Measuring Uncertainty in Neural Machine Translation with Similarity-Sensitive Entropy

Julius Cheng, Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{jncc3,av308}@cam.ac.uk

Abstract

Uncertainty estimation is an important diagnostic tool for statistical models, and is often used to assess the confidence of model predictions. Previous work shows that neural machine translation (NMT) is an intrinsically uncertain task where there are often multiple correct and semantically equivalent translations, and that well-trained NMT models produce good translations despite spreading probability mass among many semantically similar translations. These findings suggest that popular measures of uncertainty based on token- and sequence-level entropies which measure surface form diversity may not be good proxies of the more useful quantity of interest, semantic diversity. We propose to adapt similarity-sensitive Shannon entropy (S3E), a concept borrowed from theoretical ecology, for NMT. By demonstrating significantly improved correlation between S3E and task performance on quality estimation and named entity recall, we show that S3E is a useful framework for measuring uncertainty in NMT.

1 Introduction

Uncertainty estimation has a wide range of applications in neural machine translation (NMT), including unsupervised quality estimation (Fomicheva et al., 2020b), semi-supervised learning (Jiao et al., 2021; Wang et al., 2019), curriculum learning (Zhou et al., 2020), active learning (Zhao et al., 2020), interactive translation (Lam et al., 2018), and more. Many different measures exist for capturing uncertainty, each developed for the application at hand.

NMT is an intrinsically uncertain task, where a source sentence can have multiple correct translations which are equivalent in meaning (Stahlberg et al., 2022). Even large NMT models are known in practice to spread probability mass across a large number of translations. But the diffusive quality of the NMT distribution is not necessarily a problem in theory or in practice; in theory, the true data distribution may be diffuse, hence a perfect model will also be diffuse. In practice, high-probability translations are highly semantically similar to each other, and model probability correlates reasonably well with actual quality (Ott et al., 2018).

NMT models are generally evaluated on their ability to generate the desired semantics irrespective of lexical form (Freitag et al., 2021), hence the uncertainty measure used to assess the confidence of a model prediction should also measure semantic diversity rather than lexical diversity. NMT distributions are known to be highly diverse over surface forms, i.e. token sequences, but diversity over token sequences does not necessarily reflect semantic diversity. We therefore suspect that surface form uncertainty measures over the model distribution such as token- or sequence-level entropy would not be good measurements of model confidence compared to ones that accounts for semantic similarity across sequences.

With this motivation, we propose to adapt similarity-sensitive Shannon entropy (S3E) (Ricotta and Szeidl, 2006) to measure semantic uncertainty in NMT. S3E was originally proposed in theoretical ecology to quantify biodiversity while accounting for species similarity, but it is a general framework that permits flexibility in defining the similarity function, and thus has broad applicability beyond ecology.

We adapt S3E to NMT tasks by specifying appropriate similarity functions. We also show how the S3E framework relates to and generalizes previous work on uncertainty estimation, and present practical methods for estimating S3E efficiently and accurately for NMT. In quality estimation (QE) experiments, we estimate S3E using embeddings from models pretrained on large amounts of data, and show that this has higher correlation with translation quality than previously used similarity-insensitive uncertainty measures. Further, to illus-

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2115–2128
March 17-22, 2024 ©2024 Association for Computational Linguistics

trate the flexibility of S3E and the importance of matching the similarity function to the task, we perform a named entity recall task, where we find that the best correlation with task performance is achieved when specifying a similarity function focusing exclusively on named entities.

2 Background

2.1 Neural machine translation

In the conventional setup for conditional language generation problems such as NMT, a transformer encoder-decoder language model (LM) is trained to predict $p(y_{(t)}|y_{(<t)}, x; \theta)$, where $y_{(t)}$ is the next token, $y_{(<t)}$ is the sequence of previous tokens (the *prefix*), x is the source sentence, and θ are the model parameters. By the chain rule of probability, the probability of a sequence under the model is $p(y|x;\theta) = \prod_t^T p(y_{(t)}|y_{(<t)}, x; \theta)$. The model is trained with backpropagation and stochastic gradient descent to minimize the cross-entropy between the model prediction and the distribution of all token, prefix, and source sentence combinations in a dataset $\mathcal{D}$, equivalent to maximizing the log probability of $\mathcal{D}$.

At test time, a decision rule is used to produce an output. The typical choice for producing a high-quality output is beam search; however, reranking methods have been shown to consistently improve results, including noisy channel reranking (Yee et al., 2019), quality estimation (Fernandes et al., 2022), and minimum Bayes risk decoding (MBR) (Freitag et al., 2022).

Some applications utilize random samples from the model; unbiased samples can be generated by successively drawing tokens from the model distribution, appending them to the prefix, and repeating this process until an end-of-sequence token is reached, a procedure sometimes known as *ancestral sampling*. Quite often, the token distribution is truncated or reshaped in order to produce higher quality sequences at the expense of diversity (Meister et al., 2023).

2.2 Uncertainty and diversity

Uncertainty is an overloaded term but generally refers to the confidence of the prediction of a statistical model. For probabilistic models, it is sometimes formally defined in information-theoretic terms, such as the sequence- or token-level entropy of an LM (Malinin and Gales, 2021). Entropy in LMs can be measured over word alignment distributions (Jiao et al., 2021) or attention weights (Rikters and Fishel, 2017) instead.

Some works attempt to disentangle aleatoric uncertainty, i.e. ambiguity in the data, from epistemic uncertainty i.e. lack of knowledge of which parameters best model the data. When parameter uncertainty is modeled, for example in Monte Carlo dropout (Gal and Ghahramani, 2016), then epistemic uncertainty might be measured as the variance of some statistic over parameter settings (Malinin and Gales, 2021; Fomicheva et al., 2020b).

The term *diversity* often refers to similar concepts as uncertainty (diffuse distributions are both uncertain and diverse), but is usually but not exclusively applied when it is desirable in conjunction with quality, e.g. for open-ended tasks such as story generation (Alihosseini et al., 2019; Zhu et al., 2018a).

In our work, we limit our study of uncertainty to quantities derivable from the standard conditional distribution, e.g. $p(y|x;\theta)$ or $p(y_{(t)}|y_{(<t)}, x; \theta)$. S3E is applicable to any probabilistic model $p(y|x;\theta)$ including non-autoregressive (Xiao et al., 2023) and energy-based models (Bhattacharyya et al., 2021).

2.3 Intrinsic uncertainty in NMT

A task which has multiple correct outputs is said to have intrinsic uncertainty. In NMT, a source sentence may have multiple acceptable translations. This mostly occurs when there are multiple correct translations which are equivalent in meaning, but it can happen when the source sentence is ambiguous, such as when translating to a more highly inflected language, e.g., the source sentence may not specify number, tense, or gender which are required in the target language (Ott et al., 2018).

Stahlberg et al. (2022) show that NMT models spread probability mass across a much larger number of outputs compared to models trained on the less intrinsically uncertain task of grammatical error correction (Bryant et al., 2023). But this is not necessarily problematic for NMT distributions: high-probability outputs are highly semantically similar to each other, beam search with small beam size finds good translations on average (Ott et al., 2018), and various statistics derived from randomly sampled outputs match those of the data distribution well, which may explain the success of MBR decoding for NMT (Eikema and Aziz, 2020).

We posit that the NMT task is mostly intrinsically uncertain in the surface form of the target, but not in its semantics (excluding cases like ambiguity in the source sentence). Therefore, distributions that are diverse in surface forms are a natural outcome of NMT training, and a measure of uncertainty that captures semantic variation should better relate to prediction quality than

2.4 Shannon entropy

The Shannon entropy of a discrete probability distribution over a variable $\mathcal{Y}$ is defined as:

$$\begin{aligned}\mathcal{H}(\mathcal{Y}) &= -\sum_{y\in\mathcal{Y}} p(y)\log p(y) \\ &= -\,\mathbb{E}_{y\sim\mathcal{Y}}[\log p(y)].\end{aligned} \tag{1}$$

$-\log p(y)$ is the information content or surprisal of an event y, hence entropy is the expected surprisal over a distribution. Entropy can be computed exactly for the LM next-token distribution, but for the sequence-level distribution which has infinite support, it must be estimated. An unbiased and consistent estimator for the sequence-level Shannon entropy is the mean surprisal over samples:

$$\mathcal{H}(p(\cdot|x;\theta)) \approx \frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} \log p(y|x;\theta), \tag{2}$$

where $\mathcal{Y}$ is an array of samples drawn i.i.d. from $p(\cdot|x;\theta)$.

3 Similarity-sensitive entropy

In theoretical ecology, similarity-sensitive measures of biodiversity which allow for flexible specifications of similarity have been studied extensively (Rao, 1982; Ricotta, 2005). The similarity-sensitive Shannon entropy (S3E), originally proposed by Ricotta and Szeidl (2006) and for which Leinster (2022) provides a comprehensive treatment, is defined as:

$$\mathcal{H}_{\mathcal{S}}(\mathcal{Y}) = -\,\mathbb{E}_{y_i\sim\mathcal{Y}}\left[\log(\mathbb{E}_{y_j\sim\mathcal{Y}}[\mathcal{S}(y_i,y_j)])\right], \tag{3}$$

where $\mathcal{S}(y,y)=1$ and $0\leq\mathcal{S}(y,y')\leq 1$.

The difference between S3E and Shannon entropy (SE) is that the negative surprisal of an event is not log probability but log of the expected similarity between the outcome and all other outcomes. We call this the *similarity-sensitive surprisal* (SSS). Intuitively, outcomes are less surprising or informative if they are similar to other outcomes. We note a few desirable properties of S3E:

- SE is recovered when $\mathcal{S}(y,y') =$ equals 1 when $y=y'$ and 0 otherwise. Hence, SE is a special case of S3E with the strictest possible similarity function.
- For any given distribution, SE is the largest possible entropy in the family of S3Es.
- If $\mathcal{S}(y,y') = 1$ for all y, y' in the support, then the SSS is always 0, and thus $\mathcal{H}_{\mathcal{S}} = 0$. In other words, there is no uncertainty if all outcomes are the same.

Proofs for these properties and of all theoretical details in this work are in the Appendix.

The close relation between SE and S3E means that by comparing the two empirically, we study the impact of the choice of $\mathcal{S}$ on the usefulness of the uncertainty measure.

3.1 Estimation

Like SE, sequence-level S3E can also be estimated with Monte Carlo samples. Let $\mathbf{y}$ be a collection of samples $y_1, ..., y_n$ drawn i.i.d. from $p(\cdot|x;\theta)$. An unbiased estimator for Equation 3 is

$$-\frac{1}{n}\sum_{i=1}^{n}\log\big(\frac{1}{n-1}\sum_{j=1,i\neq j}^{n}\mathcal{S}(y_i,y_i)\big). \tag{4}$$

The inner summation is an unbiased estimator for the expected similarity of y_i because the chance of each y_i appearing in the samples list is independent of y_i, except for y_i itself, which always appears, so we exclude it. Alternately, we can incorporate the exact contribution of y_i in estimating its own expected similarity. Let $p(y_i)$ be shorthand for $p(y_i|x;\theta)$. Then we estimate S3E with:

$$\begin{aligned}&-\frac{1}{n}\sum_{y_i\in\mathbf{y}}\log\big(p(y_i)A + (1-p(y_i))B\big), \\ &A = \mathcal{S}(y_i,y_i) = 1, \\ &B = \frac{1}{|\mathbf{y}^{\neg y_i}|}\sum_{|\mathbf{y}^{\neg y_i}|}\mathcal{S}(y_i,y_i),\end{aligned} \tag{5}$$

where $|\mathbf{y}^{\neg y_i}|$ denotes the elements in $\mathbf{y}$ excluding those equal to y_i. We split the estimation similarity of y_i into two terms: $p(y_i)\mathcal{A}$ and $(1-p(y_i))B$. The first term, $p(y_i)A = p(y_i)\mathcal{S}(y_i,y_i) = p(y_i)$, is the contribution of y_i to the expected similarity. The second term is the contribution from the rest of the distribution $p(y_i|y_i\neq y_i)$. We refer to the estimators from Equations 4 and 5 as X and $\hat{X}$

respectively. Both are biased estimates of S3E but are nevertheless effective in practice. A detailed treatment can be found in the Appendix.

3.2 Similarity functions

S3E is not a single measure of uncertainty, but a class of uncertainties over choices of $\mathcal{S}$. We argue that $\mathcal{S}$ should be chosen to the reflect the desired type of uncertainty for the application. A common use case of uncertainty is to estimate the quality of a prediction. For NMT, where the quality of prediction is rated based mostly on where the prediction captures the desired semantics, the relevant uncertainty measure should reflect semantic diversity rather than lexical diversity. Hence, $\mathcal{S}$ should return the semantic similarity between two sentences.

Semantic similarity functions are largely evaluated on their correlation with human judgments. Early n-gram based metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are still widely used. These were followed by feature-based learned metrics (Stanojević and Sima'an, 2014). Today, state-of-the-art NMT metrics and task-agnostic sentence embedding models (Reimers and Gurevych, 2019; Gao et al., 2021) are all based on pretrained transformers such as BERT (Devlin et al., 2019).

These metrics vary in the type of semantic differences they measure. Masked language model training produces models that capture a wide range of linguistic features (Tenney et al., 2019) despite being trained on unlabeled data. SimCSE (Gao et al., 2021) is fine-tuned on natural language inference datasets. COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) are fine-tuned on human ratings of translation quality. In this work, we focus on BERT-based models and thereby inherit their strengths (Yenicelik et al., 2020) and weaknesses (Mickus et al., 2020) in modeling semantic similarity.

The scaling of similarity functions can be arbitrary or follow certain distributions, such as when using cosine similarity in BERT embedding spaces. For example, we find that the BERT cosine similarity between random samples from a well-trained model can consistently exceed 0.8. Even when the metric is good (in that higher values correspond to higher similarity), poor scaling can diminish the discriminative power of expected similarity. Therefore, we endow $\mathcal{S}$ with a scaling parameter α and define $\mathcal{S}_\alpha(x, x') = \mathcal{S}(x, x')^{\exp(\alpha)}$ which allows us to reshape the similarity function.

3.3 Connection to MBR

Minimum Bayes risk decoding (Goel and Byrne, 2000) is a decision rule that chooses the output with the lowest risk, or highest utility, over the model distribution. For conditional LMs, the MBR output sequence is:

$$\arg\max_{y} \mathbb{E}_{y' \sim p(\cdot|x;\theta)}[u(y, y')], \tag{6}$$

where u is some measure of text similarity. MBR is related to S3E as seen in Equation 3 in that it also uses the expected similarity of an output against other outputs from the model distribution. Kumar (2005) observes that if the utility function only returns 1 for identical inputs and 0 otherwise, this recovers the more common maximum a posteriori (MAP) objective which seeks the highest probability output. Analogously, equipping S3E with such a similarity function recovers SE. Also, MAP seeks the output with the highest probability and therefore minimum surprisal. If u satisfies the requirements for similarity function $\mathcal{S}$ defined in Section 3, then MBR seeks the output with the minimum SSS.

There are important differences between MBR and S3E. The utility function u, unlike $\mathcal{S}$, has no restriction on its range of output. More importantly, the magnitude of $\mathcal{S}$ should be comparable across different inputs. Let $\hat{u}(y, y') = u(y, y') + a$ where a is a real constant. MBR decoding with u or $\hat{u}$ has the same result, hence only relative utility is relevant for MBR. For uncertainty measurement, uncertainty scores across distributions conditioned on different source sentences must be comparable, hence the magnitude of $\mathcal{S}$ needs to be comparable across input pairs sampled from different conditional distributions.

3.4 Related work

Similarity-sensitive uncertainty and diversity measures have been considered recently in machine learning and NLP. Kuhn et al. (2023) measure uncertainty for question answering by clustering elements into meaning classes of semantically different outputs, then estimating the Shannon entropy over meaning classes. This turns out to be a special case of S3E where $\mathcal{S}(y, y') = 1$ if and only if y and y' are deemed equivalent by a textual entailment detector[1]. S3E can be seen as generalizing

[1] This is accurate of the basic definition of semantic entropy given in the work, excluding their length-normalization procedure.

their semantic entropy measure for soft similarity metrics, which are typical for many NLP tasks including NMT. Their uncertainty measure correlates with question answering accuracy better than previous methods, which supports our argument that the choice of similarity measure in S3E is application-dependent.

Friedman and Dieng (2023) propose the Vendi Score to score the diversity of a generative model. The Vendi Score is a function of the von Neumann entropy of a similarity matrix over model samples. Like our work, they define an information-theoretic measure of diversity that incorporates a user-specified similarity measure, but their goal is to measure the diversity over a very large similarity matrix over a dataset or samples from an unconditional generative model.

Fomicheva et al. (2020b) use a variety of uncertainty measures to predict NMT prediction quality. Our quality estimation experiments closely resemble theirs, as they use a lexical similarity metric like self-BLEU (Zhu et al., 2018b), which is also known outside of NLP as Rao's quadratic entropy (Rao, 1982). We extend their work in a number of ways: by relating lexical diversity to information-theoretic concepts, by using better similarity functions, and by introducing estimation and tuning methods which greatly improve correlation with translation quality.

4 Experiments

Our experiments are conducted on English-German (en-de), Estonian-English (et-en), and Nepali-English (ne-en), representing high, medium, and low resource language pairs respectively. We use pre-trained translation models for all experiments. For en-de, we use the ensemble translation model from Ng et al. (2019)[2]. For et-en and ne-en, we use the many-to-one multilingual model from Tang et al. (2021)[3]. In Sections 4.1.1 and 4.4,

Our S3E estimation procedure requires $O(n^2)$ calls to the semantic similarity function. In order to keep a reasonable runtime, we use sentence embedding models where similarity is computed with cosine distance. This way, the expensive embedding step is linear time, while only the faster cosine similarity computation is quadratic time. For German sentence embedding, we use a multilingual SBERT model[4]. For English, we use supervised SimCSE[5].

Cosine similarity ranges from -1 to 1. To create a valid S3E similarity function $\mathcal{S}$, we replace negative values with 0 following (vor der Brück and Pouly, 2019). Negative cosine similarity is rare in practice and may be caused by antonymy, which we ignore. Let f be a sequence embedding model. Then:

$$\mathcal{S}(y, y') = \max(0, \frac{f(y) \cdot f(y')}{\|f(y)\|\|f(y')\|}). \qquad (7)$$

As a baseline similarity function, we use the SacreBLEU (Post, 2018) implementation of chrF++ (Popović, 2017) with default settings and normalize the range to $[0, 1]$.

To obtain model predictions, we use beam search with beam size 5 for all language pairs. Whenever random samples are employed, we obtain 128 samples for each instance with ϵ-sampling (Hewitt et al., 2022) with $\epsilon = 0.02$, which was shown by Freitag et al. (2023) to perform well in MBR decoding.

In all experiments, we first tune the S3E similarity scaling parameter α for best performance on a validation set and only report results on the test set. We search for the optimal α over all integers in $[-1, 10]$. In Tables 1, 2, and 3, the optimal α found in validation and used in test is displayed in parentheses beside relevant results. Our code is publicly available[6].

4.1 Quality estimation

We show that similarity-sensitive diversity measures equipped with high-quality semantic similarity metrics correlate better with translation quality than the ones based on various similarity-insensitive entropies used in previous work (Zhao et al., 2020; Fomicheva et al., 2020a; Malinin and Gales, 2021).

In the first experiment, we measure the correlation between the various uncertainty measures against the quality of the model prediction as estimated by a supervised QE model. In the second one, we measure the correlation between various uncertainty measures against human judgments of quality, where the prediction comes from a different model than the one used to measure uncertainty.

[2] https://github.com/facebookresearch/fairseq/blob/main/examples/translation

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/multilingual

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

[5] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

[6] https://github.com/juliusc/s3e

	en-de		et-en		ne-en	
	ρ	r	ρ	r	ρ	r
Prediction-based						
Total token surprisal	0.370	0.205	0.261	0.150	0.402	0.339
Avg. token surprisal	0.352	0.282	0.356	0.333	0.333	0.357
Total token SE	0.218	0.089	0.180	0.078	0.326	0.250
Avg. token SE	0.244	0.251	0.248	0.196	0.242	0.211
SSS, BERT, $\alpha = 0$	0.369	0.344	0.591	0.606	0.573	0.510
SSS, BERT, best α	(5) 0.436	(4) 0.406	(3) 0.648	(4) **0.649**	(6) 0.623	(5) 0.547
Distribution-based						
Sequence SE	0.371	0.232	0.369	0.258	0.567	0.484
Avg. token surprisal	0.315	0.319	0.539	0.542	0.530	0.489
Avg. token SE	0.265	0.280	0.535	0.543	0.545	0.510
S3E, chrF++, $\alpha = 0$	0.138	0.176	0.399	0.417	0.440	0.473
S3E, chrF++, best α	(4) 0.390	(3) 0.332	(3) 0.523	(3) 0.493	(4) 0.591	(4) 0.556
S3E, BERT, $\alpha = 0$	0.304	0.303	0.543	0.568	0.562	0.569
S3E, BERT, best α	(6) **0.487**	(6) **0.424**	(5) **0.655**	(4) 0.647	(6) **0.676**	(5) **0.659**

Table 1: Spearman (ρ) and Pearson (r) correlations between the COMETKiwi score of the model prediction and uncertainty measures. Sections are divided between distribution-based vs. prediction-based measures,and similarity-sensitive vs. insensitive measures. S3E is presented with different choices of similarity function and optimized vs. unoptimized scaling parameter α.

In the former, uncertainty measures can be seen as a measure of prediction confidence. In the latter, uncertainty serves as a general QE method.

We do not expect S3E to outperform strong supervised methods such as COMET because uncertainty is a limited predictor of quality; a model can be confidently wrong or unconfidently right. However, S3E can be useful as a diagnostic tool or when supervised QE is unavailable.

4.1.1 Model confidence

We explore the performance of similarity-sensitive uncertainty measures against well-known similarity-insensitive ones. We additionally organize our measures into ones based on the *distribution* versus those based on the *prediction*. Distribution-based measures, unlike prediction-based ones, are unaware of the prediction and gather statistics over randomly sampled sequences. Here, we use S3E with two choices of similarity metric: BERT and chrF++. Our baselines are sequence-level SE, average token surprisal, and average token SE. Given an array of samples $\mathbf{y}$, sequence-level SE is the average negative log probability as in Equation 2. Average surprisal is taken over all tokens in all samples in $\mathbf{y}$:

$$-\frac{1}{|\mathbf{y}|}\sum_{y\in\mathbf{y}}\frac{1}{|y|}\sum_{t}^{|y|}\log p(y_{(t)}|y_{(<t)},x;\theta), \tag{8}$$

where $|y|$ denotes the length of sequence y. The average token entropy is computed similarly, except that the token surprisal is replaced by the SE of the token distribution at each step:

$$\frac{1}{|\mathbf{y}|}\sum_{y\in\mathbf{y}}\frac{1}{|y|}\sum_{t}^{|y|}\mathcal{H}(p(\cdot|y_{(<t)},x;\theta)). \tag{9}$$

For prediction-based measures, we use SSS with BERT. Recall that SSS is the log average similarity of a sequence over the model distribution. Our baselines are similarity-insensitive token-level measures on a prediction y: summed and averaged token-level surprisals $-\log(p(y_{(t)}|y_{(<t)},x;\theta)$, and summed/average token-level entropies $\mathcal{H}(p(\cdot|y_{(<t)},x;\theta))$.

We measure the correlation between these uncertainty measures with prediction quality, which we estimate with CometKiwi[7] (Rei et al., 2023). For en-de, et-en, and ne-en language pairs, we use the WMT22, WMT18, and FLORES (Team et al., 2022) validation and test sets respectively. The results are shown in Table 1.

Overall, we see that the choice of similarity metric and α has a large effect on the performance of S3E. S3E with BERT-based similarity outperforms chrF++ by a large margin. Tuning α greatly outperforms $\alpha = 0$ for all similarity functions. Between distribution-based and prediction-based measures,

[7] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

	en-de		et-en		ne-en	
	ρ	r	ρ	r	ρ	r
Total token surprisal	0.421	0.426	0.479	0.464	0.220	0.242
Avg. token surprisal	0.405	0.396	0.574	0.567	0.325	0.356
Total token entropy	0.079	0.084	0.270	0.272	0.128	0.115
Avg. token entropy	0.295	0.243	0.339	0.331	0.213	0.254
SSS, BERT, $\alpha = 0$	0.318	0.322	0.629	0.540	0.641	0.550
SSS, BERT, best α	(9) **0.436**	(9) **0.438**	(6) **0.720**	(7) **0.663**	(3) **0.648**	(3) **0.579**
CometKiwi	0.623	0.705	0.859	0.852	0.789	0.783

Table 2: Spearman (ρ) and Pearson (r) correlations between human direct assessment scores for a translation and 1) token-level surprisal/entropy statistics derived the translation and 2) SSS. CometKiwi performance is included for comparison against supervised QE methods.

it appears that former are generally better. This is true both of the similarity-sensitive and insensitive measures. Under suboptimal choices, S3E is not clearly better than the similarity-insensitive uncertainties, but the overall best result is S3E with BERT similarity and α tuning.

We notice that among the similarity-insensitive uncertainties, none is clearly preferable for all language pairs. The best choice may vary due to language, tokenization, training data size, or other factors, but this is currently poorly understood.

Note that the lower correlation scores in en-de are not necessarily due to poorer uncertainty estimates, but the increased difficulty of the task; en-de translation quality is consistently high, which reduces the variance of quality scores (Fomicheva et al., 2022).

4.1.2 General QE

We show that SSS correlates well with translation quality for predictions that come from other models, making it a competitive unsupervised QE metric. We use the MQLE-PE dataset (Fomicheva et al., 2022) which contains human-rated direct assessment scores of machine-generated translations. This experiment resembles the concurrent work of Naskar et al. (2023), which uses MBR utility as a quality estimator. Again, we use BERT as the similarity metrics and tune α on a held-out set.

The results are shown in Table 2. We see again that S3E with BERT and tuned α outperform all other uncertainty measures. The improvement over the baseline is very large for et-en and ne-en but small for en-de. For comparison, we show that S3E underperforms against CometKiwi, but this is to be expected since S3E is unsupervised, and CometKiwi is trained on datasets with direct assessments scores. Also, SSS as a QE metric is inherently limited in the following way: if the model distribution is highly semantically diverse, then a sentence can never have low SSS regardless of its quality. Nevertheless, we show that SSS outperforms these well-known unsupervised measures.

4.2 S3E estimator design choices

In all previous experiments, we estimate S3E with $\hat{X}$, use $n = 128$ samples, and tune α on the validation set prior to test time. We illustrate the impact of these choices here. Figure 2 shows the performance of S3E estimators X and $\hat{X}$ across choices of α and n as measured by Spearman ρ against COMETKiwi scores.

The choice of α has a significant effect for both estimators. At $\alpha = 6$, the performance drops less slowly for $\hat{X}$ when reducing n than it does for X. $\hat{X}$ has the overall highest performance for all settings of n.

4.3 Estimator variance

The performance of the various estimators in Section 4.1.1 is not due solely to the quality of the uncertainty metric, but to its estimator as well. In Figure 1, we examine the performance of these estimators by measuring the variance in rankings across random runs as well as the impact of using different numbers of samples. In this experiment, we run 4 random runs of each estimator on the en-de validation dataset given 8, 16, 32, 64, and 128 samples. We compute the average Spearman correlation ρ with COMETkiwi scores on predictions as per Section 4.1.1. To measure variance for a particular setting, we take the average ρ between prediction rankings on the full dataset from two random runs, which we call *self-ρ*. For S3E, we set $\alpha = 6$ and use both estimators X and $\hat{X}$.

S3E with X has much lower self-ρ than other

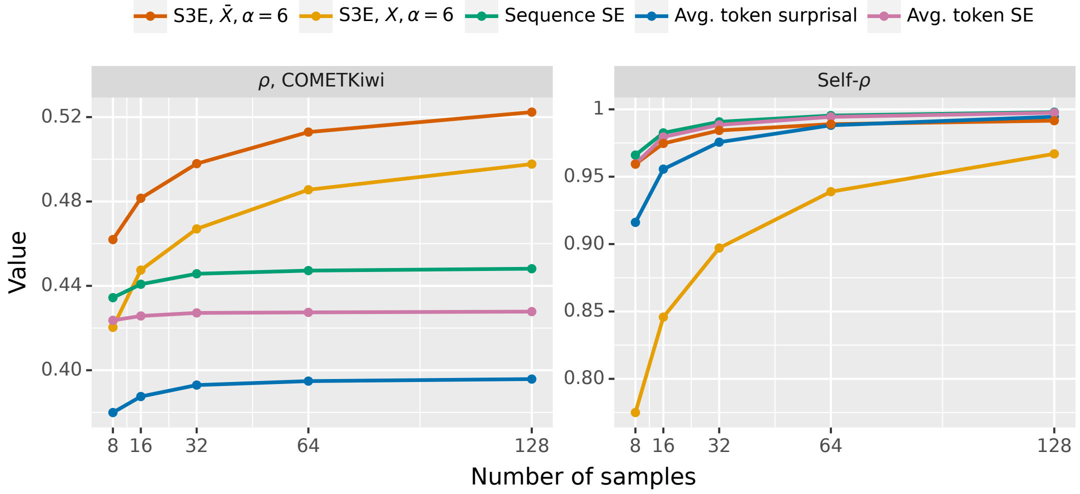

Figure 1: Given an uncertainty metric and n samples, the ρ against COMETKiwi scores (left) and self-ρ (right).

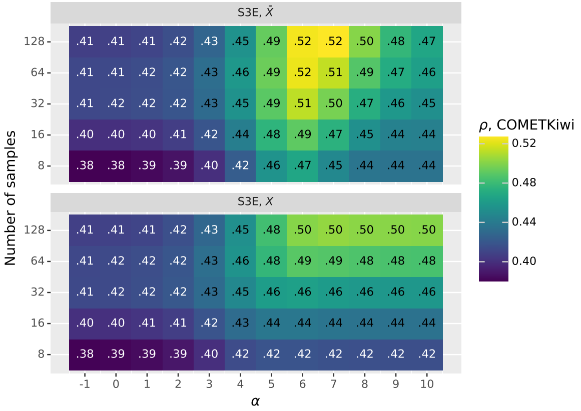

Figure 2: ρ of S3E under estimator choices (X or $\hat{X}$), scaling parameter α, and number of samples n.

estimators, and this likely explains its poor correlation with COMETKiwi scores compared to $\hat{X}$. Despite this, it outperforms the similarity-insensitive uncertainties at 32 samples, while the others do not benefit much from more than 32. S3E with $\hat{X}$ is the best performer at any number of samples and has variance comparable to the best similarity-insensitive uncertainties.

4.4 Named entity recall

To illustrate the importance of specifying the similarity function to match the task and to demonstrate the flexibility of S3E, we apply S3E to a different evaluation task: named entity recall. We use the same setup from Section 4.1.1 including predictions, samples, and datasets, but filter out instances in the validation and test sets instances which do not contain named entities.

Like in Section 4.1.1, we measure the correlation of various uncertain measures with task performance, but our evaluation metric here is the number of named entity tokens in the target sentence that occur in the prediction, or the named entity token recall (NETR). We also use NETR as a similarity function for S3E. Let $f(y)$ be the set of tokens in y which are part of a named entity. Then NETR is defined as:

$$\mathcal{S}(y, y') = \begin{cases} 1, & \text{if } |f(y)| = 0 \\ \frac{|\{y_{(t)} \in f(y) | y_{(t)} \in f(y')\}|}{|f(y)|}, & \text{otherwise}, \end{cases} \tag{10}$$

where $y_{(t)}$ is the tth token in y. This is the portion of tokens in y recalled by y'. Note that $\mathcal{S}$ here is asymmetric. Tokenization and named entity extraction are performed using spaCy[8] transformer models. We compare S3E with NETR against all uncertainty measures from Section 4.1.1. The results are shown in Table 3.

S3E with NETR significantly outperforms all other methods. S3E with BERT is more predictive than similarity insensitive uncertainties but underperforms S3E with NETR similarity by a large margin. These results further illustrate the importance of choosing an appropriate similarity function for the task.

5 Conclusion

We propose to use similarity-sensitive Shannon entropy (S3E) to measure the semantic uncertainty

[8] https://spacy.io

	et-en		ne-en	
	ρ	r	ρ	r
Shannon entropy	0.006	0.028	0.158	0.134
Avg. token surprisal	0.025	0.019	0.175	0.209
Avg. token entropy	0.019	0.025	0.196	0.233
S3E, chrF++, $\alpha = 0$	0.172		0.153	0.177
S3E, chrF++, best α	(2) 0.193	(0) 0.243	(3) 0.159	(1) 0.180
S3E, BERT, $\alpha = 0$	0.205	0.287	0.228	0.253
S3E, BERT, best α	(5) 0.239	(2) 0.296	(5) 0.256	(3) 0.274
S3E, NETR, $\alpha = 0$	0.485	0.441		0.346
S3E, NETR, best α	(1) **0.500**	(1) **0.459**	(0) **0.467**	(1) **0.375**

Table 3: Correlations between various uncertainty measures and NETR of the model prediction. Some cells are left blank to avoid displaying duplicate results.

of conditional NMT distributions. Previous work shows that NMT is an intrinsically uncertain task and that NMT model distributions in practice can vary greatly in surface without varying as much in terms of semantic content. We therefore hypothesize that S3E would outperform traditional similarity-insensitive uncertainty measures in tasks such as quality estimation for which relevant quantity is semantic diversity rather than surface form diversity.

In experiments in quality estimation and named entity recall, we show that S3E with appropriately selected similarity functions indeed correlate better with task performance than previous methods, often by large margins. We propose a sample-efficient estimator for S3E which reduces estimation variance along with a scaling parameter for similarity functions which we observe to have a significant effect on performance.

We believe that S3E is a useful framework for understanding, comparing, and developing measures of uncertainty for tasks in NLP and beyond. Important steps forward for S3E are: 1) the development of faster and/or more accurate similarity functions, 2) the application of S3E to parts of the NMT training pipeline, such as semi-supervised learning and active learning, 3) the application of S3E to other conditional language generation tasks, and 4) extensions to theory which explicitly model other sources of uncertainty, such as epistemic uncertainty.

Limitations

In this work, we use the term "semantics" in a functional sense, i.e. semantics is information that humans decode from text which is used to evaluate translation quality. We do not define semantics precisely, but doing so may provide insights on how to train similarity metrics or measure semantic similarity.

We have demonstrated that S3E and SSS are useful metrics for unsupervised QE. However, applying S3E towards QE has several additional requirements compared to simpler methods. α needs to be tuned on a validation set. Random samples are generated and embedded with advanced BERT models. S3E adds complexities compared to similarity-insensitive uncertainties which are simple functions over NMT model probabilities and require no tuning, and it may not work well when high-quality similarity functions are not available, such as for low-resource languages.

We propose the scaling parameter α which we show to have a large impact on performance in QE. In fact, Table 1 shows that for en-de, S3E with BERT and $\alpha = 0$ is worse than SE. While we have provided justification in Section 3.2 for why scaling is necessary, further understanding of scaled similarity functions is needed, and there may be better ways to apply scaling besides exponentiation.

Acknowledgements

Julius Cheng is supported by a scholarship from Huawei. The authors thank Tiago Pimentel and the anonymous reviewers for their valuable feedback.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4528–4537, Online. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

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, Minnesota. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Vishrav Chaudhary, Mark Fishel, Francisco Guzmán, and Lucia Specia. 2020a. BERGAMOT-LATTE submissions for the WMT20 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1010–1017, Online. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020b. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech Language*, 14(2):115–135.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850, Online. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Shankar Kumar. 2005. *Minimum bayes-risk techniques in automatic speech recognition and statistical machine translation*. The Johns Hopkins University.

Tsz Kin Lam, Julia Kreutzer, and Stefan Riezler. 2018. A reinforcement learning approach to interactive-predictive neural machine translation. In *European Association for Machine Translation Conferences/Workshops*.

Tom Leinster. 2022. Entropy and diversity: The axiomatic approach.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Andrey Malinin and Mark John Francis Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Clara Meister, Tiago Pimentel, Luca Malagutti, Ethan Wilcox, and Ryan Cotterell. 2023. On the efficacy of sampling adapters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1455, Toronto, Canada. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290, New York, New York. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality estimation using minimum Bayes risk. In *Proceedings of the Eighth Conference on Machine Translation*, pages 806–811, Singapore. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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, Brussels, Belgium. Association for Computational Linguistics.

C.Radhakrishna Rao. 1982. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Carlo Ricotta. 2005. Through the jungle of biological diversity. *Acta Biotheoretica*, 53:29–38.

Carlo Ricotta and Laszlo Szeidl. 2006. Towards a unifying approach to diversity measures: bridging the gap between the shannon entropy and rao's quadratic index. *Theoretical population biology*, 70(3):237—243.

Matīss Rikters and Mark Fishel. 2017. Confidence through attention. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 299–311, Nagoya Japan.

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, Online. Association for Computational Linguistics.

Felix Stahlberg, Ilia Kulikov, and Shankar Kumar. 2022. Uncertainty determines the adequacy of the mode and the tractability of decoding in sequence-to-sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8634–8645, Dublin, Ireland. 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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tim vor der Brück and Marc Pouly. 2019. Text similarity estimation based on word embeddings and matrix norms for targeted marketing. 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 1827–1836, Minneapolis, Minnesota. Association for Computational Linguistics.

Shuo Wang, Yang Liu, Chao Wang, Huanbo Luan, and Maosong Sun. 2019. Improving back-translation with uncertainty-based confidence estimation. 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 791–802, Hong Kong, China. Association for Computational Linguistics.

Yisheng Xiao, Lijun Wu, Junliang Guo, Juntao Li, Min Zhang, Tao Qin, and Tie-yan Liu. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kyra Yee, Yann Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. 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 5696–5701, Hong Kong, China. Association for Computational Linguistics.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. 2020. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1796–1806, Online. Association for Computational Linguistics.

Yikai Zhou, Baosong Yang, Derek F. Wong, Yu Wan, and Lidia S. Chao. 2020. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6934–6944, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018a. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018b. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Properties of S3E

This section provides proofs for the properties of S3E described in Section 3.

A.1 SE is a special case of S3E

SE is a special case of S3E when $\mathcal{S}(y, y')$ equals 1 when $y = y'$ and 0. Start with the definition of S3E:

$$
\begin{aligned}
& -\mathbb{E}_{y_i \sim \mathcal{Y}} \left[\log \mathbb{E}_{y_j \sim p(y_j)} \left[\mathcal{S}(y_i, y_j) \right] \right] && \text{(11a)} \\
&= -\mathbb{E}_{y_i \sim \mathcal{Y}} \Big[\log \big((p(y_i)\mathcal{S}(y_i, y_i) + \\
&\quad (1 - p(y_i)) E_{y_j \sim \mathcal{Y}|y_j \neq y_i} [\mathcal{S}(y_i, y_j)] \big) \Big] && \text{(11b)} \\
&= -\mathbb{E}_{y_i \sim \mathcal{Y}} [\log \left(p(y_i)\mathcal{S}(y_i, y_i) \right)] && \text{(11c)} \\
&= -\mathbb{E}_{y_i \sim \mathcal{Y}} [\log p(y_i)]. && \text{(11d)}
\end{aligned}
$$

(11b) is the definition of S3E. (11b) splits the expected similarity over $p(y_j)$ to the contribution of event where $y_j = y_i$ versus where $y_j \neq y_i$. (11c) follows because $\mathcal{S}(y_i, y_j) = 0$ where $y_j \neq y_i$. (11d) follows from the definition of S, and SE is recovered.

A.2 SE is the largest possible S3E

For any given distribution, SE is the largest possible entropy in the family of S3Es. Suppose there is some S' that results in a larger S3E than SE for some distribution $p(y)$. Such an S' would need to result in a larger SSS than $\mathcal{S}(y)$ for some y. Take a definition of SSS derived in a similar method as Line 11b:

$$
\begin{aligned}
& -\log \big(p(y)\mathcal{S}(y, y) + \\
&\quad (1 - p(y)) E_{y' \sim p(y'|y' \neq y)} \mathcal{S}(y, y') \big). \quad \text{(12)}
\end{aligned}
$$

In order for S' to result in a larger SSS, $S'(y, y') < \mathcal{S}(y, y')$ for some y, y'. Then it is the case that $\mathcal{S}(y, y') < 1$ when $y' = y$ or $\mathcal{S}(y, y') < 0$ when $y' \neq y$, either of which violates the definition of S3E similarity functions. A contradiction is reached, so the initial claim is proven.

A.3 Zero entropy condition

If $\mathcal{S}(y, y') = 1$ for all y, y' in the support of distribution $p(y)$, then $\mathcal{H}_{\mathcal{S}}(p) = 0$:

$$
\begin{aligned}
& -\mathbb{E}_{y_i \sim p(y_i)} \log \mathbb{E}_{y_j \sim p(y_j)} \left[\mathcal{S}(y_i, y_j) \right] \\
&= -\mathbb{E}_{y_i \sim p(y_i)} \log 1 = 0
\end{aligned}
$$

B Properties of S3E estimators

Recall the S3E estimators in Equations 4 and 5, which we call X and $\hat{X}$. Let $\mathbf{y}$ be random collection n of samples drawn from $p(y)$, and let $\mathbf{y}^{y_i}, \mathbf{y}^{\neg y_i}$ denote $\mathbf{y}$ only including or excluding elements in $\mathbf{y}$ equal to y_i.

$$
X = -\frac{1}{n} \sum_{y_i \in \mathbf{y}} \log \left(\frac{1}{n-1} \sum_{y_j \in \mathbf{y}, i \neq j} \mathcal{S}(y_i, y_i) \right) \quad \text{(13)}
$$

$$
\begin{aligned}
\hat{X} &= -\frac{1}{n} \sum_{y_i \in \mathbf{y}} \log \Bigg(p(y_i) + && \text{(14)} \\
&\quad (1 - p(y_i)) \frac{1}{|\mathbf{y}^{\neg y_i}|} \sum_{y_j \in \mathbf{y}^{\neg y_i}} \mathcal{S}(y_i, y_i) \Bigg). && \text{(15)}
\end{aligned}
$$

B.1 Bias

Let $\mathbf{S}, \hat{\mathbf{S}}$ refer to the average similarity estimators (the quantity inside the log functions of the above) for X and $\hat{X}$, and let y_i be the element for which the average similarity is estimated. From on here onwards, for simplicity, let $\mathbf{y}$ a different set of i.i.d. samples than the one y_i was drawn from. X is the sample mean and is clearly unbiased. To check the unbiasedness of $\hat{X}$:

$$
\begin{aligned}
\mathbb{E}[\mathbf{S}] &\overset{?}{=} \sum_{y_j \in \mathbf{y}} p(y_j)\mathcal{S}(y_i, y_j) && \text{(16a)} \\
&= p(y_i) + (1 - p(y_i)) \\
&\quad \mathbb{E}_{\mathbf{y}^{\neg y_i}} \left[\sum_{y_j \in \mathbf{y}^{\neg y_i}} \frac{|\mathbf{y}^{y_j}|}{|\mathbf{y}^{\neg y_i}|} \mathcal{S}(y_i, y_j) \right] && \text{(16b)} \\
&= p(y_i) + \mathbb{E}_{\mathbf{y}^{\neg y_i}} \left[\sum_{y_j \in \mathbf{y}^{\neg y}} p(y_j)\mathcal{S}(y_i, y_j) \right]. && \text{(16c)}
\end{aligned}
$$

(16c) uses the fact that $|\mathbf{y_j}|$ given $|\mathbf{y}^{\neg y_i}|$ is a binomial distribution with mean $p(y_j)/(1-(p_i))|\mathbf{y}^{\neg y_i}|$. This form appears to be an unbiased estimate of average similarity, except that $|\mathbf{y}^{\neg y_i}|$ can be 0 with probability $p(y_i)^{|\mathbf{y_i}^{\neg y}|}$, and is undefined above. In practice, we use the $p(y_i)\mathcal{S}(y_i, y_i)$ as that sample value in that case, but this results in a bias. A simple correction can be applied, but $|\mathbf{y}^{\neg y_i}| = 0$ is an extremely rare event in practice. Alternatively, if we relaxed $\hat{X}$ by guaranteeing nonzero samples drawn from $p(y_j|y_j \neq y_i)$, then it would clearly be unbiased.

While $\mathbf{S}$ is and $\hat{\mathbf{S}}$ can be turned into an unbiased estimator of similarity, $X, \hat{X}$ are biased estimators due to the log function. Due to Jensen's inequality and the concavity of logarithms, $\log(\mathbb{E}[\mathbf{S}]) \geq \mathbb{E}[\log(\mathbf{S})]$, so these estimators underestimate the log similarity on average. We leave analysis of this source of bias and its impacts on the performance of S3E to future work.

B.2 Error

Supposing again that instead of $\mathbf{y}$, $\hat{S}$ used $\hat{\mathbf{y}}$ which guarantees n samples drawn from $y_j \in \mathcal{Y} | y_j \neq y_i$. In this case, it is easy to show that this modified $\hat{\mathbf{S}}$ has lower mean squared error (MSE) than $\mathbf{S}$. For two unbiased estimators, the difference in their MSE is just the difference of variance:

$$(\mathbb{E}[\mathbf{S}^2] - \mathbb{E}[\mathbf{S}]^2) - (\mathbb{E}[\hat{\mathbf{S}}^2] - \mathbb{E}[\hat{\mathbf{S}}]^2) \geq 0 \quad (17)$$

$$\mathbb{E}[\mathbf{S}^2] \geq \mathbb{E}[\hat{\mathbf{S}}^2]. \quad (18)$$

Expanding $\mathbb{E}[\mathbf{S}^2]$, we obtain:

$$\mathbb{E}_{\mathbf{y}}\left[\left(\sum_{y_j \in \mathcal{Y}} \frac{|\mathbf{y}^{y_j}|}{|\mathbf{y}|} \mathcal{S}(y_i, y_j)\right)^2\right] \quad (19a)$$

$$\sum_{y_j \in \mathcal{Y}} \sum_{y_k \in \mathcal{Y}} \left(\mathcal{S}(y_i, y_j)\mathcal{S}(y_i, y_k) \frac{\mathbb{E}_{\mathbf{y}}\left[|\mathbf{y}^{y_j}||\mathbf{y}^{y_k}|\right]}{|\mathbf{y}|^2}\right) \quad (19b)$$

Expanding $\mathbb{E}[\hat{\mathbf{S}}^2]$ similarly, we arrive at:

$$\sum_{y_j \in \mathcal{Y}} \sum_{y_k \in \mathcal{Y}} \left(\mathcal{S}(y_i, y_j)\mathcal{S}(y_i, y_k) \frac{\mathbb{E}_{\hat{\mathbf{y}}}[\alpha_j \alpha_k]}{|\hat{\mathbf{y}}|^2}\right), \quad (20a)$$

where α_j is $p(y_i)|\mathbf{y}|$ if $y_j = y_i$, or $(1 - p(y_j))|\mathbf{y}^{y_j}|$ otherwise. To show that $\hat{\mathbf{S}}$ has lower variance, $\mathbf{S}$, it suffices to show that the individual terms in (20a) are smaller than those in (19b). When subtracting, the $\mathcal{S}$ and $|\mathbf{y}|^2$ cancel out, then we can show that $\mathbb{E}_{\hat{\mathbf{y}}}[\alpha_j \alpha_k] \leq \mathbb{E}_{\mathbf{y}}\left[|\mathbf{y}^{y_j}||\mathbf{y}^{y_k}|\right]$ for all y_j, y_k. The remaining derivation is straightforward but lengthy, so we omit it.

We have shown that a simplified version of $\hat{\mathbf{S}}$ which always uses n samples $\hat{\mathbf{y}}$ has lower MSE than $\mathbf{S}$. For the version that of $\hat{\mathbf{S}}$ we presented, a proof in either direction is challenging, owing to the facts that 1) $\hat{\mathbf{S}}$ is biased, as stated earlier, and 2) $\hat{\mathbf{S}}$ uses no more samples that $\mathcal{S}$, which increases the variance for similarity contribution estimates of elements $y_j \neq y_i$. We leave such a proof to future work, meanwhile our empirical results show that $\hat{\mathbf{S}}$ is the overall better estimator for our tasks.