On the Evaluation of Machine Translation n-best Lists

Jacob Bremerman[‡] **Huda Khayrallah**[§] **Douglas W. Oard**[‡] and **Matt Post**[§†]
[‡]University of Maryland, College Park
[§]Center for Language and Speech Processing, Johns Hopkins University
[†]Human Language Technology Center of Excellence, Johns Hopkins University
{jbrem,oard}@umd.edu, {huda,post}@cs.jhu.edu

Abstract

The standard machine translation evaluation framework measures the single-best output of machine translation systems. There are, however, many situations where n-best lists are needed, yet there is no established way of evaluating them. This paper establishes a framework for addressing n-best evaluation by outlining three different questions one could consider when determining how one would define a 'good' n-best list and proposing evaluation measures for each question. The first and principal contribution is an evaluation measure that characterizes the translation quality of an entire n-best list by asking whether many of the valid translations are placed near the top of the list. The second is a measure that uses gold translations with preference annotations to ask to what degree systems can produce ranked lists in preference order. The third is a measure that rewards partial matches, evaluating the closeness of the many items in an n-best list to a set of many valid references. These three perspectives make clear that having access to many references can be useful when n-best evaluation is the goal.

1 Introduction

Machine translation evaluation has traditionally focused on one-best translation results because many common use cases (translating a user manual, reading a news article, etc.) require only a single translation. There are, however, many scenarios in which n-best translation can be useful; examples include cross-language information retrieval, where query terms may not match in the single-best output, or language learning, where a learner is interested in whether their translation is acceptable.

Optimizing translation systems for such applications might benefit from evaluation measures that focus on choosing among systems based on which produces the best *list of translated sentences*, what we refer to here for brevity as an n-best list. Often in these n-best scenarios, researchers first select 'good' MT systems (i.e., by BLEU) in the hope that these good systems will also produce good results beyond the top translation candidate. In this paper we test that hypothesis, using a newly available dataset to measure the quality of n-best lists directly.

To look at the problem in this way we must first decide what properties of an n-best list we would consider 'good'. In this paper we explore three questions:

1. How well does an n-best list include correct translations and rank correct translations above incorrect ones? (Section 3: *Head-weighted Precision*)

2. How well does an n-best list rank translations in preference order, with the better (e.g., more commonly used) translations ahead of those that are valid, but less preferred? (Section 4: *Preference Correlation*)

3. How close are all of the translations in an n-best list to one or more reference translations? (Section 5: *Unweighted Partial Match*)

We introduce measures for each of the three questions, using a ranking quality measure already widely used in information retrieval for question 1, correlation measures to address question 2, and variants of BLEU for question 3. In this latter study, we particularly note that n-best evaluation done in this way contrasts with a current standard used for both n-best and 1-best MT evaluation, 1-best single-reference BLEU.

However, our purpose is not to argue for a single n-best evaluation measure, but rather to highlight that different measures produce different system rankings, and therefore it is crucial that researchers

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 60–68,
November 20, 2020.

target	weight
私は気分が良くなるだろう。	0.015
私は気分が良くなるでしょう。	0.008
私はいい気分になるだろう。	0.007
気分が良くなるだろう。	0.007
私は気分が良いだろう。	0.006

Table 1: The top five valid Japanese translations for the STAPLE prompt *i will feel well*.

source	JA	PT
Europarl (Koehn, 2005)	-	2,408k
GlobalVoices[1]	822k	1,585k
OpenSubtitles (Lison and Tiedemann, 2016)	13,097k	196,960k
Tatoeba (tatoeba.org)	1,537k	1,215k
WikiMatrix (Schwenk et al., 2019)	9,013k	45,147k
JW300 (Agić and Vulić, 2019)	34,325k	39,023k
QED (Abdelali et al., 2014)	9,064k	8,542k

Table 2: English word tokens for all datasets used to train the MT models.

carefully consider what questions to ask when evaluating systems. The measures we propose are illustrative as answers to our research questions, but are not the only solutions; many others might work. We aim to provide groundwork and encourage future work on the topic.

Our investigation is made possible by the recent availability of annotations created for the Duolingo Simultaneous Translation and Paraphrase for Language Education (STAPLE) shared task, which contains an extensive (although not necessarily exhaustive) set of valid translations for each of several thousand "input prompt" sentences (Mayhew et al., 2020).

2 The STAPLE Shared Task

The Duolingo STAPLE dataset consists of thousands of English prompts, with large sets of valid translations of each, often numbering in the hundreds, each labeled with the relative frequency with which each valid translation was selected by language learners. Table 1 shows the five highest-frequency Japanese translations for the prompt "I will feel well," where the weights of all 480 translations sum to one. As this example illustrates, the prompts are relatively short and simple sentences.

In the 2020 STAPLE task, participating systems were asked to produce all and only the valid translations. Doing well at this task, which was evaluated using a variant of the F_1 measure, requires both ranking translations well and deciding where to truncate the n-best list (i.e., the choice of n). Our focus in this paper is on ranking quality, leaving the question of how best to evaluate truncation to other work.

We compare systems from Khayrallah et al. (2020)'s submission to the 2020 Duolingo STAPLE Shared Task. They were built using the data described in Table 2. In total, we compare 38 Portuguese and 44 Japanese systems. This includes some bad systems, many good ones, and many incremental variations in between, especially at the top end. These systems ranked among the best for these languages on the STAPLE leaderboard.

All were variations of the following standard training procedure. We used Transformer architectures (Vaswani et al., 2017) trained with fairseq (Ott et al., 2019). Models included 6 encoder and decoder layers, a model size of 512, a feed forward layer size of 2048, and 8 attention heads. Models were trained with the ADAM optimizer (Kingma and Ba, 2015) with a dropout size of 0.1 and an effective batch size of 200k tokens. Model training was terminated when validation perplexity failed to increase for 10 consecutive epoch-level checkpoints.

Our systems varied in the following experimental parameters:

- Training on all the data in Table 2, or just the data above the midline.
- Whether or not we fine-tuned on Duolingo STAPLE training data.
- Training on just the first million lines of each corpus.
- Varying the effective batch size.
- Limiting the training data to sentences containing at most 20% of tokens outside the Duolingo STAPLE training data vocabulary.

3 Head-weighted Precision

We begin with our first question: how well does a system produce valid translations and rank them above invalid translations?

A task that might be more aligned with such a question would be one that necessitates a strict, binary score for validity and is agnostic to where a truncation of the list might occur. For example, a language learner hoping to learn several valid possible ways to express a sentence in a target language may want to peruse many translation outputs, starting at the top of the list. It would be unknown where the user would stop, and it would be important that the translations are fully valid.

Of course the space of valid translations in this framework could be enormous, so it is important to consider the effect of incompleteness in the set of valid references. We consider that in section 3.2, but first we introduce a measure for head-weighted precision with a simplifying assumption that the set of valid references is complete. Assuming we have this, we say for the purpose of question 1 that a good n-best list would have a lot of valid translations, and that it would place them near the head (i.e., the top) of the ranked list. We refer to this framework then as *head-weighted precision*.

3.1 A Head-Weighted Precision Measure

We are not the first to need a measure for the quality of a ranked list—this is a central question in evaluation of search engines that produced ranked lists of documents. The simplest setup of the evaluation task in information retrieval is that documents are either on topic or off topic (i.e., relevant or not), and that it is only the order of the documents that matters. One widely used rank quality measure is uninterpolated Average Precision (AP), computed for a ranked list $\mathbf{L}$ of length k as follows:

$$AP(\mathbf{L}) = \frac{1}{N} \sum_{i=1}^{k} \frac{T_i}{i} s_i \qquad (1)$$

where N is the number of valid items, T_i is the number of valid items at or above rank i, and s_i is the true binary relevance for the item at rank i (0 or 1). The core of the computation is T_i/i, which in information retrieval is called precision; AP is the expected value of precision, measured only at optimal stopping points (i.e., where precision is maximized by just having added one more valid item). With this measure, ranked lists earn perfect scores for ranking all valid items at the top and are punished for invalid items occurring between valid ones (with invalid items nearer the top having a more deleterious impact). Because variance in system behavior across conditions may be high, it is common to compare systems using Mean AP values (MAP) computed over a representative set of conditions. In information retrieval, conditions are topics, items are documents, and validity is relevance to the topic. In n-best MT, the conditions are a representative set of sentences to be translated (in the STAPLE task, the prompts), the items are system-produced translations, and validity is whether a translation is proper (i.e., present in the STAPLE gold translations).

3.2 Dealing With Incomplete Gold Data

MAP's reliance on binary validity rather than the preference order among valid translations simplified the generation of a gold standard, but the implicit assumption that the reference set of valid translations is complete is a potential concern. Due to the richness of human language, most sentences would admit an immense number of valid translations (Dreyer and Marcu, 2012). Even the STAPLE dataset used in this paper, which contains hundreds of valid reference translations for many sentences, is surely still not complete. This effect results in systems being penalized for false negatives, receiving lower MAP scores than they should.

However, when our goal is to compare systems, we are most interested in relative, not absolute, scores. So the question to be answered is whether missing data in the ground truth adversely affects comparisons between systems. Zobel (1998) introduced a clever way to characterize such an effect. The key idea is to ablate the ground truth, and to examine the effect of that ablation on system comparisons. If removing, say, half the ground truth resulted in few reversals in the preference order between systems, then one might reasonably assume that adding even more ground truth would have similarly small effects.

The art in this approach is to design the ablation in a way that removes things that are most like the things that are likely missing. Zobel, using this technique to study the stability of MAP in information retrieval test collections, ablated relevance judgments that would not have been available had less effort been devoted to generating such judgments; in information retrieval these are the documents that no participating system retrieved at high rank. For the STAPLE dataset, a natural choice is to ablate the least frequent translations, since it seems reasonable to presume that if Duolingo was not aware of the validity of some translation, that

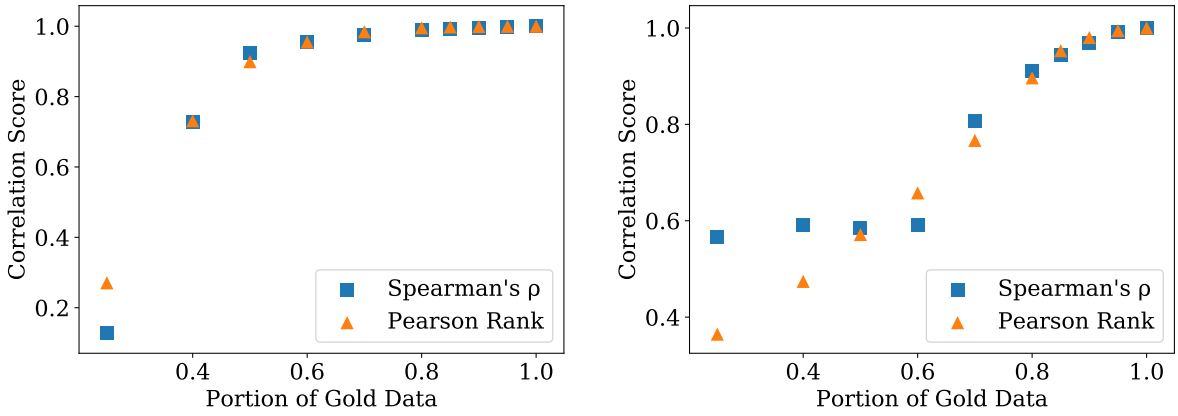

Figure 1: Spearman's ρ and Pearson Rank correlation scores for MAP system rankings for Japanese (left) and Portugese (right) under different data ablation settings. MAP results obtained for STAPLE are reliable for system ranking despite incomplete data, with results slightly more reliable for Japanese than Portuguese.

translation is likely to be rather uncommon.

Such an ablation study requires a suite of systems and a measure that characterizes the swaps between MAP scores that occur. We therefore use the aforementioned 38 Japanese MT models and 44 Portuguese MT models. From each model we generate a 1000-best list.

We can use Spearman's ρ to count the number of times the relative order of two systems is swapped. One limitation of ρ, however, is that we might care more about swaps near the top of the list of system rankings than lower down (i.e., a *head-weighted* measure). Another limitation is that we might care more about swaps between systems with very different MAP values than we do about swaps between systems with closer values (i.e., a *gap-sensitive* measure). In addition to ρ, we therefore also report Pearson Rank (Gao et al., 2016), a more recently introduced correlation measure that is head-weighted and gap-sensitive.

Figure 1 plots these correlations as progressively more common translations are ablated. The left side of the plots show how system rankings from an ablated data condition that only includes the most common translations correlate with rankings from the full data. Moving right, the correlations are compared for conditions containing more and more data, with the penultimate point representing a data condition where only the rarest translations have been removed. The flatness of the curve on the right side of the plot suggests (based on extrapolation to the right) that the presence of additional relatively uncommon translations would have been unlikely to result in many system swaps.

Pearson Rank, which accounts for head-weightedness, additionally shows that few of the system swaps are between relatively good systems. This is likely because good systems will output high-frequency translations near the top of the n-best list, so as low-frequency translations are ablated, these good systems are less likely to be affected. From this we can conclude that, at least for Japanese and Portuguese, the binarized STAPLE task ground truth is sufficiently complete to support computation of MAP scores for individual systems that can reasonably be compared, allowing us to answer Question 1 using this measure.

3.3 Comparison to the STAPLE Metric

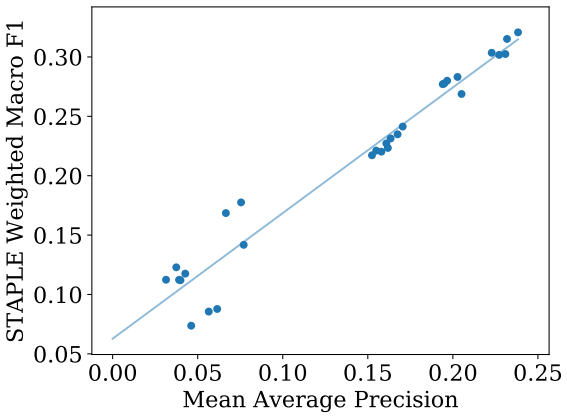

Figure 2: Correlation of system rankings based on Weighted Macro F_1 and MAP in Japanese. r^2 = 0.836, slope = 1.058.

As introduced earlier, the STAPLE task uses a weighted macro F_1 measure for evaluation. The

weighted macro F_1 measure is the same as the standard macro-averaged F_1, but recall is replaced with weighted recall. Weighted recall is calculated by using frequency weight sums provided in the gold translation data for Weighted True Positive and Weighted False Negative terms, instead of the standard raw counts.

One difference of this measure compared to MAP is that it does not evaluate a model's ability to generate n-best lists in a pure sense (agnostic to where that list may be cut off). This is because the STAPLE systems must not only generate n-best lists, but also decide where to truncate each list, in order to maximize weighted macro F_1. This means that an n-best list that outputs hundreds of valid translations only at the top of the list, but is truncated at rank 1000, would score poorly, due to precision issues. However, MAP is robust to this, since values at the very bottom of the list have only a small effect. The definition of MAP allows it to function properly for any size lists, potentially even infinite-size lists, which weighted macro F_1 does not. This is not to argue that MAP is a better measure than F_1, but simply rather that they are different measures and may be better suited for separate goals.

Despite this, we produce a correlation plot in Figure 2 to compare system MAP scores with system weighted macro F_1 scores. For these weighted macro F_1 scores, we used a thresholding technique that truncates n-best lists at a manually tuned fraction of the top hypothesis' model probability for each prompt. We find that the weighted macro F_1 values correlate very strongly with MAP.[2] We note that the correlation is particularly strong at the top-end of systems compared to the bottom-end. This is ideal since understanding and trusting evaluation measures are particularly vital for choosing among the best of systems (often not as important for choosing among the worst). From this we conclude that MAP could have been a useful formative evaluation measure when tuning n-best MT systems for the STAPLE shared task and that these two measures may actually be answering a similar question despite the differences in their properties.

4 Preference Correlation

Our second question for n-best evaluation is how well models can rank translations in preference order. Since we have model scores from the translation model and relative prevalence from the STAPLE dataset, one type of easily computed measure of quality for the model scores would be their degree of correlation with the STAPLE score for each translation (which indicates which of the translations are more commonly used; i.e., their relative prevalence).

One interesting aspect of this type of measure is that it relies on having frequency (or some other preference score) annotation information for each reference translation. Certain tasks may be better imagined to take advantage of such data. For example, a task in which models need to generate diverse translations may want to sample from valid outputs in a way that more closely reflects natural human variance. That is, it should sample a frequent translation more often than an infrequent one. Correlating a model's scores for translations with gold frequency scores may then be useful for such a case.

We consider how these Preference Correlation scores could be used for system rankings. We calculate both Spearman's ρ and Pearson's r on all of our models in Japanese and Portuguese.[3] We then construct a scatterplot of the Preference Correlation and MAP scores for each system, as shown in Figure 3. From the near-zero slope of a linear fit and the near-zero r^2 values, it is clear that both Spearman's ρ and Pearson's r are measuring something very different from MAP. That is not to say that they are not good measures; rather, it says that they are measuring something different. MAP measures how reliably systems can place valid translations early in a ranked list; correlation to the gold standard preference order measures how reliably systems can place preferred translations ahead of less preferred translations.

Both of these measures have additional limitations, which discourage their usage. First is the requirement for more nuanced gold data. While MAP only requires access to several valid translations, these measures require either a preference order or preference scores for those translations, which may be difficult to obtain. A second limitation is the handling of missing data. Both measures compare scores or rankings of translations between

[2] Spearman's $\rho = 0.956$.

[3] Spearmans's ρ considers only relative rankings, while Pearson's r additionally considers the difference in scores. Neither is head-weighted. We compute Pearson's r in log space, excluding system translations not in the STAPLE references. Another option would have been to use the model to force-decode the STAPLE references. We observe similar trends when doing so.

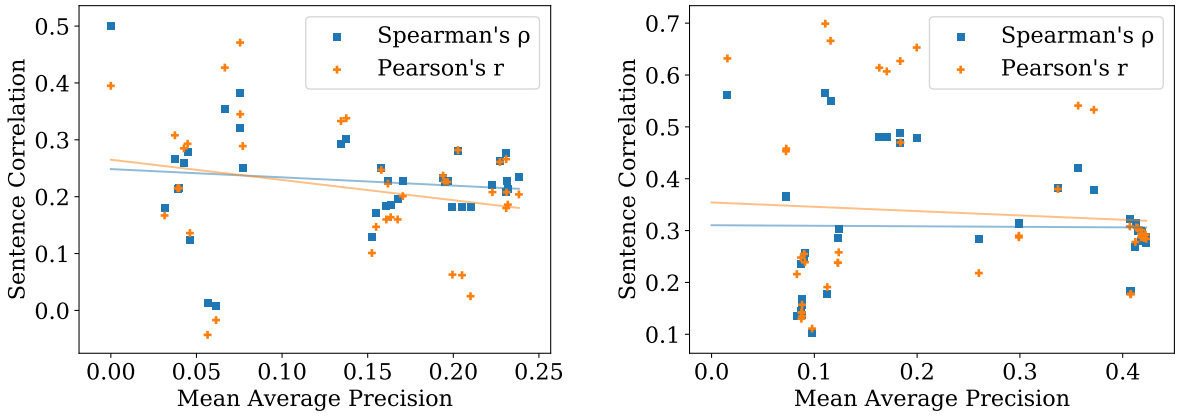

Figure 3: Correlation of system rankings based on Preference Correlation scores (Spearman's ρ and Pearson's r) vs. MAP score in Japanese (left) and Portuguese (right). Rankings do not correlate. Japanese: r^2 for Spearman's = 0.009, slope = -0.145; r^2 for Pearson's = 0.046, slope = -0.356. Portuguese: r^2 for Spearman's = -0.001, slope = -0.010; r^2 for Pearson's = -0.002, slope = -0.083. Best viewed in color.

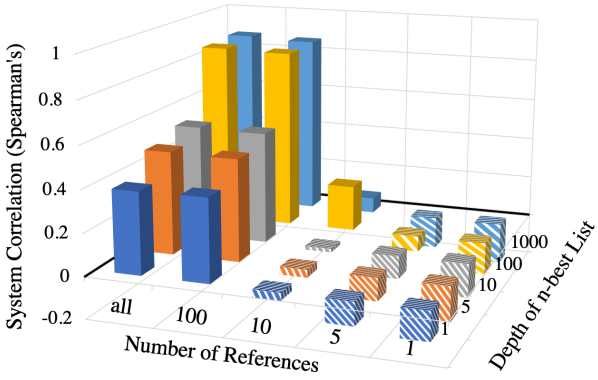

Figure 4: System ranking correlation scores (Spearman's ρ) between MAP and different configurations of BLEU (x-references and y-best outputs per source) in Japanese. Using many references and a larger depth correlates very well with MAP while 1-best 1-reference BLEU correlates poorly. Note: striped bars indicate a negative value; all references is 1536 for Japanese.

two sets. If the intersection between these two sets is very small, it limits the usefulness of the measures. Though we have many frequency scores and relative rankings in the gold translations, if the n-best lists we use to compare do not contain many of those translations, the measures could be less reliable.

5 Unweighted Partial Match

Finally, our third question is how close the translations are to the reference translations.

A task that may benefit from such a measure would be cross-language information retrieval (CLIR). In such a task, the machine translation system would serve as an upstream component of the pipeline. It is less likely for translations to need to be fully valid to be useful as compared to some other MT tasks. CLIR could benefit from combining the terms from several translation outputs regardless of if each entire sentence is perfectly valid. In this way, a measure that can assign partial credit to translations by matching n-grams as well as weighting all translations equally may be appropriate.

For this, we turn to BLEU (Papineni et al., 2002), which computes n-gram overlap between a system's translations and the available references. This raises the question of how many references we should use when we have very many available, and which of the system translations we should be using in this computation.

The STAPLE dataset provides an opportunity to explore this question. In this section, we compute BLEU measures with different numbers of references, to different depths in the n-best list. We find that at deep depths with many references BLEU ranks systems similarly to MAP, but that with fewer references its behavior is quite different.

In order to set up various configurations for our n-best BLEU measures, we perform a grid search over {1,5,10,100,1000}-best {1,5,10,100,all[4]}-reference BLEU. In what we call x-best y-reference BLEU, x refers to how many top hypotheses from the system output are used, and y refers to how many references are used. When working with multiple hypotheses in an n-best list, we simply

[4]all is up to 720 in Portuguese, and 1536 in Japanese

treat them as independent translations in a larger pseudo-corpus, pairing each with the relevant reference(s) for evaluation with BLEU.

After obtaining system scores under each of the BLEU configurations, we calculate the Spearman's ρ and Pearson Rank correlation coefficients between system rankings from BLEU compared to those from MAP. We find similar patterns for both languages and both correlation metrics, so we show Spearman's correlation for Japanese in Figure 4.

An important observation is our finding that 1-best, 1-reference BLEU does not correlate well with MAP. From this we conclude that when placing many valid translations near the top of the n-best list is important, as is the case in some applications, optimizing for 1-best 1-reference BLEU may be suboptimal.

This situation is not improved by adding hypotheses to the pseudo-corpus, so long as only a single reference is used. However, increasing the number of references does bring the correlation to moderate strength even when still only evaluating at 1-best. Once the evaluation has access to several references, evaluating deeper in the n-best list further improves the correlation with MAP, and correlations at moderately deep depths are quite substantial (e.g., $\rho = 0.86$ for 100-best 100-reference).

In Figure 5, we zoom in on two of these BLEU configurations. We choose 1-best 1-reference, which represents a standard BLEU evaluation framework, and also 100-best 100-reference, since it showed nearly the highest correlation with MAP (increasing the depth to 1000 and the references to 1000 yields only slightly stronger correlation). As the slope and r^2 of the linear fit indicate, 1-best 1-reference BLEU has little value for predicting MAP, whereas 100-best 100-reference BLEU has substantial preditive power. Moreover, this relationship is strongest for higher values of MAP; this is important, because when seeking to choose the best system for some task, a system builder would choose from among the best-performing ones.

To help explain this difference, we also performed a qualitative analysis. This revealed that systems with high 1-best 1-reference BLEU scores but low MAP produced valid translations at the top of the n-best list, but very poor translations (e.g., Latin characters in Japanese) deeper in the list. Systems with both higher MAP and 100-best 100-reference BLEU scores were better at producing reasonable sentences throughout the entire list. This makes sense as a system with a great translation at rank 1 but terrible translations between ranks 2 and 100, for example, will have a great 1-best 1-reference BLEU but will be heavily punished by MAP.

6 Discussion

Access to frequency scores in STAPLE's gold data has provided a unique chance to use correlation between those scores and model scores as a way to evaluate systems. However, we see that these measures behave quite differently from MAP, and thus we would recommend use of Preference Correlation (§4) only in cases in which fine-grained distinctions between preference scores are important for the intended application.

We also looked into using BLEU for n-best evaluation. MAP and BLEU seem like quite different ways of evaluating in terms of how they approach the problem. MAP relies on binary scores, gives no partial credit, and weighs translations at the top of the list higher. BLEU on the other hand looks at closeness, allowing for partial credit, and treats all sentences equally, no matter where they appear in the n-best list. It could be expected then that these measures would differ, as we see when comparing MAP to 1-best 1-reference BLEU. However, as we increase depth and number of references for BLEU, the correlation of the resulting system rankings increases substantially and ultimately they yield quite similar system rankings. From this we can conclude that, at least for the systems we have experimented with, and for the language learning task that the STAPLE dataset models, systems that find many good translations also tend to rank those translations well. Thus, with enough references the choice between MAP and BLEU might be made based on efficiency. We suspect that the ability of many-best many-reference to work with partial matches might give it advantages over MAP when the number of available references is more limited than in STAPLE, but we leave ablation studies to test that hypothesis to future work.

Of course, the requirement for large numbers of references, which are generally expensive to obtain, is a limitation. However, this is a separate consideration; in the Duolingo dataset that was the center of our study, they were produced organically within that task; in other settings, if evaluation of n-best lists were to be important enough, the requisite investments to create the required resources

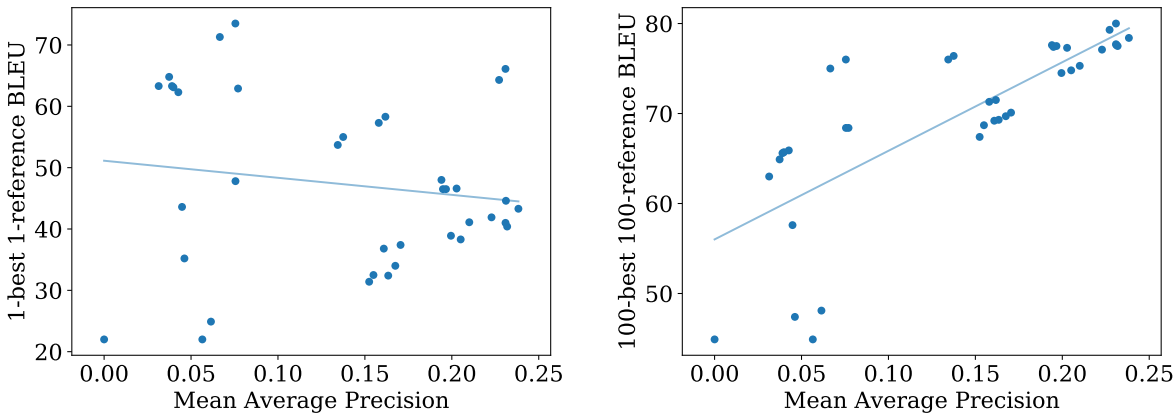

Figure 5: System scores in Japanese by MAP and BLEU (left: 1-best 1-ref, right: 100-best 100-ref). 1-best 1-reference BLEU does not correlate well with MAP ($\rho = -0.14$), but 100-best 100-reference BLEU correlates highly ($\rho = 0.86$). According to MAP, choosing the system with the highest BLEU would result in poor n-best lists in 1-best 1-ref (left) and strong n-best lists in 100-best 100-ref (right). 1-best: r^2 = 0.022, slope = -27.817 100-best: r^2 = 0.487, slope = 98.448.

could in some cases be made. In such a setting, techniques such as crowdsourcing or monolingual paraphrase generation might possibly be leveraged to reduce costs. Moreover, our ablation study indicates that the ground truth need not be completely comprehensive to be useful.

A second limitation is that our experiments were conducted on the relatively simple sentences used for the Duolingo STAPLE shared task. As with any study, it remains to be seen how well it generalizes to other settings, including other datasets. But this does not detract from our findings on the STAPLE dataset, which was after all motivated by a real language learning task that benefits large number of of people.

Perhaps our most salient general observation is that it seems that having access to more references and evaluating deeper in the list makes for better evaluation of n-best lists. Of course, this benefit must be balanced against the cost of generating the requisite number of references.

7 Conclusion

We have shown how different metrics can be used to characterize n-best list quality. In particular, we have introduced MAP as a measure for n-best list quality for machine translation systems. MAP rewards systems that place good translations near the top of the list. BLEU, computed over a pseudo-corpus built from n-best lists, and against large reference sets, ranks systems similarly to MAP. In both cases, the key distinguishing feature from typical MT system evaluation is the use of large reference sets, which yields insights unavailable with shallower evaluations using only a single reference.

MAP is but one measure among many that have been used to characterize the quality of ranked lists in other settings. As future work, we would be interested in exploring the use of measures such as inferred average precision (infAP) that are designed to be particularly robust to missing data in the gold standard (Aslam and Yilmaz, 2007), and measures such as normalized Discounted Cumulative Gain (nDCG) (Järvelin and Kekäläinen, 2002) that represent multiple degrees of utility (thus requiring more nuanced ground truth, as what we have in STAPLE).

Acknowledgments

This research has been supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies of ODNI, IARPA, or the U.S. Government.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Eval-*

uation (LREC'14), pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Javed A. Aslam and Emine Yilmaz. 2007. Inferring document relevance from incomplete information. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pages 633–642. ACM.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Ning Gao, Mossaab Bagdouri, and Douglas Oard. 2016. Pearson rank: A head-weighted gap-sensitive score-based correlation coefficient. pages 941–944.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Huda Khayrallah, Jacob Bremerman, Arya D. McCarthy, Kenton Murray, Winston Wu, and Matt Post. 2020. The JHU submission to the 2020 Duolingo shared task on simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 188–197, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the ACL Workshop on Neural Generation and Translation (WNGT)*. ACL.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Pennsylvania, USA. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *CoRR*, abs/1907.05791.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? SIGIR '98, pages 307–314, New York, NY, USA. Association for Computing Machinery.