May 2022

Metrics for Streaming Translation

Colin Cherry

Tianjun Ye

Naveen Arivazhagan

Te I

Yingjie He

See next page for additional authors

Inventor(s)
Colin Cherry, Tianjun Ye, Naveen Arivazhagan, Te I, Yingjie He, Yue Shi, Yuezhang Li, and Yuan Tian

Metrics for Streaming Translation

ABSTRACT

This disclosure describes techniques and metrics that can be utilized for evaluation of the quality of streaming translations. Per techniques of this disclosure, a time lag family of metrics and an erasure time lag family of metrics are utilized to characterize and/or compare the performance of machine translation techniques utilized in generating streaming translation. The latency of a translation session is measured by a time lag family of metrics. A time lag is the period of time by which each token in a response log lags behind a corresponding token in a query log. Erasure time lag is similar to the time lag family of metrics but changes the notion of token timing to account for output stability. Whereas time lag metrics track the timestamp of a token's first appearance, erasure time metrics track the timestamp where a token and all tokens before it became stable in the output.

KEYWORDS

- Simultaneous translation
- Streaming translation
- Translation service
- Machine translation
- Levenshtein distance
- Speech recognition
- Response latency
- Response quality
- Response stability
- Time lag

BACKGROUND

Machine translation of speech is typically performed sentence by sentence, where the speech recognition and translation from a source language to a target language occurs after a whole sentence is spoken. A complete sentence provides sufficient context to minimize

translation error. Comparison of different machine translation techniques can be made by utilizing metrics such as latency (the time taken to generate the translation) and accuracy.

Streaming translations have recently become available in various contexts. For example, a live translated transcript of a speech or video conference may be provided to participants or viewers. Further, streaming translation is available as a service from various cloud service providers. During streaming translation, partial translation results are generated and provided frequently as and when speech is detected and recognized, even prior to a complete sentence being spoken. The translated portions can sometimes change over time as larger portions of the source sentence are processed. One reason for this is structural differences between the source and target languages. Metrics designed for media translation have to take such features into account in order to accurately evaluate media translation techniques.

DESCRIPTION

This disclosure describes techniques and metrics that can be utilized for evaluation of the quality of streaming translations. Per techniques of this disclosure, a time lag family of metrics and an erasure time lag family of metrics are utilized to characterize and/or compare the performance of machine translation techniques implemented in streaming translation services. Example output from a media translation service is provided below.

Timestamp	Source_string	Target_string
000000 ms	<start_time>	<start_time>
000150 ms	Neue Arzneimittel	New medicines
000250 ms	Neue Arzneimittel könnten Eierstockkrebs	New medicines may be ovarian cancer
000400 ms	Neue Arzneimittel könnten Eierstockkrebs verlangsamen	New medicines may slow ovarian cancer

Table 1: Media translation services provide partial translations

Table 1 lists source strings and corresponding target strings generated by an example media translation service. As can be seen, at the 250 ms mark, a partial source string “Neue Arzneimittel könnten Eierstockkrebs” is translated as “New medicines may be ovarian cancer” whereas at the 400 ms mark, with the additional context obtained as the sentence gets completed, the source string “Neue Arzneimittel könnten Eierstockkrebs verlangsame” is translated as “New medicines may slow ovarian cancer.”

The user experience for streaming translation can be characterized by response latency, response quality, and response stability.

Response latency

Response latency is based on the average time taken to provide the user with quick and meaningful (e.g., non-empty, non-duplicated) partial translation results.

Response quality

Response quality can include both a final translation response quality - the translation quality after the whole sentence (or context) has been completed - as well as a partial translation

response quality. In an ideal scenario, each portion of the partial translation response appears as-is in the final response; however, in reality, as the recognition/translation service receives increased context, the partial translation response often gets adjusted accordingly.

Response stability

Response stability is based on the frequency and magnitude of the adjustment(s) made to the partial translation as the translation proceeds to the final translation output. Response stability is particularly important for evaluating streaming translation and is critical to satisfactory user experience.

For example, consider a situation where a translation service provides a 10-word partial response at a first point in time, but in the very next time interval provides a different 10-word partial response (where all 10 words are replaced by different words). Such a situation is indicative of low response stability. High magnitude(s) of adjustment of the streaming translation can lead to poor user experience and cause a loss of trust in the translation service.

As another example, consider a translation service that provides a 10-word partial response at a point in time, but provides multiple partial responses for the next few time intervals where for each partial response, the first 8 words are always the same, but where the last 2 words are changing. In this scenario, the frequency of the adjustment is high and leads to poor user experience since the user experiences a flickering screen (since the last 2 words keep changing).

Response latency, response quality, and response stability can be quantified by defining and utilizing specific metrics that are usable to measure and compare the performance of media translation services.

Time lag

The latency of a simultaneous translation session can be measured by a time lag family of metrics. A time lag is the period of time (e.g., in seconds), by which, each token (word) in a response (target) log lags behind a corresponding token in a query (source) log.

Time lag tracks the time to the first appearance of any token at a given index; that is, the timestamp for the j^{th} token is given by the state where the document first achieved a length of j tokens.

For example, in the sample log from Table 1, the fourth target token in the final document, "slow," has a timestamp of 250 ms, as that is the timestamp of the state where the fourth token ("be" at that time) first appeared. The time lag family of metrics includes multiple different time lag calculations, each of which are determined by varying the identity of the response log and query log. The time lag metrics are measured in seconds. Lower numbers are indicative of better performance.

Table 2 lists examples of time lag metrics that can be determined.

Time Lag metric type	Response Log	Query Log	Description
Source Vs Reference Source Time Lag (STL)	System source	Reference source	How many seconds does speech recognition lag behind the source speaker?
Target Vs Reference Source Time Lag (TTL)	System target	Reference source	How many seconds does machine translation lag behind the source speaker?
Target Vs Reference Target Time Lag	System target	Reference target	How many seconds does machine translation lag behind ideal streaming translation?
Target Vs Source Time Lag	System target	System source	How many seconds does machine translation lag behind speech recognition?

Table 2: Time lag family of metrics

Correspondence between tokens

Determination of time lag metrics is based on identification of corresponding tokens, across logs as well as languages. Per techniques of this disclosure, identification of corresponding tokens (both within language and across languages) assumes uniform information density and monotonic alignment within a parallel sentence pair. After adjusting for differences in sentence length, it is expected (assumed) that *n* tokens in a response sentence convey as much information as *n* tokens in the parallel query sentence. This assumption enables a definition of correspondence between tokens without introducing a dependency on a statistical word aligner.

Obtaining sentence pairs from the query and response logs can pose challenges due to unreliable sentence boundaries in logs. To facilitate accurate determination of time lag metrics, a sentence alignment that aligns reference source to reference target is included in the reference session logs. A Levenshtein projection is utilized to align the system source logs to the reference source logs, and the system target logs to the reference target logs. The time lag metrics are determined as follows.

Let *i* enumerate the response-query parallel sentences with response length r_i and query length q_i. Let *j* enumerate the tokens of a sentence, where $t_r(i, j)$ is the timestamp for the j^{th} token of response sentence *i*, and $t_q(.)$ is defined similarly for the query. The t(.) functions account for non-integer instances of *j* by interpolating between the timestamps at floor(j) and ceil(j) according to magnitude of *j*'s non-integer (fractional) part.

Time lag is defined as:

$$\frac{1}{\sum_i |r_i|} \sum_i \sum_{j=1}^{r_i} \left[t_r(i,j) - t_q\left(i, j\frac{q_i}{r_i}\right) \right]$$

Erasure time lag

Erasure time lag duplicates the time lag family of metrics, while changing the notion of token timing to account for output stability. Whereas time lag metrics track the timestamp of the first appearance of an index *j*, erasure time lags track the timestamp where the prefix ending with index *j* became stable in the output; that is, the timestamp where the token at index *j*, and all tokens before it, stopped changing.

Determination of this timestamp is made from a complete log and cannot be performed while a translation system is still running (since the outputs can continue to change). Referring to the example in Table 1, token 4 "slow" would have a timestamp of 400 ms, since that is the timestep where the prefix ending with "slow" became stable. Similarly, token 5 "ovarian" would also have a timestamp of 400 ms; even though token 5 may have stopped changing at 250 ms, but the prefix(es) ending at token 5 did not do so until 400 ms.

Table 3 lists example erasure time lag metrics that can be determined.

Erasure Time Lag metric type	Response Log	Query Log	Description
Source Vs Reference Source Erasure Time Lag (SETL)	System source	Reference source	How many seconds does stable speech recognition lag behind the source speaker?
Target Vs Reference Source Erasure Time Lag (TETL)	System target	Reference source	How many seconds does stable machine translation lag behind the source speaker?
Target Vs Reference Target Erasure Time Lag	System target	Reference target	How many seconds does stable machine translation lag behind ideal streaming translation?
Target Vs Source Erasure Time Lag	System target	System source	How many seconds does stable machine translation lag behind stable speech recognition?

Table 3: Erasure Time lag family of metrics

The time lag and erasure time lag metrics can be computed and utilized to compare the performance of different streaming translation services. The metrics provide a quantitative basis to determine the response latency, response quality, and response stability of different media translation techniques.

CONCLUSION

This disclosure describes techniques and metrics that can be utilized for evaluation of the quality of streaming translations. Per techniques of this disclosure, a time lag family of metrics and an erasure time lag family of metrics are utilized to characterize and/or compare the performance of machine translation techniques utilized in generating streaming translation. The latency of a translation session is measured by a time lag family of metrics. A time lag is the period of time by which each token in a response log lags behind a corresponding token in a query log. Erasure time lag is similar to the time lag family of metrics but changes the notion of token timing to account for output stability. Whereas time lag metrics track the timestamp of a token's first appearance, erasure time metrics track the timestamp where a token and all tokens before it became stable in the output.

REFERENCES

1. Ma, Mingbo, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang et al. "STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework." arXiv preprint arXiv:1810.08398 (2018).
2. Ma, Xutai, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. "Simuleval: An evaluation toolkit for simultaneous translation." arXiv preprint arXiv:2007.16193 (2020).

3. Iranzo-Sánchez, Javier, Jorge Civera, and Alfons Juan. "Stream-level Latency Evaluation for Simultaneous Machine Translation." arXiv preprint arXiv:2104.08817 (2021).
4. Arivazhagan, Naveen, Colin Cherry, Isabelle Te, Wolfgang Macherey, Pallavi Baljekar, and George Foster. "Re-translation strategies for long form, simultaneous, spoken language translation." In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7919-7923. IEEE, 2020.