AI-Enabled Translation Quality Assessment

Research strategy for evaluating translation quality assessment methods and validating their efficacy in the translation process

Gary Simons and Daniel Whitenack, 12 Jan 2021

Why research?

The project on AI-Enabled Translation Quality Assessment is as much a research project as it is a development project. While the primary deliverable is software tools that will address the translation checking bottleneck, we do not yet know what assessment methods are going to work well or what features need to be built into those tools. The development of those methods must therefore be driven by an approach based on research design and experimentation so that the tools are based on the methods that are proven to work best. Our end goal is not purely academic, and thus it is right to move quickly, keeping practicality and product design in mind. However, if we are not methodical in our approach to reducing our unknowns, we risk building something that will not bring value.

A major obstacle we face (even if we build something that works) is skepticism on the part of would-be users of the tools. Why should they put their trust in mysterious algorithms? Our experimentation will serve to demonstrate that the methods which emerge in the end are trustworthy, being based, at least in part, on guidance that the would-be users of the tools communicate before and during experimentation. We need to follow professional research and AI development methodologies (including peer review and shared tasks) to publish and disseminate the results of the experiments in order to substantiate the claims that will be made about the efficacy of the eventual tools.

Assessing translation quality

The purpose of this document is to explain the strategy we will be taking to evaluate and validate translation quality assessment methods in the research aspect of the project. We must first explain our vision of an AI-enabled approach to assessing translation quality. The succinct statement in the funding proposal was as follows:

Harnessing the latest artificial intelligence (AI) techniques, this project seeks to automatically and objectively assess multiple facets of translation quality and to validate the usability of those assessments with translation experts.

A key notion here is that we will tackle translation quality assessment by assessing "multiple facets of translation quality." That is, we are not treating translation quality as a single entity for which an AI algorithm will make a summary judgement. Rather, the overall quality of a translation is a judgement with respect to the many characteristics (or qualities) of a translation and how well they fit the intended audience and purpose. The same translation could be of high quality (good fit) for one purpose and of low quality (bad fit) for another. Our approach is not to use AI techniques to assign summary judgements like good and bad, but rather to assess where a translation stands on each of a variety of qualities that can be algorithmically measured. We refer to each of these measurable qualities as a *metric*, and we refer to the process of generating this metric as an assessment *method*.

There are many types of metrics and multiple candidate methods for calculating each type. We will ultimately explore many methods to discover what approach works the best. At the outset, we are working with the following taxonomy of metric types:

- *Measures of similarity* — These metrics automatically measure the similarity between a candidate translation and known reference translations.
- *Measures of readability* — These metrics estimate the reading ability needed by the reader of a candidate translation.
- *Measures of naturalness* — These metrics automatically assess the naturalness of the grammar of a candidate translation in comparison to a reference corpus that is representative of natural text.
- *Measures of comprehensibility* — These metrics assess the comprehensibility of a candidate translation based on automated answering of comprehension questions.
- *Measures of adequacy* — These metrics automatically assess the adequacy (or equivalency of conveyed information) of a candidate translation in the absence of reference translations in the target language.
- *Measures of dynamicity* — These metrics automatically assess the extent to which a candidate translation is dynamic versus literal.

The above list is not exhaustive. Other kinds of metrics are possible and it is likely that the project will develop new ones as the translation consultants involved describe additional kinds of assessment that they would like to automate. Moveover, we do NOT envision tackling all of

these kinds of metrics at once. We will begin by evaluating and validating methods in a few of these categories, and we will continue to add methods in the other categories over time.

Three stages of research

The research to prove out a candidate method will proceed in three stages. The first stage, called *evaluation*, happens in the lab. Technologists test multiple approaches and variants to discover which ones perform the best on shared datasets that include "gold standard" annotations which indicate the expected results. In the professional AI development world, these are called "shared tasks". The second and third stages, both of which are kinds of *validation*, happen in simulated real world environments and eventually in actual real world projects. Technologists must confirm that methods extend to simulated real world scenarios (in terms of available data, target languages, etc.), and translation experts must provide feedback on candidate methods and put the most promising methods to the test in real work. This validation serves to confirm that the automated results are helpful to the processes of translation and translation checking.

The evaluation process is driven by an *evaluation protocol*. The protocol defines a strategy for testing how well an implemented method performs at measuring the desired quality. The protocol is developed in advance by research designers and translation experts and involves the construction of an shared *evaluation dataset* consisting of sample data. That sample data is annotated by humans with the gold standard values for the qualities to be assessed automatically. All collaborating technologists use the shared evaluation data to experiment with methods and algorithms in search of the ones that are best at reproducing the gold standard annotations.

The validation process is similarly driven by a *validation protocol* and further shared task data. In the second overall stage, the validation protocol defines a strategy by which translation experts will confirm that a candidate method is generating valid quality metrics when applied to translations in a high-resource language that is well known to the experts. Methods that pass the second-stage validation proceed to the third and final stage, in which the proposed methods are applied to translations in existing low-resource language translations under various simulated scenarios (including for example, various levels of existing data and various language combinations with diverse characteristics). For both of these stages, research designers and translation experts develop a validation protocol in advance which specifies a mechanism for translation expert feedback and a plan for gathering and annotating shared validation data. That shared validation data should include further annotated gold standards that objectively enable researchers to understand how well the metric works in practice.

Participating in shared tasks

The first and third stages of evaluation and validation leverage the idea of "shared tasks," which have been instrumental in advancing professional AI research. Industry examples of shared tasks and corresponding methods evaluated on those shared tasks include:

- Stanford Question Answering Dataset (SQuAD 2.0)
- Various tasks and leaderboards by Allen AI
- Workshop on Machine Translation (WMT20)
- Fact Extraction and VERification (FEVER 2.0)

These shared tasks are similar in that they all provide:

- A general description of the task
- Common, gold standard evaluation and validation data sets
- A mechanism for submitting the results achieved with new methods
- Common, centrally managed evaluation scripts to process submissions
- A leaderboard showing how the various submitted methods compare to each other on a level playing field

Our own set of shared tasks (with similar components) will allow us to coordinate our work, understand what contributions already exist, build on the work of others, and objectively compare results. After working with translation experts to gather and annotate data for our various tasks, we will publish these as shared tasks on a community of practice website along with instructions for utilizing the data, submitting methods to be evaluated/validated, and a leaderboard of previously submitted methods. This community of practice website will serve as the hub of our research activities and a starting point for developing assessment methods.

To illustrate this shared task process, let's consider what an example set of shared tasks might look like for our *measure of similarity*. Remember that our methods in this category should "automatically measure the similarity between a candidate translation and known reference translations." For the first stage of evaluation we just need to determine, for a high resource language (like English), which of our methods best measures similarity in a way that is correlated to human annotations of similarity. Thus, the "gold standard" evaluation data might look like the following pairs of English texts with corresponding human annotations (numbers between 1-10) indicating similarity:

Target	Reference	Similarity
You also, be patient. Establish your hearts, for the coming of the Lord is at hand.	Meanwhile, friends, wait patiently for the Master's Arrival. You see farmers do this all the time, waiting for their valuable crops to mature, patiently letting the rain do its slow but sure work. Be patient like that. Stay steady and strong. The Master could arrive at any time.	3
And this gospel of the kingdom shall be preached in all the world for a witness unto all nations; and then shall the end come.	And this gospel of the kingdom will be preached in all the world as a witness to all the nations, and then the end will come.	9
...	...	...
Blessed is the man Who walks not in the counsel of the ungodly, Nor stands in the path of sinners, Nor sits in the seat of the scornful;	Blessed is the man who walks not in the counsel of the wicked, nor stands in the way of sinners, nor sits in the seat of scoffers;	7

This evaluation data minus the annotations would be provided on the community of practice website as part of the shared task, and it would be the job of the contributor to:

- Create a method for ranking similarity of two pieces of text on a 0-10 scale
- Evaluate the method on the evaluation data producing a set of predicted similarity labels for each of the evaluation data samples
- Submit the set of predicted similarity labels to the shared task, which will result in those predictions being compared with the gold standard labels to rank the method against other methods on the same data

In some cases, shared tasks may provide a set of training data (samples + human annotations) along with the evaluation data (just samples with annotations hidden). This scenario would promote the development of "supervised" methods. In other cases, shared tasks may only provide data without labels to promote the development of "unsupervised" methods. In either case, submissions to the shared task would create a leaderboard that ranks submissions according to predetermined measures of error. For the example shared task this might look like:

Rank	Method	Date	Average Error
1	LSDevBERT Embeddings	04/12/21	0.9
2	IDX Siamese Network	02/03/21	1.2
3	Hybrid PoS tagging	06/27/21	2.3
...	...	...	...
N	Raw BLEU score	02/01/21	4.8

This example leaderboard would represent a first stage *evaluation* of all of these methods. The individual methods listed on the community of practice website will be linked to the exact versions of code, data, and configuration that generated the results and to the corresponding natural language descriptions of the method and implementation. An experiment management system (clearML), version control system (GitHub), and wiki will help us track all of these artifacts.

To complete the second and third stages of validation, the community of practice may include other shared tasks related to the validation of methods on a selection of local languages or on a selection of data scarce scenarios. It may also include shared tasks related to the bridging of high resource language methods (e.g., word embeddings or reading comprehension models) to local language scenarios. These envisioned bridging methods include cross- or multi-lingual embedding and automated back translation.

The life cycle of research and development on a method

With the above as background, we can now describe the steps in the life cycle of working on a particular method for producing a quality metric. At the end of each description is a proposed indication of which department takes the lead and which others may also be significantly involved. The abbreviations are: IDX-DS (for the Data Science group in IDX), CRO (for the Corporate Research Office), LangTech (for Language Technology), and Trans (for Translation Services, which may also involve other departments of International Language Services in some instances).

In general, the the life cycle of researching and developing an AI-driven quality method follows these steps:

1. Preliminary Documentation
2. Shared Task Creation for Method Evaluation
3. Method Development and Evaluation
4. Structuring Validation
5. Further Development and Validation
6. Post Development Documentation

1. Preliminary Documentation

1a. Investigate literature

The process begins with a search for literature that is relevant to the kind of quality metric that is in focus. For example, what methods have researchers employed to measure similarity, comprehensibility, or naturalness in the past? What kinds of methods have proven to be successful and which of those might be most relevant to the scenarios we will encounter? A document is written which gives guidance to the technologists about what they should read and what ideas we should try. It should also identify existing code that is available, and specifies what code we should write or reuse. [Responsibility: IDX-DS, with CRO and LangTech]

1b. Explain method

A document is written which explains for our translation experts what quality metric is in focus and what methods can be employed to generate that metric. The explanation is to be aimed at a non-technologist. The purpose is to give the translation expert enough of a sense of how the method works that they can feel they have a basic understanding of what the generated metrics will be telling them. The document should also include a preview of the evaluation protocol, especially describing the kind of shared task data that may need to be created and an explanation of how it will be used in the evaluation process. The translation experts will need this kind of understanding of the method and the evaluation process in order to be able to participate in designing the evaluation and validation protocols. [Responsibility: IDX-DS, with LangTech and CRO]

2. Shared Task Creation for Method Evaluation

2a. Establish advisory team

With the above documents as a guide to understanding the method, a team of translation experts is organized and oriented to participate in the process of gathering relevant data and

annotating that gathered data as necessary for shared tasks related to evaluation. [Responsibility: Trans]

2b. Design the shared task

A protocol for evaluating methods for this quality metric using a shared task is designed, including the identification of the primary texts to be used in evaluating methods and the specification of the annotations and other data assets needed for the evaluation dataset. The other data assets commonly include some combination of: (i) human annotated data; and (ii) automatically perturbed versions of the primary texts in which an existing text is modified in specific ways to introduce labeled quality issues that a successfully implemented metric should be able to find. The protocol should specify the format for a leaderboard that tracks evaluation results, and it should include a draft of a description of the shared task that will help engage multiple teams simultaneously in friendly competition. [Responsibility: CRO, with IDX-DS and Trans]

2c. Create evaluation data

All translations and data assets specified in the evaluation protocol are pre-processed into the right format and human annotations are added as needed. Any specified perturbed texts (or computer annotations) have also been created. [Responsibility: IDX-DS, with Trans and GTIS]

2d. Add the shared task to the community of practice website

All translations and data assets are made available on the community of practice website along with the description of the shared task and instructions for making contributions. Any existing or common data pre-processing, modeling, or other tools are links to ensure that contributors are being as productive as possible. Any code used to compare contributed results to gold standard annotations is also in place and ideally automatically integrated into the submission process. [Responsibility: IDX-DS, with GTIS]

3. Method development and evaluation

The technologists develop automated methods for calculating the quality metric in focus. As they experiment with possible approaches, they are following the evaluation protocol at each iteration to record results against the evaluation data. When they are confident in a method, they submit their results (along with any related metadata and data provenance) to the shared task leaderboard on the community of practice website. In this way, contributed methods are evaluated on common ground to determine what is working the best. The code for calculating the metric and reproducing the results for each experiment is version controlled in a repository

and linked to the shared task contribution. [Responsibility: IDX-DS, LangTech, and future collaborators]

4. Structuring Validation

4a. Re-engage or establish a new advisory team

The team of translation experts from 2a is re-engaged and/or other translation experts are mustered to assist with the validation process. [Responsibility: Trans]

4b. Design validation protocol

A protocol for validating the best candidate methods is specified. This has two parts. It first defines validation tasks that translation experts will perform against high-resource translations that are understood by the expert. These tasks may include surveys and focus groups, for example. Secondly, the validation protocol must similarly define any additional shared tasks needed to confirm that methods generalize to the context of low-resource translations in which the translator understands the translation but the translation consultant does not. These shared tasks may include: (i) tasks similar to those in 2b but including low resource language translations rather than high resource language translations; and (ii) tasks similar to those in 2b but stipulating limitations related to data scarcity. [Responsibility: CRO, with IDX-DS and Trans]

4c. Create validation data and feedback mechanisms

All mechanisms for gathering translation expert feedback (e.g., surveys) are in place. All translations and data assets needed for performing the validation protocol are pre-processed into the right format and human/computer annotations are added as needed. [Responsibility: IDX-DS, with Trans and GTIS]

4d. Add the shared task(s) to the community of practice website

All translations and data assets are made available on the community of practice website along with the description of the validation shared tasks and instructions for making contributions. Any code and feedback mechanism used to gather expert feedback and compare contributed results to gold standard annotations is also in place. [Responsibility: IDX-DS, with GTIS]

5. Further development and method validation

The validation protocol is executed first against high-resource translations to gather translation expert feedback on the output of the method. This feedback should establish a baseline for the

kind of results that can be expected, and it should confirm that these kinds of results are expected to be useful in the translation process. Secondly, the validation protocol is executed using the additional shared tasks to ensure that the method extends to low-resource and local language contexts to determine how the metric performs in practice. Attempted validation of methods may reveal additional modifications that need to be made to a method. This is normal, and it is expected that contributors will need to iterate through a cycle of development, evaluation, and validation. [Responsibility: IDX-DS, LangTech, and future collaborators, with Trans]

6. Post development documentation

The evaluation and validation results for this metric are documented in a form that is accessible to the Bible translation community. [Responsibility: CRO, with Trans, LangTech and IDX-DS]

After evaluation and validation

Although we need to perform "in the lab" or "offline" evaluation and validation, we could spend an infinite amount of time in the lab over optimizing our methods. In the end, our goal is to build practical tools and product integrations that give translation teams, consultants, and administrators superpowers. We thus need to jump to tool development and live field trials after reasonable (but not overly burdensome) phases evaluation and validation.

Tools

After a particular metric has been validated, the next step within the project will be to integrate it into our own tooling and/or existing products. Three kinds of applications are envisioned:

- A set of power tools for translation consultants, which will allow them to do their checking work more thoroughly and more consistently
- Modules integrated into our translation tools (like Paratext, Scripture Forge, and Render) which alert translators to obvious problems so they can address them earlier in the translation process
- A visualization tool in the context of progress.Bible which paints the landscape of known translations by plotting their various qualities, including a way to show how specified new translations fit in against the backdrop of known and trusted translations

Field Trials / Live Translation Team Validation

We need to engage with sympathetic translation teams working on live translation to create a set of alpha users of our prototype methods and tools. In the end, these field trials will be the final confirmation of how our tooling will assist translation teams. These alpha users could include “traditional” human drafting teams, but it may also include those utilizing new technology for computer-assisted drafting.