more accurate estimates of the true labels on crowdsourced textual entailment and temporal event annotation datasets. Passonneau and Carpenter (2014) enhanced this model with weak Dirichlet priors for the worker confusion matrices and used it on word sense annotations. The priors add arithmetic stability in cases where worker responses do not include certain word senses, which can happen, for example, with infrequent word senses. This method yields a probability distribution of word senses for every instance, which can be used when evaluating computational word sense disambiguation models.

A variety of different approaches have been proposed since then. Some, like Dawid and Skene (1979), model worker reliability as a confusion matrix on each label (Raykar et al., 2010; Kim and Ghahramani, 2012). Others model annotator reliability directly (Demartini et al., 2012). More complex methods attempt to model the difficulty of instances (Whitehill et al., 2009; Raykar and Yu, 2012) or annotator expertise on individual instances (Yan et al., 2010).

Multi-Annotator Competence Estimation (MACE) (Hovy et al., 2013) was designed specifically for NLP annotations. This is a simple algorithm that models the credibility of an annotator j as the probability of not spamming θ_j. For each item i, the true class $z_i \sim \mathscr{U}$. The model assumes that the annotator response on an instance depends on true label of the instance and whether the annotator is spamming on that instance: $S_{i,j} \sim \text{Bernoulli}(1-\theta_j)$. If the annotator is not spamming (when $S_{i,j} = 0$), then the annotator label $y_{i,j}$ is copied from the true label

z_i. Otherwise, $y_{i,j} \sim \mathscr{M}_j$, where $\mathscr{M}_j$ is a multinomial distribution that describes the spamming behaviour of annotator j. These models were evaluated on three discrete-label NLP datasets (Snow et al., 2008), and are superior not just to majority voting but also the more complex models of Raykar and Yu (2012) that model instance difficulty. Further, the worker reliability parameter θ_j can be used to filter out annotators with high probability of spamming, thus increasing accuracy of the final results.

Crowdsourcing has been heavily used to collect training data for supervised machine learning systems. Recent approaches focus on training supervised machine learning models at the same time as learning the ground truth from the noisy labels. Raykar et al. (2010) pioneered this approach with EM algorithms to jointly learn a probabilistic classifier along with annotator reliabilities for binary and multi-class classification tasks. Another avenue of research is active learning, where the model iteratively chooses the item for labelling that is most useful to the classifier (for example, the item with most uncertainty) from a large pool of unlabelled data (Settles, 2009). The goal is to reduce the total cost of annotation. We can also vary the number of redundant labels collected on an instance based on the uncertainty of the label and/or of the classification model (Sheng et al., 2008).

Ordinal Data

Many NLP tasks involve ordinal data. In particular, MT quality judgements are often collected as ordinal ratings, so we next describe a key model that uses ordinal data that was used to assist with decisions on accepting papers submitted to the ACM SIGKDD 2009 conference (Flach et al., 2009). The reviewer calibration model is a probabilistic Bayesian model that was developed to infer the true underlying quality of papers. A review by judge j of a paper submission contains an ordinal rating and an expertise level e indicating reviewer confidence. Each paper submission s to the conference has a latent true quality $q_s \sim \mathscr{N}(\mu_q, v_q)$. The expertise indicates annotator precision for the review, which is drawn from a Gamma distribution $\lambda_e \sim \mathscr{G}(k_\lambda, \beta_\lambda)$.

The reviewer's latent score of a given paper submission depends on the quality of the paper and their expertise: $s_r \sim \mathcal{N}(q_s, \frac{1}{\lambda_e})$. Since the reviewer ratings are ordinal, the model also learns thresholds for each reviewer that map the continuous perceived quality with the annotator rating. Instead of directly making decisions based on the inferred quality of papers, this model was used to highlight areas in the review process that require greater scrutiny due to the variation in reviewer ratings caused by differing standards and expertise levels.

Relative Ranking

We now describe two probabilistic models that were designed to aggregate Relative Ranking data that was collected at WMT: Hopkins and May (2013) designed probabilistic models for machine translation relative ranking judgements where annotators rank translations of the same MT system of the same source sentence i. Each system j has a latent true ability μ_j which is drawn from a Gaussian prior with zero mean: $\mu_j \sim \mathcal{N}(0, \sigma_0^2)$. The quality of translation i of system j, $q_{i,j} \sim \mathcal{N}(\mu_j, \sigma_a^2)$, where the standard deviation σ_a is shared by all systems. A judge's perceived quality of a given translation $\pi_{i,j} \sim \mathcal{N}(q_{i,j}, \sigma_{obs}^2)$. When comparing two translations i and i', the annotator j would prefer the translation with higher π if the difference is greater than a fixed decision radius, i.e., if $\pi_{i,j} > \pi_{i',j} + \varepsilon$. Otherwise, the judgement is a tie.

Sakaguchi et al. (2014) adapted TrueSkill (Herbrich et al., 2007), an algorithm initially developed to estimate the true "skill" of Xbox players, for MT preference judgements. MT systems are treated as players. TrueSkill assumes an MT system's skill $q_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, where μ_j is the estimate of the MT system quality, and σ_j^2 is a system-specific variance parameter expresses the uncertainty around the current estimate of μ_j. The parameters are updated using Bayesian online learning after each observation of a win, loss or tie between two systems. Finally, TrueSkill uses active learning to select evenly matched players on Xbox live; for MT evaluation, this corresponds to active learning to select the set of systems to obtain preference judgements: it selects system with the greatest uncertainty (highest σ_j^2), then matches it with the systems closest to it in quality, maximising the information gain from each annotation. Trueskill results in more efficient estimation of MT system scores than the Hopkins and May (2013) model, which can probably be attributed to the active learning component.

Both these models offered a principled way of aggregating the relative ranking data, and Trueskill reduces the total annotations required by requesting rankings of systems that are close together in quality. But both models assume that all raters are equally competent and do not account for any variation in annotator reliability.

Continuous Labels

Efficient Annotation of Scalar Labels (EASL) was developed as an alternate to direct assessment (Graham et al., 2017) and relative ranking that aimed to combine the advantages of both methods when collecting human annotations for MT evaluation. Annotators score five translations of the same source sentence on a continuous scale. EASL models this data with bounded continuous labels as a beta distribution (Sakaguchi and Van Durme, 2018). Scores of MT system $S_i \sim \mathscr{B}(\alpha_i, \beta_i)$, with the mode of the distribution as the true quality of the system, and the variance as uncertainty. It uses an online learning algorithm to update the two parameters of the beta distribution after each annotation. Inspired by TrueSkill, they use active learning to select instances with the greatest uncertainty to annotate next. This model was evaluated on system-level data simulated from WMT 2016 DA judgements, and shown to be more efficient than DA in estimating MT system rankings. However, the reason for improvement is unclear as there were no ablation studies, and since EASL requires ratings of five translations at once, there is potential for annotators to treat it as a ranking problem. Finally, like the models for MT relative ranking (Hopkins and May, 2013; Sakaguchi et al., 2014), EASL doesn't distinguish between annotator reliability.

Raykar et al. (2010) combined the standard linear regression model with an annotator model to predict the true labels from multiply-annotated continuous data. The linear regression model with additive Gaussian noise is given by $z_i \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, 1/\gamma)$. The annotator model assumes each annotator has Gaussian noise: $y_{i,j} \sim \mathcal{N}(z_i, 1/\tau_j)$, where τ_j is the inverse variance of annotator j. These two models are combined to yield the final model $y_{i,j} \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, 1/\gamma + 1/\tau_j)$

Groot et al. (2011) extended the Gaussian process linear regression model to allow individual noise parameters for every annotator. This is a more powerful model than the model of Raykar et al. (2010) as it is non-parametric and can be applied even when individual annotators annotate only a subset of all items.

The above two methods were tested on simulated data: instead of using annotations collected from different people, they simulate annotators of varying reliabilities by adding Gaussian noise to the ground truth of real datasets. On this synthetic data, these methods perform better than models trained on all individual annotators, or trained on the mean response of all annotators.

In the MT domain, Cohn and Specia (2013) use a multi-task Gaussian process to learn to predict the estimated post-editing effort as well as the actual post editing time of different post-editors. Both tasks are subjective, and the latter is particularly challenging as different translators are faster when post-editing different translations. This model is slightly different from all previously presented models; instead of aiming to recover the one ground truth from multiple annotators, it aims to use all data to better predict the response of individual annotators. They treat the response of each annotator (whether estimated post-editing effort or actual post-editing time) as a specific task, and learn annotator-specific models that share information between each other due to explicitly modelling the covariance between the annotators. Note that these models are a generalization of the model of Groot et al. (2011) which has a noise parameter for each annotator but does not model the covariance of individual annotators.

In our use case of obtaining the true quality of the MT system outputs, the primary goal is to learn the ground truth. We need simple methods analogous to MACE for continuous data that leverage inferred annotator reliability that we can use to decrease costs of direct assessment.

3.3 Direct Assessment: Dataset and Analysis

In this section, we first review direct assessment (described in detail in Sec. 2.1.4), expanding on the quality control (QC) mechanism. We then introduce two multiply-annotated DA datasets that we will use for our experiments, and visualise this data to understand the effectiveness of QC and identify the scope for improvement from modelling annotators. We use this information to design our model, which we describe in the next section.

Direct assessment requires annotators to score translations using an analogue slider which maps to an underlying scale of 0–100. When they are crowdsourced, each HIT contains 100 translations, of which 30 are control items for quality control that are used to filter out low quality workers, for example, those who click randomly or assign the same score to all translations.

The scores of workers who pass quality control are standardised to improve worker consistency. The scores are still noisy after standardising, but this noise is expected to cancel out when a large number of scores are averaged. We confirm this in our experiments in Sec. 3.5.

The final score of an MT system is the mean standardised score of its translations after discarding scores that do not meet quality control criteria. To obtain accurate scores of individual translations, multiple (at least 15) judgements are collected and averaged.

3.3.1 Quality Control

The quality control mechanism of direct assessment relies on the assumptions that a consistent annotator would assign similar scores to the same annotations, lower scores to degraded versions of a translation compared to the original, and high scores to the reference translation.

To check for this, each HIT on MTurk contains 100 translations, of which 70 are MT system translations, and an additional 30 items are used for quality control (QC). The QC items include:

1. degraded versions of 10 MT system translations;
2. repeats of another 10 MT system translations.
3. 10 reference translations by a human expert;

Worker responses are also examined for red flags: workers who took suspiciously little time to annotate the entire HIT, workers who gave the same score to a series of translations, or appear to have scored translations at random based on the quality control items. These workers are refused payment.

For the other workers, paired statistical significance tests are used to test that the mean score of the 10 degraded translations is lower than the mean of the corresponding system translation. An arbitrary (but customary) cut off of $p = 0.05$ is used to determine "good" workers. The remaining workers are further tested to check that there is no significant difference between their scores for the repeat-pairs. The workers who do not pass both these quality control checks are paid for their efforts, but their scores are unused. If these scores have useful information, the overall cost is unnecessarily increased.

Both these tests are based on a sample-size of 10 items, and, as such, have low power. In particular, the test might not be able to detect a statistically significant difference in the scores of the degraded items. We show in the next section that about a third of the workers with moderate-to-high correlation do not pass quality control, and thus DA ends up discarding potentially useful annotations. We could increase the power by increasing the sample size of the degraded-reference-pairs, but this would be at the expense of the number of useful annotations collected.

3.3.2 Datasets

In this chapter, we use the following multiply-annotated datasets for our experiments:

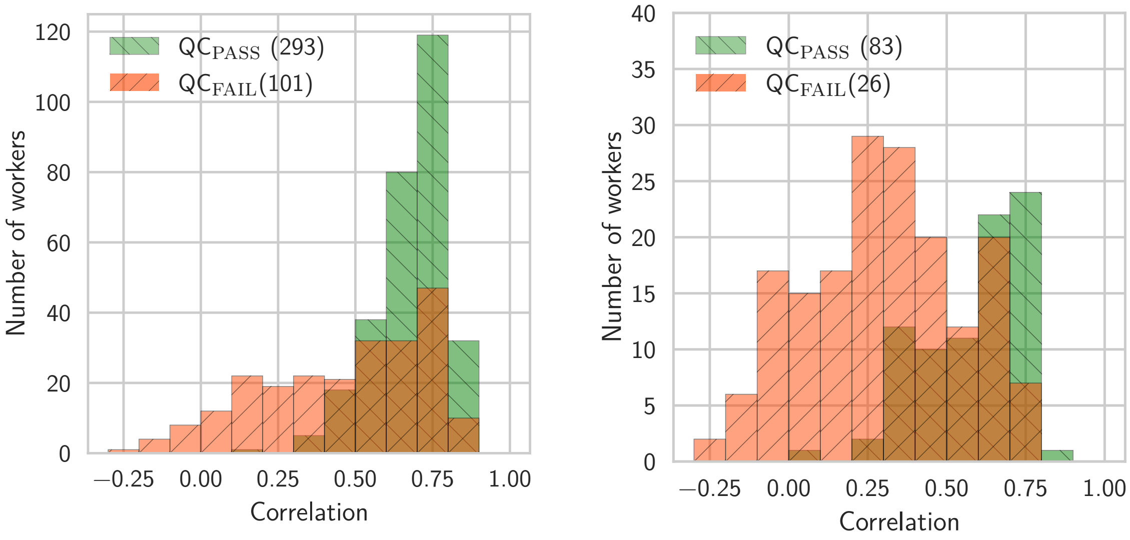
WMT13 Spanish → English (WMT13$_{\text{ES-EN}}$) Our main dataset is the Spanish → English dataset from WMT 2013 (Bojar et al., 2013) that was collected in the process of adopting direct assessment for sentence-level evaluation (Graham et al., 2015). The WMT13$_{\text{ES-EN}}$ dataset consists of 523 workers who evaluated a total of 12 HITS, of which 230 did not pass quality control. For our experiments, we use four of these HITs that were annotated by at least 80 workers who passed quality control. Having such a large dataset ensures that we can trust the ground truth and enables additional analysis while varying the quality of workers.

WMT16 Turkish → English (WMT16$_{\text{TR-EN}}$) To show that our models are effective beyond the WMT13$_{\text{ES-EN}}$ dataset, we also evaluate our models on data from the sentence-level dataset collected for the WMT 2016 metrics task (Bojar et al., 2016b). We focus on the Turkish to English dataset, which consists of 8 HITs, each annotated by at least 15 workers that pass quality control. We chose this language-pair as it has a comparatively low percentage of "good" workers compared to the rest of the language pairs: of the 256 workers that annotated the 8 HITs, only 83 workers passed quality control.

3.3.3 Analysis of Worker Scores

In this section, we highlight the variation in quality of annotators' scores, and provide evidence that the quality control mechanism can be too harsh. To do this, we visualise data from the two datasets that we use in our experiments.

Automatic metrics such as BLEU (Papineni et al., 2002) are generally evaluated using the Pearson correlation with the ground truth, which is computed as the mean of all workers that pass quality control. We similarly evaluate a worker's reliability using the Pearson correlation of the worker's scores with this ground truth over the MT system translations (excluding the quality control items). Over all the data collected for both the datasets, the group of QC$_{\text{PASS}}$ workers is, on average, more reliable than the QC$_{\text{FAIL}}$ workers (see Fig. 3.1). Importantly, we



WMT13$_{\text{ES-EN}}$

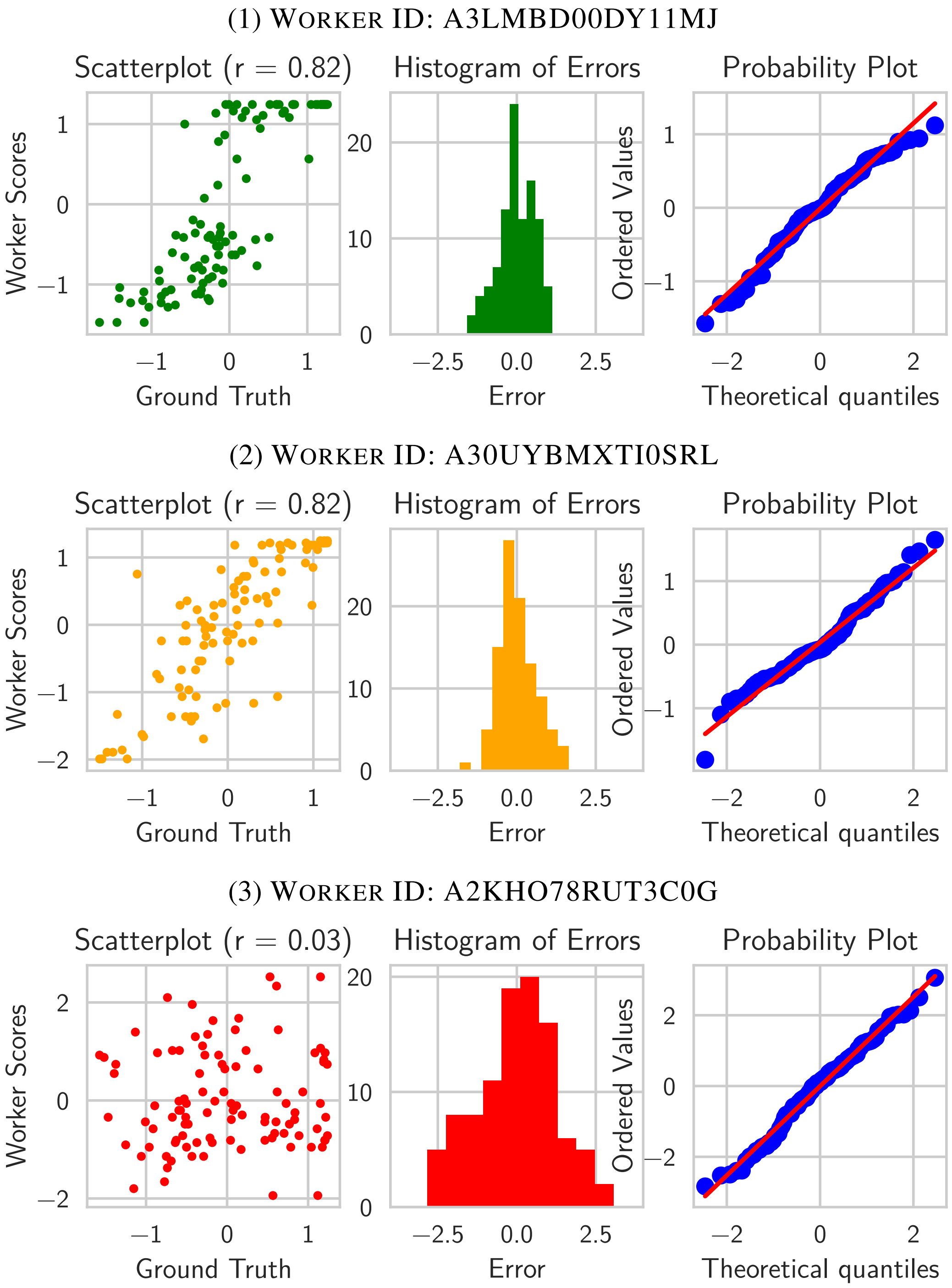
(a) WMT16$_{\text{TR-EN}}$

Fig. 3.1 Correlation of QC$_{\text{PASS}}$ vs QC$_{\text{FAIL}}$ workers with the ground truth in the WMT13$_{\text{ES-EN}}$ and WMT16$_{\text{TR-EN}}$datasets. (Note that the two figures do not have the same range on the y-axis.)

can see that almost all workers with extremely low or negative correlations were filtered out. However, there is substantial overlap in the correlations of the two groups. In particular, over the WMT16$_{\text{TR-EN}}$ dataset, the significance test was not very effective, and around a third of the workers whose scores have a correlation greater than 0.6 were discarded. Fig. 3.1a. With the WMT13$_{\text{ES-EN}}$ dataset, 10 of the 42 workers whose scores have a correlation greater than 0.8 were discarded.

When computing the mean, the scores of all workers that pass the quality control check are given equal weight, despite the variation in their reliability. Given that quality control is not always reliable, the computation of the ground truth could include scores of a few low quality worker with correlation as low as $r = 0.2$. It could be argued that this is rare, and that since we are using at least 15 workers the contribution of each individual worker is very small. A bigger problem is the discarding of potentially useful scores, as it increases the cost of the evaluation.

Fig. 3.2 shows the scatter plots of the standardised scores of three workers from the WMT13$_{\text{ES-EN}}$ dataset against the mean of all QC$_{\text{PASS}}$ workers. The first two workers both



(1) Worker ID: A3LMBD00DY11MJ

(2) Worker ID: A30UYBMXTI0SRL

(3) Worker ID: A2KHO78RUT3C0G

Fig. 3.2 Visualising the scores of three workers in the WMT13$_{\text{ES-EN}}$ dataset: The first column is a scatterplot of the worker's scores against the ground truth, the second column is a histogram or errors (the worker score − the ground truth), and the third column is a QQ plot of the errors. The first two workers have a high correlation and their errors have a low variance, whereas the third worker's scores are essentially random and the errors have a high variance. Only the first worker passes quality control. The second worker is paid, but their scores are discarded. The third worker is rejected payment.

have high correlations with the ground truth. But the second worker doesn't pass quality control; this worker is paid but the scores are unused. The third worker's scores are essentially random, and is rejected payment. We also plot a histogram of the errors (the difference between the ground truth and the worker) of each worker. The errors of all three workers have an approximately normal distribution with the mean very close to zero. The variance of the errors depends on the quality of the annotations: the errors of more reliable workers have a smaller variance. The last column contains QQ plots of the errors that compare the quantiles of the errors with the normal distribution. All points on this graph lie close to the diagonal, confirming that the errors of all workers are close to being normally distributed.

3.4 Model

To model direct assessment data, we need a model that takes in continuous data as inputs, and can take worker reliability into account to infer the true translation quality. It would also be beneficial to take advantage of the quality control items that are included in DA.

Based on the analysis from the previous section, we propose a simple model which assumes that a worker's score is normally distributed around the true quality of the translation. Each worker has a precision (inverse variance) parameter τ that models their accuracy: workers with higher τ have smaller errors.

We use the corresponding conjugate priors for both the translation rating and worker precision. The full generative process works like this:

- For each translation $i \in T$, we draw the true quality μ_i from the standard normal distribution.

$$\mu_i \sim \mathcal{N}(0,1)$$

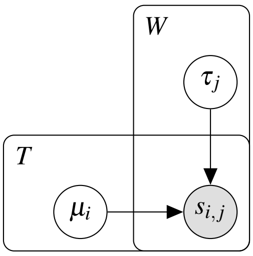


Fig. 3.3 The proposed model, where worker $j \in W$ has precision τ_j, translation $i \in T$ has quality μ_i, and worker j scores translation i with $s_{i,j}$.

- Then for each worker $j \in W$, we draw their accuracy τ_j from a shared gamma prior with shape parameter k and rate parameter θ.

$$\tau_j \sim \mathcal{G}(k, \theta)$$

- The worker's scores $s_{i,j}$ is then drawn from a normal distribution, with mean μ_i, and precision τ_j.

$$s_{i,j} \sim \mathcal{N}\left(\mu_i, \tau_j^{-1}\right) \tag{3.1}$$

Direct assessment computes the average score for each translation, and can be viewed as the Maximum Likelihood Estimate of the mean score of the same model but with a shared τ for all workers, instead of having a separate parameter τ_j for each worker.

$$s_{i,j} = \mathcal{N}\left(\mu_i, \tau^{-1}\right) \tag{3.2}$$

The joint distribution of the full model is:

$$P(\mu_i, \tau_j, s_{i,j}) = \prod_j P(\tau_j) \prod_i P(\mu_i) P(s_{i,j}|\mu_i, \tau^{-1}) \tag{3.3}$$

We want to maximise the likelihood of the observed judgements:

$$\begin{aligned} P(s) &= \int_{j=1}^{W} P(\tau_j) \int_{i=1}^{T} P(\mu_i) P(s_{i,j}|\mu_i,\tau)\, d\tau\, d\mu \\ &= \int_{j=1}^{W} \Gamma\left(\tau_j|k,\theta\right) \int_{i=1}^{T} \mathscr{N}\left(\mu_i|0,1\right) \mathscr{N}\left(s_{i,j}|\mu_i,\tau_j^{-1}\right)\, d\tau\, d\mu \end{aligned} \tag{3.4}$$

We use the Expectation Propagation algorithm (Minka, 2001) to infer posteriors over the latent variables μ and τ.[1] Expectation Propagation is a technique for Bayesian inference, which aims to find an approximate factorised distribution closest to the true distribution. It uses a message-passing algorithm to iteratively refine each factor by minimising the KL divergence from the approximate to the true distribution. We also experimented with using Gibbs Sampling and Variational Message Passing and observed similar results.

One benefit of using the above algorithms instead of expectation maximisation is that we can include additional constraints on the latent variables to help the model. We add the following constraints on the quality control items:

1. the true quality of the degraded translation is lower than the quality of the corresponding system translation
2. the true quality of the repeat items should be equal

We expect that the model will learn a high τ for good quality workers, and give their scores higher weight when estimating the mean. We believe that the additional constraints will help the model to infer the worker precision, as the model would learn that workers whose scores that have major violations of these constraints have a low precision.

Finally, note that we choose to use the standardised scores of workers as inputs to the model. If using raw scores instead, we would need additional parameters to infer the scale and offset

[1] We use the Infer.NET (Minka et al., 2018) framework to implement our models.

of the workers. In the current setup of DA where each subset of translations is scored by the same set of annotators, we believe that in the end, the scale and offset parameters this model would learn would be something very similar to the standardised scores. If, instead, we have a dataset where each annotator scores a random subset of the translations (as with the conference reviews dataset, where each reviewer rates a different set of papers (Flach et al., 2009)), then the model with an offset parameter might infer that an annotator consistently scores lower or higher relative to others. Finally, if using raw scores, the choice of Gaussian distribution to model worker scores is technically deficient as a Gaussian is unbounded. This could be remedied, for example, by using a truncated Gaussian distribution.

Relationship with Prior Work

Of the models described in Sec. 3.2 that can be applied to continuous data, EASL (Sakaguchi and Van Durme, 2018) has been directly applied to MT direct assessment data but for collecting and aggregating scores for MT systems. It models raw scores using a Beta distribution, and ignores the differences in worker scale. We showed that quality control can be aggressive with filtering out worker scores, and that there is potential in modelling annotator reliability when aggregating scores. However, EASL ignores worker reliability, and was applied only to the scores of "good" workers who pass quality control.

Our model, like many other models described in Sec. 3.2, assumes the annotator response is normally distributed around the true scores. Models like Raykar et al. (2010); Groot et al. (2011) jointly learn a regressor and the ground truth; their main goal is to learn a machine learning model, and require features. For MT evaluation, we could directly use automatic metrics or features from metrics. However, since automatic metrics are known to be biased against translations that are superficially dissimilar to the reference, using features based on these metrics (whether hand-crafted or obtained using representation learning) might potentially favour annotators who are also similarly biased. Moreover, this data is used to evaluate automatic metrics, so it would create a circular process if we were to incorporate features into the model used to construct the gold standard.

Finally, the models for relative ranking and pairwise preference data need an additional latent variable for the continuous scores perceived by the annotators, and then a mechanism to convert this to the observed score. The reviewer calibration (Flach et al., 2009) model and the model of Hopkins and May (2013) learn thresholds to convert to ordinal ratings and pairwise decisions respectively. Trueskill (Sakaguchi et al., 2014) uses custom equations to update the latent true score once a win, loss or tie is observed. Since we are directly dealing with continuous data, our model doesn't require this additional step.

3.5 Experiments

We evaluate our model on two multiply-annotated datasets that were collected for sentence-level metric evaluation: $\text{WMT13}_{\text{ES-EN}}$ and $\text{WMT16}_{\text{TR-EN}}$ presented in Sec. 3.3.2. When crowdsourcing annotations, the quality of workers that complete our HITs varies. For each dataset, we run our model on the scores of different random subsets of workers that represent the possible variation in quality. We evaluate our models based on their Pearson correlation[2] with the ground truth as we vary the number of workers per translation, comparing the translation quality inferred by our model with the following baselines:

1. raw-mean: the mean of the raw scores of all the workers input to the model,

2. z-mean: the mean of the standardised scores of all the workers input to the model, and

3. DA: the mean of the standardised scores of the subset of workers input to the model that pass quality control. This is the best practice recommended by Graham et al. (2015) and is followed by WMT to compute the translation quality scores (Bojar et al., 2016b)

[2] We compute the correlation only on the MT system translations and do not include the quality control items.

The WMT13 SPANISH → ENGLISH dataset contains 12 HITS in total, of which we use 4 HITS in our experiments as they were annotated by a large number of workers. We use the remaining 8 HITS as a development set to set the shape and rate parameters for the Gamma prior for annotator precision. We obtain $k = 5$ and $\theta = 0.3$ based on maximum likelihood fit of the distribution of worker precisions on these 8 HITs: we first compute errors of each worker from the "ground truth" (shown in Fig. 3.2), then compute the inverse variance of these errors. Note that the results are stable to small changes in the priors.

3.5.1 WMT13 SPANISH → ENGLISH

Our main experiments are on the four HITs of the WMT13$_{\text{ES-EN}}$ dataset which were annotated by at least 80 workers that passed quality control. We use the average score of 60 QC$_{\text{PASS}}$ workers as the ground truth, and use three different subsets of the remaining workers to test our models: (a) a mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers, (b) workers who fail quality control, and (c) workers who pass quality control. Of the workers that fail quality control, those whose scores were egregiously bad were rejected payment based on manual inspection. We refer to these workers as UNPAIDand the remaining workers (who were paid for their effort but whose scores were discarded) as PAIDQC$_{\text{FAIL}}$.

For each worker cohort, Fig. 3.5 shows

1. a scatterplot of the model's inferred precision of the subset of all 20 workers and the correlation of the workers with the ground truth, and

2. the correlation of the model's inferred quality compared to the baselines as we increase the number of workers per HIT.

We now analyse the results in detail:

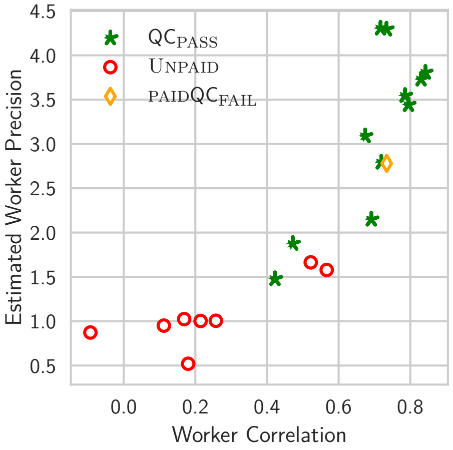
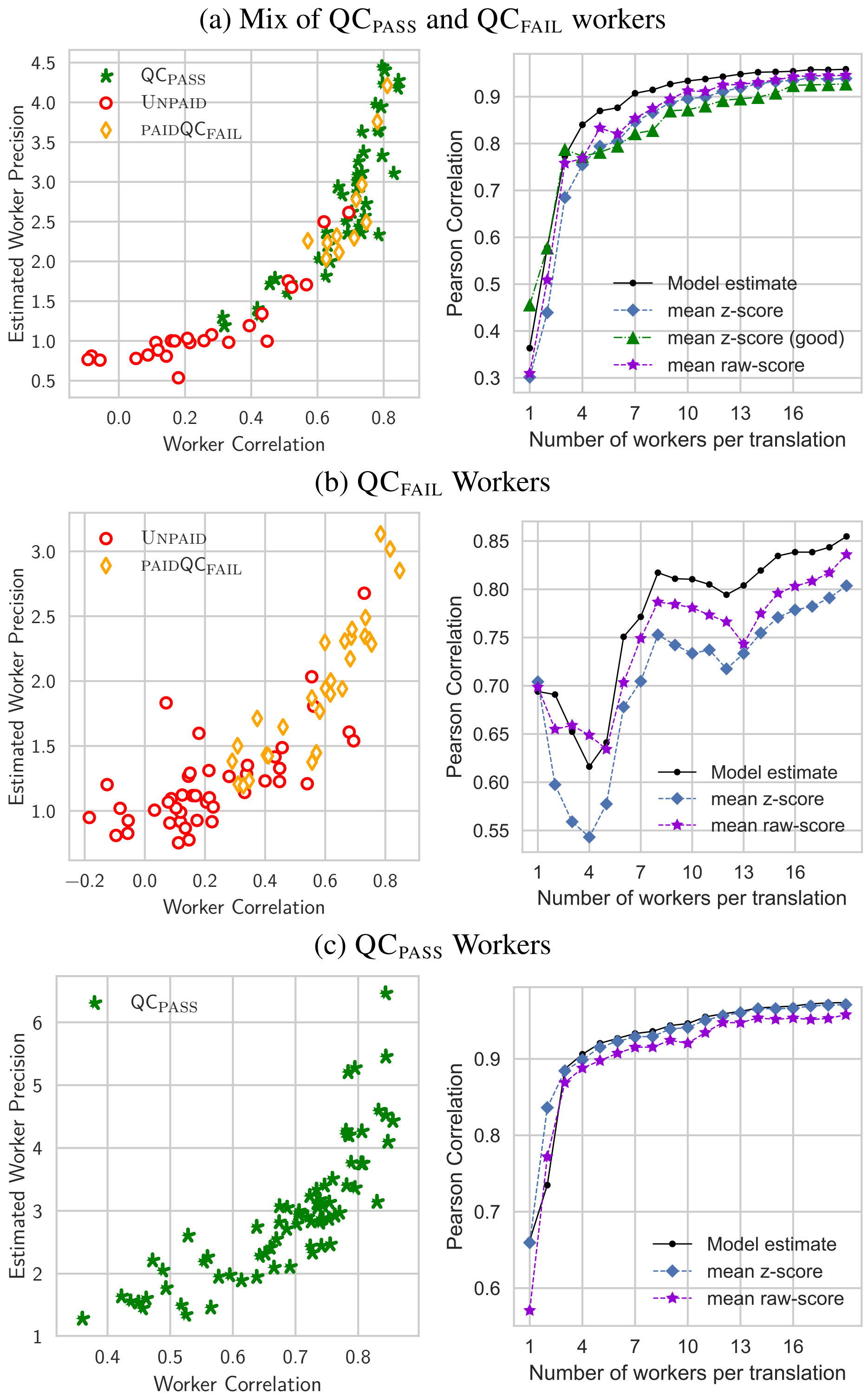


Fig. 3.4 Scatter plot of worker precision inferred by the model with five workers per translation against the correlation of the worker score with the "ground truth" on the WMT13$_{\text{ES-EN}}$ dataset.

Mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ Workers

Surprisingly, the mean of the raw scores has a slightly higher correlation than the mean of the standardised scores, particularly when there are fewer than six workers. One possible explanation for this is that workers that use only a small part of the scale essentially receive a lower "weight" when computing the mean. In an extreme example of a worker that scores randomly, adding these scores has little effect on the mean if their range is small. This might occur if they click in the same part of the analogue scale for all translations in the HIT. But if the scores are standardised, then the scores of this spammer have an equal influence on the mean as the other workers, and the correlation of the mean decreases. As we increase the number of workers, each worker has a smaller influence on the mean whether using raw scores or standardised, so a few low-quality scores do not bring down the correlation of the mean as long as there are enough high quality scores.[3]

[3] We find that workers who utilise only a small range of the scale are likely to have a low quality, probably because it is difficult to be consistent, or because they are deliberately spamming by randomly clicking on the same area of the scale.



(a) Mix of $\mathrm{QC_{PASS}}$ and $\mathrm{QC_{FAIL}}$ workers

(b) $\mathrm{QC_{FAIL}}$ Workers

(c) $\mathrm{QC_{PASS}}$ Workers

Fig. 3.5 Results on the $\mathrm{WMT13_{ES\text{-}EN}}$ dataset:
(left) Scatterplot of inferred precision against correlation with ground truth with 20 workers per HIT; (right) Correlation of the model's inferred quality with the ground truth as we increase the number of workers.

As many high quality workers fail to pass quality control, $r_{raw-mean}$ and r_{z-mean} (the correlation of the raw and standardised mean) is higher than r_{DA} (the correlation of the standardised mean of the subset of QC_{PASS} workers).

With as few as 5 worker scores per HIT, the model's inferred precision correlates highly with worker correlation (Fig. 3.4). Of the 20 workers that annotated these four HITs, 11 pass quality control, and the correlation of these QC_{PASS} workers with the ground truth ranges between 0.4 and 0.85. There are three other workers that do not pass quality control but have a moderate correlation ($r > 0.5$). The model also recognises that the remaining workers have a low quality, and learns a low precision for these workers. The model's inferred estimate has a higher correlation than all three baselines. The correlation of the model estimate, r_{model}, rapidly reaches 0.9 at 7 workers per translation. As we increase the number of workers to 20, r_{model} continues to increase even when we add low-quality workers. The gap between the correlation of the model's inferred quality and the baselines decreases, but the model is always slightly better.

Most UNPAIDworkers are weakly correlated with the ground truth. Unfortunately, some of these workers have a moderate correlation, and the correlations of QC_{PASS} and UNPAIDworkers have an overlap. Finally, most PAIDQC_{FAIL}workers are moderately correlated with the ground truth; the model takes advantage of these workers, and its estimate of the model quality is clearly better than the mean of the subset of QC_{PASS} workers.

QC_{FAIL} Workers

We next look at only the subset of workers that have not passed quality control, to check how much valuable information is lost in the process of quality control. This mix contains both workers who were rejected payment (UNPAIDworkers) and PAIDQC_{FAIL}workers who were paid but whose scores were discarded due to failing the significance tests.

The quality of the mean of raw scores is noticeably higher than the mean of z-scores, and our model quality is typically even higher. This is consistent with the observations on the previous cohort that includes a mix of QC_{PASS} and QC_{FAIL} workers, and can be attributed to low-quality workers that utilise only a small range of the scale.

In this cohort of workers, the first annotators of all four HITs coincidentally have a reasonably high quality ($r \geq 0.6$). The correlation of all three methods drops when we add low-quality workers, but the decrease in model quality is much smaller as it gives a higher weight to the high-quality workers. As we add more workers, r_{model} increases markedly; it follows the trend of r_{z-mean}, but maintains the strong improvement over both baselines. At 20 workers per translation, r_{model} is just above 0.85, whereas r_{z-mean} is only around 0.8. Note that r_{model} is the best performing method when we have more than four workers.

QC_{PASS} Workers

Finally, we look at only the subset of workers that have passed quality control, to confirm that the model does not fail in the (unlikely) scenario where all annotations have a high quality. Although some of these annotators only have a moderate correlation with the ground truth, this doesn't negatively affect r_{z-mean} due to the presence of many high quality annotators, and r_{z-mean} steadily increases as we add more annotators per HIT. The mean of the standardised scores is clearly better than the mean of raw scores, replicating the results shown in Graham et al. (2015).

At just two workers per HIT, the model is unable to correctly discern which worker is better, and so r_{model} is lower than r_{z-mean}. With low-quality workers, the model uses the constraints on the quality control items, but when both workers pass quality control, this information is less helpful. Once we add a third worker, the model can now reasonably infer worker accuracy, and r_{model} is very slightly higher than r_{z-mean}. It remains at least as high as r_{z-mean} as we increase the number of workers to 20.

3.5.2 WMT16 TURKISH → ENGLISH

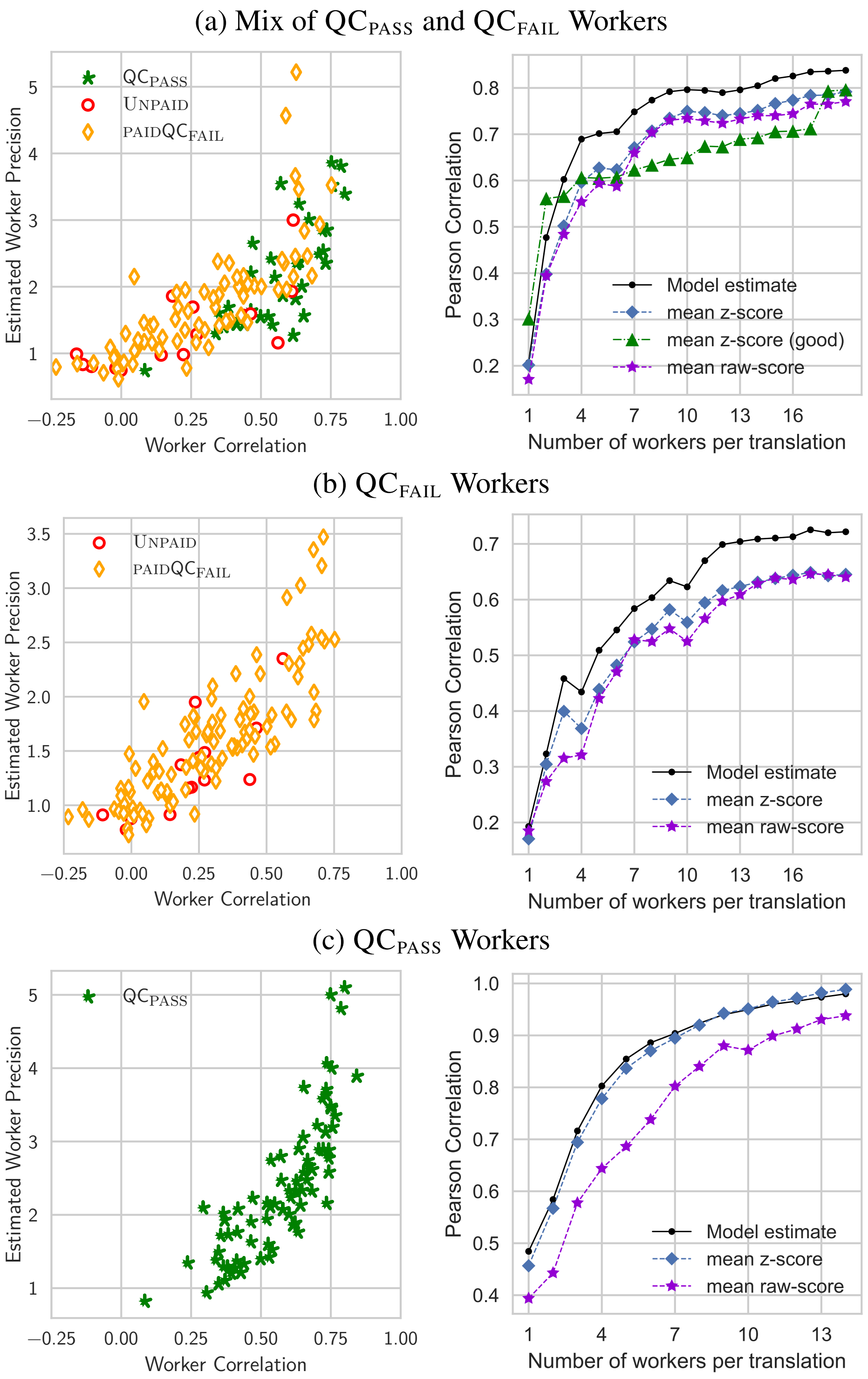
This dataset consists of 8 HITS, with at least 15 $\mathrm{QC}_{\mathrm{PASS}}$ annotations for each translation. Of the 256 workers that completed these HITs, about two thirds (67.58%) fail the quality control measures. We choose this dataset to show that our model performs well even when there is a large proportion of low-quality workers.

We use the mean of all $\mathrm{QC}_{\mathrm{PASS}}$ workers as the ground truth. This number is much smaller than the 60 worker scores used in the $\mathrm{WMT13}_{\mathrm{ES\text{-}EN}}$ dataset. Moreover, as we showed in Sec. 3.3.3, some of the $\mathrm{QC}_{\mathrm{PASS}}$ workers have a low correlation with the mean. Thus, this “ground truth” is not as reliable as with the $\mathrm{WMT13}_{\mathrm{ES\text{-}EN}}$ dataset, and we can expect our model to have lower correlations.

In this dataset, very few workers were outright rejected payment. We do not see the clear demarcation between PAID$\mathrm{QC}_{\mathrm{FAIL}}$and UNPAIDworkers; the correlation of the rejected UNPAIDis spread between −0.3 and 0.6, overlapping significantly with the PAID$\mathrm{QC}_{\mathrm{FAIL}}$workers.

As with the $\mathrm{WMT13}_{\mathrm{ES\text{-}EN}}$, we test our models on (a) a mix of $\mathrm{QC}_{\mathrm{PASS}}$ and $\mathrm{QC}_{\mathrm{FAIL}}$ workers, (b) $\mathrm{QC}_{\mathrm{FAIL}}$ workers, and (c) workers who pass quality control. Note that the mix of workers in sets (a) and (c) overlaps with the $\mathrm{QC}_{\mathrm{PASS}}$ workers used to compute the ground truth, unlike the $\mathrm{WMT13}_{\mathrm{ES\text{-}EN}}$ dataset where we had a separate set of $\mathrm{QC}_{\mathrm{PASS}}$ workers for the gold standard. This provides an advantage to the baselines when evaluating over sets (a) and (c). As the ground truth is the mean z-score of all $\mathrm{QC}_{\mathrm{PASS}}$ workers, and we have 15-17 $\mathrm{QC}_{\mathrm{PASS}}$ workers per HIT, $r_z - mean \lessapprox 1$ for set (c) at n = 15 workers.

Fig. 3.6 shows the results. With any of the three subsets of workers, the correlation of the mean standardised scores is always higher than that of the mean raw scores. This is particularly true for the set of $\mathrm{QC}_{\mathrm{PASS}}$ workers, where r_{z-mean} is clearly higher than $r_{raw-mean}$. This might be attributed to the fact that the “ground truth” is the mean of all z-scores, but r_{z-mean} is higher than $r_{raw-mean}$ even with the set of $\mathrm{QC}_{\mathrm{FAIL}}$ workers, where there is no overlap at all with the scores used to compute the ground truth.



(a) Mix of QC_{PASS} and QC_{FAIL} Workers

(b) QC_{FAIL} Workers

(c) QC_{PASS} Workers

Fig. 3.6 Results on the $WMT16_{TR-EN}$ dataset:
(left) Scatterplot of inferred precision against correlation with ground truth with 20 workers per HIT; (right) Correlation of the model's inferred quality with the ground truth as we increase the number of workers

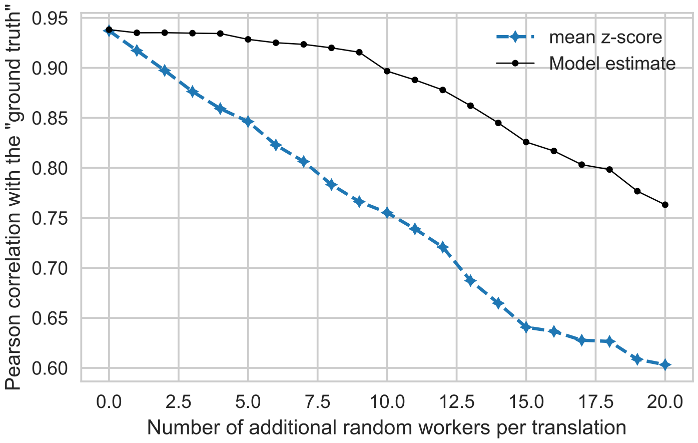


Fig. 3.7 Pearson's r of the estimated true score with the "ground truth" of five high quality workers and an additional N random workers per translation, where N ranges from 0 to 20. The model estimate degrades much slower than the baseline (the mean of the standardised scores of these workers).

Our model is clearly better than the baselines for the cohort of QC_{PASS} + QC_{FAIL} workers and the cohort of QC_{FAIL} workers, with a stronger improvement over the latter subset. On the set of QC_{PASS} workers, the model estimate is slightly better than the standardised mean when we have less than 10 workers per HIT. This is mainly due to one low quality worker (correlation less than 0.1), whose scores managed to pass quality control. After that, the z-mean gains a slight advantage over the model estimate, which might be attributed to the gold standard being computed as the mean of z-scores of all QC_{PASS} workers.

3.5.3 Adversarial Settings

In addition to the experiments above on MTurk data, we generate synthetic data to understand how well the model performs under adversarial conditions. We simulate a scenario where we have five high-quality scores per translation, and test whether the model can recover the true quality despite the presence of additional workers whose scores contain no information about the quality of the translation.

For each of the four HITS from the WMT13$_{\text{ES-EN}}$, we sample five QC$_{\text{PASS}}$ workers with high correlation ($r \geq 0.8$) with the true quality, and generate scores of additional random workers by sampling from the standard normal distribution. Fig. 3.7 shows the performance of the model as we add up to 20 random workers to the five QC$_{\text{PASS}}$ workers.

With only the five QC$_{\text{PASS}}$ workers, the correlation of the mean (r_{z-mean}) and the model's inferred quality (r_{model}) with the ground truth is 0.94. The model doesn't have much scope for improvement over the mean because all workers are very high quality. As we add random workers to the mix, r_{z-mean} drops sharply. But the correlation of the model's inferred quality is still above 0.9 even after adding up to ten random workers. The correlation of the model decreases steadily after that, but the rate of decrease is much smaller compared to the mean, and the difference between r_{model} and r_{z-mean} steadily increases as we add more random workers. After adding 20 random workers, r_{model} is 0.76, which is an improvement of 0.16 over r_{z-mean}.

3.5.4 Spammer Removal

While the model down-weights scores from annotators with a low inferred precision, these scores still have some influence on the inferred translation quality. If the scores are random (or worse, negatively correlated with the true quality), then this might negatively affect the inferred translation quality as the model attempts to fit these scores. Li (2019) likens spammer removal to feature selection for Machine Learning models, where workers can be considered as features, and noisy features can decrease performance.

Following Raykar and Yu (2012) and Rahimi et al. (2019), we exclude low-quality annotators (called "spammers") from the model's input data. More specifically, we rank annotators of each HIT based on the inferred precision of the model trained with all 20 workers in the cohort. Next, we retrain the model with the top k workers from each HIT, and plot the correlation of the models' estimated translation quality. If the model is being hindered by spammers and its estimate of worker accuracy is correct, we expect that the model estimate gets steadily better as we discard low quality scores. But as we remove more annotators, the model will begin to lose valuable information, and the model quality will begin to decrease. On the other hand, if the estimate of worker reliability is flawed, the model's correlation with the ground truth will be unstable as we discard workers.

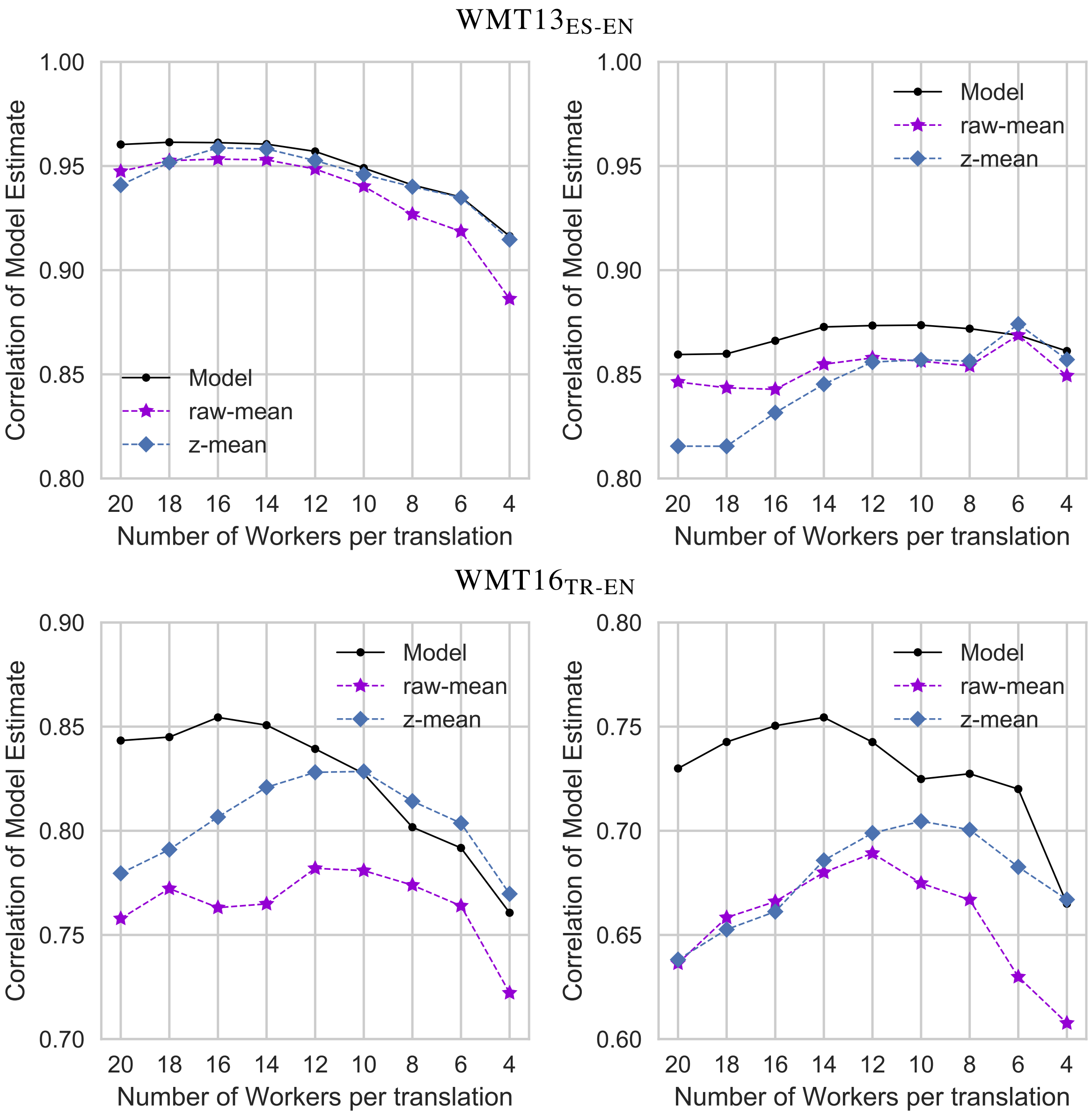
We compare spammer removal with two unsupervised baselines that correspond to the raw and standardised mean. We compute the mean raw score of all workers in the cohort, then sort workers based on their correlation with this. To simulate spammer removal on the raw scores, we then compute the mean score of the k workers with the highest correlation with the mean of all workers. We do the same process with the standardised scores.

Fig. 3.8 shows the results of removing spammers on four subsets of the data:

Mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers, WMT13$_{\text{ES-EN}}$: With all 20 workers, r_{model} is around 0.9, which is higher than r_{z-mean} and $r_{raw-mean}$. The model appears to be robust to having a few low quality workers as inputs, and r_{model} stays constant as we remove the least precise workers. The correlation then begins to decrease after we remove more than 8 workers. At the top 16 workers, r_{z-mean} surpasses $r_{raw-mean}$ and is almost as high as r_{model}, and stays approximately equal to r_{model} as we decrease k. At just the best four workers, r_{z-mean} and r_{model} are still above 0.9.

QC$_{\text{FAIL}}$ workers, WMT13$_{\text{ES-EN}}$: This subset has some low quality workers and using the top 14 workers results in an increase in r_{model} compared to using all 20 workers. Both $r_{raw-mean}$ and r_{z-mean} increase as we remove the least correlated workers, showing that the model is much more resilient to spammers compared to the mean.

Mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers, WMT16$_{\text{TR-EN}}$: Here, r_{model} increases as we remove the 4 worst workers, then decreases sharply as we remove more workers. The quality of the best workers in these subsets is not as high as with the WMT13$_{\text{ES-EN}}$ dataset, so more workers



(a) Mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers

(b) QC$_{\text{FAIL}}$ Workers

Fig. 3.8 Results of spammer removal: correlation of the model's inferred quality with the ground truth, as we decrease the number of annotators per HIT. For the model estimate, we use the top k workers, ranked by the inferred precision of these workers. For the baselines, we sort workers based on their correlation with the mean of all 20 workers.
(Column 1 is cohort of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers, and column 2 is cohort of QC$_{\text{FAIL}}$ workers.)

are needed to achieve a high correlation. Again, r_{z-mean} is clearly higher than $r_{raw-mean}$. Interestingly, the correlation of the baselines peak at the top 12 workers instead of top 16, and it appears that the model is not very effective at selecting the best workers as it is outperformed by the mean of standardised scores when k is between 4 and 12. Note that the best correlation of the model is still higher than the best correlation of the baselines.

QC$_{\text{FAIL}}$ workers, WMT16$_{\text{TR-EN}}$: Spammer removal results the highest improvement in this dataset: r_{model} increases from 0.73 to 0.76 on discarding four workers, then decreases. The model also maintains a clear improvement over both baselines. However, the quality is still low, and we clearly need to collect more annotations.

3.5.5 Analysis of Individual HITs

The quality of the model estimate depends on the set of workers who annotated the translations. In this section, we look at some representative HITS to help determine (a) what is the ideal number of spammers to remove and (b) whether we need to collect more annotations.

For each HIT, Fig. 3.9 shows

1. A heatmap showing the pairwise correlation of the scores of the workers who annotated the HIT, along with the model's inferred quality and the standardised mean. The workers are sorted based on the model's inferred precision. We use a diverging colour scheme, with a darker hue representing a stronger relationship. Instead of being centered at zero, the heatmap is centered at 0.198[4]. Thus, a random worker would have mostly blue hues in their row and column, and a strong worker would have dark browns when compared to other strong workers. If the model correctly estimates the precision of the workers, we would expect the top left block to have dark browns, with a slow transition to blue as we move to the bottom right.

[4] This represents the 95% confidence interval for the correlation of two variables of length 70 drawn from a random normal distribution.

2. a scatterplot of the model's inferred precision of the subset of all 20 workers and the correlation of these workers with the ground truth, and
3. the correlation of the model's inferred quality with the ground truth of the top k workers per HIT, based on the model's inferred accuracy with all 20 workers as input.

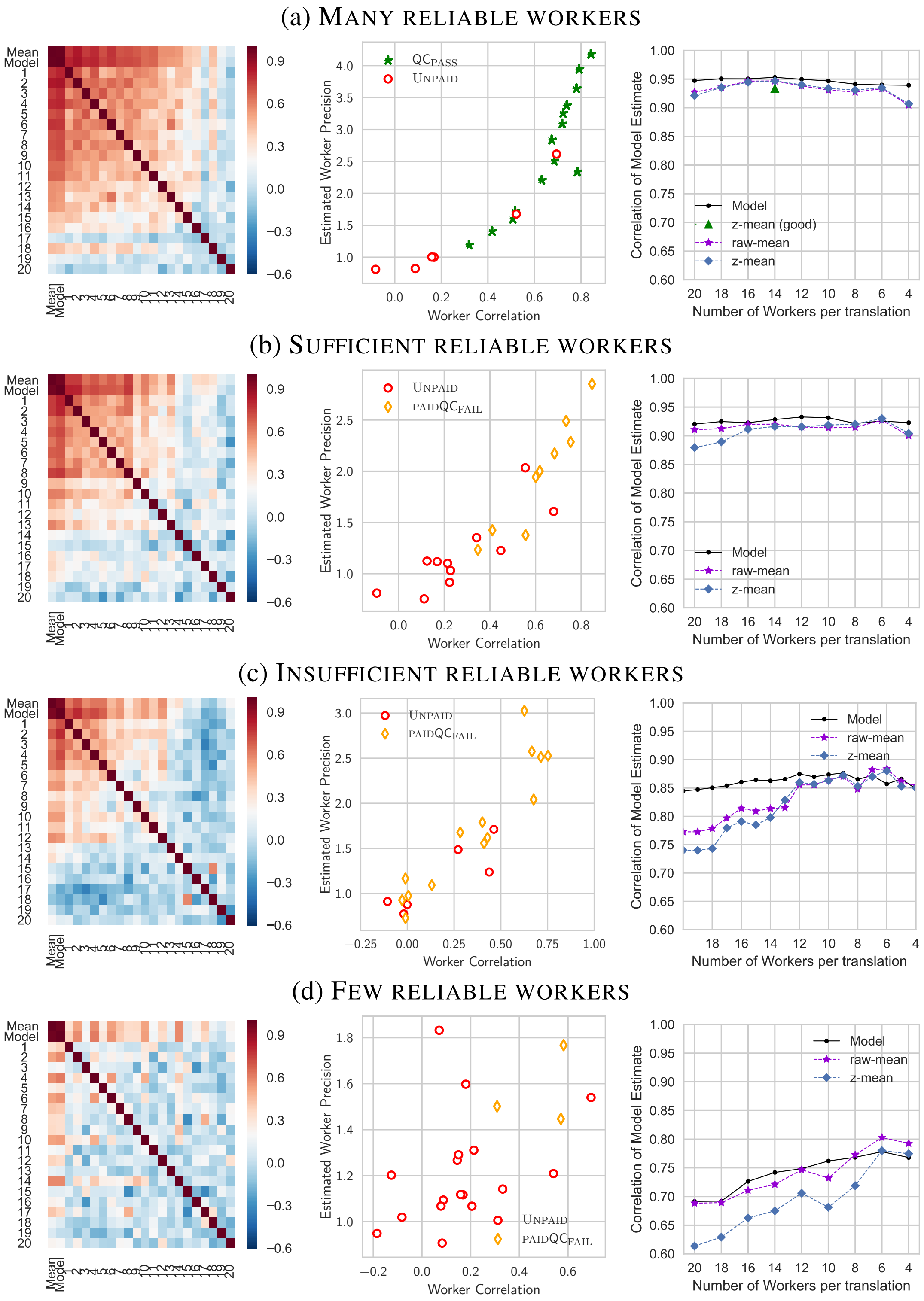
In all the heatmaps in Fig. 3.9, the workers with higher inferred precision (the top rows) correlate higher with the model than the mean score. The trend gradually reverses, and the model places less weight on scores with low inferred precision. Here are observations specific to individual HITs:

(a) This is an example of a HIT from WMT13$_{\text{ES-EN}}$ (mix of QC$_{\text{PASS}}$ and QC$_{\text{FAIL}}$ workers) that has attracted many reliable workers. The correlation heatmap has a large and intense brown block at the top left corner, and just a few rows of light blue at the bottom. The model obtains a high correlation ($r = 0.95$), and is resilient to the obvious spammers based on the heatmap.

(b) This HIT contains workers from the WMT13$_{\text{ES-EN}}$ dataset that fail quality control. The heatmap appears similar to HIT (a), except that the block of highly correlated workers is smaller. But this is sufficient to gain a correlation above 0.9 with all 20 workers. The maximum correlation is achieved using the top 12 workers, which corresponds to removing workers whose rows are predominantly blue in the heatmap.

(c) This HIT was annotated by workers from the WMT16$_{\text{TR-EN}}$ dataset that fail quality control. Like HIT (b), the heatmap indicates that the correlation is highest when using the top 12 workers. However, many of these workers are only moderately correlated with each other, indicating that these workers are less reliable, and more annotations should be collected.

(d) At the other extreme is an example from WMT13$_{\text{ES-EN}}$: QC$_{\text{FAIL}}$ workers. Here, the heatmap is predominantly blue, with scattered light brown cells, which would ideally


(a) MANY RELIABLE WORKERS

(b) SUFFICIENT RELIABLE WORKERS

(c) INSUFFICIENT RELIABLE WORKERS

(d) FEW RELIABLE WORKERS

Fig. 3.9 Analysis of Selected HITs:
Column 1: Pairwise correlation of the model estimate, standardised mean and all workers sorted by inferred accuracy. See text on page 82 for details.
Column 2: Scatterplot of the model estimate of worker precision against worker correlation.
Column 3: Correlation of the model's inferred quality with the ground truth as we remove spammers.

be present at the top left corner. This indicates that the model has failed to infer the top workers. The highest correlation is obtained at $k = 6$ workers, which is impossible to determine based on the heatmap. But the heatmap does highlight the need to collect more annotations, and with better quality annotations, the model might do better at inferring worker quality.

When collecting annotations, these heatmaps can be used as a diagnostic tool to help determine how many spammers to remove and whether we have collected enough annotations.

3.6 Conclusion

In this chapter, we looked at better methods to aggregate direct assessment data to estimate accurate scores for individual translations.

Direct assessment uses statistical significance tests over a small set of quality control items to determine "good" workers, and computes the true translation quality as the mean standardised scores of at least fifteen QC_{PASS} workers. However these tests are too conservative, and can lead to discarding potentially useful data. Workers are rejected payment only when their scores are found to be egregiously bad based on manual examination, so discarding useful scores leads to an increase in the cost of annotation.

The errors of the standardised annotator scores are normally distributed with a zero mean, with better annotators having a smaller variance. Based on this, we proposed a simple probabilistic model to aggregate data instead of using the mean of QC_{PASS} workers. The model learns the precision (inverse variance) of individual annotators based on patterns of agreement with other annotators. We also supply additional constraints on the quality control items to help with this. The model returns an estimate of the true quality that places a higher weight on more reliable workers.

We tested this model on two datasets: the WMT13$_{\text{ES-EN}}$ which has four HITs with at least 80 workers that pass quality control, and additionally, the WMT16$_{\text{TR-EN}}$ dataset that contains a large percentage of workers that do not pass quality control. It typically returns a more reliable estimate of the translation quality using fewer annotations per HIT.

We next explored spammer removal, where we rerun the model with the top k workers based on their inferred accuracy. We found that the model was a lot more resilient to spammers (random workers) compared to the mean of raw or standardised scores. When there are enough high-quality workers present, the model is not adversely affected by the presence of a few spammers and removing these workers doesn't help the model. In other scenarios, for example, with many spammers, using the scores of the top k workers resulted in an improvement over the using all 20 workers.

A major limitation of this model is that it does not directly inform us of how to choose how many spammers to remove or when to stop collecting judgements. We presented a potential direction for solving this, by plotting pairwise correlations of the worker scores. Excellent workers are highly correlated with the true quality of the translations, which, in turn, means that they must correlate with each other. On the other hand, random workers would have low pairwise correlations with the other workers (irrespective of their quality). Based on this, we can develop heuristics that answer both questions. We can filter out workers who have low correlation with others, and collect more annotations until we have at least t workers who have a correlation higher than a threshold with the other workers, where t can be empirically determined, and then tested on unseen data. These ideas, if further developed, could reduce costs further.

Finally, this work assumes that the annotator scores are unbiased and that the noise is random. In the next chapter, we show that this assumption is not always true: humans are susceptible to cognitive biases, and we present evidence for one such bias in MT direct assessment data.

Chapter 4

Sequence Effects in Crowdsourced Human Annotations

This chapter builds on the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Sequence effects in crowdsourced annotations. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2860–2865, Copenhagen, Denmark, September 2017.

4.1 Introduction

Human evaluation is regarded as the gold standard for machine translation evaluation: humans can intuitively identify how much of the meaning of the original sentence has been preserved when scoring MT outputs. In particular, when comparing MT output with a reference translation, human evaluators can quickly understand the impact of any difference between the two sentences. In contrast, automatic metrics struggle to do so, and can be biased towards translations that are superficially similar to the reference translation, which leads to sub-par performance when evaluating MT systems that often look superficially similar to the reference translation but hide semantic inadequacies.

In the previous chapter, when we inferred the true translation quality from multiply-annotated data, we assumed that errors in individual annotators' ratings are random, i.e., their ratings are unbiased, and so aggregating these ratings would yield robust and accurate estimates of the true quality.

However, humans are affected by subconscious cognitive biases. The design of the annotation task can influence the decisions made by annotators in subtle ways: besides the actual features of the instance being annotated, annotators are also influenced by factors such as the user interface, wording of the question, and familiarity with the task or domain. Accordingly, the assumption of unbiasedness is not always justified.

In this chapter, we explore one particular source of cognitive biases, and show that annotator judgements are affected by sequence bias, whereby the order of presentation can affect individuals' assessment of an item. We focus on crowdsourced data, where annotations are collected by a number of anonymous crowd workers through platforms such as Amazon Mechanical Turk.

To mitigate the cost and time requirement of obtaining annotations from experts, multiple studies have been conducted to test the feasibility of using crowdsourcing in NLP, and crowdsourcing is now regularly used to obtain NLP annotations, including MT human judgements[1]. (See Sec. 2.1.3 for a review of various attempts at crowdsourcing MT evaluation.) In particular, the official evaluation at the annual Conference on Machine Translation (abbreviated as WMT for historical reasons) has been partly crowdsourced between WMT 2010 to WMT 2013 (Callison-Burch et al., 2008; Bojar et al., 2013), and then since 2017 (Bojar et al., 2017b) to the present year.

[1] As MT systems improve in quality, there has been recent work that show that expert evaluations do not always agree with crowdsourced evaluations (Toral et al., 2018; Läubli et al., 2020; Freitag et al., 2021), but crowdsourcing is still commonly used to evaluate MT.

The items required to be annotated are broken down into small groups of instances called HITs and are presented to crowd workers. As individual anonymous workers are expected to be less accurate than experts, crowdsourcing requires more redundancy than expert annotation. Typically, multiple annotations are collected per item, which are then aggregated to yield the final labels.

In most annotation exercises, the order of presentation of instances is randomised to remove bias due to similarities in topic, style and vocabulary (Koehn and Monz, 2006; Bojar et al., 2016a). When crowdsourcing judgements, the normal practise (as used in the datasets we analyse) is for the item ordering to be randomised when creating HITs, and then to have each HIT annotated by multiple workers.[2] All workers who complete the HIT generally see the items within the HIT in the same order (Snow et al., 2008; Graham et al., 2017).

We begin the chapter with a review of the literature on cognitive biases, focusing on sequence effects. We then describe the method we use to detect sequence effects, before moving on to our experiments. We investigate the presence of sequence effects on the multiply-annotated MT direct assessment dataset (Graham et al., 2015) that we used in our experiments in Chapter 3. While direct assessment is a primary method for human evaluation of MT today, many researchers still use other methods, for example, obtaining preference judgements over two translations of the same source sentence. To show that sequence effects can also be present in other kinds of annotations, we explore additional crowdsourced NLP tasks where raw crowdsourced data is publicly available along with expert judgements: the binary tasks of textual entailment and temporal ordering (Snow et al., 2008). Finally, we analyse a crowdsourced affective text analysis dataset (Snow et al., 2008) to check if sequence effects can be detected in settings where workers annotate each sentence on different aspects (in this case, valence and five emotions), before moving on to the next sentence.

[2] This can be easily done in Amazon Mechanical Turk, for example, by specifying the maximum number of assignments per HIT.

4.2 Background: Decision making and Cognitive biases

We can regard annotations as a sequence of decisions made by the annotators. Psychologists distinguish between two modes of thought in the mind that are responsible for decision making called system 1 and system 2 (Kahneman, 2011). System 1 is fast, automatic, relies on associations and memories. This system is responsible for everyday decisions and intuitive judgements such as the distance to a given object, or simple items that have been learned though repetition such as $6 \times 7 = 42$.

System 2 requires attention and cognitive effort, and is typically activated when system 1 requires assistance, for example, in new and unfamiliar situations such as computing complex mathematical calculations.

System 1 is usually correct and justified in its decisions, but sometimes, intuitions and heuristics can lead to systematic errors or biases. System 2, on the other hand, is slow and analytical, and is more resistant to these biases.

Ideally, NLP annotators would rely completely on System 2 when making decisions. However, even in scenarios where people intend to make deliberate, careful decisions, it can be difficult to “turn off” the automatic judgements of system 1, and subconscious heuristics can factor into decisions that are intended to be completely rational. In this chapter, we focus on the biases that arise when making decisions in a sequence: instead of evaluating each instance independently, annotators might be influenced by their response to previous instances. Possible explanations for sequence effects include:

Gambler’s fallacy: Once annotators have developed an idea of the distribution of scores or labels, they can come to expect even small sequences to follow the distribution. For example, in binary annotation tasks, if they expect that positive (1) and negative (0) items are equally likely, then they believe the sequence 11100001 is less likely than the sequence 01011010, even though both contain the same proportion of positive and negative items (50% False and

50% True). When asked to generate random sequences, experimental subjects refrain from producing strings that contain patterns or streaks, and when judging sequences, any sequences that happen to contain these patterns are judged as non-random (Bar-Hillel and Wagenaar, 1991). This can affect the decisions they make when annotating a sequence of items: if they assign a negative label to an item, they may approach the next item with a prior belief that it is more likely to be a positive. In this way, the distribution of their judgements can match their prior more closely than it really should. Evidence for gambler's fallacy has been found in lab settings as well as real world decisions.

In the lab, to demonstrate gambler's fallacy, undergraduate students were given the choice of gambling or playing it safe when predicting the outcome of a series of coin tosses (Gold and Hester, 2008). When they chose to gamble, they receive 100 points if the result of the fair coin toss is a preselected "winning" side, say heads. If they choose the safe option, they receive 70 points irrespective of the outcome of the coin toss. The rational choice would be to always choose the safe option, as the expected value on gambling on a fair coin is 50 points. After a streak of three heads, as expected, participants were more likely to choose the safe option. However, they were more likely to gamble on the 100 points after three tails, showing that they were subconsciously engaging in the Gambler's fallacy. In another experiment, they found that this bias disappears when switching coins or allowing the coin a day to "rest" before tossing the coin again.

Additionally, real world high-stakes decision-makers have been shown to be prone to the gambler's fallacy. Chen et al. (2016) found that refugee asylum judges are more likely to reject an application when the previous application was accepted. Similarly, they showed evidence of gambler's fallacy in the decisions of baseball empires when calling whether a pitch was a ball or a strike, and loan officers when deciding whether to grant a loan. This bias is weaker if decision makers are more experienced or given strong incentives to be right.

Sequential contrast effects: Sequential contrast effects appear when the perceived quality of the current item is affected by the quality of the items observed immediately before. When scoring a series of items, a high quality item may raise the bar for the next item. On the other hand, a bad item may make the next item seem better in comparison. Sequential contrast effects have been observed in scenarios such as judgements in physical attractiveness (Kenrick and Gutierres, 1980; Bhargava and Fisman, 2014) and financial markets, where the earnings of publicly traded firms are perceived as more or less impressive by investors depending on the previous day's earnings (Hartzmark and Shue, 2018).

Assimilation effects: When faced with a sequence of decisions, annotators might show assimilation effects where there is a positive bias towards the previous decision(s) made. In continuous tasks such as rating the quality of items, this means that the score of one item is closer to the previously assigned score, compared to a scenario where the two decisions were made independently. Assimilation effects have been reported in real world sequential decisions, for example, in the scores of expert gymnastics judges (Damisch et al., 2006) and jury decisions in criminal courts (Bindler and Hjalmarsson, 2018). One explanation for this is anchoring and insufficient adjustment: the annotator uses their score of the previous item as an anchor, and adjusts the score of the current item from this anchor based on perceived similarities and differences with the previous item (Tversky and Kahneman, 1974). Anchoring effects may decrease as people gain experience and expertise in the task (Wilson et al., 1996).

Cognitive biases in NLP annotations

NLP research relies heavily on annotated datasets for training and evaluation. When collecting annotations for any NLP task, care is usually taken to ensure that the annotations are of high quality, through careful design of label sets, annotation guidelines and training of annotators (Hovy et al., 2006), and intuitive user interfaces (Stenetorp et al., 2012).

There has been much work on the potential of errors due to cognitive biases. Schoch et al. (2020) caution against the influence of positive or negative framing of the questions, and anchoring effects formed based on examples presented to annotators. In a more concrete example, annotators are shown to exhibit anchoring effect when asked to revise predictions from machine learning models which serve as the initial anchors; the labels of annotators correcting the POS tags and dependency parser predictions by the machine learning models are biased towards the initial tags, and tend to have more errors compared to annotations made from scratch (Berzak et al., 2016).

To the best of our knowledge, there are no mentions in the NLP literature of the cognitive biases arising from the sequential nature of annotating items. During annotation tasks requiring discrete labels, we might expect annotators to develop prior heuristics on the distribution of the labels, resulting in errors when the sequence of actual items do not conform to these priors. When assessing the quality of NLP systems, there is potential for sequential contrast or assimilation effects. In the next section, we present a simple model for detecting sequence effects in annotations.

4.3 Methodology

Ideally, the response of a given annotator for a given item can solely be explained by the true label/score of the item. If this is the case, we should not be able to predict the decision on current item given previous decisions. If, however, the annotator is influenced by the previous item, there would be either a positive or negative relationship between consecutively annotated items.

Following Chen et al. (2016), we use a simple linear model to test whether the annotation of an instance is correlated with the annotation on previous instances, conditioned on control variables such as the gold standard (i.e. expert annotations[3]):

$$\mathrm{Y}_{i,t} = \beta_0 + \beta_1 \mathrm{Y}_{i,t-1} + \beta_2 \mathrm{Gold}_t + \eta \tag{4.1}$$

where $\mathrm{Y}_{i,t}$ is the annotation given by an annotator i to an instance t, and η is Gaussian noise with zero mean. To fit predictions to the data, we use linear regression for continuous tasks and logistic regression for binary tasks. If there is no dependence between consecutive instances, and annotators assign labels/scores based only on the aspects of the current instance, then the data can be explained from the gold score (learning a positive β_2 value) and bias term (β_0), and the autocorrelation coefficient β_1 will be close to zero.

If we do not use the ground truth as a control variable, then a positive or negative autocorrelation might arise by chance, particularly if the sentences are not randomised. But since we do include this control, a non-zero value of β_1 is evidence of mistakes being made by annotators due to sequential bias. A positive value of β_1 can be explained by priming or anchoring, and a negative value with sequential contrast effects or the gambler's fallacy. Accordingly, we test the statistical significance of the $\beta_1 \neq 0$ to determine whether sequence effects are present in crowdsourced text corpora.

Note that this is a first order model that assumes linearity; it doesn't consider higher order interactions, and is unable to describe annotator decisions that are influenced by earlier responses.

4.4 Experiments

In this section, we present evidence of sequence effects in the machine translation adequacy dataset. We also explore other influential crowdsourced datasets including both binary and

[3] For the Machine Translation dataset described in Sec. 4.4.1, we use the mean of at least fifteen crowd workers as a proxy for expert annotations.

continuous annotation tasks: recognising textual entailment, event ordering, and affective text analysis.

4.4.1 Machine Translation Direct Assessment

Our main dataset is based on direct assessment (Graham et al., 2017) — the current best-practise, as adopted by WMT since 2017 (Bojar et al., 2017b). We test for autocorrelation in this data, exploring the difference in how workers with different quality annotations are affected by sequence bias. We also analyse whether workers get more or less biased as they annotate more items. Finally, we investigate whether the presence of sequence effects can influence aggregate scores.

We briefly summarise direct assessment (DA), which we described in more detail in Sec. 3.3, before presenting results. Annotators are asked to judge the adequacy of translations using a 100-point sliding scale which is initialised at the mid point. There are 3 marks on the scale dividing it into 4 quarters to aid workers with internal calibration. They are given no other instructions or guidelines on how to use the scale, and what quality each part of the scale represents[4].

Each HIT contains 100 items, and is designed to include quality control items (10 repeat MT translations, 10 reference translations, and 10 deliberately degraded translations) to filter out poor quality scores. Worker who submitted scores of clearly bad quality were rejected payment based on manual inspection of the quality control items and additional key indicators like the time taken to complete the annotations. We refer to these workers as “bad”. The scores on the quality control items of the remaining workers are further checked using statistical significance tests, and only workers who fulfil these criteria are used to compute the final results. We refer to these workers as “good” or “moderate”, based on whether they fulfil these additional criteria. (See Sec. 3.3.1 for more details on quality control.)

[4] In contrast, the FLORES setup of DA (Guzmán et al., 2019), presents additional guidelines to annotators, for example, "the 0–10 range represents a translation that is completely incorrect and inaccurate".

Finally, to eliminate differences due to different internal scales, every individual worker's scores are standardised. Following Graham et al. (2015), we use the average of standardised scores of at least 15 "good" workers as the ground truth.

In this chapter, we base our analysis on the adequacy dataset on Spanish-English newswire data from WMT 2013 (Graham et al., 2015; Bojar et al., 2013). The dataset consists of 12 HITS of 100 sentence pairs each; each HIT is annotated by at least 30 workers that pass quality control (we describe the dataset more fully in Sec. 4.4.1). In the original datasets, the scores of translations that are not MT system outputs are discarded once they have been used to determine the quality of the crowd workers. In our experiments, we consider all the instances that were rated by the annotators to preserve continuity. We refer to the final dataset as "MT$_{adeq}$".

Results

As this is a continuous output, we use a linear regression model, whereby the current score is predicted based on the previous score with the mean of all worker scores as control. As the scores were standardised, we do not control for overall annotator bias towards positive or negative values.

As seen in Tab. 4.1, we see a small but significant positive autocorrelation for "good" and "moderate" workers, which increases if we remove the mean scores as a control variable. The autocorrelation is much higher for bad (rejected) workers. We used an ANOVA to determine if the difference in β_1 between the three groups is statistically significant. We found no significant difference in β_1 between the good and moderate workers, but these two groups have a significantly lower autocorrelation than the bad workers.

Since scores of individual workers contain assimilation effects, it is highly likely that the mean itself is biased. Using this as the gold standard means that we are explaining away some of the positive autocorrelation by this control variable. Thus, if we used an unbiased

	Good	**Moderate**	**Bad**
N items	48216	24696	17738
β_1 (autocorrelation)	0.030^{***}	0.038^{***}	0.193^{***}
β_2 (gold)	0.741^{***}	0.661^{***}	0.256^{***}
β_1' (autocorrelation)	0.066^{***}	0.053^{***}	0.196^{***}

Table 4.1 MT_{adeq} dataset: Coefficients of the linear model showing sequence bias of good, moderate and bad workers. β_1 is the autocorrelation coefficient using the mean of good workers as a control, and β_1' is the autocorrelation coefficient of a model with no controls. Stars denote statistical significance: * = 0.05, ** = 0.01, and *** = 0.001.

estimate of the true quality instead, the value of the autocorrelation coefficient would probably be somewhere in between the two values when we include and exclude the mean as a control.

In the rest of our experiments, we include the mean as a control variable.

Learning effect: In this dataset, each annotator scores a sequence of a 100 translations. As workers score more translations, do they become more or less prone to sequential bias? If the task is new and unfamiliar, there might be a learning effect as annotators get more proficient at rating translations, resulting in a decrease in the bias. On the other hand, if the task is too long or monotonous, annotators might get fatigued and this might result in a higher propensity to be influenced by cognitive biases. We divide the dataset into 3 equal sized buckets based on the position of the translation in the HIT, and test for autocorrelation in each subset.

As shown in Tab. 4.2, for good and moderate workers, the bias is stronger in the first group of sentences annotated, decreases in the second, and is much smaller in the last. This could be because workers are familiarising themselves with the task earlier on, and calibrating their scale. The autocorrelation of moderate workers is comparatively higher than that of the good workers in the first tertile, but this difference is negligible in the next two tertiles.

On the other hand, there is no such trend with bad quality scores. In fact, the bias is highest in the last tertile, possibly because the workers are not putting in sufficient effort to produce

Position	Good	Moderate	Bad
1st Tertile	0.043 534 ***	0.062 781***	0.179 367***
2nd Tertile	0.032 168 8***	0.034 176***	0.172 99 ***
3rd Tertile	0.015 494 **	0.013 832	0.224 88 ***

Table 4.2 MT$_{adeq}$ dataset: Autocorrelation coefficient β_1 of worker scores for translations in the first, second or third tertile based on the position of the sentence of the HIT.

accurate scores, or are becoming more efficient at subverting the process such as by forming click patterns between similar positions on the screen.

Worst case scenario Next we assess the potential impact of sequence effects in the worst case situation.

When collecting annotations for evaluating MT systems, translations are sampled randomly across all MT systems included in the evaluation. If the translations are not randomised across MT systems, is there a possibility for the rankings to be affected due to sequence effects? We divide the dataset into 3 equal sized buckets based on the "gold" score of the previous sentence, which we discretise into low, middle and high based on equal-frequency binning. As shown in Tab. 4.3, when each translation is annotated by a single worker, we can see that the sentences in the "low" partition and the "high" partition have a difference of 0.18, which is highly significant;[5] moreover, this difference is likely to be sufficiently large to alter the rankings of systems in an evaluation. The bias decreases when we increase the number of annotations per translation and use the average score, but remains significant because all workers scored the translations in the same order. This shows that the mean is also affected by sequence bias.

This means that if were to order a HIT such that a specific system's output is seen consistently immediately after a bad (or good) output, then this could deflate (or inflate) the aggregate score of the system, and potentially change the system's rank.

[5] $p < 0.001$ using Welch's two-sample t-test

N	**All**	**Low**	**Middle**
1	0.013 838 974 23	−0.092 821 641 04	0.050
5	0.004 058 806 497	−0.053 533 566 18	−0.017
10	−0.000 782 429 078	−0.048 042 367 97	−0.039
15	−0.000 895 762 842 9	−0.049 623 619 21	−0.023

Table 4.3 Impact of sequence effects on MT system scores in the worst case scenario: Translations following a low quality translation receive a lower score than those following a good translation: "All" is the mean score of all sentences in the dataset, where each sentence score is calculated as the average of *N* (standardised) worker scores. "Low", "Middle", and "High" are mean scores of sentences where the previous sentence annotated is of low, medium and high quality, resp. "H − L" is the difference between the average high and low scores.

4.4.2 Sequence Effects in NLP

In this section, we explore the potential for sequence effects in different kinds of annotations. First, we look at the binary annotation tasks: recognising textual entailment ("RTE") and event temporal ordering ("TEMPORAL"). Next, we look at the affective text analysis dataset which, like direct assessment, is carried out on a scale of 0-100, but instead of a single quality score, contains annotations for different emotions for each sentence. All three datasets are taken from the crowdsourcing study of Snow et al. (2008), and include crowdsourced annotations, as well as expert annotations which we use as the true label in our experiments. As with the MT_{adeq} dataset, we look for autocorrelation between workers' scores, and explore how this is influenced by the quality of workers' annotations.

Recognising Textual Entailment (RTE) and Event Temporal Ordering

In the RTE task, annotators are presented with two sentences and are asked to judge whether the second text can be inferred from the first. With the TEMPORAL dataset, they are shown two sentences describing events and asked to indicate which of the two events occurred first. The RTE dataset contains 800 unique instances with 20 instances per MTurk HIT and the TEMPORAL dataset contains 462 unique instances with 10 instances per HIT. For both tasks,

Task	All	Non-extreme	Good
RTE	-0.10221	-0.19198^{**}	-0.16943^{*}
TEMPORAL	0.1977	$-0.511\ ^{***}$	-0.5671^{***}

Table 4.4 Autocorrelation coefficient β_1 for RTE and TEMPORAL data. Stars denote statistical significance: * = 0.05, ** = 0.01, and *** = 0.001.

each HIT was annotated by 10 workers, all of whom see the instances in the same sequential order.

Results As this is a discrete task, we use logistic regression on worker labels against labels on the previous instance in the current HIT, with the expert judgements as a control variable.[6] We also add an additional control, namely the percentage of True labels assigned by the worker overall, which accounts for the overall annotator bias. Without this control, the value of the previous label can indicate the tendency of a given worker to give positive labels, leading to a positive autocorrelation.[7] To calculate the percent of positive labels, we use all labels by the worker in the current HIT excluding the current label to avoid giving the model any information about the current instance.

As shown in Tab. 4.4, over all workers ("All"), we find a small negative autocorrelation for both the RTE and TEMPORAL tasks. One possibility is that this is biased by opportunistic workers who assign the same label to all instances in the HIT, for which we can't expect any first order sequence effects. Following Chen et al. (2016), we calculate autocorrelation for the subgroup of "Non-extreme" workers whose rate of positive decisions lies between 0.2 and 0.8. The autocorrelation of this subgroup is higher, and is statistically significant. Finally, we also show results for workers with at least 60% accuracy when compared to expert annotations ("Good"). We observe a significantly negative value of β_1, showing that even good workers are prone to assimilation effects.

[6] We see similar results when using linear regression instead of logistic regression.

[7] This is not relevant to MT_{adeq} dataset as all worker's scores are standardised.

	All	Good	Bad
β_1 (autocorrelation)	-0.02797^{*}	-0.00615	-0.03876^{*}
β_2 (gold)	0.44533^{***}	0.65632^{***}	0.23330^{***}

Table 4.5 Autocorrelation coefficient β_1 for the AFFECTIVE dataset. Stars denote statistical significance: * = 0.05, ** = 0.01, and *** = 0.001.

Affective Text Analysis

In the affective text analysis task ("AFFECTIVE"), annotators are asked to rate news headlines for anger, disgust, fear, joy, sadness and surprise on a continuous scale of 0–100. In addition to these six emotions, they rate sentences for (emotive) valence, i.e., how strongly negative or positive they are (-100 to $+100$). The original dataset consisted of 1000 headlines, each annotated by 6 experts (Strapparava and Mihalcea, 2007). In the crowdsourced dataset, 100 headlines were randomly sampled from the original dataset and divided into 10 HITs, with 10 workers annotating each HIT (Snow et al., 2008). We test for autocorrelation of scores of each aspect individually, controlling for the expert scores and worker correlation with the mean of expert scores. We also look separately at datasets of good and bad workers, based on whether the correlation with the expert annotations is greater than 0.5.

Results For individual emotions, we do not observe any significant autocorrelation ($p \geq 0.05$). As there are only 1000 annotations per emotion, we also look at results when pooling data for all aspects. Though we find a statistically significant negative autocorrelation for scores of the full dataset (Tab. 4.5), this disappears when we filter out bad workers. Given the difficulty of this very subjective task, it is likely that many of workers considered 'bad' might have simply found this task too difficult, and thus become more prone to sequence effects. On the other hand, good workers were providing six other scores between rating the same emotion of two consecutive sentences. This interruption might decrease the chance of being influenced by the annotation for the previous sentence.

4.5 Discussion and Conclusions

We have shown significant sequence effects across several independent crowdsourced datasets: a negative autocorrelation in the RTE and TEMPORAL datasets, and a positive autocorrelation in the MT_{adeq} dataset. The negative autocorrelation can be attributed either to sequential contrast effects or the gambler's fallacy, and the positive autocorrelation to assimilation effects.

These effects were not significant for the AFFECTIVE dataset, perhaps due to the nature of the annotation task, whereby annotations of one emotion are separated by six other annotations, thus limiting the potential for sequencing effects. This is consistent with previous research: Chen et al. (2016) found that the autocorrelation is smaller between judicial cases that are separated by a longer time period.

These effects were typically more significant in the subset of inaccurate annotators, and people who are annotating data seemingly at random are very prone to cognitive biases. However, there is evidence of sequence effects even in crowdworkers who provide a reasonably good quality of annotations. Since all workers see the instances in the same order, this affects any other inferences made from the data, including aggregated assessment.

With MT evaluation using direct assessment, the judgements are subjective, and when people are asked to rate them on a continuous scale, they need time to calibrate their scale. We show that the sequential bias decreases for better workers as they annotate more sentences in the HIT, indicating a learning effect. This is in line with experiments on rating physical attractiveness where sequence effects attenuated as their experimental subjects gained more experience (Bhargava and Fisman, 2014).

When collecting DA annotations for evaluating MT systems, the original proposal for DA ensures that the translations are selected randomly from the MT systems evaluated (Graham et al., 2017). The order of translations in the HIT is randomized, so any bias in individual scores of a system's translations would be cancelled when averaging scores of all its translations, assuming a sufficiently large sample. Thus we do not expect sequential bias to have a marked effect on system rankings or other macro-level conclusions on the basis of this data. However, when collecting multiple judgements for accurate sentence-level scores, all workers see the scores in the same order. Thus, the scores of individual translations remain biased, which augurs poorly for the use of these annotations at the sentence level, such as when used in error analysis, or for training and evaluating automatic metrics.

In cases where the instances are independent, sequence problems can be easily addressed by adequate randomisation — providing each individual worker with a separate dataset that has been randomised such that no two workers see the same ordered data. In this way sequence effects can be considered as independent noise sources, rather than a systematic bias, and consequently the aggregate results over several workers will remain unbiased.

In recent years, as the quality of MT increases, it is now recommended to evaluate MT at the document level. For WMT, HITs are now structured such that annotators see the translation of a given document by an MT system in order, before moving on to another document translated by another MT system (see Sec. 2.1.1 for details). In this case, we want annotators to evaluate each sentence of the document in context, thus it would be desirable to see assimilation effects within the scores of a given document. But we still need each document to be evaluated independently, without any influence of sequential contrast or assimilation effects. If the quality of translated documents vary across HITs, it is possible that workers will calibrate differently. For example, the presence of a document translated by a very low quality MT system in the HIT might result in the remaining documents receiving comparatively higher scores. Thus Direct Assessment evaluations might run in to the same issues as relative ranking, where an MT system's final rank is influenced by luck in being compared more often with low or high quality systems Bojar et al. (2011). Knowles (2021) demonstrated that system rankings at WMT 2020 were indeed affected by the selection of the MT systems in the HIT; systems that were paired with low quality systems had an unfair advantage over systems that are paired more often with higher quality systems.

The analysis in this chapter has shown that cognitive biases can distort evaluation and annotation exercises; in particular, we demonstrated sequence effects in annotations by crowd-workers. We limited our scope to binary and continuous responses, however it is likely that sequence effects are prevalent for multinomial and structured outputs, e.g., in discourse and parsing, where priming is known to have a significant effect (Reitter et al., 2006).

Another important question for future work is whether sequence bias is detectable in expert annotators, not just crowd workers. We showed that autocorrelation decreases as crowd workers annotate more MT outputs. Chen et al. (2016) report that more experienced judges are less susceptible to the Gambler's fallacy, and Reilly et al. (1998) show that assimilation effects were smaller if their experimental subjects were familiar with the task. While we can hope that data annotated by experts is less biased, it is still important to empirically determine that this is so.

So far, this thesis explored ways to improve robustness of human evaluation of MT. In the next chapter, we move to automatic evaluation, proposing new metrics that compare MT with reference translation. In Chapter 6, where we explore evaluation of automatic metrics, we touch on some of the themes of this chapter: extremely high or low scoring MT systems can not just bias human annotations, but also have a major negative impact on the results of automatic metric evaluation.

Chapter 5

Automatic Metrics

This chapter builds on the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* pages 2799–2808, Florence, Italy, July 2019.

5.1 Introduction

The previous two chapters explored two different aspects of improving human evaluation of MT. While human evaluation is considered more reliable than automatic methods, it is often impractical to obtain human judgements due to the cost and time required.

Automatic metrics are an indispensable part of machine translation ("MT") evaluation, serving as a proxy for human evaluation. They provide immediate feedback during MT system development to validate ideas, tune parameters of systems, and perform neural model architecture selection. Thus, the reliability of metrics is critical to progress in MT research.

Automatic MT metrics attempt to automatically predict the quality of a translation by comparing it to a reference translation of the same source sentence. BLEU (Papineni et al., 2002), which measures the precision of *n*-grams between the Machine Translation output and a human reference translation, has been the chosen measure of evaluating research hypotheses since it was introduced. BLEU has many flaws which have been extensively studied (Callison-Burch et al., 2006), and many new metrics have been developed that address these flaws. Sec. 2.2.2 presents a detailed review of of approaches to designing metrics, which we briefly summarise here:

Character-level variants such as BEER, CHRF and CHARACTER overcome the problem of harshly penalising morphological variants (Stanojević and Sima'an, 2014; Popović, 2015; Wang et al., 2016). In order to allow for variation in word choice and sentence structure, other metrics use information from linguistic tools such as POS taggers, lemmatizers, synonym dictionaries, dependency and constituency parsers, and semantic role analysers (Banerjee and Lavie, 2005; Snover et al., 2006; Liu et al., 2010; Giménez and Màrquez, 2007b; Castillo and Estrella, 2012; Guzmán et al., 2014). More recently, metrics have adopted word embeddings to capture the semantics of individual words (Lo, 2017).

However, classic word embeddings are independent of word context, and context is captured instead using hand-crafted features or heuristics. This chapter aims to improve over existing automatic MT evaluation methods, through developing a series of new metrics based on contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), a technique which captures rich and portable representations of words in context and have been shown to provide important signal to many other NLP tasks.

We begin with an introduction to contextual word embeddings. We propose simple pre-trained metrics that use off-the-shelf contextual embeddings to approximate the precision, recall and F-score when comparing an automatic translation with a reference. We next develop a series of trained models, which aim to learn sentence representations of the translation and the reference that take into account similarities to words in the other sentence. We also apply these models in a reference-free setting, where we use multilingual word embeddings to compare the

MT system translation directly with the source sentence.[1] We evaluate these metrics on the WMT 2017 metrics shared task dataset (Bojar et al., 2017a), which was the most recent publicly available dataset when we completed this work. We then analyse the two different strategies for the datasets to train these models – using a small set of reliable, multiply-annotated dataset compared to a larger, noisier singly-annotated dataset and find that the latter strategy is more efficient. Finally, we present a qualitative analysis of our metrics compared to BLEU.

We completed this work in early 2019, and our metrics use BERT (Devlin et al., 2019) embeddings which were released in late 2018. Since then, there has been an explosion of research in understanding BERT, as well as the new models that improve on BERT, and we talk of the implications on automatic metric evaluation at the end of the chapter in Sec. 5.6.

5.2 Background: Contextual Word Embeddings

Word embeddings are real-valued vector representations of words in a high-dimensional vector space. These embeddings are learned from large monolingual corpora, and capture semantic and syntactic relationships between words (Liu et al., 2019a; Rogers et al., 2020; Lin et al., 2019; Tenney et al., 2019). By mapping words to embeddings, we can capture a soft-similarity between two words which is helpful when computing similarity in meaning between the MT output and the reference.

Classic word embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) map each word to a single vector irrespective of context. More recent methods for creating contextual word embeddings such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) compute representations for a word that depend on the sentence-context of the word. These embeddings can be very useful with MT evaluation: when comparing embeddings of the same word in two sentences, their similarity will be higher if the context is similar.

[1] The code of our metrics is available online at https://github.com/nitikam/mteval-in-context

These are obtained from deep neural models that are trained on self-supervised tasks on large text corpora. ELMo uses a two-layer biLSTM that is pre-trained on a bi-directional language modelling task. ELMo embeddings are a concatenation of the hidden states of the forward and backward LSTM. BERT uses a bi-directional transformer decoder, which captures context using self-attention on all words of the sentence. BERT is trained using a masked language model and the next sentence prediction task.

Both ELMo and BERT were developed on English data. Additionally, Multilingual BERT was trained on a concatenation of data from 102 languages. Multilingual BERT has no explicit cross-lingual objective and was not trained on parallel data; these are not true cross-lingual vectors where words from all languages are in a shared embedding space. Nevertheless, multilingual BERT has been successfully applied to tasks that require cross-lingual semantic understanding such as natural language inference, part of speech tagging, named entity recognition, paraphrase detection and question answering (Wu and Dredze, 2019; Hu et al., 2020; Lewis et al., 2020; Lauscher et al., 2020). This is true even with languages that do not have lexical overlap (Pires et al., 2019).

5.3 Metrics

We wish to predict the score of a translation t of length l_t against a human reference r of length l_r. For all models, we use fixed pre-trained contextualised word embeddings $\mathbf{e}_k$ to represent each token in the MT output and reference translation, in the form of matrices $\mathbf{W}_t$ and $\mathbf{W}_r$.

5.3.1 Pre-trained Metrics

These metrics compute the soft-similarity between the embeddings of the tokens in the MT output and the reference, and are inspired by metrics proposed by Corley and Mihalcea (2005) that use the Wu & Palmer method on WordNet (Fellbaum, 1998) to compute semantic similarity between two words.

We first use cosine similarity to measure the pairwise embedding similarity between the tokens in t and r. We approximate the precision of a token in t with its maximum similarity with any token in r. Our first metric, BERTP is the average precision of all tokens in t:

$$\text{precision}_i = \max_{j=1}^{l_r} \text{cosine}(\mathbf{e}_i, \mathbf{e}_j) \tag{5.1}$$

$$\text{BERT}\textsc{p} = \sum_{i=1}^{l_t} \frac{\text{precision}_j}{l_t} \tag{5.2}$$

Similarly, BERTR is the average recall of the reference, where the approximate recall of a token in r with its maximum similarity with any token in t.

$$\text{recall}_j = \max_{i=1}^{l_t} \text{cosine}(\mathbf{e}_i, \mathbf{e}_j) \tag{5.3}$$

$$\text{BERT}\textsc{r} = \sum_{j=1}^{l_r} \frac{\text{recall}_j}{l_r} \tag{5.4}$$

Finally, we compute the F1-score as the harmonic mean of the precision and recall.

$$\text{BERT}\textsc{f} = 2\frac{\text{BERT}\textsc{p} \cdot \text{BERT}\textsc{r}}{\text{BERT}\textsc{p} + \text{BERT}\textsc{r}} \tag{5.5}$$

These metrics use a greedy matching between the tokens of the MT output and the reference. We also experimented with computing the similarity over the optimal one-to-one alignment between the tokens using the Hungarian algorithm, but found this decreased performance on the development set. This is in line with the experiments of Rus and Lintean (2012).

5.3.2 Supervised Metrics

In theory, BERT uses context from the entire sentence to compute token representations. However, in the end, these metrics compute the average similarity (precision or recall) over all tokens in the sentence, and computing the average can hide small but critical errors. So we look at learning similarity between sentences representations. The metric RUSE (Shimanaka et al., 2018) uses pre-trained sentence embeddings to independently compute representations of the MT and the reference, then predicts the score using a neural regressor that is trained on MT human evaluation data; can we improve on this by computing sentence representations that are aware of the pairwise similarities between the words of the two sentence? We first describe a BiLSTM baseline that replicates RUSE, except that it learns sentence representations from scratch on MT human evaluation data, then explore models that use attention to compute conditional representations of the two sentences.

Trained BiLSTM Our first model learns independent representations of the MT output and the reference translation, then predicts the quality of the MT based on the interactions between the two sentence representations.

We first encode the embeddings of the translation and reference with a bidirectional LSTM, and concatenate the max-pooled and average-pooled hidden states of the BiLSTM to generate $\mathbf{v_t}$ and $\mathbf{v_r}$, respectively:

$$\mathbf{v}_{s,max} = \max_{k=1}^{l_s} \mathbf{h}_{s,k}, \quad \mathbf{v}_{s,avg} = \sum_{k=1}^{l_s} \frac{\mathbf{h}_{s,k}}{l_s} \tag{5.6}$$

$$\mathbf{v}_s = [\mathbf{v}_{s,max}; \mathbf{v}_{s,avg}] \tag{5.7}$$

To get the predicted score, we run a feed-forward network over the concatenation of the sentence representations of t and r, and their element-wise product and difference (a useful heuristic first proposed by Mou et al. (2016)). We train the model by minimizing mean squared error with respect to human scores.

$$\mathbf{m} = [\mathbf{v}_t; \mathbf{v}_r; \mathbf{v}_t \odot \mathbf{v}_r; \mathbf{v}_t - \mathbf{v}_r] \tag{5.8}$$

$$y = \mathbf{w}^\mathsf{T}\text{ReLU}(\mathbf{W}^\mathsf{T}\mathbf{m} + b) + b' \tag{5.9}$$

Trained BiLSTM + attention When sentence embeddings are learned independently, it is difficult to encode all the relevant information that is required for comparing the two sentences. It can be very useful for the model to also take into account pairwise similarity between the tokens of the two sentences. Accordingly, our next model uses the attention mechanism to compute sentence representations that capture similarities with the tokens in the other sentence.

The attention mechanism was first proposed for neural MT models to solve the bottleneck of representing the meaning of the entire sentence in a single vector (Bahdanau et al., 2015), and allows the decoder to focus on the hidden representations of relevant parts of the input. The attention mechanism has been showed to be useful a multitude of other tasks, including sentence-pair tasks such as natural language inference. Models that include attention to model pairwise word interactions when learning sentence representations have a higher accuracy than systems that process the two sentences independently (Rocktäschel et al., 2016; Chen et al., 2017; Wang et al., 2017).

To obtain a sentence representation of the translation which is conditioned on the reference, we compute the attention-weighted representation of each word in t. The attention weights are obtained by running a softmax over the dot product similarity between the hidden state of the translation and reference BiLSTM. Similarly, we compute the MT-aware representation of the reference:

$$a_{i,j} = \mathbf{h_r}_i^\mathsf{T}\mathbf{h_t}_j \tag{5.10}$$

$$\tilde{\mathbf{h}}_r = \sum_{j=1}^{l_t} \frac{\exp(a_{i,j})}{\Sigma_i \exp(a_{i,j})} \cdot \mathbf{h}_t \tag{5.11}$$

$$\tilde{\mathbf{h}}_t = \sum_{i=1}^{l_r} \frac{\exp(a_{i,j})}{\Sigma_j \exp(a_{i,j})} \cdot \mathbf{h}_r \tag{5.12}$$

We then use $\tilde{\mathbf{h}}_t$ and $\tilde{\mathbf{h}}_r$ as our sentence representations in Eq. (5.6) – (5.9) to compute the final scores.

Enhanced Sequential Inference Model (ESIM): We also directly adapt ESIM (Chen et al., 2017), a high-performing model on the natural language inference (NLI) task (Bowman et al., 2015), to the MT evaluation setting. The NLI task, which requires a model to predict whether the premise entails the hypothesis, is closely related to the MT evaluation task (Padó et al., 2009). A good translation entails the reference and vice-versa: missing content in the MT output would break the entailment of the reference, and hallucinated content in the MT would break entailment of the MT output. Further, any inaccuracy in the MT output would result in a contradiction. An additional complexity that is not present in the NLI task is that MT output is not always fluent.

We use the human reference translation r and the MT output t as inputs to the ESIM model. The model first encodes r and t with a BiLSTM, then computes the attention-weighted representations of each with respect to the other (Eq. (5.10) – (5.12)). This model next "enhances" the representations of the translation (and reference) by capturing the interactions between $\mathbf{h}_t$ and $\tilde{\mathbf{h}}_t$ (and $\mathbf{h}_r$ and $\tilde{\mathbf{h}}_r$):

$$\mathbf{m}_r = [\mathbf{h}_r; \tilde{\mathbf{h}}_r; \mathbf{h}_r \odot \tilde{\mathbf{h}}_r; \mathbf{h}_r - \tilde{\mathbf{h}}_r] \tag{5.13}$$

$$\mathbf{m}_t = [\mathbf{h}_t; \tilde{\mathbf{h}}_t; \mathbf{h}_t \odot \tilde{\mathbf{h}}_t; \mathbf{h}_t - \tilde{\mathbf{h}}_t] \tag{5.14}$$

These representations are passed through a feed-forward layer to project them back to the model dimensionality, and then through a second BiLSTM to compose local sequential information. The final representation of each pair of reference and translation sentences is the concatenation of the average-pooled and max-pooled hidden states of this BiLSTM. To compute the predicted score, we apply a feed-forward regressor over the concatenation of the two sentence representations.

$$\mathbf{p} = [\mathbf{v}_{r,avg}; \mathbf{v}_{r,max}; \mathbf{v}_{t,avg}; \mathbf{v}_{t,max}] \tag{5.15}$$

$$y = \mathbf{w}^{\intercal}\mathrm{ReLU}(\mathbf{W}^{\intercal}\mathbf{p} + b) + b' \tag{5.16}$$

5.3.3 Reference-free Metrics

The above metrics compare the MT output with the reference translation in a monolingual setting. This requires a one-time investment in obtaining a high quality reference translation. We also explore evaluating MT outputs in a reference-free setting, where we compare them directly with the source sentence. We compute our metric scores in the exact same way as the reference-based metrics, just replacing the reference translation with the source sentence, and using multilingual embeddings to encode the source sentence and reference translation.

5.4 Experimental Setup

5.4.1 Data

To train and evaluate our models, we use human evaluation data from the Conference on Machine Translation (WMT) (Bojar et al., 2016a, 2017b), which is based on the direct assessment ("DA") method (Graham et al., 2017). Here, MT system outputs are evaluated by humans in comparison to a human reference translation on a continuous scale (Graham et al., 2015, 2017). Each annotator assesses a set of 100 items, of which 30 are quality control items used to filter out low-quality annotations. Individual worker scores are first standardised, and then the final score of an MT system is computed as the average score across all translations in the test set. To obtain an accurate score for individual translations, the average score is calculated from scores of at least 15 "good" annotators that pass quality control. This data is then used to evaluate automatic metrics at the sentence level (Graham et al., 2015). See Sec. 2.1.4 for more details on DA.

Training sets: We train on the crowdsourced human evaluation data of news domain of WMT 2016, which includes the following language pairs:

- Czech, Finnish, German, Romanian, Russian and Turkish → English[2]
- English → Russian.

We use data collected in two settings:

1. TRAINS: This dataset consists of accurate multiple-annotated scores for 560 translations per language pair sampled randomly from the outputs of all MT systems participating in the WMT 2016 translation task. This data was collected for evaluating metrics at the sentence level.

2. TRAINL: This dataset consists of mostly singly-annotated[3] DA scores for around 125k translations from six source languages into English, and 12.5k translations from English-to-Russian. This data was collected to obtain human scores for MT systems participating in the WMT 2016 translation task; an average of 2,666 translations were evaluated per MT system across all the language pairs.

Development set For the validation set, we use the accurate sentence-level DA judgements collected for WMT 2015 data (Bojar et al., 2015) for four to-English language pairs (Czech,

[2] We used the Turkish → English dataset in Chapter 3 to demonstrate the efficacy of our probabilistic model to aggregate DA judgements.

[3] About 15% of the translations have a repeat annotation collected as part of quality-control.

German, Finnish and Russian), and English-to-Russian. The dataset consists of 500 translation-reference pairs that were randomly sampled from each language pair.

Test sets We evaluate on in-domain data from the WMT 2017 (Bojar et al., 2017a) news task in the following settings:

- Sentence-level DA judgements in 7 to-English and 2 from-English language pairs.
- Sentence-level DARR judgements in 5 from-English language pairs. Due to insufficient annotations collected, DA annotations collected were converted to preference judgements in the following manner: if at least two MT system translations of a source sentence were evaluated, and the average score for System A is reasonably greater than the average score of System B, then this is interpreted as a Relative Ranking judgement (denoted as DARR; relative ranking from DA scores) where Sys A is better than Sys B.
- System-level DA scores in 7 to-English and 7 from-English language pairs.

We also evaluate on out of domain, system-level data for five from-English language pairs from the WMT 2016 IT task. This data was collected using the relative ranking (RR) method where annotators rank the quality of 5 translations of the same source sentence (Sec. 2.1.2). These RR judgements were aggregated using the Bayesian probabilistic model Trueskill (Herbrich et al., 2007) to obtain scores for MT systems (Sakaguchi et al., 2014) (as described in Sec. 3.2).

5.4.2 Implementation Details

We implement our models using AllenNLP in PyTorch. We experimented with both ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) embeddings, and found that BERT consistently performs as well as, or better than ELMo, thus we report results using only BERT embeddings in this chapter.

For BERTP, BERTR and BERTF, we use the top layer embeddings of the tokens of the MT and Reference translations. We use the `bert_base_uncased` model for all to-English language pairs, the `bert_base_chinese` model for English-to-Chinese and the `bert_base_multilingual_cased` model for the remaining from-English language pairs.[4]

For the trained metrics, we learn a weighted average of all layers of BERT embeddings. On the to-English test sets, we use `bert_base_uncased` embeddings and train on the WMT 2016 to-English data. On all other test sets, we use the `bert_base_multilingual_cased` embeddings and train on the concatenation of the WMT 2016 English-to-Russian and all to-English data.

For the reference-free setting, we train a single model for all language pairs; we use `bert_base_multilingual_cased` embeddings and train on all available data from the WMT 2016 news dataset.

For all our trained neural metrics, we fix the dimension of the BiLSTM hidden state to 300 and set the Dropout rate to 0.5, based on recommendations of the original ESIM paper. We use the Adam optimizer with an initial learning rate of 0.0004 and batch size of 32, and use early stopping on the validation dataset.

Training the ESIM model on the full dataset takes around two hours on a single V100 GPU, and all metrics take less than two minutes to evaluate a standard WMT dataset of 3000 translations.

5.5 Results

We report the Pearson correlation of our proposed metrics against existing metrics on the WMT 2017 to-English news dataset in Tab. 5.1. MEANT_2.0 (Lo, 2017), which uses pre-trained

[4] At the time this work was done, there were no publicly available monolingual models for languages other than English and Chinese.

		cs–en	de–en	fi–en	lv–en	ru–en	tr–en	zh–en	**AVE.**
Baselines	BLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
	CHRF	0.514	0.531	0.671	0.525	0.599	0.607	0.591	0.577
	MEANT_2.0	0.578	0.565	0.687	0.586	0.607	0.596	0.639	0.608
	RUSE	0.614	0.637	0.756	0.705	0.680	0.704	0.677	0.682
PreTr	BERTP	0.641	0.667	0.807	0.695	0.701	0.711	0.658	0.697
	BERTR	0.655	0.650	0.777	0.671	0.680	0.702	0.687	0.689
	BERTF	0.659	0.671	0.805	0.695	0.702	0.718	0.686	0.705
TrainS	BiLSTM	0.517	0.556	0.735	0.672	0.606	0.619	0.565	0.610
	BiLSTM + attention	0.611	0.603	0.763	0.740	0.655	0.695	0.694	0.680
	ESIM	0.534	0.546	0.757	0.704	0.621	0.632	0.629	0.632
TrainL	BiLSTM	0.628	0.621	0.774	0.732	0.689	0.682	0.655	0.682
	BiLSTM + attention	0.704	0.710	0.818	0.777	0.744	0.753	0.737	0.749
	ESIM	0.692	0.706	0.829	0.764	0.726	0.776	0.732	0.746
Source	BERTP_SRC	0.161	0.390	0.121	0.072	0.314	0.155	0.457	0.238
	BERTR_SRC	0.134	0.357	0.034	0.013	0.255	0.082	0.445	0.188
	BERTF_SRC	0.148	0.376	0.077	0.041	0.287	0.118	0.455	0.215
	ESIM_SRC	0.549	0.536	0.714	0.661	0.604	0.646	0.304	0.573

Table 5.1 Pearson's r on the WMT 2017 sentence-level evaluation data. PreTr: Unsupervised metric that relies on pre-trained embeddings; TrainS: trained on accurate 3360 instances; TrainL: trained on noisy 125k instances; Source: reference-free metrics, where ESIM_SRC is trained on the TrainL dataset. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold (William's test; Graham and Baldwin, 2014).

		en–cs	en–de	en–fi	en–lv	en–ru	en–tr	en–zh
		τ	τ	τ	τ	r	τ	r
Baselines	SENT-BLEU	0.274	0.269	0.446	0.259	0.468	0.377	0.642
	CHRF	0.388	**0.339**	0.549	**0.432**	0.605	**0.490**	0.608
	BEER	0.398	**0.336**	**0.557**	**0.420**	0.569	**0.490**	0.622
	MEANT_2.0-NOSRL	0.395	0.324	**0.565**	**0.425**	0.636	**0.482**	0.705
	MEANT_2.0	–	**0.350**	–	–	–	–	0.727
P	BERTP	0.395	**0.349**	**0.554**	**0.407**	0.627	**0.530**	0.754
	BERTR	**0.429**	**0.388**	**0.580**	**0.435**	0.665	**0.530**	**0.798**
	BERTF	**0.422**	**0.373**	**0.580**	**0.424**	0.658	**0.547**	**0.796**
T	ESIM	0.387	**0.393**	0.542	0.392	**0.723**	**0.571**	0.725
Source	BERTP_SRC	0.219	0.151	0.395	0.078	0.426	0.206	0.473
	BERTR_SRC	0.249	0.189	0.413	0.127	0.426	0.255	0.449
	BERTF_SRC	0.238	0.171	0.407	0.104	0.431	0.247	0.466
	ESIM_SRC	0.272	0.326	0.418	0.239	0.594	0.506	0.575

Table 5.2 Pearson's r and Kendall's τ on the WMT 2017 from-English sentence-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our pre-trained metrics, followed by our supervised metric, and our reference-free metrics. Note than ESIM and ESIM_SRC are both trained on the TrainL dataset. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold (William's test (Graham and Baldwin, 2014) for Pearson's r and Bootstrap (Efron and Tibshirani, 1993) for Kendall's τ).

word2vec embeddings, is the best pre-trained metric. And RUSE (Shimanaka et al., 2018) is the best supervised metric. We also include SENT-BLEU and CHRF baselines.

Our pre-trained metrics, BERTP, BERTR and BERTF, surpass or equal the correlation of all metrics participating in the WMT 2017 metrics shared task. Of the three, BERTF has the highest correlation on average.

The architecture of our BiLSTM baseline is similar to RUSE, except that we learn the sentence representation instead of using pre-trained sentence embeddings. In the TRAINS setting (the sentence-level data, as with RUSE), the BiLSTM baseline does not perform well, indicating that it is difficult to learn sentence representations from such a small dataset. However adding attention makes it competitive with RUSE, proving the value of computing sentence representations that are aware of token-level similarities with the other sentence. The ESIM model — which also uses attention to compute conditional representations of both the MT and the reference, but has many more parameters than the previous model — improves on the BiLSTM model but underperforms compared to the BiLSTM model with attention.

The performance of all models improves substantially when these metrics are trained on the larger, singly-annotated training data (“TrainL”), i.e., using data from only those annotators who passed quality control. Clearly the additional input instances make up for the increased noise level in the prediction variable. The simple BiLSTM model performs as well as RUSE, and both the models with attention substantially outperform this benchmark.

Finally, the performance of the pre-trained metrics suffers in the reference-free setting, as Multilingual BERT does not provide a true cross-lingual signal. All three metrics have a much lower correlation than BLEU. However, when trained on top of Multilingual BERT, ESIM_SRC has a higher correlation than BLEU, and is even competitive with CHRF.

We now evaluate the pre-trained BERTP, BERTR and BERTF metrics and the ESIM model (trained on the large dataset) in the other settings. In the sentence-level tasks out-of-English (Tab. 5.4), BERTR and BERTF (based on BERT-Chinese) significantly outperform all

		cs–en	de–en	fi–en	lv–en	ru–en	tr–en	zh–en
		4	11	6	9	9	10	16
Baselines	BLEU	0.971	0.923	0.903	0.979	0.912	0.976	0.864
	chrF	0.939	0.968	0.938	0.968	0.952	0.944	0.859
	CharacTer	0.972	0.974	0.946	0.932	0.958	0.949	0.799
	BEER	0.972	0.960	0.955	0.978	0.936	0.972	0.902
	RUSE	0.990	0.968	0.977	0.962	0.953	0.991	0.974
PreTr	BERTp	0.980	0.939	**0.992**	**0.994**	0.920	0.985	0.902
	BERTr	**0.996**	**0.971**	0.948	0.980	**0.949**	**0.994**	**0.967**
	BERTf	0.989	**0.958**	**0.979**	**0.992**	0.935	**0.994**	**0.949**
T	ESIM	0.983	0.949	0.985	0.974	0.921	0.986	0.901
Source	BERTp_src	**0.442**	0.931	0.007	0.195	**0.918**	0.322	0.844
	BERTr_src	**0.031**	**0.965**	0.177	0.146	**0.875**	0.486	0.830
	BERTf_src	**0.283**	**0.965**	0.100	0.169	**0.921**	0.417	0.865
	ESIM_src	0.964	0.839	0.904	0.826	0.781	0.876	0.720

Table 5.3 Pearson's r on the WMT 2017 to-English system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our pre-trained metrics, followed by our supervised metric trained in the TrainL setting: noisy 125k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

metrics in the English-to-Chinese test set. These metrics (based on multilingual BERT) are highly competitive with other metrics in the remaining language pairs. BERTp has a slightly lower correlation than BERTr and BERTf. ESIM, which was trained on a concatenation of to-English and to-Russian data, outperforms all metrics on the English-to-Russian data. But the results are mixed on other language pairs. ESIM is trained only on to-English and to-Russian data, so its performance is still impressive. Further, five of these language pairs are evaluated on Kendall's Tau over pairs of translations on the same source sentence. Our training method using squared error as part of regression loss is better suited to Pearson's r — and performance might be increased through a different loss, such as hinge loss over pairwise preferences (Stanojević and Sima'an, 2014) which would better reflect Kendall's Tau.

		en–cs	en–de	en–fi	en–lv	en–ru	en–tr	en–zh
		14	16	12	17	9	8	11
Baselines	BLEU	0.956	0.804	0.920	0.866	0.898	0.924	0.981
	BEER	0.970	0.842	0.976	0.930	0.944	0.980	0.914
	CharacTer	0.981	0.938	0.972	0.897	0.939	0.975	0.933
	chrF	0.976	0.863	0.981	0.955	0.950	0.991	0.976
PreTr	BERTp	0.965	0.796	0.968	**0.955**	0.942	**0.987**	0.977
	BERTr	**0.984**	**0.889**	**0.980**	**0.951**	**0.971**	**0.988**	**0.991**
	BERTf	**0.979**	0.845	0.975	**0.954**	0.959	**0.988**	**0.991**
T	ESIM	**0.967**	0.839	**0.962**	**0.942**	**0.966**	0.960	0.974
Source	BERTp_src	0.685	0.485	0.933	0.417	0.735	0.144	0.867
	BERTr_src	0.766	0.592	0.921	0.547	0.366	0.500	0.704
	BERTf_src	0.731	0.537	0.929	0.495	0.579	0.350	0.819
	ESIM_src	0.846	0.779	0.888	0.834	0.869	0.868	0.867

Table 5.4 Pearson's r on the WMT 2017 from-English system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our pre-trained metrics, followed by our trained metric, and our reference-free metrics. Note than ESIM and ESIM_src are both trained on the TrainL dataset. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

		en–cs	en–de	en–es	en–nl	en–pt
	num systems	5	10	4	4	4
Baselines	BLEU	0.750	0.621	0.976	0.596	0.997
	chrF	0.845	0.588	0.915	0.951	0.967
	BEER	0.744	0.621	0.931	0.983	0.989
	CharacTer	0.901	0.930	0.963	0.927	0.976
PreTr	BERTp	0.521	0.591	0.900	0.966	0.973
	BERTr	0.602	0.810	0.910	0.984	0.988
	BERTf	0.599	0.696	0.905	0.977	0.981
T	ESIM	0.745	0.838	0.984	0.828	0.997

Table 5.5 Pearson's r on the WMT 2016 IT domain system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our pre-trained metrics, followed by our supervised metric trained on the TrainL dataset. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

In the reference-free setting, the pre-trained metrics all have lower correlations than SENT-BLEU. ESIM_SRC has a higher correlation than SENT-BLEU on the English-to-German, English-to-Russian and English-to-Turkish data, is competitive on English-to-Czech data, and is outperformed on the remaining three language pairs. This could be attributed to the similarities between English and German, and the inclusion of English-to-Russian data in the training set.

On the system-level evaluation of the news domain, all our reference-based metrics are competitive with all other metrics in all language pairs both to- and out-of-English (see Tab. 5.3 and Tab. 5.4). The results of the reference-free metrics are surprising: the pre-trained metrics are not outperformed by any of the reference-based metrics on German-to-English and Russian-to-English data, despite their low sentence-level correlations. But in many other language pairs, these metrics have very low correlations; in the worst case, BERTP has a correlation of 0.007 over Finnish-to-English data. The correlation of ESIM_SRC is always above 0.7, but never outperforms BLEU, even in language pairs where it was substantially better than SENT-BLEU at the sentence-level.

In the IT domain, we have mixed results (Tab. 5.5). CHARACTER is the only metric that has a correlation above 0.9 on the English-to-German language pair. BERTR and ESIM have a higher correlation than the rest of the metrics. There is no consistent winner on the other language pairs, and we note that ESIM is competitive with the other metrics, despite the change in domain and the lack of training data in any of these language pairs.

5.5.1 Training Efficiency

When trained on the large, singly-annotated dataset (TrainL), our supervised metrics improved substantially over training on the small, multi-annotated dataset (TrainS) when evaluating our reference-based metrics on to-English translations. The total number of annotations in the TrainS setting are much smaller than in the TrainL setting. In this section, we present a direct comparison between the quality of ESIM trained using the two strategies while keeping the total number of annotations equal.

Fig. 5.1 shows how the average sentence-level correlation of ESIM improves as we increase the number of annotations collected, using two different strategies:

1. TrainS: collecting additional annotations on the same set of 3360 training instances, and
2. TrainL: collecting annotations on a new set of 3360 training instances.

The former strategy decreases noise in the training data, while the latter improves diversity of training instances.

We find that on the same number of training instances (3360), the model performs better on cleaner multiply-annotated data compared to singly-annotated data ($r = 0.57$ vs 0.64). In the TrainS setting, when we collect more than five annotations per translation, the gain in correlation is negligible. In the TrainL setting, increasing the size of the training set is clearly beneficial even after collecting annotations on more than 50000 instances.

Therefore, when we have a choice between collecting multiple annotations for the same instances vs collecting annotations for additional instances, the second strategy leads to more gains.

5.5.2 Qualitative Analysis

Automatic metrics may overestimate translation quality because of superficial similarities between the MT system translation and reference output. On the other hand, they might assign low scores to perfectly valid translations that deviate from the reference. We manually inspect translations in the validation set and, in this section, present examples that illustrate how BLEU, BERTR and ESIM score them.

Tab. 5.6 shows examples of translations that receive high human scores. In the first three examples, our proposed metrics correctly recognise synonyms and minor word re-orderings, unlike SENT-BLEU which relies on *n*-gram matching. In example 4, there are no exact matches

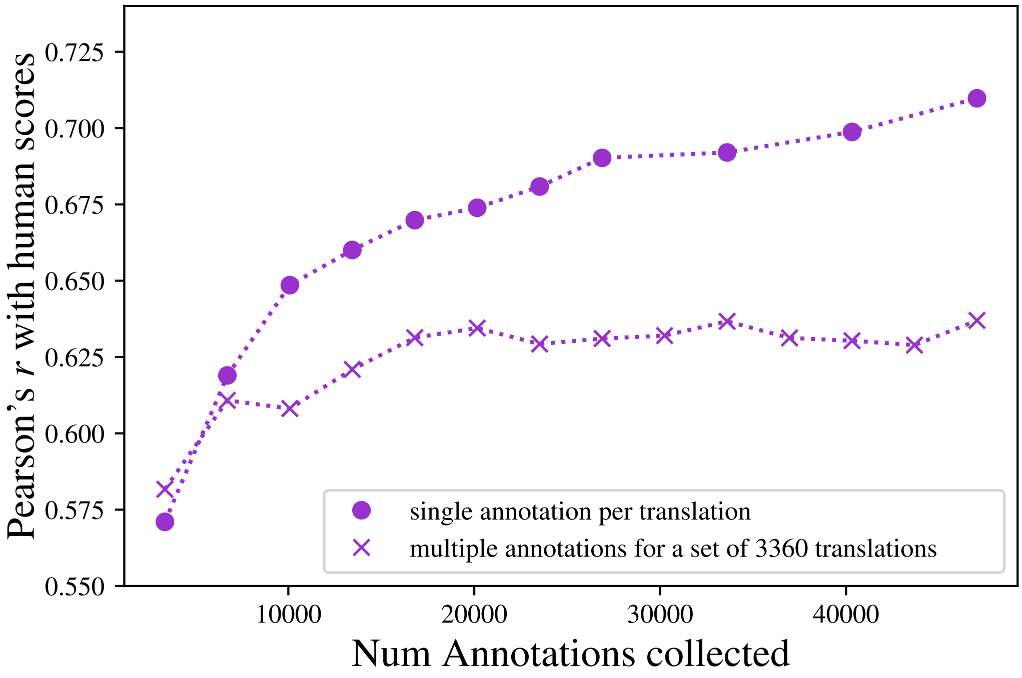


Fig. 5.1 Average Pearson's r for ESIM over the WMT 2017 to-English sentence-level dataset vs. the total number of annotations in the training set. We contrast two styles of collecting data: (1) the circles are trained on a single annotation per instance; and (2) the crosses are trained on the mean of N annotations per instance, as N goes from 1 to 14. The first strategy is more data-efficient.

between the words of the MT output and the reference, so it is unsurprising that SENT-BLEU assigns it a low score. While ESIM recognises the similarity between the two sentences (*working together* and *cooperation*), BERTR does not, possibly because it measures recall and penalises the MT output for being more concise. Finally, the last example shows that none of the metrics recognise a completely different way of expressing the same meaning.

Tab. 5.7 presents examples of low quality MT outputs. The first three examples show that SENT-BLEU gives high scores to translations with high partial overlap with the reference. In example 1, the first half of the MT output is identical to the reference, but the second half is non-sensical. In examples 2 and 3, key phrases of the MT output are wrong, resulting in an overall confusing translation. BERTR gives low scores to the first two sentences, but not the

		Translations with High Human scores	ESIM	BERTr	Sent-BLEU
1.	ref:	The negotiations have been scheduled to take place next Saturday, the Russian Minister of Energy, Alexander Nowak, said on Monday.	High	High	Low
	sys:	The negotiations are scheduled for coming Saturday, said the Russian energy minister Alexander Nowak on Monday.			
2.	ref:	Lesotho military says no coup planned; PM stays in South Africa			
	sys:	Lesotho-military member says that no coup is planned; Prime Minister remains in South Africa			
3.	ref:	In September 2011, Abbott's condition worsened again, and his consultant took his CT scans and X-rays to a panel of experts.			
	sys:	In September 2011 Abbotts state worsened again and his family doctor brought his CT-Scans and X-rays to an expert group.			
4.	ref:	The boardroom is now contemplating the possibility of working together.	High	Low	Low
	sys:	Now the boards are thinking about a possible cooperation.			
5.	ref:	He ended up spending a month off work.			
	sys:	In the end, he could not go to work for a month.	Low	Low	Low

Table 5.6 Examples of good translations in the WMT 2015 sentence-level DA dataset and whether ESIM, BERTr and Sent-BLEU correctly give them high scores

third which is a very long sentence and a few words with low similarity are lost when computing the average similarity. Note that ESIM correctly recognises all three as low quality translations. However, in some cases, ESIM can be too permissive of bad translations which contain closely related words (example 4). Finally, we show examples where humans recognize that seemingly minor differences are actually unacceptable. In example 5, a single word substitution (replacing *shrapnel* with *garnet*) changes the meaning of the entire sentence. Example 6 contains another single word substitution (*raced* instead of *hit*) and the MT output is missing a key verb (*blew*

		Translations with Low Human scores	ESIM	BERTr	Sent-BLEU
1.	ref:	The military plays an important role in Pakistan and has taken power by force several times in the past.	Low	Low	High
	sys:	The military plays an important role in Pakistan and has already more frequently geputscht.			
2.	ref:	For the benefit of the school, Richter nurtured a good relationship with the then Mayor, Ludwig Götz (CSU).			
	sys:	For the good of the school of judges as rector of a good relationship with the former mayor Ludwig Götz (CSU)			
3.	ref:	Behind much of the pro-democracy campaign in Hong Kong is the Occupy Central With Love and Peace movement, whose organizers have threatened to shut down the financial district if Beijing does not grant authentic universal suffrage.	Low	High	High
	sys:	Behind the pro-democracy campaign in Hong Kong is the movement Occupy Central With Love and Peace, whose organizers have threatened the acupuncture, off, if Beijing allows no real universal suffrage.			
4.	ref:	Foreign goods trade had slowed, too.	High	Low	Low
	sys:	Foreign trade also slowed the economy.			
5.	ref:	Some shrapnel pieces are still in my knee.	High	High	High
	sys:	Some garnet fragments are still in my knee.			
6.	ref:	Stewart hit the wall for the second time after his right front tire blew out on lap 172, ending his night.			
	sys:	Stewart raced for the second time against the wall after his right front tire on lap 172 and ended his evening.			

Table 5.7 Examples of bad quality translations in the WMT 2015 sentence-level DA dataset and whether ESIM, BERTr and Sent-BLEU correctly give them low scores

out). Unfortunately, all the metrics assign high scores to these MT outputs, highlighting a key challenge for automatic metrics.

5.6 Conclusion and Discussion

In this chapter, we showed that pre-trained contextual embeddings are very useful for automatic MT evaluation metrics. We proposed simple pre-trained metrics that essentially compute the precision, recall and F-score of the MT output and the reference translation. We found that these metrics are highly effective, and that we can further improve on them when we train on previous MT human evaluation data.

We also run our metrics in a reference-free setting, where we compare the MT output directly with source. The pre-trained metrics have moderate correlations for some language pairs, and non-significant correlations in others. ESIM_SRC outperforms SENT-BLEU at the sentence-level in all the to-English language pairs as well as three from-English language pairs. However, the success at the sentence level does not translate to the system level, where it never has a higher correlation than BLEU.

Finally, we evaluated our metrics on the WMT 2016 data in the IT domain, showing that ESIM, which is trained on news data, is robust to change in domain.

Since this work was done in 2019, there has been plenty of research on probing ELMo and, to a larger extent, BERT of what information is encoded in these embeddings. Our pre-trained metrics were independently proposed as BERTscore (Zhang et al., 2020) when our paper was under review. The authors found that the embeddings from intermediate layers of BERT outperformed the top layer, and so empirically determined the best layer, resulting in a small boost in correlation. They also tested the utility of IDF-weighting on tokens based on their frequency in the reference set; this doesn't always improve performance, so it is left as optional. Finally, they investigated replacing BERT with other pre-trained contextual embedding models that improve on BERT, for example, by enhancing the architecture or pre-training scheme. When their paper was published in 2020, the best model was RoBERTa$_{\text{large}}$ (Liu et al., 2019b) on the to-English language pairs, but at the time of writing, DeBERTa (He et al., 2020) has the highest correlation.

YiSi-1 was updated in 2019 to use the optimal layer of BERT embeddings (Lo, 2019). In 2020, YiSi-1 was improved further when evaluating translations where the target language is not English (Lo, 2020), by replacing multilingual BERT with monolingual BERT models evaluating translations into French and Finnish, and with XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2019) for other non-English languages.

More recently, automatic metric performance has been improved further by trained metrics that build on the success of RUSE and ESIM. BLEURT (Sellam et al., 2020) first pre-trains BERT on synthetic data generated by applying a variety of perturbations to using automatic metrics (BERTscore, SENT-BLEU and ROUGE) as a soft signal, before finetuning on MT human evaluation data. COMET (Rei et al., 2020a) is trained on top of XLM-RoBERTa$_{\text{large}}$ (Conneau et al., 2019) and incorporates the source sentence in addition to the MT output and reference translation.

There has also been tremendous progress in reference-free evaluation in recent years. COMET-QE (Rei et al., 2020b), like ESIM_SRC, is a neural model that encodes the source and MT output using cross-lingual embeddings(XLM-RoBERTa$_{\text{large}}$), and predicts translation quality from the learned representations. PRISM-src (Thompson and Post, 2020) adopts a completely new approach and uses a multilingual neural machine translation model to score the MT outputs given the source sentence.

In this chapter, we proposed new metrics and showed that they are highly correlated with human scores at both the sentence- and system- level. In the next chapter, we explore system-level metric evaluation in more detail.

Chapter 6

Meta Evaluation: Reevaluating Automatic Metric Evaluation

This chapter builds on the paper:

> Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020.

6.1 Introduction

In the previous chapter, we presented a family of new automatic metrics, and evaluated them on the system and sentence-level following standard practices established at the metrics shared task at the annual Conference of Machine Translation (WMT). At the system-level, metrics are evaluated based on the Pearson's correlation coefficient of metric scores against human scores. The stronger the correlation, the more we can trust that an improvement in metric scores will also lead to an improvement in human scores, thus justifying the use of automatic metrics as a proxy for human evaluation. In this chapter, we revisit system-level evaluation of automatic metrics, and the implications of using automatic metrics to evaluate MT systems.

Our intuitive understanding of the strength of the association between two variables given a particular value of Pearson correlation assumes that the data is from a bivariate normal distribution. However, the same value of Pearson's r can be the result of different underlying relationships between the two variables (in our case, the metric and human scores). Anscombe (1973) showed an example of four bivariate datasets that all have the same summary statistics (mean and standard deviation of X, mean and standard deviation of Y, and Pearson's correlation between X and Y), but a quick glance at a scatterplot of the data reveals that the datasets have very different distributions and tell different stories.[1]

The findings of the WMT 2019 metrics shared task (Ma et al., 2019) indicate the presence of two such patterns in their data:

1. **Outliers**: these are MT systems whose quality differs markedly from the rest of the systems in the dataset. When we are evaluating metrics on the entire set of systems, these outlier systems can have a disproportionate influence on the Pearson correlation between metric and human scores. This is problematic as the presence of outliers can result in high metric correlations even when the metric makes major errors with scoring the rest of the systems, resulting in a false confidence in metric reliability.

2. **Heteroskedasticity**: the variance of metric errors depends on the quality of the MT systems. The value of Pearson correlation can be misleading when there is heteroskedasticity in the data. If metrics make more errors when scoring high quality MT systems, then the Pearson correlation computed on the entire set of MT systems overestimates their reliability on high-quality systems. The findings of the WMT 2019 metrics task (Ma

[1] https://janhove.github.io/teaching/2016/11/21/what-correlations-look-like contains examples of sixteen different patterns that clearly illustrate the extent of this phenomenon

et al., 2019) show that metrics get less reliable when we evaluate them only on strong MT systems, indicating the presence of heteroskedasticity.

In this chapter, we first have a closer look at Pearson's correlation coefficient and review how it is affected by outliers and heteroscedasticity. Then, after describing the data and metrics we use for our analysis, we revisit the findings of the WMT 2019 metrics task to investigate the influence of these two patterns on metric evaluation: in Sec. 6.4, we show that outliers can have a disproportional influence on Pearson correlation, and in Sec. 6.5, we find no empirical evidence for heteroskedasticity in WMT 2019 metrics data.

As we mentioned earlier, a strong correlation between metric and human scores is taken to indicate that the metric can be reliably used as a cheaper alternative to human evaluation. If metric scores have a positive correlation with human scores, then many incremental improvements in metric scores on a particular test set will ultimately lead to better MT systems over time. Thus, if our goal is improvement in MT quality in the long-term, then a strong correlation between an automatic metric with human scores can validate the use of these metrics.

However, evaluating metrics using a correlation measure ignores the fact that important decisions are made using these metrics. Metrics are used to compare pairs (or small sets) of systems: deciding between different architectures of a model, deciding whether a given feature improves quality, deciding whether the new system beats the existing state of the art, or deciding whether an idea is worth publishing or if a paper should be accepted. Ideally, conclusions based on a difference in metric scores should agree with conclusions by humans, or if they differ, it should be due to a small margin. While Pearson's r implicitly takes these errors into account when computing the value of the correlation (larger errors lead to smaller correlations), we believe it is beneficial to explicitly test how well automatic metrics satisfy these criteria.

In the second part of the chapter, we investigate the utility of MT metrics when comparing two systems (Sec. 6.6). More concretely, we seek to quantify the extent of improvement required under an automatic metric such that the ranking reliably reflects human assessment. In doing so, we consider both type I and II errors when evaluating two systems A and B. Type I errors occur when a metric concludes that A is significantly better than B, when humans either judge them to be insignificant or judge B to be significantly better than A. Type II errors correspond to errors where metrics do not find a statistically significant difference between the two systems, but humans do. Both types of errors have the potential to stunt progress in the field: if we make decisions solely based on (flawed) automatic metric scores, this leads to spurious "improvements" that can not be replicated later. On the other hand, falsely rejecting ideas that result in true improvement might potentially shut down exploration of promising research directions.

6.2 Background: Pearson Correlation Coefficient

The Pearson correlation coefficient, denoted by r, is a measure of the strength of the linear relationship between two variables. Pearson's r is the covariance of the two variables (metric scores $\mathbf{m}$ and human scores $\mathbf{h}$) divided by the product of their standard deviations.

$$r = \frac{\text{Cov}(\mathbf{h}, \mathbf{m})}{\sigma_h \sigma_m}, \tag{6.1}$$

where

$\text{Cov}(\mathbf{h}, \mathbf{m})$ is the covariance of human and metric scores

σ_h and σ_m are the standard deviation of human and metric scores respectively.

When we insert the formulae for computing the sample covariance and sample standard deviations, the resulting formula for the sample Pearson correlation is:

$$r = \frac{\sum_{i=1}^{n}(h_i - \overline{h})(m_i - \overline{m})}{\sqrt{\sum_{i=1}^{n}(h_i - \overline{h})^2 (m_i - \overline{m})^2}}, \tag{6.2}$$

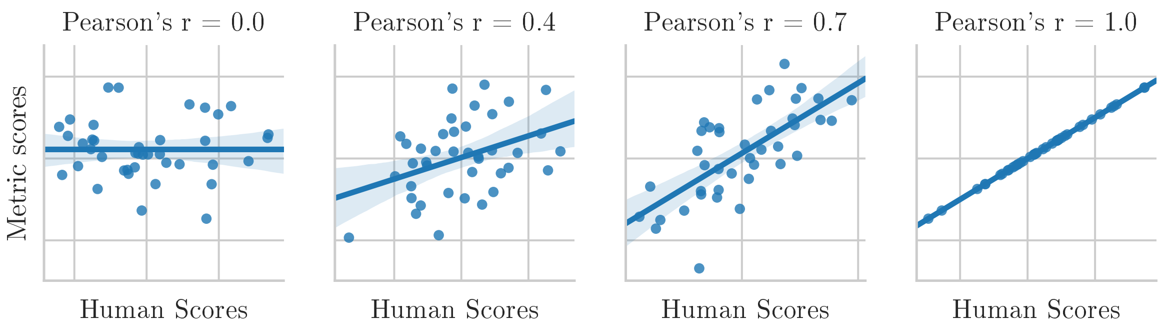


Fig. 6.1 Scatter plots of simulated human and metric scores as the strength of the relationship increases.

where

h_i and m_i are the human and metric score respectively of system i

$\overline{h}$ and $\overline{m}$ are the mean human and metric scores respectively.

Pearson's r ranges from -1 to 1. The sign indicates the direction of the relationship: a value greater than zero implies a positive relationship, less than zero is a negative relationship. The absolute value indicates the strength of the relationship: a value of ± 1 indicates a perfect relationship, and the relationship grows weaker as the value approaches zero. Fig. 6.1 shows scatter plots of examples of simulated human and metric scores as the correlation between the two increases from 0 to 1.

The value of the Pearson correlation can mask the presence of different patterns in the data (Fig. 6.2), and we next describe two such patterns: (a) the presence of outliers and (b) heteroskedasticity.

6.2.1 Influence of Outliers

An outlier is "an observation (or subset of observations) which appears to be inconsistent with the remainder of the dataset" (Barnett and Lewis, 1974). Pearson's correlation is sensitive to outliers in the sample, as it is calculated using the mean and standard deviation which are both

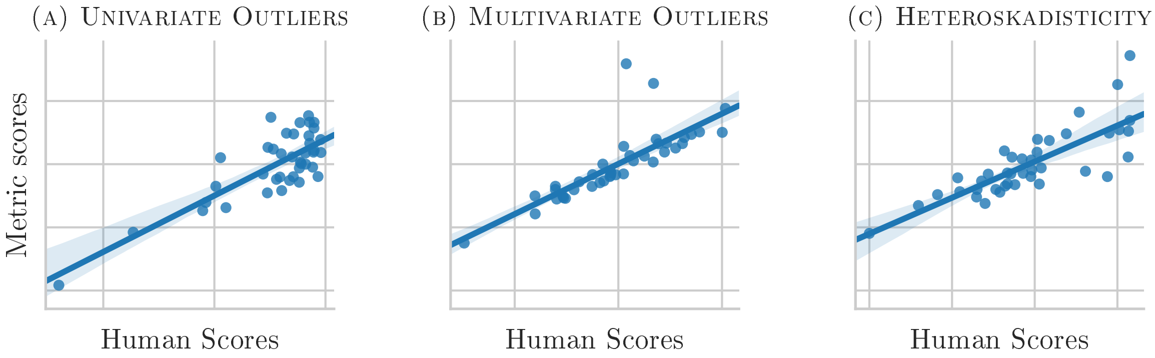


Fig. 6.2 Scatter plots of simulated human and metric scores with the same correlation ($r = 0.8$), but different patterns in the data.

highly influenced by outliers. Even a single outlier can have a drastic impact on the value of the correlation coefficient. In the extreme case, outliers can give an illusion of a strong correlation when there is none, or mask the presence of a true relationship.

The WMT metrics task datasets consist of MT systems that were submitted as a part of the annual WMT news translation task, and anonymous online services such as Google Translate and Microsoft Translator. The number of MT systems included is generally small, and can include both univariate outliers (Fig. 6.2 A), where the value of a single variable (human or metric scores) is exceptionally large or small compared to the values of that variable, and multivariate outliers (Fig. 6.2 B), where the combination of human and metric scores is unusual. These outliers can arise in metric evaluation data for two reasons:

- Univariate outliers caused by MT system quality: typically, the quality of most systems lies in a small range, but some systems can be either much better or much worse than the rest of the systems in the cohort. These systems are generally worse than the others, for example, because they are experimenting with different approaches or data conditions, or because they are submissions by students who are new to developing MT systems. In rare cases, the dataset might include exceptional MT systems that clearly surpass the quality of the other systems included in the evaluation. These outlier systems will have extremely low or high human scores, and we refer to them as low outliers and high outliers respectively. If metrics correctly score these MT systems (both high and low outliers), this results in a high value of Pearson's r, even if the metric makes major errors in scoring the rest of the systems.

- Multivariate outliers due to metric errors: these outliers arise in the metric evaluation datasets when metrics make major errors scoring MT systems. Automatic metrics can be biased towards systems that have a superficial similarity to the reference translation, resulting in low scores for high quality translations that are phrased differently. On the other hand, if an MT system was overfit to a metric, it might lead to high metric scores but low human scores. The correlation between metric and human scores increases if we remove these MT systems.

Spearman correlation coefficient and Kendall's Tau correlation are robust alternatives to Pearson correlation: they are based on ranks and pairwise relationships respectively between the two points, and as such, are not disproportionately influenced by outliers. However, they are not a good fit for metric evaluation as they do not take into account the differences in scores. In particular, they harshly penalise metrics that have a different ordering of systems that humans have judged to be of similar quality (§ 2.2).

Another robust solution is to recompute Pearson correlation after removing the outliers. Standard methods for computing robust correlations involve removing both univariate and multivariate outliers in the joint distribution of the two variables: the metric and human scores in our case. However, multivariate outliers include system pairs that indicate metric errors, and we believe they should not be removed because they provide important data about the errors and possible biases of the metric. Thus, we only look towards detecting univariate outliers caused by human scores.

Univariate Outlier Detection

A common method of detecting outliers is to simply standardise human scores, and remove systems with scores that lie 2.5 (or some other predefined cut-off) standard deviations away from the mean. However, standardising depends on the mean and standard deviation, which are themselves affected by outliers. A more robust alternative is to use the median instead of the mean as the measure of central tendency, and the Median Absolute Deviation (MAD) instead of the standard deviation as the estimator of scale.

For MT systems with human scores **h**, we use the following steps to detect outlier systems (Rousseeuw and Hubert, 2011; Leys et al., 2013):

1. Compute MAD, which is the median of all absolute deviations from the median of human scores

$$\text{MAD} = b \times \text{median}(|\mathbf{h} - \tilde{h}|),$$

   where $\tilde{h}$ is the median of human scores and b is the consistency constant that ensures that the value of MAD is consistent with the standard deviation of the underlying distribution. This value is computed as the inverse of the 0.75 quartile of the distribution, and is equal to 1.483 for the Gaussian distribution.

2. compute robust scores:

$$\mathbf{z} = (\mathbf{h} - \tilde{h})/\text{MAD}$$

   A value of z_i indicates that the score h_i is at a distance of $z_i * \text{MAD}$ from $\tilde{h}$

3. discard systems that are distant from the median in either direction: more precisely, discard system i if $|z_i|$ exceeds a cut-off (we use 2.5 as recommended by Leys et al. (2013))

6.2.2 Heteroskedasticity

Heteroskedasticity occurs when the variance of the error terms is not constant across all values of the independent variable, human scores in our case (Fig. 6.2 C). The presence of heteroskedasticity in metric scores implies that metric reliability is dependent on the quality of the MT systems. The findings of the WMT 2019 metrics task show that metric correlations break down when we restrict the set of MT systems to the best MT systems of the language pair. If metrics are less reliable when evaluating strong MT systems, then reporting the Pearson's correlation coefficient for the whole set of MT systems means that we are underestimating the reliability for low-quality MT systems, while over-estimating it for the best systems.

We can informally detect heteroskedasticity by inspecting the scatter plot of human and metric scores; the errors around the line of best fit will form a cone-like pattern, as their variance increases with an increase in human scores (Fig. 6.2 C).

We next describe the datasets and selected metrics, before exploring the influence of these two patterns in these datasets.

6.3 Data

In this chapter, we primarily use data from the WMT 2019 metrics task (Ma et al., 2019) for our analysis, and supplement this with analysis on the WMT 2017 (Bojar et al., 2017a) and WMT 2018 (Ma et al., 2018) datasets.

In WMT 2019, metrics were evaluated on the following 18 language pairs (7 language pairs translating to English, 8 translating out of English, and 3 that do not include English):

- English (en) ↔ Czech (cs), German (de), Finnish (fi), Gujarati (gu), Kazakh (kk), Lithuanian (lt), Russian (ru), and Chinese (zh)
- English (en) → Czech (cs), German → French, German → Czech, and French → German

6.3.1 Human scores: Direct Assessment (DA)

For the ground truth for evaluating metrics, we use direct assessment (DA) (Graham et al., 2017) scores collected as part of the human evaluation at WMT 2019 (Barrault et al., 2019). Annotators are asked to rate the adequacy of a set of translations compared to the corresponding source/reference sentence on a slider which maps to a continuous scale between 0 and 100. Low-quality annotations are filtered out based on quality control items included in the annotation task. Each annotator's scores are standardised to account for different scales. The score of an MT system is computed as the mean of the standardised score of all its translations. (See Sec. 2.1.4 for details on direct assessment.)

In WMT 2019, typically around 1500–2500 annotations were collected per system for language pairs where annotator availability was not a problem. To assess whether the difference in scores between two systems is not just chance, the Wilcoxon rank-sum test is used to test for statistical significance.

6.3.2 Metrics

Automatic metrics compute the quality of an MT output (or set of translations) by comparing it with a reference translation by a human translator. For the WMT 2019 metrics task, participants were also invited to submit metrics that rely on the source instead of the reference. In this chapter, we focus on the following metrics that were included in evaluation at the metrics task at WMT 2019:

Baseline metrics

- BLEU (Papineni et al., 2002) is the precision of n-grams of the MT output compared to the reference, weighted by a brevity penalty to punish overly short translations. BLEU has high variance across different hyper-parameters and pre-processing strategies, in response to which sacreBLEU (Post, 2018) was introduced to create a standard implementation for all researchers to use; we use this version in our analysis.

- TER (Snover et al., 2006) measures the number of edits (insertions, deletions, shifts and substitutions) required to transform the MT output to the reference.
- CHRF (Popović, 2015) uses character *n*-grams instead of word *n*-grams to compare the MT output with the reference. This helps with matching morphological variants of words.

Best metrics across language pairs

- YISI-1 (Lo, 2019) computes the semantic similarity of phrases in the MT output with the reference, using contextual word embeddings (BERT: Devlin et al. (2019)).
- ESIM (Chen et al., 2017) is a trained neural model that first computes sentence representations from BERT embeddings, then computes the similarity between the two strings. We adapted this model to MT evaluation; see Sec. 5.3 for details.[2]

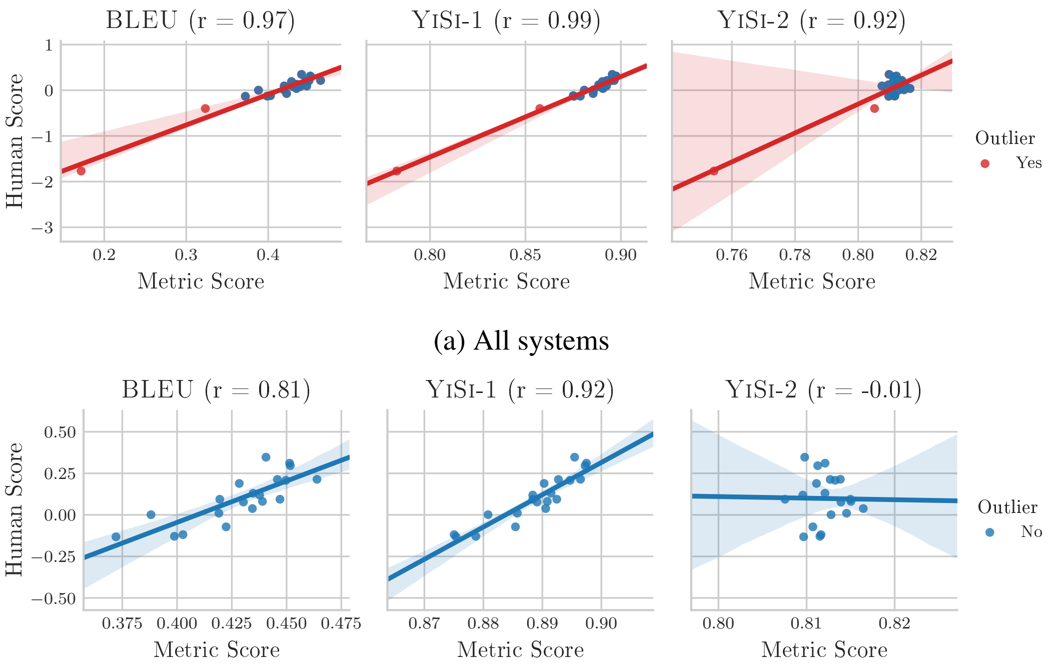
Reference-free metric

- YISI-2 (Lo, 2019) is the same as YISI-1, except that it uses cross-lingual embeddings to compute the similarity of the MT output with the source.

The baseline metrics, particularly BLEU, were designed to use multiple references. However, in practice, they have only have been used with a single reference in recent years.

We describe the baselines and YISI-1 in more detail in Sec. 2.2.4, and ESIM in Sec. 5.3.

[2] ESIM's submission to WMT shared task does not include scores for the language pairs en-cs and en-gu. In this chapter, we use scores obtained from the same trained model that was used in the original submission.



(a) All systems

(b) After discarding outliers

Fig. 6.3 Scatter plots (and Pearson's *r*) for metrics (a) with all systems and (b) without outliers for the WMT19 English → German language pair. We also show the line of best fit and the confidence interval around the line. Systems further away from the best-fit line indicate errors.

6.4 The Influence of Outliers on the Correlation of Automatic Metrics

When there are systems that are generally much worse (or much better) than the majority of the systems, metrics are usually able to correctly assign low (or high) scores to these systems. This can result in an inflated value of the Pearson correlation, giving misleading estimates of the relationship between human and metric scores of other systems.

Based on a visual inspection of the data, we can see there are two outlier systems in the English → German language pair (Fig. 6.3a). To illustrate the influence of these systems on Pearson's *r*, we repeatedly sub-sample ten systems from the 22 MT systems in the English → German data (see Fig. 6.4). When the most extreme outlier (`en-de-task`) is present in the sample, the correlation of all reference-based metrics is greater than 0.97. The selection of

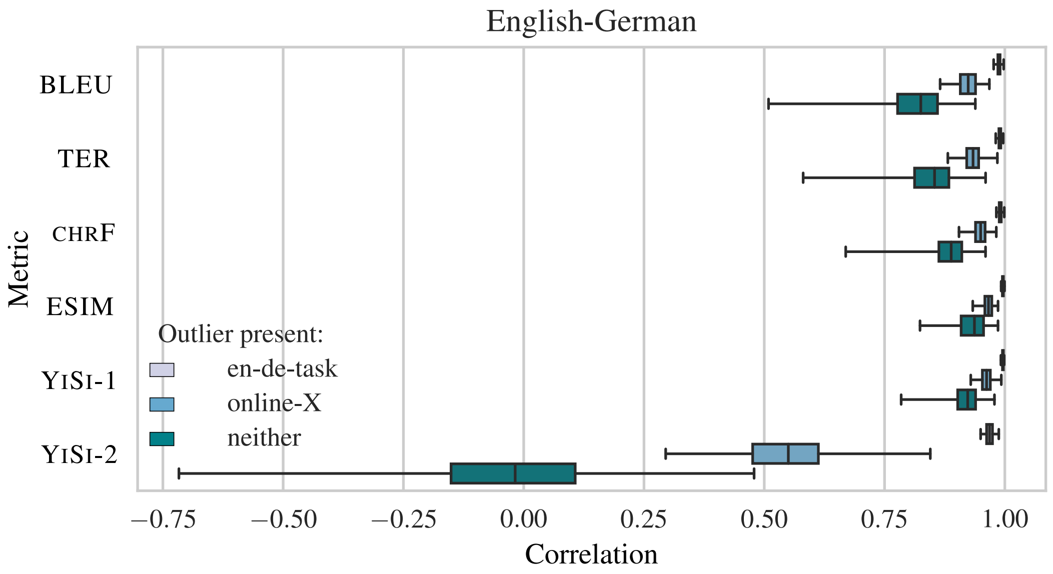


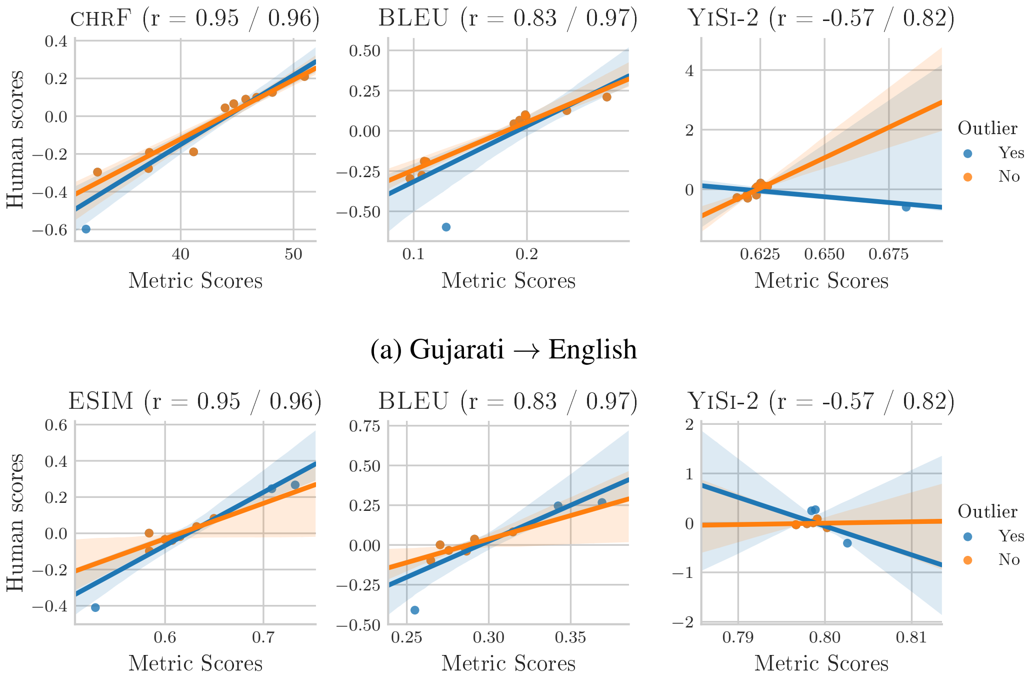
Fig. 6.4 Pearson's r for metrics, when sub-sampling systems from the English → German language pair from WMT19. We group the samples in the presence of the two outliers ("`en-de-task`" and "`online-X`"), and when neither is present.

	de–en		gu–en		kk–en		lt–en		ru–en		zh–en	
	All	−out	All	−out	All	−out	All	−out	All	−out	All	−out
#sys	16	15	11	10	11	9	11	10	14	13	15	13
BLEU	0.81	0.79	0.83	0.97	0.95	0.91	0.96	0.97	0.87	0.81	0.90	0.81
chrF	0.92	0.86	0.95	0.96	0.98	0.77	0.94	0.93	0.94	0.88	0.96	0.84
ESIM	0.94	0.90	0.88	0.99	0.99	0.95	0.99	0.99	0.97	0.95	0.99	0.96
YiSi-1	0.95	0.91	0.92	1.00	0.99	0.92	0.98	0.98	0.98	0.95	0.98	0.90
YiSi-2	0.80	0.61	−0.57	0.82	−0.32	0.66	0.44	0.35	−0.34	0.71	0.94	0.62

Table 6.1 Correlation of metrics with and without outliers ("All" and "−out", resp.) for the to-English language pairs from WMT19 that contain outlier systems

	de–cs		en–de		en–fi		en–kk		en–ru		fr–de	
	All	−out	All	−out	All	−out	All	−out	All	−out	All	−out
#sys	11	10	22	20	12	11	11	9	12	11	10	7
BLEU	0.87	0.74	0.97	0.81	0.97	0.94	0.85	0.58	0.98	0.95	0.87	0.85
chrF	0.97	0.97	0.98	0.88	0.99	0.97	0.97	0.90	0.94	0.97	0.86	0.80
ESIM	0.98	0.99	0.99	0.93	0.96	0.93	0.98	0.90	0.99	0.99	0.94	0.83
YiSi-1	0.97	0.98	0.99	0.92	0.97	0.94	0.99	0.89	0.99	0.98	0.91	0.85
YiSi-2	0.61	0.12	0.92	−0.01	0.70	0.48	0.34	0.69	−0.77	0.13	−0.53	0.07

Table 6.2 Correlation of metrics with and without outliers ("All" and "−out", resp.) for the language pairs into languages other than English from WMT19 that contain outlier systems.



(a) Gujarati → English

(b) French → German

Fig. 6.5 Scatter plots (and Pearson's *r*) for metrics with and without outliers for: (a) French → German, and (b) Gujarati → English data from WMT19. We also show the line of best fit and the confidence interval around the line. Systems further away from the best-fit line indicate errors.

systems has a higher influence on the correlation when neither outlier is present, and we can see that YISI-1 and ESIM have stronger correlations than BLEU. The impact of the presence of the outliers is even higher for the reference-free metric YISI-2.

For each language pair, we use the median absolute deviation (MAD) estimator (see Sec. 6.2.1) to detect outlier systems that have either much better or much worse direct assessment scores than the rest of the MT systems. In WMT 2019, these are low quality systems, with the exception of French → German, where the MAD estimator identifies the top two systems as well as the lowest ranked system as outliers (Fig. 6.5b).

Tables 6.1 and 6.2 show Pearson's *r* with and without outliers for the language pairs that contain outliers. Some interesting observations are as follows:

- For some language pairs, Lithuanian → English and English → Finnish for example, the correlation between the reference based metrics and DA is high irrespective of the presence of the outlier;
- the correlation of BLEU with DA drops sharply from 0.85 to 0.58 for English → Kazakh when outliers are removed;
- for English → German, the correlation of BLEU and TER appears to be almost as high as that of YISI-1 and ESIM. However, when we remove the two outliers, there is a much wider gap between the metrics. This suggests that YISI-1 and ESIM are more reliable and should be used in place of BLEU.
- if metrics wrongly assign a high score to a low outlier, removing these systems increases correlation, and only reporting the correlation after discarding outliers is not ideal. For instance, CHRF correctly scores system `JU-Saarland` as the worst system of Gujarat → English, but most other metrics give a relatively high score compared to the next best systems (Fig. 6.5a). Thus, we suggest reporting correlations over all systems as well as without outliers.

Finally, Fig. 6.6 shows the correlation of selected metrics on all language pairs, when computed over all systems and after discarding outliers. The graph of YISI-2 highlights the major impact these outliers can have on the Pearson correlation. We also note that once we remove outliers, ESIM and YISI-1 clearly outperform the baselines BLEU and TER, which wasn't apparent when considering the correlation over all systems.

Tables 6.3, 6.4, and 6.5 show the results for all metrics, when outliers are removed.[3]

[3] The code to detect outlier MT systems and generate these tables is available at https://github.com/nitikam/tangled

	en-cs	en-de		en-fi		en-gu	en-kk		en-lt	en-ru		en-zh
	All	All	-out	All	-out	All	All	-out	All	All	-out	All
	11	22	20	12	11	11	11	9	12	12	11	12
BEER	**0.990**	0.983	0.869	**0.989**	**0.978**	0.829	0.971	0.826	**0.982**	0.977	0.947	0.803
BERTr	0.983	**0.991**	**0.913**	0.960	0.913	**0.898**	**0.985**	**0.902**	0.971	0.982	0.964	**0.925**
BLEU	0.897	0.921	0.419	**0.969**	0.943	0.737	0.852	0.576	**0.989**	0.986	0.967	0.901
CDER	0.985	0.973	0.849	**0.978**	**0.957**	0.840	0.927	0.668	**0.985**	**0.993**	**0.981**	0.905
CharacTER	**0.994**	**0.986**	**0.886**	0.968	0.939	**0.910**	0.936	**0.895**	0.954	**0.985**	**0.982**	0.862
chrF	0.990	0.979	0.881	**0.986**	**0.972**	**0.841**	**0.972**	**0.900**	**0.981**	0.943	0.968	0.880
chrF+	**0.991**	0.981	0.883	**0.986**	**0.970**	0.848	**0.974**	**0.907**	**0.982**	0.950	0.973	0.879
EED	**0.993**	0.985	**0.894**	**0.987**	**0.978**	**0.897**	**0.979**	**0.883**	0.975	0.967	**0.984**	0.856
ESIM	–	**0.991**	**0.928**	0.957	0.926	–	**0.980**	**0.900**	**0.989**	**0.989**	**0.986**	0.931
ESIM_src	0.793	0.962	0.702	0.892	0.790	**0.744**	0.916	0.305	0.876	0.254	0.677	**0.883**
hLEPORa_baseline	–	–		–		0.841	0.968	**0.852**	–	–		–
hLEPORb_baseline	–	–		–		0.841	0.968	0.852	0.980	–		–
ibm1-morpheme	-0.871	0.870	0.198	0.084	-0.254	–	–		-0.810	–		–
ibm1-pos4gram	–	0.393	0.449	–		–	–		–	–		–
LASIM	–	0.871	0.007	–		–	–		–	-0.823	-0.336	–
LP.1	–	-0.569	0.558	–		–	–		–	-0.661	0.178	–
NIST	0.896	0.321	-0.246	0.971	0.936	0.786	0.930	0.611	**0.993**	**0.988**	**0.973**	0.884
PER	0.976	0.970	0.815	**0.982**	**0.961**	0.839	0.921	0.545	0.985	0.981	0.955	0.895
sacreBLEU-BLEU	**0.994**	0.969	0.806	**0.966**	0.939	0.736	0.852	0.576	**0.986**	0.977	0.946	0.801
sacreBLEU-chrF	0.983	0.976	0.874	0.980	0.958	0.841	0.967	0.840	0.966	**0.985**	**0.988**	0.796
TER	0.980	0.969	0.841	**0.981**	**0.960**	**0.865**	0.940	0.547	**0.994**	**0.995**	**0.985**	0.856
UNI	0.028	0.841	-0.251	0.907	0.808	–	–		–	0.919	0.760	–
UNI+	–	–		–		–	–		–	0.918	0.746	–
USFD	–	-0.224	-0.301	–		–	–		–	0.857	0.514	–
USFD-TL	–	-0.091	-0.212	–		–	–		–	0.771	0.177	–
WER	0.982	0.966	0.831	**0.980**	**0.958**	**0.861**	0.939	0.525	**0.991**	**0.994**	**0.983**	0.875
YiSi-0	**0.992**	0.985	0.869	**0.987**	**0.977**	0.863	0.974	0.840	0.974	0.953	**0.967**	0.861
YiSi-1	0.962	**0.991**	**0.917**	0.971	0.937	**0.909**	**0.985**	**0.892**	0.963	**0.992**	**0.978**	**0.951**
YiSi-2	0.324	0.924	-0.014	0.696	0.478	0.314	0.339	**0.685**	0.055	-0.766	0.134	-0.097
YiSi-2_srl	–	0.936	0.155	–		–	–		–	–		-0.118

Table 6.3 Correlation of metrics for the from-English language pairs at WMT19. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

	de-en		fi-en	gu-en		kk-en		lt-en		ru-en		zh-en	
	All	-out	All	All	-out	All	-out	All	-out	All	-out	All	-out
	16	15	12	11	10	11	9	11	10	14	13	15	13
BEER	0.906	0.852	**0.993**	0.952	0.982	0.986	**0.930**	0.947	0.948	0.915	0.819	0.942	0.806
BERTR	**0.926**	**0.897**	0.984	0.938	**0.995**	0.990	0.829	0.948	0.959	**0.971**	**0.933**	0.974	0.911
BLEU	0.849	0.770	0.982	0.834	0.975	0.946	**0.912**	0.961	**0.980**	0.879	0.830	0.899	0.807
CDER	0.890	0.827	**0.988**	0.876	0.975	0.967	0.843	**0.975**	**0.981**	0.892	0.875	0.917	0.847
CHARACTER	0.898	0.852	**0.990**	0.922	0.978	0.953	0.833	0.955	0.963	0.923	0.828	0.943	0.845
CHRF	**0.917**	0.862	**0.992**	0.955	0.962	0.978	0.775	0.940	0.933	0.945	0.876	0.956	0.841
CHRF+	**0.916**	0.860	**0.992**	0.947	0.961	0.976	0.769	0.940	0.934	0.945	0.878	0.956	0.851
EED	0.903	0.853	**0.994**	0.976	0.988	0.980	0.779	0.929	0.930	0.950	0.872	0.949	0.840
ESIM	**0.941**	**0.896**	0.971	0.885	0.986	0.986	**0.945**	**0.989**	**0.990**	**0.968**	**0.946**	**0.988**	**0.961**
ESIM_SRC	0.828	0.753	0.931	0.730	0.906	0.835	0.117	0.899	0.855	0.692	0.741	0.903	0.842
HLEPORA_BASELINE	–		–	–		0.975	0.855	–		–		0.947	0.879
HLEPORB_BASELINE	–		–	–		0.975	0.855	0.906	0.930	–		0.947	0.879
IBM1-MORPHEME	-0.345	-0.223	0.740	–		–		0.487	0.638	–		–	
IBM1-POS4GRAM	-0.339	-0.137	–	–		–		–		–		–	
LASIM	0.247	0.334	–	–		–		–		-0.310	0.260	–	
LP.1	-0.474	-0.279	–	–		–		–		-0.488	0.168	–	
METEOR++_2.0(S)	0.887	0.844	**0.995**	0.909	0.939	0.974	0.859	0.928	0.935	**0.950**	0.878	0.948	0.836
METEOR++_2.0(S+C)	0.896	0.850	**0.995**	0.900	0.930	0.971	0.871	0.927	0.931	**0.952**	**0.890**	0.952	0.841
NIST	0.813	0.705	0.986	0.930	0.985	0.942	0.837	0.944	**0.963**	0.925	**0.878**	0.921	0.722
PER	0.883	0.808	**0.991**	0.910	0.948	0.737	0.533	0.947	0.933	0.922	0.880	0.952	0.884
PREP	0.575	0.452	0.614	0.773	0.967	0.776	**0.817**	0.494	0.397	0.782	0.685	0.592	0.111
SACREBLEU-BLEU	0.813	0.794	0.985	0.834	0.975	0.946	**0.912**	0.955	0.967	0.873	0.813	0.903	0.807
SACREBLEU-CHRF	**0.910**	0.852	**0.990**	0.952	0.937	0.969	0.750	0.935	0.923	0.919	0.874	0.955	0.846
TER	0.874	0.812	**0.984**	0.890	0.947	0.799	0.566	0.960	0.975	0.917	0.896	0.840	0.717
UNI	0.846	**0.809**	0.930	–		–		–		0.805	0.666	–	
UNI+	0.850	**0.805**	0.924	–		–		–		0.808	0.669	–	
WER	0.863	0.803	0.983	0.861	0.926	0.793	0.579	0.961	**0.981**	0.911	0.885	0.820	0.716
WMDO	0.872	**0.857**	**0.987**	0.983	0.981	**0.998**	**0.953**	0.900	0.923	0.942	0.844	0.943	0.851
YISI-0	0.902	0.847	**0.993**	**0.993**	**0.990**	0.991	0.876	0.927	0.933	**0.958**	**0.889**	0.937	0.782
YISI-1	**0.949**	**0.914**	**0.989**	0.924	**0.997**	0.994	**0.920**	**0.981**	**0.978**	**0.979**	**0.947**	**0.979**	0.899
YISI-2	0.796	0.612	0.642	-0.566	0.820	-0.324	0.662	0.442	0.346	-0.339	0.708	0.940	0.622
YISI-2_SRL	0.804	0.630	–	–		–		–		–		0.947	0.675

Table 6.4 Correlation of metrics for the to-English language pairs of WMT19. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

	de-cs		de-fr	fr-de	
	All	-out	All	All	-out
	11	10	11	10	7
BEER	**0.978**	0.976	**0.941**	0.848	**0.794**
BERTr	0.961	**0.956**	**0.964**	0.889	**0.812**
BLEU	**0.941**	0.922	0.891	0.864	**0.821**
CDER	0.864	0.734	**0.949**	0.852	**0.794**
CharacTER	0.965	0.959	0.928	0.849	**0.848**
chrF	**0.974**	**0.970**	0.931	0.864	**0.796**
chrF+	0.972	0.967	0.936	0.848	**0.785**
EED	**0.982**	**0.984**	**0.940**	0.851	0.792
ESIM	**0.980**	**0.986**	**0.950**	**0.942**	**0.825**
ESIM_src	0.720	0.346	0.888	0.707	0.140
hLEPORa_baseline	0.941	0.903	0.814	–	
hLEPORb_baseline	**0.959**	**0.951**	0.814	–	
ibm1-morpheme	0.355	0.009	-0.509	-0.625	-0.357
ibm1-pos4gram	–		0.085	-0.478	-0.719
NIST	**0.954**	0.944	**0.916**	0.862	**0.800**
PER	0.875	0.757	0.857	**0.899**	0.427
sacreBLEU-BLEU	0.869	0.742	0.891	0.869	**0.846**
sacreBLEU-chrF	**0.975**	**0.980**	**0.952**	0.882	**0.815**
TER	0.890	0.787	**0.956**	**0.895**	**0.673**
WER	0.872	0.749	**0.956**	**0.894**	**0.657**
YiSi-0	**0.978**	0.972	**0.952**	0.820	**0.836**
YiSi-1	0.973	**0.980**	**0.969**	**0.908**	**0.846**
YiSi-2	0.606	0.122	0.721	-0.530	**0.066**

Table 6.5 Correlation of metrics for the language pairs of WMT19 that do not include English. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

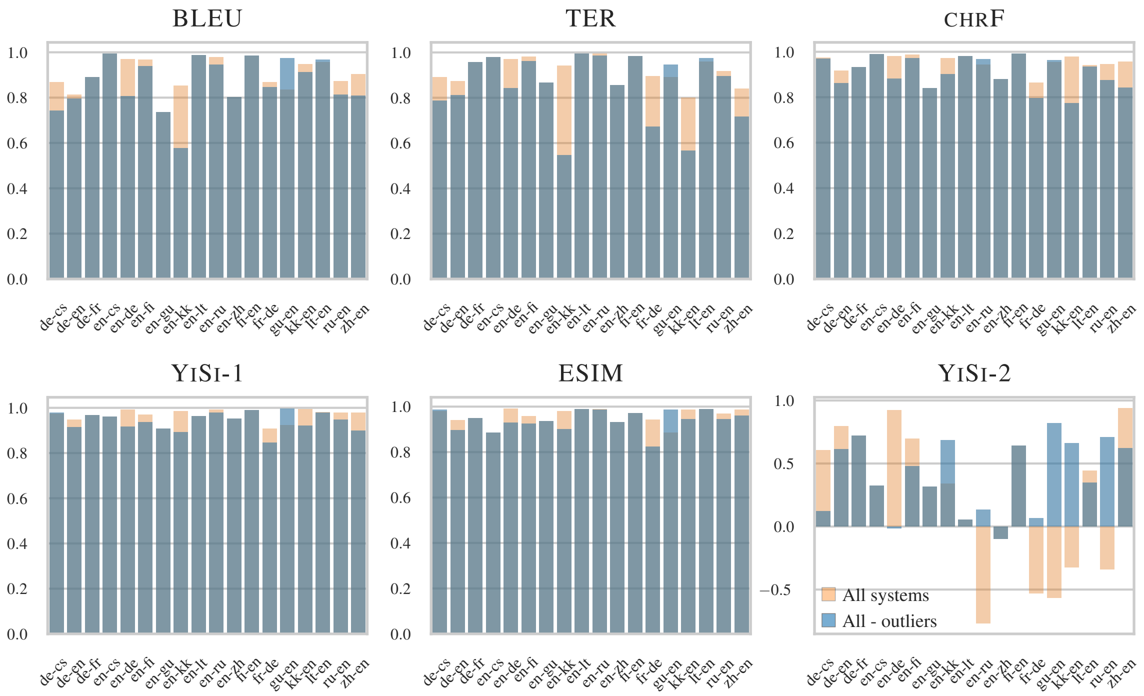


Fig. 6.6 Correlation of metrics with all systems (orange) and after discarding outliers (blue) over all language pairs of WMT19

Influence of Outliers on Statistical Significance Tests

The WMT metrics task uses the William's test (Williams, 1959) to detect statistical significance when comparing the correlation of two metrics. This test is applicable when comparing the difference between two dependent correlations that share a variable, and was recommended for automatic metrics by Graham and Baldwin (2014). Metrics that are not significantly outperformed by any other metric are declared the "winners" of the task. (See Sec. 2.3.3 for details on the William's test.) In this section, we explore the effect of outlier removal on the William's test and on the winning metrics of a language pair.

The power of the William's test depends on the sample size and correlation between the two metrics: the test is more likely to detect statistical significance between the metrics if they are highly correlated. While the sample size is constant for a language pair, the correlations of any given pair of metrics depends on the similarity of their approaches, and so the test does not have the same differentiating power for all pairs of metrics.

Removing outliers decreases sample size and also influences the correlation of two metrics:

- if both metrics give low scores to the low outliers, the resulting high correlation between the two metrics means the test has a high power to distinguish between the metrics. When we discard outliers, the correlation between metrics decreases, and the power of the test decreases.
- if two metrics differ in how they score the outlier (one metric gives a low score, and the other a relatively high score), then removing the outlier might increase the correlation between the two metrics on the remaining MT systems, potentially increasing the power of the William's test.

In the German → English language pair, the reference-free metric UNI (Yankovskaya et al., 2019) is a not outperformed by other metrics after removing outliers. With the outlier present, the correlation of UNI with other metrics is high (as they all give low scores to the outlier), so the power of the test is high. However, since UNI is less correlated with the other metrics once the outlier is removed, the test has low power and is unable to differentiate between UNI and the remaining metrics.

In a more extreme example, after removing the 3 outliers, there are only 7 MT systems left in the French → German language pair. The correlation of YiSi-2 drops from 0.530 to just 0.066. Since YiSi-2 also has low correlation with the other metrics, and the sample size is so small, the William's test doesn't detect a statistically significant difference between YiSi-2 and any other metric, even though the other metrics are highly correlated with human scores. So despite a negligible correlation with human scores, YiSi-2 is classified as a winner in this language pair.

BERTR and YISI-1 both incorrectly assign a relatively high score to the low outlier in the Gujarati → English language pair. They are competent when scoring the remaining systems, and these metrics are not outperformed by any other metric on discard that outlier, and are now “winner”. This is not ideal, and reporting correlation only after discarding outliers can result in a loss of important information about metrics.

Finally, YISI-1 is still the best performing metric with the most wins across all language pairs, with ESIM coming in second.

Previous Years

We computed results without outliers for the system-level metrics task of WMT 2017 (Tables 6.6 and 6.7) and WMT 2018 (Tables 6.8 and 6.9).

In the WMT 2017 data, we detect the presence of outliers in six out of the 14 language pairs. As with 2019, the outliers are systems with low DA scores, with the exception of English → Russian, where the best system scores much higher than the rest and is clearly an outlier.

Of these six language pairs, the correlation of all metrics remains high after outlier removal for English → Chinese. For German → English, English → Latvian and English → Finnish, the decrease in correlation of BLEU is a little higher compared to other metrics like BERTR and CHRF. Finally, with English → Russian and Russian → English, there is a sharp drop in correlation of BLEU compared to other metrics. Surprisingly, the correlation of ESIM is even smaller than BLEU for Russian → English, so ESIM is not uniformly better than BLEU.

In the WMT 2018 metrics task, all participating metrics have a high correlation with DA scores for all language pairs except Turkish → English. For most of these, the correlation stays high after discarding outliers. However, we see a decrease in correlation for three language pairs. For English → Turkish and German → English, we see that the drop in the correlation of BLEU is much higher than in YISI-1 and CHRF. Finally, the story is unique for Chinese → English. Of the 14 MT systems, the DA scores of the top nine are close together, and the MAD estimator detects the remaining five systems as outliers. When we compute correlation after discarding these five outliers, all metrics have a correlation between 0.5 and 0.7. When we look

	cs-en	de-en		fi-en	lv-en	ru-en		tr-en	zh-en
	All	All	-out	All	All	All	-out	All	All
	4	11	10	6	9	9	7	10	16
AutoDA	**0.438**	0.959	**0.948**	0.925	0.973	0.907	0.536	0.916	0.734
BEER	0.972	**0.960**	**0.951**	**0.955**	**0.978**	0.936	0.725	0.972	0.902
BERTr	**0.996**	**0.971**	**0.972**	**0.948**	**0.980**	**0.949**	**0.689**	**0.994**	**0.967**
Blend	0.968	**0.976**	**0.961**	**0.958**	**0.979**	**0.964**	**0.828**	**0.984**	0.894
BLEU	0.971	0.923	0.905	0.903	**0.979**	0.912	0.612	0.976	0.864
bleu2vec_sep	**0.989**	0.936	0.911	0.888	0.966	0.907	0.414	0.961	0.886
CDER	0.989	–		**0.927**	**0.985**	–		0.973	0.904
CharacTER	-0.972	-0.974	-0.952	-0.946	-0.932	-0.958	-0.697	-0.949	-0.799
chrF	0.939	**0.968**	**0.956**	**0.938**	0.968	0.952	**0.776**	0.944	0.859
chrF++	0.940	**0.965**	**0.957**	0.927	**0.973**	0.945	0.716	0.960	0.880
ESIM	0.983	**0.949**	**0.960**	**0.985**	**0.974**	**0.921**	**0.437**	**0.986**	0.901
ESIM_src	0.964	0.839	0.675	0.904	0.826	0.781	-0.188	0.876	0.720
MEANT_2.0	0.926	0.950	0.906	**0.941**	**0.970**	**0.962**	**0.866**	0.932	0.838
MEANT_2.0-nosrl	0.902	0.936	0.879	0.933	0.963	**0.960**	**0.875**	0.896	0.800
ngram2vec	0.984	0.935	0.908	0.890	0.963	0.907	0.426	0.955	0.880
NIST	**1.000**	0.931	0.924	0.931	0.960	0.912	**0.743**	0.971	0.849
PER	0.968	0.951	0.952	**0.896**	0.962	0.911	0.711	0.932	0.877
TER	0.989	0.906	0.859	**0.952**	**0.971**	0.912	0.661	0.954	0.847
TreeAggreg	0.983	0.920	0.906	**0.977**	**0.986**	0.918	**0.776**	**0.987**	0.861
UHH_TSKM	**0.996**	0.937	0.937	0.921	**0.990**	0.914	0.619	**0.987**	0.902
WER	0.987	0.896	0.826	**0.948**	**0.969**	0.907	0.603	0.925	0.839

Table 6.6 Correlation of metrics for the to-English language pairs of WMT17. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

	en-cs	en-de	en-fi		en-lv		en-ru		en-tr	en-zh	
	All	All	All	-out	All	-out	All	-out	All	All	-out
	14	16	12	11	17	16	9	7	8	11	10
AutoDA	**0.975**	0.603	0.879	0.416	0.729	0.471	0.850	-0.119	0.601	0.976	0.964
AutoDA-tecto	**0.969**	–	–		–		–		–	–	
BEER	**0.970**	0.842	0.976	0.882	**0.930**	**0.834**	0.944	0.603	0.980	0.914	0.882
BERTr	**0.982**	**0.877**	**0.979**	**0.918**	**0.949**	**0.889**	**0.971**	**0.820**	**0.996**	**0.992**	**0.989**
Blend	–	–	–		–		**0.953**	**0.658**	–	–	
BLEU	0.956	0.804	0.920	0.825	0.866	0.804	0.898	0.360	0.924	0.981	0.971
bleu2vec_sep	**0.963**	0.810	0.942	0.802	0.859	0.774	0.903	0.330	0.911	–	
CDER	0.968	0.813	–		–		–		0.957	0.983	0.975
CharacTER	-0.981	-0.938	-0.972	-0.894	-0.897	-0.790	-0.939	-0.716	-0.975	-0.933	-0.935
chrF	**0.976**	**0.863**	**0.981**	**0.912**	**0.955**	**0.895**	0.950	0.664	**0.991**	–	
chrF+	**0.976**	0.855	**0.980**	**0.903**	**0.956**	**0.899**	0.948	0.655	**0.988**	–	
chrF++	**0.974**	**0.852**	**0.979**	**0.894**	**0.956**	**0.897**	0.945	0.634	0.986	–	
ESIM	**0.974**	**0.861**	**0.971**	**0.851**	**0.954**	**0.911**	**0.968**	**0.906**	0.978	0.970	**0.963**
ESIM_src	0.846	**0.779**	0.888	0.502	0.834	0.587	0.869	0.205	0.868	0.867	0.786
MEANT_2.0	–	**0.858**	–		–		–		–	0.956	0.933
MEANT_2.0-nosrl	**0.976**	0.770	**0.972**	0.851	**0.959**	**0.904**	**0.957**	**0.726**	**0.991**	0.943	0.915
ngram2vec	–	–	0.940	0.809	0.862	0.781	–		–	–	
NIST	0.962	0.769	0.957	0.793	**0.935**	**0.884**	0.920	0.447	**0.986**	0.976	0.966
PER	0.954	0.687	0.949	0.726	0.851	0.809	0.887	0.177	0.963	0.934	0.936
TER	0.955	0.796	0.961	0.781	**0.909**	**0.862**	0.933	0.490	0.967	0.970	0.954
TreeAggreg	0.947	0.773	0.965	0.805	0.927	0.841	0.921	0.450	0.983	0.938	0.908
WER	0.954	0.802	0.960	0.779	**0.906**	**0.862**	**0.934**	0.490	0.956	0.954	0.930

Table 6.7 Correlation of all metrics for the from-English language pairs at WMT17. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

	cs-en	de-en		et-en		fi-en		ru-en		tr-en	zh-en	
	All	All	-out	All	-out	All	-out	All	-out	All	All	-out
	5	16	12	14	12	9	7	8	7	5	14	9
BEER	0.958	0.994	0.907	**0.985**	0.958	**0.991**	**0.985**	0.982	0.929	0.870	0.976	**0.591**
BERTR	**0.993**	0.990	0.896	0.981	**0.980**	**0.993**	0.969	**0.989**	**0.986**	0.504	**0.986**	**0.646**
BLEND	**0.973**	0.991	0.868	0.985	0.970	**0.994**	**0.985**	**0.993**	**0.982**	**0.801**	**0.976**	**0.596**
BLEU	**0.970**	0.971	0.831	**0.986**	0.942	0.973	0.954	0.979	0.924	-0.657	**0.978**	**0.661**
CDER	**0.972**	0.980	0.862	**0.990**	0.943	0.984	**0.980**	0.980	0.939	-0.664	**0.982**	0.609
CHARACTER	**0.970**	**0.993**	0.920	0.979	0.897	0.989	0.908	**0.991**	0.964	-0.782	0.950	**0.610**
CHRF	0.966	0.994	0.908	0.981	0.960	0.987	0.969	**0.990**	0.954	0.452	0.960	**0.620**
CHRF+	0.966	0.993	0.896	0.981	0.959	0.989	0.970	**0.990**	0.955	0.174	0.964	**0.632**
ESIM	**0.984**	**0.995**	**0.987**	**0.991**	**0.980**	0.980	**0.943**	**0.989**	**0.994**	**0.969**	**0.983**	**0.554**
ESIM_SRC	0.948	0.970	0.895	0.965	0.678	0.759	0.785	0.961	0.851	-0.847	0.963	**0.527**
ITER	**0.975**	0.990	0.892	0.975	0.828	**0.996**	0.952	0.937	0.682	-0.861	**0.980**	**0.638**
METEOR++	**0.945**	0.991	0.947	0.978	0.966	0.971	**0.986**	**0.995**	**0.979**	0.864	0.962	0.534
NIST	**0.954**	0.984	0.861	0.983	0.954	0.975	**0.966**	0.973	0.898	**0.970**	0.968	0.422
PER	**0.970**	0.985	0.844	**0.983**	**0.966**	**0.993**	**0.961**	0.967	0.922	0.159	0.931	**0.676**
RUSE	**0.981**	**0.997**	0.958	**0.990**	**0.966**	**0.991**	**0.975**	**0.988**	**0.974**	**0.853**	**0.981**	**0.583**
TER	0.950	0.970	0.856	**0.990**	0.948	0.968	0.938	0.970	0.891	0.533	0.975	0.338
UHH_TSKM	0.952	0.980	0.817	**0.989**	0.947	0.982	**0.976**	0.980	0.942	0.547	**0.981**	**0.372**
WER	0.951	0.961	0.857	**0.991**	0.942	0.961	0.932	0.968	0.873	0.041	**0.975**	0.351
YISI-0	0.956	**0.994**	0.946	0.975	0.964	0.978	**0.980**	**0.988**	0.952	**0.954**	0.957	**0.600**
YISI-1	**0.950**	0.992	0.960	0.979	**0.967**	0.973	**0.986**	**0.991**	**0.970**	**0.958**	0.951	**0.583**
YISI-1_SRL	**0.965**	**0.995**	0.957	0.981	**0.964**	0.977	**0.982**	**0.992**	**0.976**	**0.869**	0.962	**0.559**

Table 6.8 Correlation of metrics for the to-English language pairs of WMT18. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

	en-cs	en-de		en-et		en-fi	en-ru		en-tr		en-zh	
	All	All	-out	All	-out	All	All	-out	All	-out	All	-out
	5	16	12	14	12	12	9	8	8	7	14	11
BEER	**0.992**	**0.991**	**0.929**	**0.980**	**0.972**	**0.961**	**0.988**	0.958	**0.965**	**0.936**	0.928	**0.912**
BERTR_MULT	0.994	**0.988**	0.870	**0.971**	0.952	**0.971**	**0.990**	**0.964**	0.922	0.774	–	
BERTR_ZH	–	–		–		–	–		–		**0.969**	0.884
BLEND	–	–		–		–	**0.988**	0.954	–		–	
BLEU	0.995	0.981	**0.910**	0.975	0.969	**0.962**	0.983	**0.964**	0.826	0.528	0.947	0.903
CDER	0.997	0.986	0.900	**0.984**	**0.974**	**0.964**	**0.984**	0.946	0.861	0.584	0.961	0.897
CHARACTER	**0.993**	**0.989**	**0.900**	0.956	0.926	**0.974**	0.983	0.937	0.833	0.513	**0.983**	**0.933**
CHRF	**0.990**	**0.990**	0.913	**0.981**	**0.974**	**0.969**	**0.989**	**0.962**	0.948	0.856	0.944	0.870
CHRF+	**0.990**	0.989	0.906	**0.982**	**0.974**	**0.970**	0.989	0.956	0.943	0.843	0.943	0.857
ESIM	0.960	**0.988**	**0.933**	**0.978**	**0.958**	**0.971**	**0.979**	**0.956**	0.908	0.742	**0.978**	0.909
ESIM_SRC	0.852	0.946	0.827	0.902	0.732	0.917	0.938	0.730	0.746	0.321	0.917	0.705
ITER	0.915	**0.984**	0.835	**0.981**	**0.982**	**0.973**	0.975	0.906	0.865	0.597	–	
NIST	**0.999**	0.986	**0.920**	**0.983**	**0.982**	0.949	**0.990**	**0.972**	0.902	0.720	0.950	0.923
PER	0.991	0.981	0.860	0.958	0.968	0.906	**0.988**	**0.966**	0.859	0.608	0.964	**0.931**
TER	**0.997**	**0.988**	**0.917**	**0.981**	0.973	**0.942**	**0.987**	**0.962**	0.867	0.597	**0.963**	**0.945**
WER	**0.997**	**0.986**	**0.911**	**0.981**	**0.972**	**0.945**	0.985	0.960	0.853	0.556	0.957	0.935
YISI-0	0.973	0.985	**0.921**	0.968	0.955	0.944	**0.990**	**0.960**	**0.990**	**0.972**	0.957	0.883
YISI-1	**0.987**	0.985	**0.922**	**0.979**	**0.967**	0.940	**0.992**	**0.967**	**0.976**	**0.938**	0.963	**0.891**
YISI-1_SRL	–	**0.990**	**0.931**	–		–	–		–		0.952	0.900

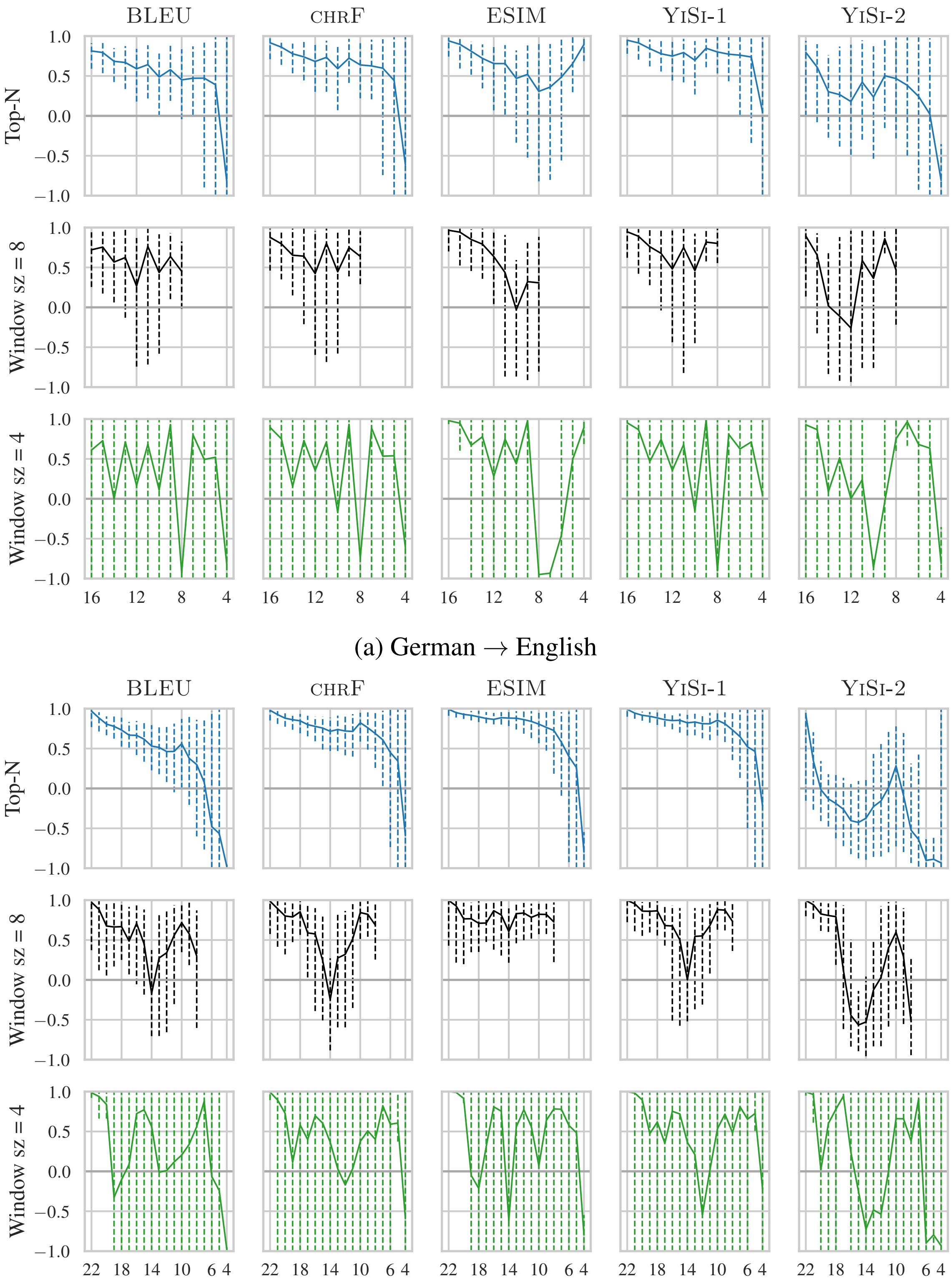
Table 6.9 Correlation of metrics for the from-English language pairs of WMT18. For language pairs that contain outlier systems, we also show correlation after removing outlier systems. Values in bold indicate that the metric is not significantly outperformed by any other metric under the Williams Test.

closer at the DA scores, these systems are all in one big cluster with no significant difference found between any two systems. It would be unreasonable to expect high correlation from the metrics on these MT systems.

6.5 The Influence of the Quality of MT Systems on Metric Reliability

In general, reference-based metrics have a high correlation with human scores across all language pairs. Typically, the correlation of reference-based metrics is greater than $r = 0.8$ in all language pairs at WMT 2019, and we can infer that it is reasonable to use these metrics in place of human evaluation. However, the correlation is dependent on the systems that are being evaluated, and as the quality of MT increases, we want to be sure that the metrics evaluating these systems stay reliable. To estimate the validity of the metrics for high-quality MT systems, Ma et al. (2019) sorted the systems based on their direct assessment scores, and plotted the correlation of the top N systems, with N ranging from all systems to the best four systems. They found that for seven out of 18 language pairs, the correlation between metric and human scores decreases as we decrease N, and tends towards zero or even negative when $N = 4$.

Of the nine language pairs that included a human translation as a part of the evaluated systems, there are four language pairs (German → English, English → German, English → Russian, and English → Chinese) where the quality of the best MT systems is close to human performance (Barrault et al., 2019). If metrics are unreliable for strong MT systems, we would expect to see a sharp degradation in correlation for these language pairs. But as we look at the top N systems for these language pairs, the pattern is mixed: correlation of the top N systems decreases as we decrease N for German → English and English → German, stays the same for English → Russian, and actually increases for English → Chinese. On the other hand, the



(a) German → English

(b) English → German

Fig. 6.7 Pearson correlation coefficient computed over the top-N systems, or over a rolling window of 4 or 8 systems on English → German and German → English datasets of WMT19. Systems are sorted by DA quality score, and the x axis shows the index of the starting system. The error bars indicate 95% confidence intervals computed using bootstrap resampling.

correlation decreases when computed over the top N English → Kazakh MT systems, where the human scores indicate that best system's quality is far from the quality of human translation.

Is there another explanation for these results? Pearson's r between metrics and DA scores is unstable for small samples, where we see major fluctuations in the value of Pearson's r as we add or decrease a single MT system. This is particularly true when the systems are very close in terms of quality. The low correlation over top-N systems (when N is small) could just be an artefact of this instability. We add error bars to the top-N plots of Ma et al. (2019) by computing the 95% confidence intervals using the percentile bootstrap method: we compute the correlation between metric and human scores of 1000 bootstrap samples of the MT systems in the subset, sort these correlations, and use the values at the 2.5 and 97.5 percentiles as the confidence limits. In general, the error bars get wider as N decreases, indicating less confidence in the results.

We also visualise the correlation of a rolling window of systems, starting with the worst N systems, and moving forward by one system until we reach the top N systems. The number of systems stays constant for all points in these graphs, which makes for a more valid comparison than the original setting where the sample size varies. If the metrics are indeed less reliable for strong systems, we should see the same pattern on these graphs as with the top N systems.

Fig. 6.7 shows the correlations of the top-N, and rolling window of correlations with window size of 4 and 8 on the German → English and English → German data. For both language pairs, the results of the top-N analysis shows decreasing correlation as N decreases. However, this decrease in correlation is also accompanied by decreasing confidence in the value of the correlations (wider error bars).

For the German → English language pair (Fig. 6.7 (a)), the correlation of most metrics is very unstable when $N = 4$. Both BLEU and CHRF perfectly correlate with human scores for systems ranked 2–5, but then the correlation drops to a large negative value when considering the top 4 systems. On the other hand, ESIM, which shows an upward trend when looking at the top-N systems, exhibits the opposite behaviour. Even worse, for English → German (Fig. 6.7 (b)), YISI-2 obtains a perfect correlation at some values of N, when in fact its correlation with human scores is negligible once outliers are removed (Sec. 6.4). We observe similar behaviour across all language pairs: the correlation is more stable as N increases (that is, we see fewer fluctuations in the value of r, and narrower error bars), but there is no consistent trend in the correlation that depends on the quality of the systems in the sample.

If we are to trust Pearson's r at small sample sizes, then the reliability of metrics doesn't really depend on the quality of the MT systems. Given that the sample size is small to begin with (typically 10–15 MT systems per language pair), we believe that we do not have enough data to use this method of sub-sampling systems to assess whether metric reliability decreases with the quality of MT systems. A possible explanation for the low correlation of subsets of MT systems is that it depends on how close these systems are in terms of quality. In the extreme case, the difference between the DA scores of all the systems in the subset can be statistically insignificant, so metric correlation over these systems can be attributed to chance. In the next section, we look at individual pairs of MT systems to examine whether the conclusions of metrics agree with human evaluation (DA), taking statistical significance of the score difference into account.

Metric reliability depends on the quality of the references (Freitag et al., 2020), and perhaps, as MT systems get better, automatic metrics require better references. But based on the data from the WMT 2019 metrics task, there is no empirical proof that the metric reliability decreases as MT system quality increases, and we believe that metrics can be unreliable irrespective of MT system quality.

6.6 Beyond Correlation: Metric Decisions for System Pairs

So far, we have looked at evaluating metrics using Pearson correlation with human scores, and looked closer at whether the high value of the correlation was hiding important patterns. In this

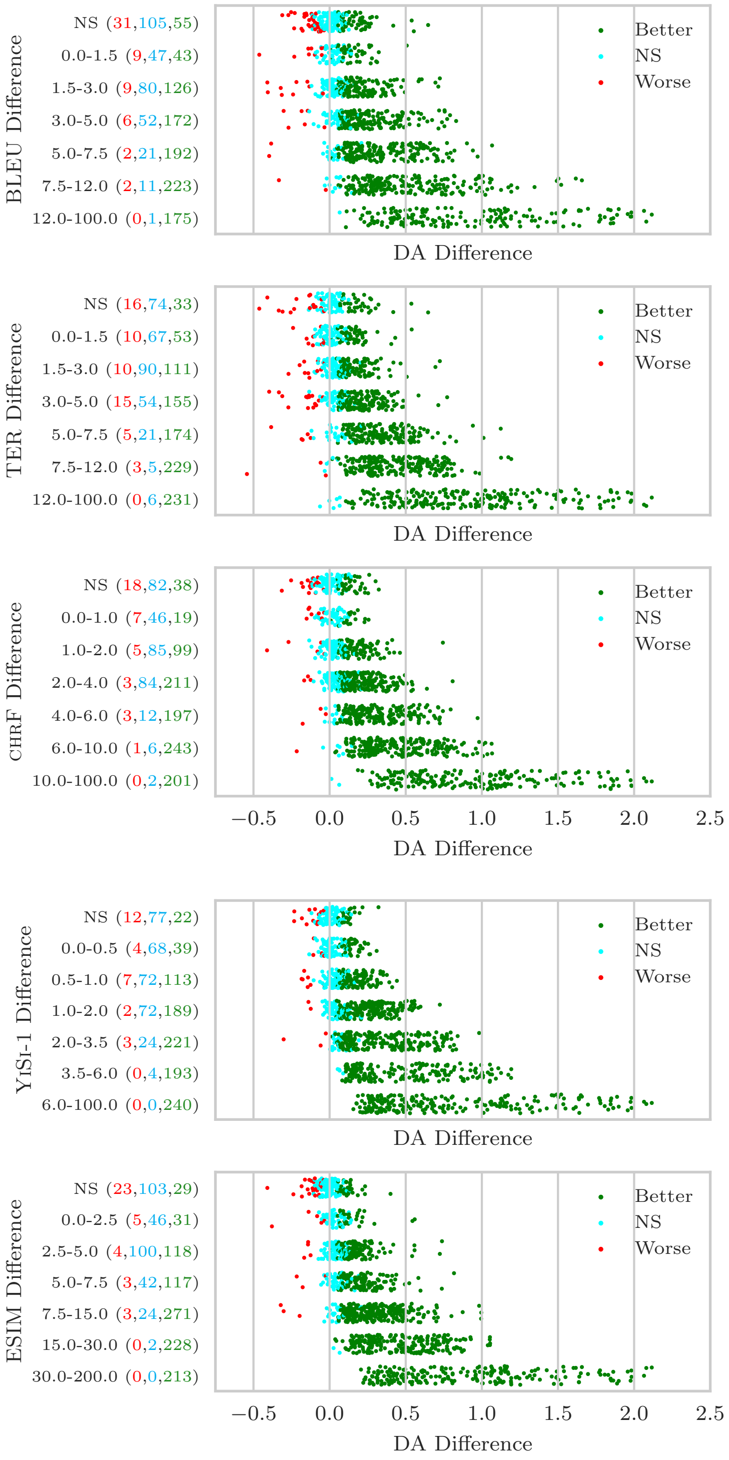


Fig. 6.8 Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference) computed across all language pairs in WMT 2019. The colours indicate pairs judged by humans to be insignificantly different (cyan/light gray), significantly worse (red/dark gray on the left) and significantly better (green/dark gray on the right), and the numbers next to the metric differences on the x-axis indicate the number of system pairs in the category corresponding to the colour.

section, we switch focus to how much we can trust automatic metrics when using them as a proxy for human judgements to make decisions about MT systems.

When we are comparing two systems, a metric might conclude that system A is better than system B with a certain score difference, or, if the score difference is not statistically significant, that the two systems are of similar quality. Basing important decisions on metric score alone runs the risk of making wrong decisions with respect to the true gold standard of human judgements. That is, while a change may result in a significant improvement in a metric, this may not be judged to be an improvement by human assessors (type I error). On the other hand, the metric might not detect a true improvement in the two systems (type II error).

Graham et al. (2014) computed accuracy of metric conclusions compared to human decisions on all pairs of MT systems in the WMT13 Spanish ↔ English language pairs. We extend their methodology by dividing metric errors into type I and type II errors, and then further examining how this depends on the value of the metric score difference. Finally, instead of computing these for individual language pairs, we combine data across all language pairs used in recent iterations of the WMT metrics shared task, focussing on WMT 2019.

For computing the statistical significance of human scores, we apply the Wilcoxon rank-sum test which is used by WMT when ranking systems. We use the bootstrap method (Koehn, 2004) to test for statistical significance of the difference in BLEU and TER between two systems. YISI-1 and ESIM compute the system score as the average of sentence scores, so we use the paired t-test to compute significance. Although CHRF is technically the macro-average of n-gram statistics over the entire test set, the online implementation also allows us to optionally compute the micro-average, and the two methods are highly correlated. We treat CHRF as a micro-average when computing significance such that we can use the more powerful paired t-test over sentence scores.

Figure 6.8 visualises the agreement between metric score differences and differences in human DA scores. Ideally, only differences judged as truly significant would give rise to significant and large magnitude differences under the metrics; and when metrics judge differences to be insignificant, ideally very few instances would be truly significant. However, this is not the case: there are substantial numbers of insignificant human differences even for very high metric differences (cyan, for higher range bins); moreover, the “NS” category — denoting an insignificant difference in metric score — includes many human significant pairs (red and green, top bin).

Considering BLEU (the top plot in Figure 6.8), for insignificant BLEU differences, humans judge one system to be better than the other for half of these system pairs. This corresponds to a type II error. It is of concern that BLEU cannot detect these differences. Worse, the difference in human scores has a very wide range. Conversely, when the difference in BLEU scores is small (between 0–3) but significant, more than half of these systems are judged to be insignificantly different in quality (corresponding to a type I error). For higher BLEU deltas, these errors diminish, however, even for a BLEU difference between 3 and 5 points, about a quarter of these system pairs are of similar quality. This paints a dour picture for the utility of BLEU as a tool for gate keeping (i.e., to define a ‘minimum publishable unit’ in deciding paper acceptance on empirical grounds, through bounding the risk of Type I errors), as the unit would need to be whoppingly large to ensure only meaningful improvements are accepted. Were we to seek to minimise Type II errors in the interests of nurturing good ideas, the threshold would need to be so low as to be meaningless, effectively below the level required for acceptance of the bootstrap significance test.

TER scores also contain major errors: the metric can wrongly conclude that a system is much better than another when humans have judged them similar, or even worse, drawn the opposite conclusion.

CHRF, YISI-1 and ESIM have fewer errors compared to BLEU and TER. When these metrics mistakenly fail to detect a difference between systems, the human score difference is considerably lower than for BLEU. Accordingly, they should be used in place of BLEU.

However the above argument is likely to still hold true as to their utility for gate keeping or nurturing progress, in that the thresholds would still be particularly punitive or permissive, for the two roles, respectively.

Impact of Target Language

Most metrics are designed primarily for English. Metrics evaluating in English have better available resources, such as word embeddings trained on a larger monolingual corpus. Other languages can also be harder for metrics compared to English, for example, if they are morphologically more complex (Bouamor et al., 2014; Guzmán et al., 2016). Finally, supervised models have a lot more training data in English as there are multiple language pairs every year that have English as the target language. In this section, we check whether this impacts the correctness of metric decisions.

When we split the system pairs based on whether the target language is English (Fig. 6.9), we find that all metrics make fewer errors when comparing the MT and reference translations in English compared to other languages, particularly the errors where metrics and humans both come to significant but opposite conclusions. In addition, metrics have a greater discriminative power when evaluating English translations, as there are very few type II errors where metrics do not detect a significant difference.

Impact of MT system Diversity

One possible reason for metric errors is the diversity in approaches of MT systems evaluated. For example, the metrics are known to be biased against rule based systems (Callison-Burch et al., 2006). But in recent years, the MT systems submitted to WMT were dominated by neural systems (recurrent models in 2017, and transformer models in 2018 and 2019) (Bojar et al., 2017b, 2018; Barrault et al., 2019). There were very few rule-based systems submitted[4].

[4] just three in WMT 2019

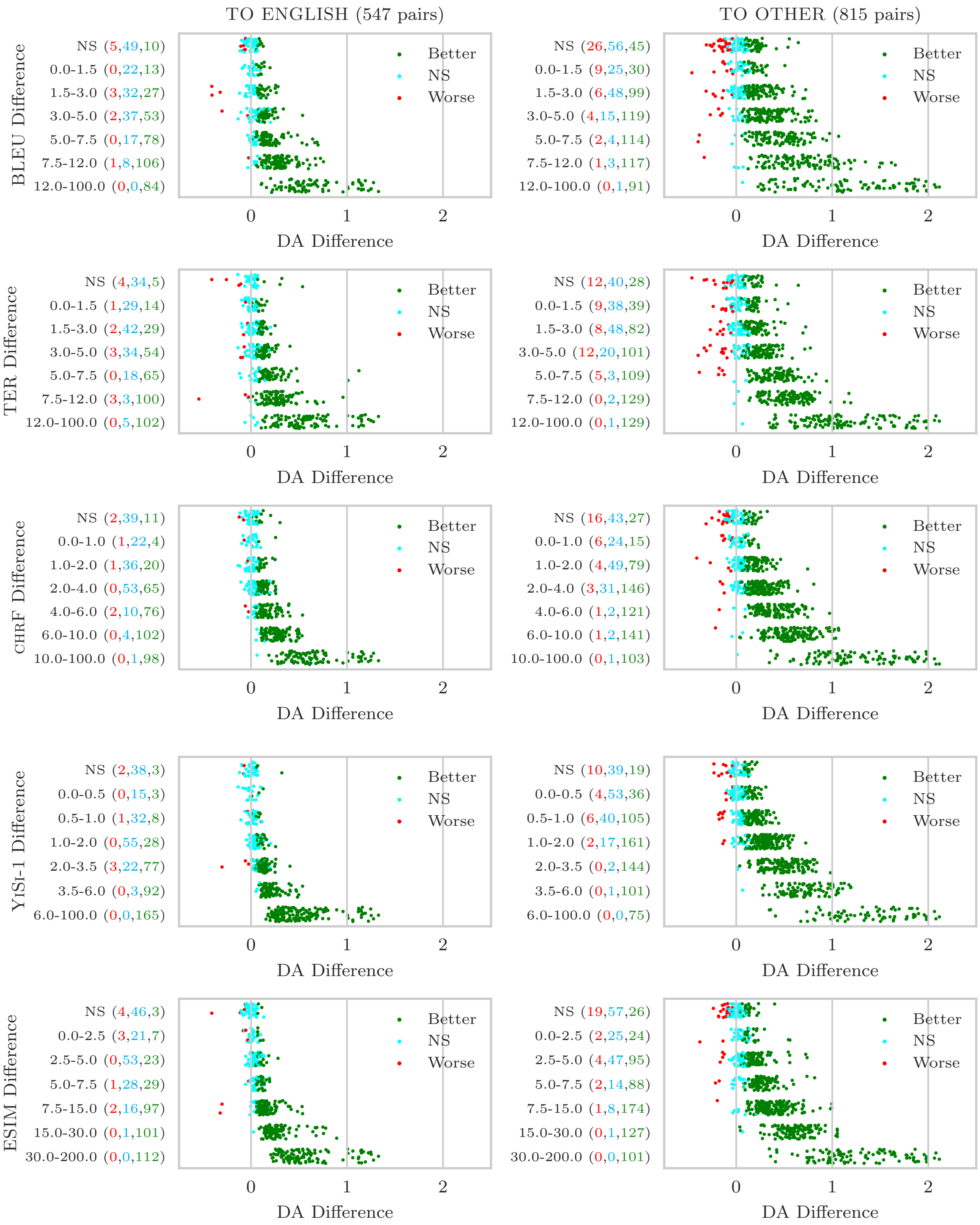


Fig. 6.9 Comparing metrics on to-English vs other-than-English language pairs in WMT 2019: Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference). The colours indicate pairs judged by humans to be insignificantly different (cyan/light gray), significantly worse (red/dark gray on the left) and significantly better (green/dark gray on the right).

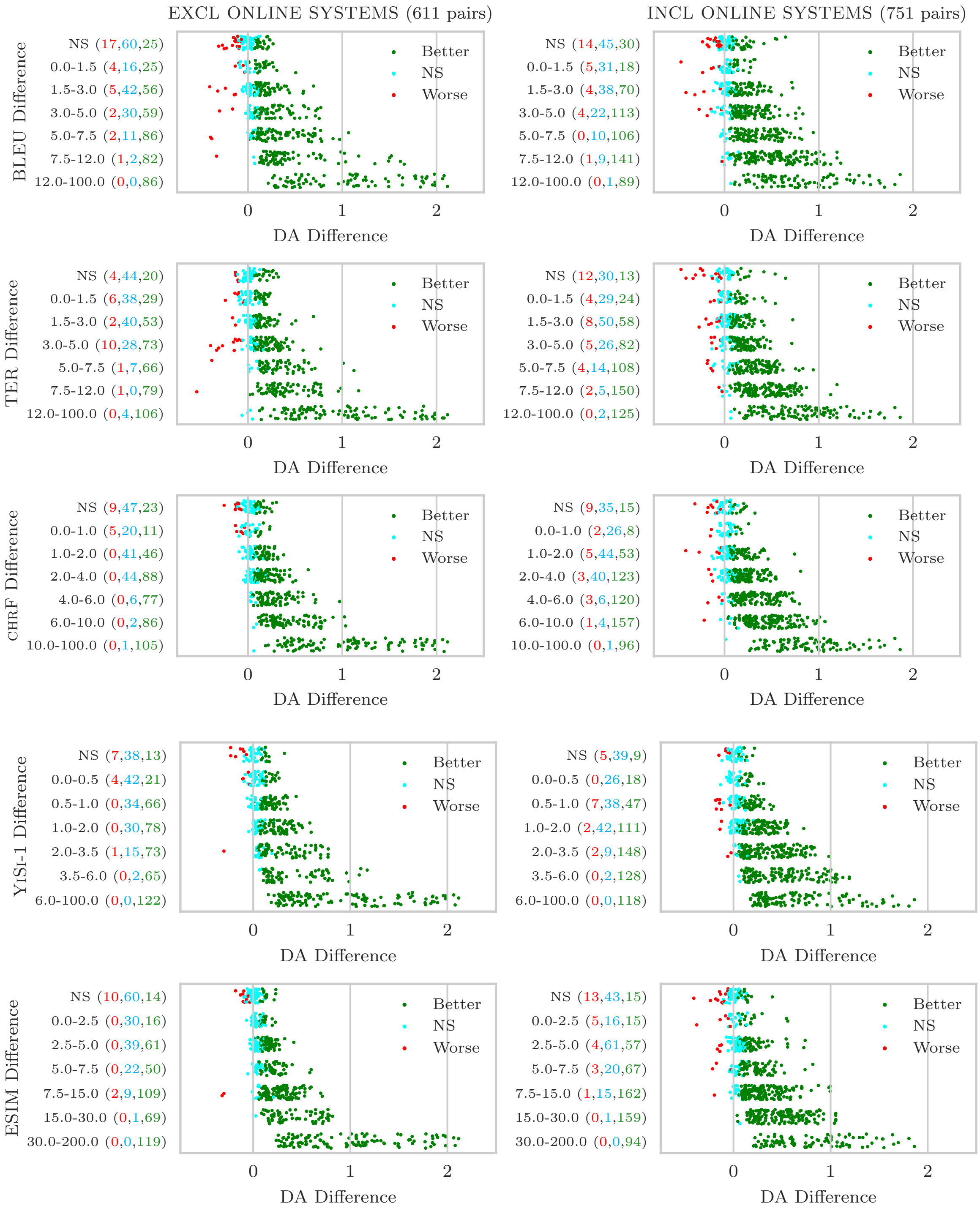


Fig. 6.10 WMT 2019 system pairs that (a) exclude vs (b) include online systems: Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference). The colours indicate pairs judged by humans to be insignificantly different (cyan/light gray), significantly worse (red/dark gray on the left) and significantly better (green/dark gray on the right).

These were mostly outperformed by neural systems in the human evaluation, and didn't pose a challenge to automatic metrics.

The systems evaluated consist of a mix of systems submitted by researchers (mostly neural models) and anonymous online systems such as Google and Amazon translation systems. The identity of these systems is hidden, and their approach, architecture, training data, and any data pre- or post-processing is unknown. Unlike most submissions to the translation task, these systems were not iteratively tuned on BLEU on the official WMT validation sets. Thus, we can expect the online systems to be more diverse than the submissions to the WMT translation task. Given that most academic papers do not compare with online systems, it is useful to divide the system pairs into two categories: EXCLONLINE, which excludes online systems, i.e., where both systems are submissions to the translation task; and INCLONLINE, where at least one system is an anonymous online system.

Fig. 6.10 shows that CHRF, YISI-1 and ESIM have fewer errors in EXCLONLINE than INCLONLINE, but this doesn't hold for BLEU and TER. The percentage of total errors in EXCLONLINE is greater than in INCLONLINE across all metrics and all years. This increase can be mostly attributed to type I errors: system pairs where the metric score difference is significant but humans didn't detect a significant difference. This is because there are more system pairs that are close together in quality, possibly because these systems were less diverse.

When we consider only type I errors where the metric score difference is significant, but human score-differences are significant in the opposite direction, we see that ESIM, CHRF and YISI-1 have significantly fewer errors when excluding online systems, but this is not true with BLEU and TER.

The increase in the total errors of TER and BLEU compared to the other metrics is striking when we restrict the comparisons to MT system submissions. This is far from ideal given that these metrics are widely used in academia both during system development (model selection on the validation set, for instance) and when benchmarking against other research systems when presenting final results.

Previous Years

When we compare metric errors on system pairs from WMT 2017 and 2018 (Fig. 6.11), we find that the overall "difficulty" of the metrics task varies between the years. In WMT18, the high correlations of all metrics with human scores (which mostly stay high even after removing outliers) translate to fewer errors (both type I and type II). In particular, there are remarkably few errors where both metric and humans score differences are significant but reach opposite conclusions, and these major errors are restricted to system-pairs where at least one online system is included. For WMT17 data, we see more errors compared to WMT18. Finally, even as the total percentage of errors varies between the years, the relative patterns between metrics stays the same. BLEU is clearly outperformed by CHRF, YISI-1, and ESIM in all cases.

6.6.1 Agreement between Metrics

While MT experiments are typically reported using BLEU as an evaluation measure, sometimes BLEU is used in combination with other metrics such as TER and METEOR (Banerjee and Lavie, 2005), and CHRF. So far, we investigated the reliability of using individual metrics to compare MT systems. We now investigate whether reporting a combination of metrics results in more reliable decisions.

Fig. 6.12 shows the agreement between metric decisions when comparing MT systems in 2019. BLEU and TER are both lexical metrics that compare the MT and the reference at the word-level. They are highly correlated, and when BLEU is wrong, TER contradicts BLEU only 20% of the times.

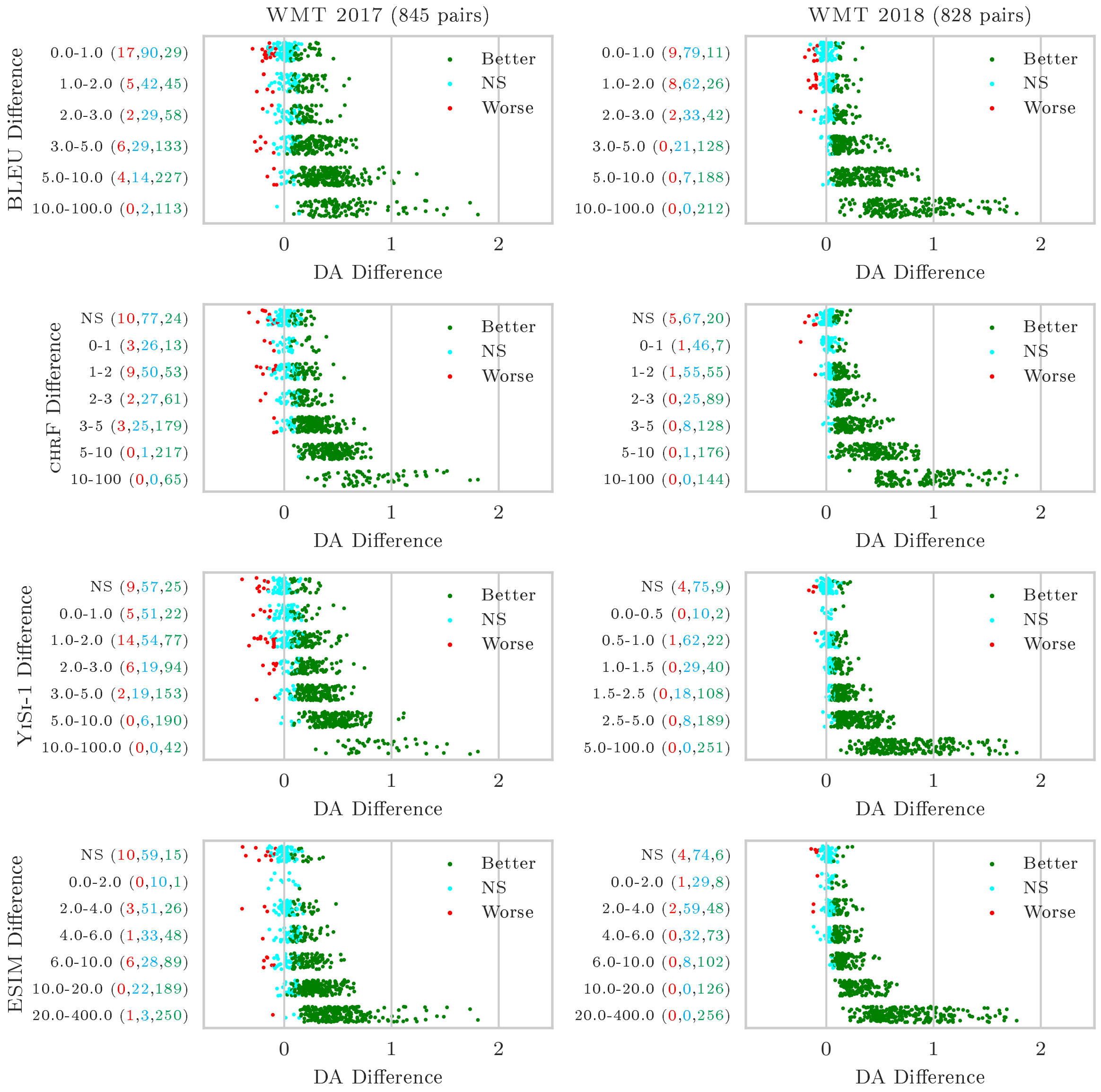
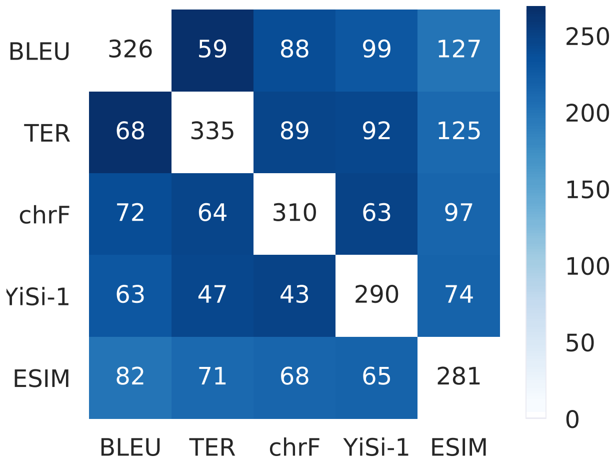


Fig. 6.11 Pairwise differences in human DA evaluation (x-axis) compared to difference in metric evaluation (binned on y-axis; NS means insignificant metric difference) computed across all language pairs in WMT 2017 and WMT 2018. The colours indicate pairs judged by humans to be insignificantly different (cyan/light gray), significantly worse (red/dark gray on the left) and significantly better (green/dark gray on the right).



	BLEU	TER	chrF	YiSi-1	ESIM
BLEU	326	59	88	99	127
TER	68	335	89	92	125
chrF	72	64	310	63	97
YiSi-1	63	47	43	290	74
ESIM	82	71	68	65	281

Fig. 6.12 The agreement between metric errors over all 1362 system-pair comparisons at WMT19. The values in the diagonal indicate the total number of type I and type II errors for the metric. The off-diagonal cells show the total number of errors made by the row-metric where the column-metric is correct.

Surprisingly, the conclusions of CHRF have a higher agreement with YISI-1 than BLEU, even though CHRF and BLEU are both lexical metrics whereas YISI-1 computes similarity based on contextual word embeddings.

The decisions of ESIM, a trained neural model, diverge a little more from the other metrics. ESIM has the fewest errors of all metrics, but we see in Fig. 6.8 that the errors of YISI-1 have a smaller magnitude.

Overall, despite the variety of approaches towards the task, all five metrics have common biases: over half of all erroneous decisions made by a particular metric are made in common with all other metrics.

This could be taken positively: when all five metrics agree on their decision, they are twice as likely to be right. But what conclusions can we draw when the metrics disagree? As expected, when BLEU or TER disagree with CHRF, YISI-1 and ESIM, the former are more likely to be wrong. We need further investigation as to which metrics to trust in this situation.

6.7 Conclusion and discussion

In this chapter, we first revisited the findings of the metrics task at WMT 2019, which flagged potential problems in the current best practises for assessment of evaluation metrics: (a) metric correlations are affected by outlier MT systems, and (b) the metrics are less reliable over strong MT systems.

We presented a robust way to detect outliers using the median and the median absolute deviation (MAD), and found that outlier systems can have a disproportionate influence on the Pearson correlation coefficient of metrics with human scores. In some cases, the correlation doesn't change much when outliers are removed, indicating that metrics are reliable when scoring all systems. However, even if the metric is not completely reliable when scoring most systems, it can be easy for metrics to correctly detect that the outlier system is much worse than others, and removing outliers can lead to a drop in Pearson's r. In extreme cases, the presence of outliers can give the illusion of a strong correlation when there is none. Finally, if a metric's assessment of the outlier is wrong, then dropping the outlier will lead to an increase in correlation, which ignores metric errors. Accordingly, we recommend future evaluations to report metric correlations with all systems, as well as after outlier removal (Leys et al., 2019).

The outlier detection method presented here, the MAD estimator, relies on the assumption that the distribution is symmetric. The distribution of human scores is typically reasonably close to satisfying this assumption, but there are some examples where this assumption is not true. This might result in removing systems that, on visual inspection, do not appear to be outliers. One avenue for future work is to investigate outlier detection methods that do not require this assumption.

We next investigated the relationship between metric reliability and the quality of the MT systems evaluated. We showed that the decrease in correlation when evaluating only the best MT systems can be attributed to comparing Pearson's r computed over different sample sizes. Pearson's correlation coefficient is known to be unstable for small sample sizes, and this instability is exacerbated when the systems in consideration are very close in quality. This goes some way to explaining the findings whereby strong correlations between metric scores and human judgements evaporate when considering small numbers of strong systems. We show that the same could be true for any small set of similar quality systems. Thus, this effect can, in some cases, be attributed to noise, rather than true shortcomings in the metrics themselves.

We note that it can still be true that metrics become more unreliable as the MT systems get better. However, we believe that the number of systems evaluated for each language pair in the WMT metrics tasks is too small (typically less than 15) to yield a conclusive empirical proof.

In common use, metrics are used to compare two systems, and accordingly we have also investigated the real meaning encoded by a difference in metric score, in terms of what this indicates about human judgements of the two systems. Most published works report BLEU differences of 1-2 points, however at this level we show this magnitude of difference only corresponds to true improvements in quality as judged by humans about half the time. Although our analysis assumes the human evaluation method "direct assessment" to be a gold standard, and clearly it has shortcomings, our analysis does suggest that the current rule of thumb for publishing empirical improvements based on small BLEU differences has little meaning.

Overall, this chapter adds to the case for retiring BLEU as the de facto standard metric, and instead using other metrics such as CHRF, YISI-1 or ESIM in its place during system development. These have higher correlations with human scores, particularly when we discard outliers, and are more powerful in assessing empirical improvements. However, human evaluation must always be the gold standard, and for continuing improvement in translation, to establish significant improvements over prior work, all automatic metrics make for inadequate substitutes.

Chapter 7

Conclusion and Future work

Reliable evaluation is crucial for progress in any task, and the work in this thesis forms a part of the movement to improve evaluation in NLP. We looked at improving robustness of three major aspects of machine translation evaluation.

We proposed methods to improve the collection and aggregation of human annotations of translation quality, which have the potential to decrease the cost of collecting annotations and improve data quality.

We developed a family of automatic metrics that advanced the existing state of the art on automatic evaluation. In concurrent work, our pre-trained metrics were independently proposed as BERTscore (Zhang et al., 2020), which has been widely adopted in the research community when evaluating machine translation and other natural language generation tasks such as summarization. The success of our supervised metrics inspired metrics such as COMET (Rei et al., 2020a) and BLEURT (Sellam et al., 2020) that have further improved the quality of MT metrics.

Finally, we revisited the evaluation of these automatic metrics. We present definitive evidence that the character-based metric CHRF and metrics that rely on pre-trained contextual embeddings are superior to BLEU, and our research is a part of the movement shift away from using BLEU when automatic evaluation is necessary. The work has also had a high impact on the running of evaluation campaigns for automatic metrics. We argued that metric evaluation needs to focus on the primary use-case of comparing MT systems. Kocmi et al. (2021) conducted a large-scale analysis of the performance of selected metrics over 3000 pairs of MT systems, and provided additional evidence to our conclusions in Chapter 6: neural embedding-based metrics (including ESIM) are far more trustworthy compared to lexical metrics like BLEU.

This chapter summarises the findings of the thesis, then presents avenues for future work.

7.1 Summary

Human Evaluation

In **Chapter 3**, we investigated the potential of probabilistic models to aggregate crowdsourced direct assessment (DA) data. We showed that the quality control mechanism of DA, which tests for internal consistency of annotators, often filters out useful data, thus increasing the overall cost of the evaluation. We proposed a simple Bayesian unsupervised model to aggregate DA scores. The model assumes that the annotator scores are normally distributed around the true quality of the translation, with an annotator-specific precision. The model infers reliable estimates of annotator precision, with help from additional constraints to the model based on the quality control items. The model effectively weights the scores of annotators based on the inferred precision to come up with a better estimate of the translation quality compared to the average scores, even when restricted to annotators that pass quality control. We found that we can further improve the model's estimate of translation quality by re-running the model after discarding scores of annotators with the lowest model precision. Finally, we showed that we can use pairwise correlation heat maps of annotator scores as a diagnostic tool to help decide (a) how many low-quality annotators to remove and (b) whether we need to collect more annotations.

Human judgements are, in theory, unbiased, and are considered to be the most reliable method to evaluate machine translation systems. However, people are susceptible to cognitive biases when making decisions. In **Chapter 4**, we looked at one specific source of cognitive bias that can arise when making a sequence of decisions. When annotators evaluate a set of translations, ideally, they would score each translation on its own merits. However, we used a simple linear regression model to show that crowdsourced annotator scores are positively autocorrelated with the score of the previous translation. When collecting multiple judgements of translation quality, if all annotators see these translations in the same order, any aggregate score will also contain this bias. To mitigate this, we suggest randomising the translations such that no two annotators see the translations in the same order.

Automatic Evaluation

Chapter 5 focused on automatic evaluation metrics. We proposed new automatic metrics that use contextual word embeddings to encode the MT output and the reference translation. Our first metrics approximate the precision, recall and F-score between the two sentences using a greedy maximum-matching of the embeddings of the MT output and the reference translation: namely, computing the maximum cosine similarity between the embedding of a reference token against any token in the MT output.

We then explored a series of supervised models that are built on top of contextualised word embeddings, including ESIM, a model that was first developed for natural language inference. These models uses cross-sentence attention and sentence matching heuristics to generate a representation of the translation and the reference. We investigated the tradeoff between the number of instances in the training set and the number of annotations per instance. We find that training ESIM on a large, singly-annotated set of human evaluation judgements clearly outperforms training the model on a smaller, multiply-annotated dataset, thus adding to the literature that diversity of training instances is more important than accuracy.

Our pre-trained metrics, though simple in formulation, are highly effective and rival or surpass previous metrics on the WMT 2017 dataset. Our supervised metrics further improve on these results.

Meta evaluation

In **Chapter 6**, we took a closer look at evaluating our metrics. In recent years, they have been evaluated based on their Pearson Correlation with human scores. We showed that Pearson correlation is highly sensitive to outlier MT systems that are markedly different in quality from the rest of the MT systems. These systems have a disproportionate influence on the value of the correlation: if the metrics score these systems correctly, it leads to high correlations even when they make errors scoring the remaining systems, thus leading to false confidence in the utility of these metrics. We thus propose to also include the correlation without outliers when evaluating metrics at the system level. This analysis was included in the results of the WMT 2020 metrics shared task, and the findings showed that outliers also inflate correlation at the sentence-level.

We next investigated whether metric reliability decreases with an increase in MT system quality. Findings from the WMT 2019 metrics task indicate that metric correlation sharply decreases when evaluated on smaller subsets of the best MT systems. We show that this can be attributed more to the instability of Pearson correlation on small sample sizes, and that there is no empirical evidence in the data to indicate that metrics are more likely to make mistakes when evaluating strong MT systems. The machine translation researchers have always been aware that automatic metrics are flawed, and relying on these metrics has always carried the risk of making wrong conclusions. Our findings suggest that this risk is independent of the quality of the MT systems being evaluated.

Finally, we investigated how much we can trust metric conclusions when comparing pairs of MT systems, which is the most common use case of these metrics. Small differences in

BLEU scores are often used in academia to claim a new state of the art over existing systems. We find that such a small difference in BLEU scores is essentially a coin toss: about half the time, there is no statistically significant difference in the human scores of the two systems. And in rare cases, humans even find a difference in the opposite direction. We recommend using more sophisticated metrics than BLEU when automatic evaluation is required during system development, and basing any final conclusions on human evaluation. We acknowledge that direct assessment is not perfect, that it is possible that the discrepancies between the metric and human decisions can actually be errors of human evaluation and not metrics. In fact, many recent studies have questioned the validity of crowdsourced human judgements as untrained annotators often miss errors in the MT output (Castilho et al., 2017; Läubli et al., 2020; Freitag et al., 2021). Note that our method of looking at metric score distributions is independent of the method of human evaluation, and the same analysis can be repeated with higher quality human scores.

7.2 Future Work

Refining the evaluation of MT is still an active research ares, and we present some avenues for future work that follow from the work in this thesis.

Human evaluation

In this thesis, we proposed a probabilistic model to aggregate multiply-annotated translation quality judgements. We show that the model's inferred quality outperforms the current best practice, which is using the mean of workers who pass quality control, particularly when we remove spammers (the least reliable annotators). The model does not directly tell us how many spammers to remove to obtain the highest accuracy, and we propose heurstics to do so based on looking at pairwise correlations between all annotators. A more principled option is to extend the model to include a parameter for whether workers are spamming, which determines whether the worker's scores are uniformly distributed or distributed normally around the true quality. This would allow the model to decide on the utility of a given annotator's scores when fitting the data.

This is useful for getting accurate estimates of the quality of individual translations. When evaluating MT systems, direct assessment computes the score of an MT system as the average score of all (or a sufficiently large number) its translations in the test set, typically collecting only a single annotation per translation. We could extend our model to aggregate these scores to learn the true quality of MT systems. In this case, we model MT system scores as a Gaussian distribution centred around the scores of its translations. Since we will have multiple systems translating each source sentence, the model could also be enhanced to model translation difficulty of a sentence.

Automatic metrics

We train our metrics with a large, noisy set of judgements; our regression model uses a squared error loss that assumes a constant Gaussian noise for each instance in the training set. In reality, we showed in Chapter 3 that annotator reliability varies, and we can get better aggregation of human scores when we model their reliability. We can apply this idea when training metrics, by giving more weight to reliable annotators when computing the loss. The neural model would jointly learn the parameters of the regression model and annotator precision.

Meta-evaluation

In Chapter 6, we looked at how much we can trust metric decisions when comparing pairs of MT systems. More specifically, we compared metric agreement with human decisions on pairs of systems included in the WMT evaluation, which are a mix of anonymous online systems and researcher submissions to the shared task. We found that metrics are more likely to make major errors when one or both systems is an online systems, and hypothesise that this might be due to diversity of the online services compared to WMT submissions. We need to test this hypothesis, for example, by comparing the pairwise similarity of MT system translations, perhaps looking at similarities in lexical choice and word order.

We use automatic metrics in two broad scenarios(Resnik and Lin, 2010), which are: (a) *formative evaluation*: making small changes to the systems and evaluate whether this improves the quality; and (b) *summative evaluation*: comparing an MT system with a completely different MT system developed by others, for example to claim state of the art on a given dataset. The WMT metrics shared task is an accurate reflection of the second scenario, and we used data from this task to recommend a careful human evaluation to support any final conclusions when comparing MT systems. But obtaining human judgements is not practical for formative evaluation. Are metrics more or less reliable when measuring incremental improvements on a single MT system?

To this end, we need to carefully apply a diverse set of ideas for improvement on a single MT system. Once we have the conducted a careful human evaluation, we could evaluate metrics using both the traditional method of computing correlation as well as analysis over the system pairs to understand the validity of individual decisions and how they depend on metric score differences. This test set would then either validate the continued use of BLEU over the years, or give decisive evidence for switching to a different metric. Finally, this will also yield a rich dataset to train future metrics.

This is an exciting time for machine translation research; MT quality has improved significantly in recent years, and better evaluation methods will accelerate progress in the quality of MT systems.

Bibliography

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171, Online. Association for Computational Linguistics.

Joshua Albrecht and Rebecca Hwa. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 880–887, Prague, Czech Republic. Association for Computational Linguistics.

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 1998. Automatic acquisition of hierarchical transduction models for machine translation. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Francis J Anscombe. 1973. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan. Association for Computational Linguistics.

Maya Bar-Hillel and Willem A Wagenaar. 1991. The perception of randomness. *Adv. Appl. Math.*, 12(4):428–454.

Petra Barančíková. 2014. Parmesan: Meteor without paraphrases with paraphrased references. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 355–361, Baltimore, Maryland, USA. Association for Computational Linguistics.

Vic Barnett and Toby Lewis. 1974. *Outliers in Statistical Data.* Wiley.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. Anchoring and agreement in syntactic annotations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2224, Austin, Texas. Association for Computational Linguistics.

Saurabh Bhargava and Raymond Fisman. 2014. Contrast effects in sequential decisions: Evidence from speed dating. *The Review of Economics and Statistics*, 96(3):444–457.

Anna Bindler and Randi Hjalmarsson. 2018. Path Dependency in Jury Decision-Making. CEPR Discussion Papers 13012, C.E.P.R. Discussion Papers.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Michael Bloodgood and Chris Callison-Burch. 2010. Using Mechanical Turk to build machine translation evaluation sets. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 208–211, Los Angeles. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017a. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016b. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017b. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 286–295, Singapore.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Andy Way, Panayota Georgakopoulou, Maria Gialama, Vilelmini Sosoni, and Rico Sennrich. 2017. Crowdsourcing for nmt evaluation: Professional translators versus the crowd. *Translating and the Computer*, 39.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio. Association for Computational Linguistics.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Boxing Chen and Hongyu Guo. 2015. Representation based translation evaluation metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 150–155, Beijing, China. Association for Computational Linguistics.

Daniel Chen, Tobias J. Moskowitz, and Kelly Shue. 2016. Decision-making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Linguistic Data Consortium. 2002. Linguistic data annotation specification:assessment of fluency and adequacy in arabic-english and chinese-english translations. Technical report.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.

Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of Machine Translation Summit IX*, pages 63–70, New Orleans.

Christopher Culy and Susanne Z Riehemann. 2003. The limits of n-gram translation evaluation metrics. In *Proceedings of the Ninth Machine Translation Summit*.

Lysann Damisch, Thomas Mussweiler, and Henning Plessner. 2006. Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied*, 12(3):166.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 469–478, New York, NY, USA. Association for Computing Machinery.

Michael Denkowski, Hassan Al-Haj, and Alon Lavie. 2010. Turker-assisted paraphrasing for english-arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70.

Michael Denkowski and Alon Lavie. 2010a. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *AMTA 2010: Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*.

Michael Denkowski and Alon Lavie. 2010b. Exploring normalization techniques for human judgments of machine translation adequacy collected using amazon mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 57–61, Los Angeles, USA.

Michael Denkowski and Alon Lavie. 2010c. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, USA.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*, volume 57. CRC press.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. TerrorCat: a translation error categorization-based MT quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 64–70, Montréal, Canada. Association for Computational Linguistics.

Peter A. Flach, Sebastian Spiegler , Bruno Golenia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. 2009. Novel tools to streamline the conference review process: Experiences from sigkdd'09. *SIGKDD Explorations*, 11.

Marina Fomicheva, Núria Bel, Iria da Cunha, and Anton Malinovskiy. 2015. UPF-cobalt submission to WMT15 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 373–379, Lisbon, Portugal. Association for Computational Linguistics.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *arXiv preprint arXiv:2104.14478*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. 2007a. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. 2007b. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264.

Eric Gold and Gordon Hester. 2008. The gambler's fallacy and the coin's memory. In Joachim I. Krueger, editor, *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes.*, chapter 2. Taylor & Francis Group.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 1183–1191, Denver, USA.

Perry Groot, Adriana Birlutiu, and Tom Heskes. 2011. Learning from multiple annotators with Gaussian processes. In *Proceedings of the 21st International Conference on Artificial Neural Networks - Volume Part II*, ICANN'11, pages 159–164, Espoo, Finland.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal.

Francisco Guzmán, Houda Bouamor, Ramy Baly, and Nizar Habash. 2016. Machine translation evaluation for Arabic using morphologically-enriched embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1398–1408, Osaka, Japan. The COLING 2016 Organizing Committee.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–698, Baltimore, Maryland. Association for Computational Linguistics.

Samuel M. Hartzmark and Kelly Shue. 2018. A tough act to follow: Contrast effects in financial markets. *The Journal of Finance*, 73(4):1567–1613.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Mark Hopkins and Jonathan May. 2013. Models of translation competitions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1424, Sofia, Bulgaria. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 1120–1130, Atlanta, USA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York, USA.

Jeff Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, 1 edition. Crown Publishing Group, New York, NY, USA.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 402–408, Baltimore, Maryland, USA. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

Douglas T. Kenrick and Sara E. Gutierres. 1980. Contrast effects and judgments of physical attractiveness: When beauty becomes a social problem. *Journal of Personality and Social Psychology*, 38(1):131.

Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–498, San Diego, California. Association for Computational Linguistics.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 619–627, La Palma, Canary Islands. PMLR.

Rebecca Knowles. 2021. On the stability of system rankings at wmt. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Philipp Koehn and Christof Monz. 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. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Christophe Leys, Marie Delacre, Youri L Mora, Daniël Lakens, and Christophe Ley. 2019. How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32(1).

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Hang Li. 2011. Learning to rank for information retrieval and natural language processing. *Synthesis lectures on human language technologies*, 4(1):1–113.

Maoxi Li, Chengqing Zong, and Hwee Tou Ng. 2011. Automatic evaluation of Chinese translation output: Word-level or character-level? In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 159–164, Portland, Oregon, USA. Association for Computational Linguistics.

Yuan Li. 2019. *Probabilistic models for aggregating crowdsourced annotations*. Ph.D. thesis, 'The University of Melbourne'.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 354–359.

Chang Liu and Hwee Tou Ng. 2012. Character-level machine translation evaluation for languages with ambiguous word boundaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–929, Jeju Island, Korea. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor, Michigan. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 589–597, Copenhagen, Denmark.

Chi-kiu Lo. 2019. YiSi — a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Chi-kiu Lo, 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. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2013. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 422–428, Sofia, Bulgaria.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 1–9.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a novel combined MT metric based on direct assessment — CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 598–603, Copenhagen, Denmark.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

Benjamin Marie and Marianna Apidianaki. 2015. Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391, Lisbon, Portugal. Association for Computational Linguistics.

Dennis N. Mehay and Chris Brew. 2007. BleuÂtre: Flattening syntactic dependencies for mt evaluation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 122–131, Skövde,Sweden.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

T. Minka, J.M. Winn, J.P. Guiver, Y. Zaykov, D. Fabian, and J. Bronskill. 2018. Infer.NET 0.3. Microsoft Research Cambridge. http://dotnet.github.io/infer.

Thomas P. Minka. 2001. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, Seattle, USA.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136.

John J. Neill and Olive Jean Dunn. 1975. Equality of dependent correlation coefficients. *Biometrics*, 31(2):531–543.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

S. Nießen, S. Vogel, H. Ney, and C. Tillmann. 1998. A dp based search algorithm for statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98/COLING '98, page 960–967, USA. Association for Computational Linguistics.

NIST. 2002. The nist 2002 machine translation evaluation plan (mt-02). Technical report.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 161–168, Boston, Massachusetts, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Joseph Olive, Caitlin Christianson, and John McCary. 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, 1st edition. Springer Publishing Company, Incorporated.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 297–305.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181–188.

Kishore Papineni, 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 Association for Computational Linguistics*, pages 311–318, Philadelphia, USA.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

John R. Pierce and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, Washington, DC, USA.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović and Hermann Ney. 2009. Syntax-oriented evaluation measures for machine translation output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13(16):491–518.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Susanne P. Reilly, James W. Smither, Michael A. Warech, and Richard R. Reilly. 1998. The influence of indirect knowledge of previous performance on ratings of present performance: The effects of job familiarity and rater training. *Journal of Business and Psychology*, 12(4):421–435.

David Reitter, Frank Keller, and Johanna D Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York, USA.

Philip Resnik and Jimmy Lin. 2010. 11 evaluation of nlp systems. *The handbook of computational linguistics and natural language processing*, 57.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Peter J Rousseeuw and Mia Hubert. 2011. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Vasile Rus and Mihai Lintean. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162, Montréal, Canada. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16, Online (Dublin, Ireland). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Christophe Servan, Alexandre Bérard, Zied Elloumi, Hervé Blanchon, and Laurent Besacier. 2016. Word2Vec vs DBnary: Augmenting METEOR using vector representations or lexical resources? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1159–1168, Osaka, Japan. The COLING 2016 Organizing Committee.

Burr Settles. 2009. Active learning literature survey.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 614–622, New York, NY, USA.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 764–771, Belgium, Brussels.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Matthew Snover, 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 Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level MT evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 123–129, Edinburgh, Scotland. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. Eed: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, Maryland, USA. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401, Lisbon, Portugal. Association for Computational Linguistics.

Miloš Stanojević. 2017. *Permutation Forests for Modeling Word Order in Machine Translation*. Ph.D. thesis, University of Amsterdam.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Computational Linguistics and Intelligent Text Processing*, pages 341–351, Berlin, Heidelberg. Springer Berlin Heidelberg.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation*, pages 619–622, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Haozhou Wang and Paola Merlo. 2016. Modifications of machine translation evaluation metrics by using word embeddings. In *Proceedings of the Sixth Workshop on Hybrid Approaches to Translation (HyTra6)*, pages 33–41, Osaka, Japan. The COLING 2016 Organizing Committee.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 764–773, Prague, Czech Republic. Association for Computational Linguistics.

John S. White and Theresa A. O'Connell. 1994. Evaluation in the ARPA machine translation program: 1993 methodology. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

John S. White, Theresa A. O'Connell, and Francis E. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043. Curran Associates, Inc.

Evan James Williams. 1959. *Regression analysis*. Wiley.

Timothy D. Wilson, Christopher E. Houston, Kathryn M. Etling, and Nancy Brekke. 1996. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yan Yan, Romer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 932–939, Chia Laguna Resort, Sardinia, Italy. JMLR Workshop and Conference Proceedings.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy. Association for Computational Linguistics.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU participation in WMT2015 metrics task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421, Lisbon, Portugal. Association for Computational Linguistics.

Hui Yu, Xiaofeng Wu, Jun Xie, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2014. RED: A reference dependency based MT evaluation metric. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2042–2051, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Author/s:
Mathur, Nitika

Title:
Robustness in Machine Translation Evaluation

Date:
2021

Persistent Link:
http://hdl.handle.net/11343/295770

File Description:
Final thesis file