## **Distributional Properties of Subword Regularization**

Marco Cognetta Tokyo Institute of Technology cognetta.marco@gmail.com Vilém Zouhar ETH Zürich vzouhar@inf.ethz.ch

Naoaki Okazaki Tokyo Institute of Technology okazaki@c.titech.ac.jp

#### Abstract

Subword regularization, used widely in NLP, improves model performance by reducing the dependency on exact tokenizations, augmenting the training corpus, and exposing the model to more unique contexts during training. BPE and MaxMatch, two popular subword tokenization schemes, have stochastic dropout regularization variants. However, there has not been an analysis of the distributions formed by them. We show that these stochastic variants are heavily biased towards a small set of tokenizations per word. If the benefits of subword regularization are as mentioned, we hypothesize that biasedness artificially limits the effectiveness of these schemes. Thus, we propose an algorithm to uniformly sample tokenizations that we use as a drop-in replacement for the stochastic aspects of existing tokenizers, and find that it improves machine translation quality.

#### **1** Introduction

Tokenization is the first stage in almost all natural language processing pipelines, where raw text is transformed into a format that is understood by the model. Modern neural models use subword tokenization, which represents text as a sequence of subword units drawn from a subword vocabulary (e.g., decompositional  $\rightarrow$  de composition al). Popular subword tokenization schemes are BPE (Sennrich et al., 2016), MaxMatch/WordPiece (Wu et al., 2016), and UnigramLM (Kudo, 2018). Unintentionally, the downstream models are thus not conditioned on the raw text, but rather the exact tokenization of the text. During training, subword regularization (where static tokenizations are replaced with sampled tokenizations) is often used to break the dependency on the exact tokenization. It also serves as data augmentation, and improves performance in a variety of downstream tasks.

| <b>BPE-Dropout</b> $p = 0.1$ |         | MaxMatch-Dropout | p = 0.3 |
|------------------------------|---------|------------------|---------|
| to ken ization               | 97.77%  | to ken ization   | 34.29%  |
| to ke n ization              | 1.89%   | t oken ization   | 14.66%  |
| to k en ization              | 0.25%   | to ke n ization  | 10.48%  |
| to ken iz ation              | 0.04%   | to ken iz ation  | 7.21%   |
| t oken ization               | 0.03%   | t oke n ization  | 4.39%   |
| to k en iz ation             | 0.01%   | to k en ization  | 3.15%   |
| to ke n iz ation             | 0.01%   | t oken iz ation  | 3.05%   |
| to ken i z ation <           | < 0.01% | to ke n iz ation | 2.14%   |

Example 1: The most frequently observed tokenizations of the word tokenization and their empirical frequencies with BPE-Dropout and MaxMatch-Dropout.

There are two main types of stochastic tokenizers: those which learn a distribution from text (e.g., UnigramLM) and those which inject randomness by corrupting the tokenization scheme (e.g., BPE-Dropout, Provilkov et al., 2020 and MaxMatch-Dropout, Hiraoka, 2022). In our work, we focus on the latter, for which no prior study of the resulting distributions exists. BPE- and MaxMatch-Dropout add randomness *post hoc* into the underlying deterministic tokenization, and the distributions they produce are essentially unrelated to the text distribution. We find that these distributions are heavily biased, in that they do not produce uniform tokenization distributions (see Example 1).

While BPE- and MaxMatch-Dropout empirically perform well in practice, there is no reason to believe they would be the best distribution for improving model performance, especially because their distributions are unrelated to the underlying corpus statistics. However, there are reasons to believe that a different strategy, uniform sampling, would be *better* for training, as it would increase the amount of regularization and augmentation injected into the training process. We experiment with replacing the stochastic aspects of BPE- and MaxMatch-Dropout with one which samples uniformly at random from all possible tokenizations, and find that it improves modeling quality on several translation tasks.

<sup>0</sup>Our code is available at github.com/mcognetta/ distributional-properties-of-subword-regularization.

## 2 Motivation

Though stochastic tokenization is known to improve model quality, it remains unclear which tokenization distribution is the best. BPE- and MaxMatch-Dropout, which induce *unlearned* distributions (probabilities that are not chosen by a learning algorithm), are a natural way of injecting randomness into the underlying tokenization algorithm *post hoc*. However, empirically, we see that they both induce heavily *biased* distributions, and hypothesize that an *unbiased* stochastic tokenizer would be universally better. This hypothesis is based on three explanations for subword regularization's effectiveness:

1) **Regularization** Subword regularization regularizes the model by breaking the dependency on a single, canonical tokenization. As shown in Example 1, BPE- and MaxMatch-Dropout allocate most of their probability mass to only a few tokenizations for a given input. A tokenizer that uniformly samples from the distribution will expose the model to a greater variety of unique tokenizations of the same input text during training.

**2)** Augmentation Subword regularization acts as data augmentation by increasing the number of unique inputs that are seen during model training. An unbiased tokenization sampler will produce more unique tokenizations of the same input than a biased sampler.

**3) Efficiency** Subword regularization increases the tokenizer's *efficiency* in the information-theoretic sense,<sup>1</sup> which is a quality shown to be well correlated with downstream task performance (Gutierrez-Vasques et al., 2021; Zouhar et al., 2023). A tokenizer with unbiased sampling will generally have higher efficiency than a biased one.

## 3 Subword Tokenization

Subword tokenizers are typically deterministic in that the same character sequence will result in the same tokenized output sequence. *Stochastic* variants were developed to allow for sampling tokenizations, which has been shown to improve model quality and robustness in a variety of NLP tasks. We briefly introduce three common subword tokenization schemes and their stochastic variants, which all share the same formalization.

| <b>Inputs</b> : Word $w \in \Sigma^+$ , (Ordered) Merg                  | ges $\mu$  |
|---|--|
| <b>Output</b> : Tokenized sequence $t \in \mathcal{V}^+$                |  |
| 1: $\varphi \leftarrow \langle (w_i, w_{i+1})   (w_i, w_{i+1}) \in \mu$ | $\wedge \operatorname{Rand}() > p \rangle$           |
| 2: for $\varphi \neq \emptyset$ do                                      |  |
| 3: $(x, y) \leftarrow \arg \max_{\mu} \varphi$                          | $\triangleright$ Ordered by $\mu$                    |
| 4: $w \leftarrow \text{REPLACE}((x, y) \rightarrow xy, w)$              |  |
| 5: $\varphi \leftarrow \langle (w_i, w_{i+1})   (w_i, w_{i+1}) \in$     | $\equiv \mu \land \operatorname{Rand}() > p \rangle$ |
| 6: <b>return</b> <i>w</i>   |  |
|   | · 1 1 · · ·  |

| Algorithm 1 | <b>BPE</b> Inference | (with dropout) |
|-------------|----------------------|----------------|
|-------------|----------------------|----------------|

**Formalization.** Deterministic tokenization maps words  $w \in \Sigma^+$  to a sequence of subwords from a finite vocabulary  $\Sigma \subseteq \mathcal{V} \subset \Sigma^*$  as  $t(w) \in \mathcal{V}^+$ . An important part of tokenization is that it is lossless—a tokenization of an input can be inverted to recover the original word. For example,  $t(\text{tokenization}) = \text{to ken ization but impor$  $tantly } t^{-1}(\text{to ken ization}) = \text{tokenization}.$ 

In contrast, stochastic tokenization is not a oneto-one mapping, but rather a probability distribution function  $T_w$  for each word w. This assigns each tokenization  $\bar{w}$  a probability  $T_w(\bar{w}) \in [0, 1]$ . Continuing Example 1,  $T_w(\text{to ken ization}) =$ 0.98 and  $T_w(\text{to ke n ization}) = 0.02$ . During application of the tokenizer, the specific tokenization of w is sampled from the distribution  $T_w$ .

**BPE and Dropout.** BPE forms a vocabulary by iteratively merging the most frequently cooccurring pair of tokens in the corpus (Sennrich et al., 2016), see Appendix B, Algorithm 4. This forms a vocabulary  $\mathcal{V}$  and an ordered list of merges  $\mu$ . During inference, the merges are applied greedily to the input (Algorithm 1). Until no more merges are available, the highest priority available merge is found and applied to the sequence. To implement dropout, a probability p is introduced, and the highlighted statement randomly removes candidate merges during each iteration (Provilkov et al., 2020).

**MaxMatch and Dropout.** Given a subword vocabulary,<sup>2</sup> MaxMatch tokenizes text from left to right by iteratively selecting the longest matching subword, shown in Algorithm 2. MaxMatch-Dropout randomly discards matching subwords and falls back to shorter ones via the condition on Line 6 (Hiraoka, 2022).

**UnigramLM.** UnigramLM (Kudo, 2018) introduced the concept of subword regularization. It learns a vocabulary and unigram probabilities for each token in the vocabulary according to some

<sup>&</sup>lt;sup>1</sup>Rényi efficiency is defined as  $H_{\alpha}(p_{\mathcal{V}})/\log(|\mathcal{V}|)$ , where  $H_{\alpha}$  is Rényi entropy,  $\mathcal{V}$  is a subword vocabulary, and  $p_{\mathcal{V}}(w)$  is the unigram probability of subword in the tokenized corpus.

<sup>&</sup>lt;sup>2</sup>We use the standard WordPiece training algorithm as described by (Schuster and Nakajima, 2012).

```
Inputs: Word w \in \Sigma^+, Vocabulary \mathcal{V}
Output: Tokenized sequence t \in \mathcal{V}^+
 1: t \leftarrow []
 2: i \leftarrow 1
 3: while i \leq |w| do
 4:
            z \leftarrow w_i
            for j \in 1 \dots \max_{v \in \mathcal{V}} |v| do
 5:
                  if w_{i \dots i+j} \in \mathcal{V} \land \text{RAND}() > p then
 6:
 7:
                        z \leftarrow w_{i \dots i+i}
            t \xleftarrow{\text{append}}
 8:
                         ~
 9:
            i \leftarrow i + |z|
10: return t
```

Algorithm 2: MaxMatch Inference (with dropout)

loss function over the training corpus. The tokenization of an input text is sampled from the probability distribution induced by the model and some temperature  $\alpha$ .<sup>3</sup> We do not explore the distributional properties of UnigramLM here, since they are highly corpus dependant.

**Dropout Distributions.** Example 1 and Appendix D show empirical probabilities for certain values of p in BPE- and MaxMatch-Dropout. While BPE- and MaxMatch-Dropout were not designed to form (or even claimed to be) unbiased distributions, here we concretely show that their distributions are biased, under mild conditions.

**Lemma 3.1.** Let  $\mathcal{B} = (\mathcal{V}, \mu)$  be a BPE tokenizer such that there exists  $(a, b), (b, b), (b, c) \in \mu$  with  $(a, b) >_{\mu} (b, b) >_{\mu} (b, c)$  and  $abb, bbc, abbc \notin \mathcal{V}$ . Then, there exists a word  $w \in \Sigma^+$  for which the distribution of the dropout tokenizer  $\mathcal{B}'(w)$  is nonuniform for any p.

*Proof.* Since  $abb, bbc, abbc \notin \mathcal{V}$ , then  $(ab, b), (a, bb), (b, bc) \notin \mu$ . Consider the word abbc. There are 5 possible tokenizations [a, b, b, c], [a, b, bc], [a, bb, c], [ab, b, c], [ab, bc]. We proceed by case analysis and compute the probability of each, given a dropout probability p.

- case  $[a, b, b, c] \rightarrow p^3$ Merges (a, b), (b, b), and (b, c) must be dropped which has probability  $p \times p \times p = p^3$ .
- case  $[a, b, bc] \rightarrow p^2(1-p)$ Merges (a, b), and (b, b) must be dropped, but not (b, c).
- case  $[a, bb, c] \rightarrow p(1-p)$ Merge (a, b) must be dropped but not (b, b). Merge (b, c) is irrelevant as  $(b, b) >_{\mu} (b, c)$ .

case  $[ab, b, c] \rightarrow (1-p)p$ case  $[ab, bc] \rightarrow (1-p)^2$ 

To be uniform,  $p^3 = p^2(1-p) = p(1-p) = (1-p)^2 = \frac{1}{5}$ , which does not exist. Hence, there is no p such that  $\mathcal{B}'(abbc)$  is uniform.

**Lemma 3.2.** Let  $\mathcal{M}$  be a MaxMatch tokenizer over vocabulary  $\mathcal{V}$ , such that  $\Sigma \subset \mathcal{V}$  and there exists a token  $v \in \mathcal{V} \setminus \Sigma$  which is a proper prefix of some other token  $z = vy \in \mathcal{V}$ . Then, there exists a word  $w \in \Sigma^+$  for which the distribution of the dropout tokenizer  $\mathcal{M}'(w)$  is non-uniform for any p.

*Proof.* Consider the distribution of tokenizations  $\mathcal{M}'(z)$ , under which the probability of z being the final tokenization is (1 - p). Let the total number of tokenizations be n, and assume the distribution is unbiased. The probability of the tokenization  $[v, y_1, y_2, \ldots, y_k]$  is  $(1 - p)p^k$ . Thus, the distribution is only unbiased if  $(1 - p) = (1 - p)p^k = \frac{1}{n}$ . Since there are at least 3 tokenizations  $[z], [z_1, z_2, \ldots, z_n]$ , and  $[v, y_1, y_2, \ldots, y_k]$ , this is a contradiction.

A consequence of Lemmas 3.1 and 3.2 is that it is not even possible to form an unbiased sampler from BPE- or MaxMatch-Dropout by simply picking a different p for each input (contrary to the typical implementation, which uses a fixed p for all inputs), and so another approach must be used.



(a) An automaton  $\mathcal{A}$  representing ababc.



(b) A transducer  $\mathcal{T}$  for the subword vocabulary {a, b, c, ab, #a, #b, #c, #ab, #bc}.



(c) A lattice, A ∘ T, of all possible tokenizations of ababc.
 Figure 1: Uniformly sampling tokenizations from A ∘ T.

<sup>&</sup>lt;sup>3</sup>Note that  $\alpha = 0$  yields the uniform distribution, but this would sample a tokenization of the entire sentence, rather than word-by-word, and has been shown to harm model quality (Kudo 2018; Figure 1).

#### 4 Uniformly Sampling Tokenizations

Given a subword vocabulary, we first produce a character-to-subword finite-state transducer representing it. Encoding an input word as a linear finite-state automaton and composing it with this transducer produces a lattice which encodes all possible tokenizations. Since the word length is finite, this lattice must be acyclic, and we can sample paths from it using Algorithm 6 (Lavrov, 2018). An example of the transducer construction, composition, and sampling is shown in Figure 1.

| <b>Inputs</b> : Word $w$ , Tokenizer $T$ , Probability $p$<br><b>Output</b> : Tokenized sequence $t \in \mathcal{V}^+$ |  |  |  |
|--|--|--|--|
| 1: if RAND() $\leq p$ then<br>2: return UNIFORMSAMPLE(w)<br>3: else  |  |  |  |
| 4: return $T(w)$   |  |  |  |
| Algorithm 3: Uniform Sampling Tokenization.  |  |  |  |

Given a baseline BPE or MaxMatch tokenizer, we implement our uniform sampling tokenizer by constructing a subword transducer from its subword vocabulary and selecting a dropout probability p. During training, a word is tokenized via uniform sampling with probability p and via the deterministic tokenizer with probability 1 - p, as shown in Algorithm 3.

One of the reasons for the success of subword regularization is that they expose the model to a more diverse set of tokenizations (Section 2). Figure 2 shows that across any choice of p, even with far fewer samples, a much more diverse set of tokenizations for a given word is observed when using Uniform Sampling compared to Dropout, indicating that a model will be exposed to a far greater number of unique contexts during training.



Figure 2: The number of unique, observed tokenizations of a word with N samples and dropout p.

#### **5** Experiments

We use English↔German, English↔Romanian, and English↔French as our translation tasks. For each language pair, we train a baseline BPE and MaxMatch tokenizer with the same vocabulary size and use them to build Dropout and Uniform Sampler variants so that the vocabulary between a baseline tokenizer and its stochastic variants is exactly the same and only the tokenization distribution is different. We include a UnigramLM tokenizer with the same vocabulary size as a learned-distribution baseline.We use the same underlying transformer model (Appendix A) for each language pair, and only change the embedding and decoding layers, according to the choice of tokenizer. We compare tokenizer efficiency (via tokenization-scorer), BLEU (Papineni et al., 2002; Post, 2018), CHRF (Popović, 2015), and COMET<sub>DA-22</sub> (Rei et al., 2022) by averaging the results of three experimental runs per model. We use p = 0.1 for BPE-Dropout, p = 0.3 for MaxMatch-Dropout, and  $\alpha = 0.3$  for UnigramLM. These values were taken from the literature where they have been shown to perform well (Provilkov et al., 2020; Hiraoka, 2022; Kudo, 2018).<sup>4</sup>

For our smallest dataset,  $EN\leftrightarrow DE$ , we experiment with the use of dropout with both the sourceand target-side tokenizers. However, for the largest,  $EN\leftrightarrow FR$ , we use dropout only on the source side, in line with other work which has shown targetside dropout to not be helpful (and sometimes even harmful) in high-resource settings (Provilkov et al., 2020). For  $EN\leftrightarrow RO$ , which is in between the size of the other two, we experiment with both settings.

For Uniform Sampling, we use p = 0.1 and 0.25, which were chosen as an estimate of the frequency that a non-canonical tokenization of word in BPE- and MaxMatch-Dropout was sampled, respectively. Thus, we should expect Uniform Samplers to have roughly the same amount of non-canonically-tokenized-words in a corpus as BPE-and MaxMatch-Dropout, so the salient difference is the variety of tokenizations.

A subset of the results are shown in Table 1, with the complete set appearing in Appendix C, Table 5. In nearly every translation metric, Uniform Sampling outperforms BPE- and MaxMatch-

<sup>&</sup>lt;sup>4</sup>We performed a small hyperparameter sweep of p for MaxMatch- and BPE-Dropout found that, as p increases, performance quickly degrades. This result matches what was found in (Provilkov et al. 2020; Figure 2) and (Hiraoka 2022; Figure 2).

| Tokenizer                     | Efficiency | BLEU  | CHRF  | COMET |
|-------------------------------|------------|-------|-------|-------|
| BPE                           | 0.4636     | 28.44 | 55.80 | 76.22 |
| BPE + Dropout $(p=0.1)$       | 0.4747     | 29.37 | 56.63 | 77.51 |
| BPE + Uniform $(p=0.1)$       | 0.4731     | 30.05 | 56.37 | 78.12 |
| BPE + Uniform $(p=0.25)$      | 0.4719     | 30.16 | 56.47 | 78.08 |
| MaxMatch                      | 0.4584     | 28.41 | 55.97 | 76.57 |
| MaxMatch + Dropout $(p=0.3)$  | 0.4530     | 29.13 | 56.43 | 77.38 |
| MaxMatch + Uniform $(p=0.1)$  | 0.4657     | 29.18 | 56.43 | 77.76 |
| MaxMatch + Uniform $(p=0.25)$ | 0.4633     | 29.43 | 56.57 | 77.62 |
| Unigram ( $\alpha$ =1)        | 0.4452     | 28.40 | 55.93 | 76.66 |
| Unigram ( $\alpha$ =0.3)      | 0.3796     | 28.97 | 56.33 | 77.44 |

(a) English→German (source+target dropout)

| Tokenizer                       | Efficiency | BLEU  | CHRF  | COMET |
|---------------------------------|------------|-------|-------|-------|
| BPE                             | 0.4524     | 23.56 | 53.20 | 81.03 |
| BPE + Dropout $(p=0.1)$         | 0.4614     | 23.98 | 53.70 | 81.90 |
| BPE + Uniform $(p=0.1)$         | 0.4594     | 23.83 | 53.67 | 82.00 |
| BPE + Uniform $(p=0.25)$        | 0.4647     | 24.13 | 53.73 | 82.20 |
| MaxMatch                        | 0.4476     | 23.52 | 53.23 | 81.17 |
| MaxMatch + Dropout $(p=0.3)$    | 0.4578     | 23.95 | 53.70 | 81.98 |
| MaxMatch + Uniform (p=0.1)      | 0.4528     | 24.32 | 53.90 | 82.11 |
| MaxMatch + Uniform ( $p=0.25$ ) | 0.4563     | 24.10 | 53.87 | 82.06 |
| Unigram ( $\alpha$ =1)          | 0.4338     | 23.68 | 53.37 | 81.28 |
| Unigram ( $\alpha$ =0.3)        | 0.4284     | 24.17 | 53.87 | 82.00 |

(b) English $\rightarrow$ Romanian (source only dropout)

Table 1: Experimental results for  $EN \rightarrow DE$  and  $EN \rightarrow RO$ . In each block, we compare a baseline tokenizer with its dropout and uniform sampling variants. Each group has the same vocabulary and differs only in the tokenization distribution. The best performing model for each baseline and metric is **bolded**. The full results for all languages is in Appendix C.

Dropout. However, curiously, Uniform Sampling does not always have higher efficiency than BPEor MaxMatch-Dropout (but is always higher than the baseline), as Uniform Sampling guarantees maximum entropy at the *word*-tokenization level, which does not necessarily translate to the *global*tokenization entropy.

There is only one metric (EN $\rightarrow$ DE, BPE, CHRF) where a Uniform Sampling model is not the best. However, in that same case, the Uniform Sampler improved upon the BPE-Dropout model by 0.8 BLEU, which is nearly as much as the BPE-Dropout improved upon the BPE baseline. In addition, the +0.61 increase in COMET<sub>DA-22</sub> corresponds to a 82% agreement accuracy with humans (Kocmi et al., 2024). In the EN $\rightarrow$ RO pair, Uniform Sampling models were the best across all metrics and underlying tokenizers. Further, Uniform Sampling consistently outperforms UnigramLM both in terms of raw translation quality metrics and improvement over the deterministic baseline. In line with other research, the effects of subword tokenization lessen as the dataset size grows (Provilkov et al., 2020). Still, the observed improvements extend to our full experimental results (Appendix C) and support our hypothesis that an unbiased tokenizer should generally outperform biased dropout tokenizers.

#### 6 Conclusion

We investigate the distributions induced by BPEand MaxMatch-Dropout, two popular subword regularization schemes. We hypothesize and show that BPE- and MaxMatch-Dropout are suboptimal in that they form heavily biased distributions. We introduce a Uniform Sampler tokenizer, which guarantees uniform distributions and consistently outperforms BPE- and MaxMatch-Dropout on machine translation tasks.

**Future work.** Uniform Sampling is uniform at the *word* level, but past research suggests that uniformity at the *global* unigram level is desired (Gutierrez-Vasques et al., 2021; Zouhar et al., 2023). Therefore, algorithms could be designed to directly optimize global uniformity. Further investigations should reconcile how both Uniform Sampling and UnigramLM improve performance despite their opposing motivations (higher/lower entropy).

## Limitations

We did not establish statistical significance for our results, but note that the trend holds across language pairs, tokenizers, and metrics. We did not do substantial hyperparameter searching for vocabulary size or dropout rates, but rather used values that commonly appear in the literature. It is possible that some trends in our results may change with different choices of tokenization hyperparameters.

We also did not experiment with extremely-low resource settings (our smallest setting,  $EN \leftrightarrow DE$ has 150k sentence pairs), or very large settings (our largest,  $EN \leftrightarrow FR$ , has 2M sentence pairs). Additionally, in our largest case, the improvement seen by Uniform Sampling are less consistent and less significant. However, this is in line with prior research that shows the diminishing effectiveness of subword regularization as the corpus size increases.

#### Acknowledgements

These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT), Japan.

### References

- Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In Proceedings of the First Conference on Machine Translation, pages 131–198, Berlin, Germany. Association for Computational Linguistics.
- Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign.*
- Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.
- Tatsuya Hiraoka. 2022. MaxMatch-dropout: Subword regularization for WordPiece. In *Proceedings of the* 29th International Conference on Computational Linguistics, pages 4864–4872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*.
- Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple

subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Misha Lavrov. 2018. A procedure for sampling paths in a directed acyclic graph. Mathematics Stack Exchange.
- 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.
- Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- 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.
- Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1882–1892, Online. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In 2012 IEEE International

*Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012,* pages 5149–5152. IEEE.

- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.
- Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

## **A** Training Details

We use fairseq (Ott et al., 2019) for language modeling, HuggingFace's tokenizers library for our underlying BPE and MaxMatch tokenizers, and OpenFst (Allauzen et al., 2007) for the subword lattice construction. For UnigramLM, we used SentencePiece (Kudo and Richardson, 2018). We used fairseq's transformer\_iwslt\_de\_en architecture for EN $\leftrightarrow$ DE, and the baseline transformer architecture for EN $\leftrightarrow$ RO and EN $\leftrightarrow$ FR. The hyperparameters and optimizer configuration are given in Tables 2, 3, and 4. Our datasets were:

- EN↔DE: 160k, IWSLT14 (Cettolo et al., 2014)
- EN↔RO: 600k, WMT16 (Bojar et al., 2016)
- $EN \leftrightarrow FR: 2M$ , Europarl (Koehn, 2005)

| $EN \leftrightarrow DE: (10k, 10k)$ |
|-------------------------------------|
| 512                                 |
| 1024                                |
| 4                                   |
| 6                                   |
| 0.3                                 |
|                                     |

Table 2: The transformer\_iwslt\_en\_de architecture, used for the English⇔German task.

|                             | EN $\leftrightarrow$ RO: (14k, 14k) |
|-----------------------------|-------------------------------------|
| Vocabulary Sizes (src, tgt) | $EN \leftrightarrow FR: (30k, 30k)$ |
| Embedding Dimension         | 512                                 |
| FFN Dimension               | 2048                                |
| Number of Heads             | 6                                   |
| Number of Layers            | 8                                   |
| Dropout                     | 0.1                                 |

Table 3: The transformer architecture, used for the English $\leftrightarrow$ Romainan and English $\leftrightarrow$ French tasks.

| Optimizer          | ADAM                        |
|--------------------|-----------------------------|
| $\beta_1, \beta_2$ | (0.9, 0.98)                 |
| Learning Rate      | $5 \times 10^{-4}$          |
| Warmup             | 4000 steps                  |
| Scheduler          | Inverse Square Root         |
| Tokens-per-batch   | 8192                        |
|                    | $EN \leftrightarrow DE: 8$  |
| Patience           | $EN \leftrightarrow RO: 10$ |
|                    | $EN \leftrightarrow FR: 5$  |

Table 4: The optimizer parameters, used for all tasks.

## **B** Algorithms

Algorithms 5 and 6 are simply a reference for sampling uniformly from a DAG. As a rejection sampling scheme, their runtime depends heavily on the underlying DAG. In our implementation, we use a more efficient linear time (in the input length) sampling algorithm. We were unable to find a reference for this algorithm in the literature, and including it here it is outside of the scope of the paper.

| <b>Inputs</b> : Corpus $C$ , Alphabet $\Sigma$ , Target size $n$ ,             |  |  |  |
|--|--|--|--|
| <b>Outputs</b> : Vocabulary $\mathcal{V}$ , Merges $\mu$                       |  |  |  |
| 1: $\mathcal{V} \leftarrow \Sigma$   |  |  |  |
| 2: for $i \in 1 n$ do  |  |  |  |
| 3: $(x, y) \leftarrow \arg \max \text{COUNT}((a, b), C)$                       |  |  |  |
| $a,b\!\in\!\mathcal{V}$  |  |  |  |
| 4: $\mathcal{V} \leftarrow \mathcal{V} \cup \{xy\}$                            |  |  |  |
| 5: $\mu \leftarrow \mu \cup \langle (x,y) \rangle$                             |  |  |  |
| 6: $\mathcal{C} \leftarrow \text{Replace}((x, y) \rightarrow xy, \mathcal{C})$ |  |  |  |
| 7: return $\mathcal{V}, \mu$   |  |  |  |
|  |  |  |  |

Algorithm 4: BPE Training.

| Inputs: Directed Acyclic Graph D,                      |
|--|
| <b>Outputs</b> : Path $\pi$ , Path-probability $p$     |
| 1: $\pi \leftarrow \langle \rangle$                    |
| 2: $p \leftarrow 1$                                    |
| 3: $CUR \leftarrow q_{start}$                          |
| 4: while CUR is not final do                           |
| 5: $(w,q) \sim \text{Uniform}(\text{Adj}(\text{cur}))$ |
| 6: APPEND $(\pi, (CUR, w, q))$                         |
| 7: $p \leftarrow p \times \frac{1}{\text{Deg(CUR)}}$   |
| 8: $CUR \leftarrow q$                                  |
| 9: return $\pi, p$                                     |
|  |

Algorithm 5: Biased DAG Sampling.

| Inputs: Directed Acyclic Graph D,               |
|---|
| <b>Output</b> : Path $\pi$                      |
| 1: $p_{min} = \prod_{q \in D} \frac{1}{DEG(q)}$ |
| 2: $(\pi, p) \sim \text{BiasedSample}(D)$       |
| 3: while RAND() > $\frac{p_{min}}{p}$ do        |
| 4: $(\pi, p) \sim \text{BIASEDSAMPLE}(D)$       |
| 5: return $\pi$                                 |
|   |

Algorithm 6: Unbiased DAG Sampling.



Figure 3: Distribution uniformity measured by Shannon Efficiency (higher=more uniform; excludes the canonical form, which takes up most of the probability mass). Our Uniform Sampling versions *guarantee* balanced sampling of tokenizations, which is not true for the standard Dropout versions whose balance depends non-linearly on the dropout rate p.

# C Full Experiments

| Efficiency   | BLEU   | CHRF   | COMET   | Tokenizer  | Efficiency  | BLEU  | CHRF  | COMET  |  |  |
|--|--|--|---|--|---|---|---|--|--|--|
| 0.4636   | 33.66  | 57.10  | 79.30   | BPE  | 0.4636  | 28.44   | 55.80   | 76.22  |  |  |
| 0.4747   | 35.06  | 58.07  | 80.51   | BPE + Dropout $(p=0.1)$  | 0.4747  | 29.37   | 56.63   | 77.51  |  |  |
| 0.4731   | 35.03  | 57.97  | 80.46   | BPE + Uniform $(p=0.1)$  | 0.4731  | 30.05   | 56.37   | 78.12  |  |  |
| 0.4719   | 35.22  | 58.13  | 80.57   | BPE + Uniform $(p=0.25)$   | 0.4719  | 30.16   | 56.47   | 78.08  |  |  |
| 0.4584   | 33.85  | 57.17  | 79.48   | MaxMatch   | 0.4584  | 28.41   | 55.97   | 76.57  |  |  |
|  |  |  |   |  | 1   |   |   | 77.38  |  |  |
|  |  |  |   |  |   |   |   | 77.76  |  |  |
|  |  |  |   |  |   |   |   | 77.62  |  |  |
|  |  |  |   | 0  |   |   |   | 76.66  |  |  |
| 0.3796   | 34.24  | 57.70  | 80.31   | Unigram ( $\alpha$ =0.3)   | 0.3796  | 28.97   | 56.33   | 77.44  |  |  |
| (a) German→English (source+target dropout)               |  |  |   |  |   | (b) English→German (source+target dropout)  |   |  |  |  |
| Efficiency   | BLEU   | CHRF   | COMET   | Tokenizer  | Efficiency  | BLEU  | CHRF  | COMET  |  |  |
| 0.4034   | 40.86  | 64.57  | 86.39   | BPE  | 0.4034  | 41.27   | 65.77   | 86.86  |  |  |
| 0.4137   | 40.96  | 64.57  | 86.50   | BPE + Dropout ( $p=0.1$ )  | 0.4137  | 41.45   | 65.90   | 87.08  |  |  |
| 0.4139   | 41.10  | 64.70  | 86.52   | BPE + Uniform $(p=0.1)$  | 0.4139  | 41.54   | 65.93   | 87.04  |  |  |
| 0.4259   | 40.86  | 64.57  | 86.36   | BPE + Uniform $(p=0.25)$   | 0.4259  | 41.35   | 65.83   | 87.03  |  |  |
| 0.4003   | 41.02  | 64.70  | 86.48   | MaxMatch   | 0.4003  | 41.38   | 65.87   | 87.00  |  |  |
| 0.4186   | 40.88  | 64.57  | 86.47   | MaxMatch + Dropout $(p=0.3)$   | 0.4186  | 41.24   | 65.80   | 86.95  |  |  |
| 0.4094   | 41.04  | 64.70  | 86.54   | MaxMatch + Uniform (p=0.1)   | 0.4094  | 41.44   | 65.93   | 87.07  |  |  |
| 0.4194   | 40.80  | 64.50  | 86.38   | MaxMatch + Uniform $(p=0.25)$  | 0.4194  | 41.22   | 65.77   | 86.93  |  |  |
| 0.3801   | 40.59  | 64.43  | 86.27   | Unigram ( $\alpha$ =1)   | 0.3801  | 40.47   | 65.27   | 86.36  |  |  |
| 0.3773   | 40.71 64.53 86.36 Unigram  |  | 86.36   | Unigram ( $\alpha$ =0.3)   | 0.3773  | 40.15   | 65.00   | 86.08  |  |  |
| (c) French $\rightarrow$ English (source only dropout)   |  |  |   |  | (d) English→French (source only dropout)  |   |   |  |  |  |
| Efficiency   | BLEU   | CHRF   | COMET   | Tokenizer  | Efficiency  | BLEU  | CHRF  | COMET  |  |  |
| 0.4524   | 30.81  | 57.30  | 77.77   | BPE  | 0.4524  | 23.56   | 53.20   | 81.03  |  |  |
| 0.4614   | 32.13  | 58.17  | 79.50   | BPE + Dropout $(p=0.1)$  | 0.4614  | 23.98   | 53.70   | 81.90  |  |  |
| 0.4594   | 31.92  | 58.13  | 79.64   | BPE + Uniform $(p=0.1)$  | 0.4594  | 23.83   | 53.67   | 82.00  |  |  |
| 0.4647   | 31.85  | 58.23  | 79.54   | BPE + Uniform $(p=0.25)$   | 0.4647  | 24.13   | 53.73   | 82.20  |  |  |
| 0.4476   | 31.01  | 57.23  | 78.03   | MaxMatch   | 0.4476  | 23.52   | 53.23   | 81.17  |  |  |
|  |  |  |   |  | 1   |   |   | 81.98  |  |  |
|  |  |  |   |  | 1   |   |   | 82.11  |  |  |
|  |  |  |   | · · · · · · · · · · · · · · · · · · ·  |   |   |   | 82.06  |  |  |
| 0.4338   | 30.34  | 56.97  | 77.74   | Unigram $(\alpha = 1)$   | 0.4338  | 23.68   | 53.37   | 81.28  |  |  |
| I  |  |  |   |  | 1   |   |   | 82.00  |  |  |
| (e) Romanian $\rightarrow$ English (source only dropout) |  |  |   |  | (t) English $\rightarrow$ Romanian (source only dropout)  |   |   |  |  |  |
| Efficiency   | BLEU   | CHRF   | COMET   | Tokenizer  | Efficiency  | BLEU  | CHRF  | COMET  |  |  |
| 0.4524   | 30.81  | 57.30  | 77.77   | BPE  | 0.4524  | 23.56   | 53.20   | 81.03  |  |  |
| 0.4672   | 32.48  | 58.87  | 80.29   | BPE + Dropout $(p=0.1)$  | 0.4672  | 24.85   | 54.30   | 82.71  |  |  |
| 0.4615   | 32.46  | 58.87  | 80.26   | BPE + Uniform $(p=0.1)$  | 0.4615  | 24.78   | 54.30   | 83.03  |  |  |
| 0.4562   | 32.83  | 59.07  | 80.71   | BPE + Uniform $(p=0.25)$   | 0.4562  | 24.77   | 54.00   | 82.67  |  |  |
| 0.4502   |  |  |   |  |   |   |   | 81.17  |  |  |
| 0.4302   | 31.01  | 57.23  | 78.03   | MaxMatch   | 0.4476  | 23.52   | 53.23   | 01.17  |  |  |
|  | 31.01<br>32.89   | 57.23<br>59.03   | 78.03<br><b>80.69</b>   | MaxMatch<br>MaxMatch + Dropout (p=0.3)   | 0.4476<br>0.4465  | 23.52<br>25.02  | 53.23<br>54.20  | 82.67  |  |  |
| 0.4476   |  |  |   |  | 1   |   |   |  |  |  |
| 0.4476<br>0.4465   | 32.89  | 59.03  | 80.69   | MaxMatch + Dropout $(p=0.3)$   | 0.4465  | 25.02   | 54.20   | 82.67  |  |  |
| 0.4476<br>0.4465<br>0.4544                               | 32.89<br>32.83   | 59.03<br>58.97   | <b>80.69</b><br>80.36   | MaxMatch + Dropout $(p=0.3)$<br>MaxMatch + Uniform $(p=0.1)$   | 0.4465<br>0.4544  | 25.02<br>24.77  | 54.20<br><b>54.40</b>   | 82.67<br><b>83.08</b>  |  |  |
|  | 0.4636<br>0.4747<br>0.4747<br>0.4731<br>0.4719<br>0.4584<br>0.4530<br>0.4657<br>0.4633<br>0.4452<br>0.3796<br>sh (source<br>Efficiency<br>0.4034<br>0.4137<br>0.4137<br>0.4139<br>0.4259<br>0.4003<br>0.4186<br>0.4094<br>0.4137<br>0.4139<br>0.4259<br>0.4003<br>0.4186<br>0.4094<br>0.4194<br>0.3801<br>0.3773<br>ish (source<br>Efficiency<br>0.4524<br>0.4614<br>0.4524<br>0.4647<br>0.4528<br>0.4528<br>0.4563<br>0.4338<br>0.4284<br>glish (source | 0.4636         33.66           0.4747         35.06           0.4747         35.06           0.4747         35.06           0.4719         35.22           0.4584         33.85           0.457         35.17           0.4657         35.17           0.4633         35.32           0.4452         33.37           0.3796         34.24           sh (source+target of 0.4137         40.86           0.4137         40.96           0.4137         40.96           0.4137         40.96           0.4139         41.10           0.4259         40.86           0.4003         41.02           0.4186         40.88           0.4094         41.04           0.4194         40.80           0.3801         40.59           0.3773         40.71           ish (source only dr           Efficiency BLEU           0.4524         30.81           0.4614         32.13           0.4578         31.90           0.4528         32.02           0.4563         31.83           0.4338         30.34 | 0.4636         33.66         57.10           0.4747         35.06         58.07           0.4731         35.03         57.97           0.4719 <b>35.22 58.13</b> 0.4584         33.85         57.17           0.4530         34.92         57.87           0.4657         35.17         58.10           0.4633 <b>35.32 58.13</b> 0.4452         33.37         56.77           0.3796         34.24         57.70           sh (source+target dropour         Efficiency         BLEU         CHRF           0.4034         40.86         64.57         0.4137         40.96         64.57           0.4137         40.96         64.57         0.4139 <b>41.10 64.70</b> 0.4139 <b>41.04 64.70</b> 0.4186         40.88         64.57           0.4003         41.02 <b>64.70</b> 0.4194         40.80         64.50           0.3801         40.59         64.43         0.3773         40.71         64.53           0.4194         40.80         64.50         0.4614 <b>32.13</b> 58.17           0.4524         3 | 0.463         33.66         57.10         79.30           0.4747         35.06         58.07         80.51           0.4747         35.03         57.97         80.46           0.4731         35.03         57.97         80.46           0.4719 <b>35.22 58.13 80.57</b> 0.4584         33.85         57.17         79.48           0.4530         34.92         57.87         80.37           0.4657         35.17         58.10         80.60           0.4633 <b>35.32 58.13 80.71</b> 0.4452         33.37         56.77         79.43           0.3796         34.24         57.70         80.31           sth (source+target dropout)           Efficiency         BLEU         CHRF         COMET           0.4034         40.86         64.57         86.36           0.4137         40.96         64.57         86.36           0.4139         41.10         64.70         86.48           0.4186         40.88         64.57         86.36           0.4186         40.89         64.50         86.38           0.3801         40.59 | 0.4636         33.66         57.10         79.30         BPE           0.4747         35.06         58.07         80.51         BPE + Dropout (p=0.1)           0.4731         35.03         57.97         80.46         BPE + Uniform (p=0.1)           0.4719 <b>35.22 58.13 80.57</b> BPE + Uniform (p=0.25)           0.4584         33.85         57.17         79.48         MaxMatch + Dropout (p=0.3)           0.4657         35.17         58.10         80.60         MaxMatch + Uniform (p=0.1)           0.4633 <b>35.32 58.13 80.71</b> MaxMatch + Uniform (p=0.25)           0.4452         33.37         56.77         79.43         Unigram (α=1)           0.3796         34.24         57.70         80.31         Unigram (α=0.3)           sh (source+target dropout)         (b) English→Germ         (b) English→Germ           0.4137         40.96         64.57         86.50         BPE + Dropout (p=0.1)           0.4259         40.86         64.57         86.50         BPE + Uniform (p=0.1)           0.4139         41.10         64.70         86.54         MaxMatch + Uniform (p=0.25)           0.4003         41.02         64.50         86.3 | 0.4636         33.66         57.10         79.30         BPE         0.4636           0.4747         35.06         58.07         80.51         BPE + Uniform (p=0.1)         0.4747           0.4731         35.03         57.97         80.46         BPE + Uniform (p=0.1)         0.4731           0.4719         35.22         58.13         80.57         BPE + Uniform (p=0.25)         0.4719           0.4584         33.85         57.17         79.48         MaxMatch         0.4553         0.4553           0.4530         34.92         57.87         80.37         MaxMatch + Uniform (p=0.1)         0.4657           0.4633         35.32         58.13         80.71         MaxMatch + Uniform (p=0.25)         0.4633           0.4452         33.37         56.77         79.43         Unigram (α=1)         0.4452           0.3796         34.24         57.70         80.31         Unigram (α=1)         0.4133           0.4137         40.96         64.57         86.50         BPE + Dropout (p=0.1)         0.4137           0.4139 <b>11.10</b> 64.70         86.52         BPE + Uniform (p=0.25)         0.4259           0.4034         41.04         64.70         86.54         MaxMatch + | 0.4636         33.66         57.10         79.30         BPE         0.4636         28.44           0.4747         35.05         58.07         80.51         BPE + Dropout (p=0.1)         0.4747         29.37           0.4731         35.03         57.97         80.46         BPE + Uniform (p=0.1)         0.4719         30.16           0.4584         33.85         57.17         79.48         MaxMatch         0.4530         29.43           0.4633         35.22         58.13         80.57         MaxMatch + Dropout (p=0.3)         0.4532         29.43           0.4452         33.37         56.77         79.43         Unigram (α=1.)         0.4452         28.40           0.3796         34.24         57.70         80.31         Unigram (α=0.3)         0.3796         28.97           sh (source+target dropout)         (b) English→German (source+target dropout (p=0.1)         0.4137         41.45           0.4137         40.96         64.57         86.59         BPE + Uniform (p=0.1)         0.4134         41.27           0.4137         41.06         64.77         86.54         BPE + Uniform (p=0.1)         0.4134         41.27           0.4134         40.26         64.57         86.56         BPE + U | 0.4636         33.66         57.10         79.30<br>80.51         BPE<br>BPE         0.4636<br>Dependence         0.4636         28.44         55.80<br>0.4747           0.4717         35.06         58.07         80.46         BPE         Uniform (p=0.1)         0.4717         30.05         56.37           0.4719         35.22         58.13         80.57         BPE         Uniform (p=0.25)         0.4719         30.16         56.47           0.4530         34.92         57.87         80.37         MaxMatch         0.4530         29.43         56.57           0.4533         35.17         58.10         80.67         MaxMatch         Unigram (\alpha=1)         0.4633         29.43         56.57           0.4452         33.37         56.77         79.43         Unigram (\alpha=0.3)         0.3796         28.40         55.93           0.4132         34.24         57.70         80.31         Unigram (\alpha=0.3)         0.3796         28.40         55.93           0.4134         40.86         64.57         86.59         BPE         Propout (p=0.1)         0.4137         41.45         65.93           0.4139         41.10         64.70         86.52         BPE         Propout (p=0.1)         0.4134         41.55 |  |  |

(g) Romanian→English (source+target dropout)

(h) English $\rightarrow$ Romanian (source+target dropout)

Table 5: The main results of machine translation performance (average across 3 seeds). In almost all cases the Uniform sampling yields the best results.

| BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ something1.60%something14.79%start ted2.15%start ed14.88%so mething1.52%something4.35%start ted0.01%start ed2.13%so mething0.06%something4.35%start ed0.07%start ed4.37%so mething0.09%something0.31%something0.93%start ed0.02%start ed0.91%so mething0.03%some thing0.91%start ed0.02%start ed0.91%so mething0.03%some thing0.42%start ed0.00%start ed0.39%so mething0.01%somet hing0.38%start ed0.00%start ed0.39%something0.01%somet hing0.29%start ed0.00%start ed0.39%something0.00%somet ting0.29%start ed0.00%  |                                  |        |                                |        |                                  |        |                        |        |  |
|--|----------------------------------|--------|--------------------------------|--------|----------------------------------|--------|------------------------|--------|--|
| some thing1.60%somet hing14.79%star ted2.15%start ed14.88%so met hing1.52%somet hing4.54%start ed0.11%start ed6.21%so me thing0.09%somet hing1.31%start ed0.07%start ed4.37%so me thing0.03%somet hing0.93%start ed0.02%start ed0.91%so me thing0.03%some thing0.91%start ed0.02%start ed0.91%so me thing0.01%some thing0.42%start ed0.00%start ed0.39%some thing0.00%some thing0.29%start ed0.00%start ed0.39%percent73.54%percent64.05%different8.244%different  | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout p=0.3         |        | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout p=0.3 |        |  |
| so met hing $1.52\%$ somet hing $4.54\%$ star ted $0.11\%$ start ed $6.21\%$ so me thing $0.09\%$ some thing $4.35\%$ start ted $0.07\%$ start ted $4.37\%$ so me thing $0.03\%$ some thing $0.93\%$ start ted $0.02\%$ start ted $1.31\%$ so me thing $0.03\%$ some thing $0.91\%$ start ted $0.02\%$ start ted $0.91\%$ s ome thing $0.03\%$ some thing $0.91\%$ start ted $0.02\%$ start ted $0.91\%$ s om thing $0.01\%$ some thing $0.42\%$ start ted $0.00\%$ start ted $0.42\%$ some thing $0.01\%$ some thing $0.42\%$ start ted $0.00\%$ start ted $0.39\%$ some thing $0.00\%$ some thing $0.29\%$ start ted $0.00\%$ start ted $0.39\%$ some thing $0.00\%$ some thing $0.29\%$ start ted $0.00\%$ start ted $0.29\%$ some thing $0.00\%$ somet thing $0.29\%$ start ted $0.00\%$ start ted $0.29\%$ some thing $0.00\%$ somet thing $0.29\%$ start ted $0.00\%$ start ted $0.29\%$ some thing $0.00\%$ somet thing $0.29\%$ start ted $0.00\%$ start ted $0.29\%$ something $0.00\%$ somet thing $0.29\%$ start ted $0.00\%$ start ted $0.29\%$ percent $7.54\%$ percent $69.93\%$ different $0.20\%$ differ ent <th>something</th> <th>96.50%</th> <th>something</th> <th>69.85%</th> <th>started</th> <th>97.56%</th> <th>started</th> <th>69.97%</th>   | something                        | 96.50% | something                      | 69.85% | started                          | 97.56% | started                | 69.97% |  |
| so m et hing $0.16\%$ some thing $4.35\%$<br>some th ingstart ed $0.07\%$ start ted $4.37\%$<br>start tedso m et hing $0.03\%$ some th ing $0.93\%$ start ted $0.02\%$ start ted $0.91\%$<br>start ted $0.02\%$ start ted $0.02\%$<br>start ted $0.91\%$<br>start ted $0.02\%$ start ted $0.02\%$<br>start ted $0.91\%$<br>start ted $0.02\%$<br>start ted $0.91\%$<br>start ted $0.02\%$<br>start ted $0.02\%$<br>start ted $0.02\%$<br>start ted $0.39\%$<br>start ted  | some thing                       | 1.60%  | somet hing                     | 14.79% | star ted                         | 2.15%  | start ed               | 14.88% |  |
| so me thing $0.09\%$ something $1.31\%$ st ar ted $0.05\%$ star ted $1.33\%$ some thing $0.03\%$ something $0.93\%$ st ar ted $0.02\%$ st ar ted $0.91\%$ so me thing $0.03\%$ something $0.91\%$ st ar ted $0.02\%$ st ar ted $0.91\%$ so me thing $0.03\%$ something $0.42\%$ st ar ted $0.02\%$ st ar ted $0.39\%$ something $0.01\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.39\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.29\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.29\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.29\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.29\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.00\%$ st ar ted $0.29\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.20\%$ $0.20\%$ something $0.00\%$ something $0.29\%$ st ar ted $0.29\%$ $0.29\%$ something $0.00\%$ something $0.29\%$ $0.00\%$ st ar ted $0.29\%$ percent $73.54\%$ percent $4.60\%$ different $0.20\%$ different $1.27\%$ percent $0.18\%$ per cent   | so met hing                      | 1.52%  | somet hi n g                   | 4.54%  | star te d                        | 0.11%  | start e d              | 6.21%  |  |
| $\begin{array}{c c c c c c c c c c c c c c c c c c c $   | so m et hing                     | 0.16%  | some thing                     | 4.35%  | start ed                         | 0.07%  | star ted               | 4.37%  |  |
| s ome thing $0.03\%$ som eth ing $0.91\%$ s tar ted $0.02\%$ star ted $0.42\%$ som eth ing $0.01\%$ some thing $0.42\%$ star ted $0.01\%$ s tar ted $0.39\%$ somet hing $0.01\%$ somet hing $0.29\%$ star ted $0.00\%$ star ted $0.39\%$ somet hing $0.00\%$ somet hing $0.29\%$ star ted $0.00\%$ star ted $0.39\%$ some th ing $0.00\%$ somet hing $0.29\%$ star ted $0.00\%$ star ted $0.39\%$ somet hing $0.00\%$ somet hing $0.29\%$ star ted $0.00\%$ star ted $0.39\%$ somet hing $0.00\%$ somet hing $0.29\%$ star ted $0.00\%$ star ted $0.29\%$ percent $73.54\%$ percent $69.93\%$ different $82.44\%$ differ ent $14.74\%$ per cent $8.12\%$ percent $64.0\%$ differ ent $8.74\%$ differ ent $14.74\%$ per cent $8.12\%$ per cent $6.40\%$ differ ent $8.02\%$ differ ent $1.36\%$ per cent $0.18\%$ per cent $0.39\%$ differ ent $0.29\%$ differ ent $1.27\%$ per cent $0.18\%$ per cent $0.33\%$ differ ent $0.03\%$ differ ent $0.94\%$ per cent $0.18\%$ per cent $0.33\%$ differ ent $0.03\%$ differ ent $0.94\%$ per cent $0.19\%$ per cent $0.33\%$ differ ent $0.03\%$ differ ent $0.94\%$   | so me thing                      | 0.09%  | somet h ing                    | 1.31%  | st ar ted                        | 0.05%  | star te d              | 1.33%  |  |
| so m eth ing $0.03\%$ some t hing $0.42\%$ star t ed $0.01\%$ s tar t ed $0.39\%$ somet hing $0.01\%$ somet h ing $0.38\%$ star t ed $0.00\%$ star t ed $0.39\%$ some th ing $0.00\%$ some th ing $0.29\%$ s t ar t ed $0.00\%$ star t ed $0.39\%$ some th ing $0.00\%$ some th ing $0.29\%$ s t ar t ed $0.00\%$ st ar t ed $0.39\%$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ percent $73.54\%$ percent $69.93\%$ different $82.44\%$ different $69.95\%$ per cent $8.12\%$ percent $6.40\%$ diff er ent $8.74\%$ differ ent $14.74\%$ per cent $7.84\%$ per cent $4.53\%$ diff er ent $0.29\%$ diff er ent $3.05\%$ per cent $0.18\%$ per cent $0.90\%$ diff er ent $0.29\%$ diff er ent $1.36\%$ p er cent $0.18\%$ per cent $0.39\%$ diff er ent $0.20\%$ diff er ent $0.94\%$ p er cent $0.11\%$ per cent $0.38\%$ diff er ent $0.03\%$ differ ent $0.94\%$ p er cent $0.10\%$ per cent $0.37\%$ diff er ent $0.03\%$ differ ent $0.94\%$ p er cent $0.10\%$ per cent $0.37\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.10\%$ per cent $0.37\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.1$   | som eth ing                      | 0.03%  | some th ing                    | 0.93%  | st art ed                        | 0.02%  | st art ed              | 0.91%  |  |
| somet hing $0.01\%$ somet hing $0.38\%$<br>some things tart ted $0.00\%$ start ed $0.39\%$<br>s tart edBPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ per cent $73.54\%$<br>per centpercent $69.93\%$<br>differentBPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ per cent $8.83\%$<br>per centpercent $14.61\%$<br>differ ent $8.74\%$<br>differ ent $8.74\%$<br>differ ent $14.74\%$<br>differ entper cent $7.84\%$<br>per centper cent $4.53\%$<br>differ ent $0.20\%$<br>differ ent $0.20\%$<br>differ ent $14.74\%$<br>differ entper cent $0.18\%$<br>per centper cent $0.39\%$<br>diff er ent $0.20\%$<br>diff er ent $0.20\%$<br>diff er ent $14.74\%$<br>diff er entper cent $0.18\%$<br>per centper cent $0.39\%$<br>diff er ent $0.20\%$<br>diff er ent $0.20\%$ <b< td=""><td>s ome thing</td><td>0.03%</td><td>som eth ing</td><td>0.91%</td><td>s ta r ted</td><td>0.02%</td><td>star t ed</td><td>0.42%</td></b<> | s ome thing                      | 0.03%  | som eth ing                    | 0.91%  | s ta r ted                       | 0.02%  | star t ed              | 0.42%  |  |
| some th ing $0.00\%$ some th ing $0.29\%$ s t art ed $0.00\%$ st art ed $0.26\%$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ GPC ent73.54%percent69.93%different8.2.44%different69.95%per cent8.38%per cent8.2.44%different69.95%per cent8.2.44%differ ent6.9.95%per cent8.2.44%differ ent14.74%per cent8.33%per cent6.3.3%per cent0.13%differ ent8.74%differ ent4.44%per cent0.33%per cent0.33%differ ent0.20%differ ent0.20%differ ent0.20%per cent0.33%per cent0.33%per cent0.33%per cent0.33%per cent0.33%differ ent0.03%per cent <th< td=""><td>so m eth ing</td><td>0.03%</td><td>some t hing</td><td>0.42%</td><td>star t ed</td><td>0.01%</td><td>s ta r ted</td><td>0.39%</td></th<>  | so m eth ing                     | 0.03%  | some t hing                    | 0.42%  | star t ed                        | 0.01%  | s ta r ted             | 0.39%  |  |
| BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ percent         73.54%         percent         69.93%         different         82.44%         different         69.95%           per cent         8.83%         percent         14.61%         different         8.74%         different         69.95%           per cent         8.12%         percent         6.40%         different         8.74%         different         4.44%           per cent         0.88%         per cent         1.33%         different         0.29%         different         1.36%           per cent         0.18%         per cent         0.39%         different         0.13%         different         0.94%           per cent         0.11%         per cent         0.38%         different         0.03%         different         0.94%           per cent         0.10%         per cent         0.37%         different         0.03%         different         0.41%           per cent         0.26%         different         0.01%         different         0.56%           per cent         0.26%         together         0.46%         happening         9.511%         happeni  | somet hing                       | 0.01%  | somet h in g                   | 0.38%  | s ta r t ed                      | 0.00%  | st art e d             | 0.39%  |  |
| percent73.54%<br>percentpercent69.93%<br>69.93%<br>percentdifferent82.44%<br>differentdifferent69.95%<br>69.95%per cent8.83%<br>percentpercent14.61%<br>differentdifferent82.44%<br>differentdifferent69.95%<br>differentper cent8.12%<br>percentpercent6.40%<br>differentdifferent8.74%<br>differentdifferent14.74%<br>differentper cent7.84%<br>per centper cent4.53%<br>differentdifferent8.05%<br>differentdifferent14.74%<br>differentper cent0.18%<br>per centper cent0.29%<br>differentdifferent1.36%<br>differentdifferent3.05%<br>differentper cent0.12%<br>per centper cent0.39%<br>different0.13%<br>different0.13%<br>different0.14%<br>differentper cent0.11%<br>per centper cent0.38%<br>differentdifferent0.03%<br>different0.44%per cent0.10%<br>per centper cent0.26%different0.03%<br>different0.44%per cent0.09%<br>per centper cent0.26%different0.03%<br>different0.14%per cent0.09%<br>per centper cent0.26%different0.03%<br>different0.14%per cent0.09%<br>per centper cent0.26%different0.03%<br>different0.14%per cent0.09%<br>per centper cent0.26%different0.03%<br>different0.14%per cent <th>some th ing</th> <th>0.00%</th> <th>some th in g</th> <th>0.29%</th> <th>s t art ed</th> <th>0.00%</th> <th>st ar ted</th> <th>0.26%</th>   | some th ing                      | 0.00%  | some th in g                   | 0.29%  | s t art ed                       | 0.00%  | st ar ted              | 0.26%  |  |
| per c ent $8.83\%$ perce nt $14.61\%$ dif fe rent $8.74\%$ differ ent $14.74\%$ per cent $8.12\%$ perce nt $6.40\%$ diff erent $8.05\%$ differ ent $14.74\%$ per cent $7.84\%$ per cent $4.53\%$ diff erent $0.29\%$ differ ent $3.05\%$ per cent $0.88\%$ per cent $1.33\%$ diff erent $0.29\%$ diff erent $3.05\%$ per cent $0.18\%$ per cent $0.90\%$ dif ferent $0.20\%$ diff erent $1.36\%$ p er cent $0.11\%$ per cent $0.39\%$ dif ferent $0.07\%$ diff erent $0.94\%$ p er cent $0.11\%$ per cent $0.37\%$ diff erent $0.03\%$ differ ent $0.94\%$ p er cent $0.10\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.44\%$ p er cent $0.10\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.44\%$ p er cent $0.99\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.44\%$ p er cent $0.99\%$ tog ether $14.68\%$ happening $95.11\%$ happening $70.05\%$ to ge ther $0.95\%$ tog ether $3.05\%$ happening $1.93\%$ happen ing $14.80\%$ to ge ther $0.08\%$ tog ether $1.35\%$ happening $0.91\%$ happening $4.26\%$ to ge ther $0.03\%$ tog ether $1.33\%$ happening $0.91\%$ happening $0.94\%$  | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout <i>p</i> =0.3 |        | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout p=0.3 |        |  |
| per c ent $8.83\%$ perce nt $14.61\%$ dif fe rent $8.74\%$ differ ent $14.74\%$ per cent $8.12\%$ per cent $6.40\%$ diff erent $8.05\%$ differ ent $4.44\%$ per cent $7.84\%$ per cent $4.53\%$ differ ent $0.29\%$ differ ent $4.44\%$ per cent $0.88\%$ per cent $1.33\%$ differ ent $0.20\%$ differ ent $1.36\%$ p er c ent $0.18\%$ per cent $0.90\%$ dif fer ent $0.13\%$ differ ent $1.27\%$ p er cent $0.11\%$ per cent $0.39\%$ differ ent $0.07\%$ differ ent $0.94\%$ p er cent $0.11\%$ per cent $0.37\%$ differ ent $0.03\%$ differ ent $0.94\%$ per cent $0.10\%$ per cent $0.37\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.10\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.10\%$ per cent $0.37\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.25\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.03\%$ <td>percent</td> <td>73.54%</td> <td>percent</td> <td>69.93%</td> <td>different</td> <td>82.44%</td> <td>different</td> <td>69.95%</td>   | percent                          | 73.54% | percent                        | 69.93% | different                        | 82.44% | different              | 69.95% |  |
| per cent $8.12\%$ percent $6.40\%$ diff erent $8.05\%$ differ ent $4.44\%$ per cent $7.84\%$ per cent $4.53\%$ differ ent $0.29\%$ differ ent $3.05\%$ per cent $0.18\%$ per cent $1.33\%$ differ ent $0.20\%$ differ ent $3.05\%$ per cent $0.18\%$ per cent $0.90\%$ differ ent $0.20\%$ differ ent $1.36\%$ per cent $0.12\%$ per cent $0.39\%$ differ ent $0.13\%$ differ ent $0.94\%$ per cent $0.11\%$ per cent $0.38\%$ differ ent $0.03\%$ differ ent $0.94\%$ per cent $0.10\%$ per cent $0.37\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.01\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ differ ent $0.03\%$ differ ent $0.41\%$ per cent $0.09\%$ per cent $0.26\%$ happening $95.11\%$ happening $70.05\%$ together $9.94\%$ together $14.68\%$ happen ing $1.93\%$ happen ing $14.80\%$ together $0.08\%$ together $3.05\%$ happen ing $0.91\%$ happen ing $4.26\%$ together $0.03\%$ together $1.33\%$ happen ing $0.08\%$ happen ing $0.94\%$ <t< td=""><td>per c ent</td><td>8.83%</td><td>perce nt</td><td>14.61%</td><td></td><td>8.74%</td><td></td><td>14.74%</td></t<>  | per c ent                        | 8.83%  | perce nt                       | 14.61% |                                  | 8.74%  |                        | 14.74% |  |
| per ce nt $0.88\%$ pe r cent $1.33\%$ dif fer ent $0.20\%$ differ ent $1.36\%$ p er c ent $0.18\%$ per ce nt $0.90\%$ dif fer ent $0.13\%$ diff er ent $1.27\%$ p er c ent $0.12\%$ per ce n t $0.39\%$ dif fer ent $0.07\%$ diff er ent $0.94\%$ p er cent $0.11\%$ p er cent $0.38\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.10\%$ per c ent $0.37\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.10\%$ per c ent $0.26\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.09\%$ per c ent $0.26\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.09\%$ per cent $0.26\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.09\%$ per cent $0.26\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.09\%$ per cent $0.26\%$ diff er ent $0.03\%$ diff er ent $0.94\%$ p er cent $0.09\%$ per cent $14.68\%$ happen ing $1.93\%$ happen ing $14.80\%$ to ge ther $0.95\%$ to ge ther $1.35\%$ happen ing $0.91\%$ happen ing $4.26\%$ to g ether $0.03\%$ to g ether $1.33\%$ happen ing $0.08\%$ happen ing $0.94\%$ to g ether $0.03\%$ to g ether $0.95\%$ happen ing $0.03$  | per cent                         |        | perce n t                      | 6.40%  | diff ere nt                      | 8.05%  | differ en t            | 4.44%  |  |
| p er c ent $0.18\%$<br>p er ce n tper ce n t $0.90\%$<br>p er ce n tdif f ere n t $0.13\%$<br>dif f ere n tdiff ere n t $1.27\%$<br>dif f er en tp er cent $0.11\%$<br>p er centp er cent $0.38\%$<br>d iff er entdiff er ent $0.07\%$<br>diff er ent $0.94\%$<br>diff er entp er cent $0.10\%$<br>p er centper cent $0.37\%$<br>p er centdiff er ent $0.03\%$<br>diff er ent $0.94\%$<br>diff er entBPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ toge ther $9.94\%$<br>tog ethertog ether $14.68\%$<br>happen inghappen ing $1.93\%$<br>happen ing $14.80\%$<br>happen ingtoge ther $0.08\%$<br>tog ethertog ether $3.05\%$<br>happe ning $0.91\%$<br>happen ing $1.42\%$<br>happen ingtog e ther $0.03\%$<br>tog ethertog ether $1.35\%$<br>happe ning $0.08\%$<br>happen ing $0.91\%$<br>happen ingto g e ther $0.03\%$<br>to g ethertog ether $1.33\%$<br>happen ing $0.08\%$<br>happen ing $0.94\%$<br>happen ingto g ether $0.03\%$<br>to g etherto g ether $0.95\%$<br>to g ether $0.95\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ing  | perce nt                         | 7.84%  | per cent                       | 4.53%  | differ ent                       | 0.29%  | diff ere nt            | 3.05%  |  |
| p er ce n t $0.12\%$<br>p er ce n tper ce n t $0.39\%$<br>p er centdif fer ent $0.07\%$<br>dif fer entdif fer ent $0.94\%$<br>diff er entp er cent $0.11\%$<br>p er centper cent $0.38\%$<br>p er centdiff er ent $0.03\%$<br>diff er entdiff er ent $0.94\%$<br>diff er entBPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ toge ther $9.94\%$<br>to ge thertoge ther $14.68\%$<br>happen inghappen ing $1.93\%$<br>happen inghappen ing $14.80\%$<br>happen ingto ge ther $0.03\%$<br>to ge therto ge ther $1.35\%$<br>happen ing $0.16\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>diff er entto ge ther $0.03\%$<br>to ge therto ge ther $1.35\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>diff er entto ge ther $0.03\%$<br>to ge therto ge ther $0.03\%$<br>happen ing $0.91\%$<br>happen ing $0.94\%$<br>diff er entto ge ther $0.03\%$<br>to ge therto ge ther $0.95\%$<br>happen ing $0.95\%$<br>happen ing $0.08\%$<br>happen ing $0.94\%$<br>happen ingto ge ther $0.03\%$<br>to ge therto ge ther $0.95\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ingto ge ther $0.03\%$<br>to ge therto ge ther $0.95\%$<br>happen ing $0.03\%$<br>happen ing $0.89\%$   | per ce nt                        | 0.88%  | pe r cent                      | 1.33%  | dif fe re nt                     | 0.20%  | differ e nt            | 1.36%  |  |
| p er cent $0.11\%$<br>per centp er cent $0.38\%$<br>per c entd iff er ent $0.03\%$<br>d iff er entdiff er ent $0.94\%$<br>d iff er entper cent $0.10\%$<br>per c entper c ent $0.37\%$<br>per centd iff er ent $0.03\%$<br>d iff er entdiff er ent $0.94\%$<br>d iff er entBPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ toge ther $9.94\%$<br>to ge thertog ether $14.68\%$<br>happen inghappen ing $1.93\%$<br>happen inghappen ing $14.80\%$<br>happen ingto ge ther $0.03\%$<br>to ge therto ge ther $3.05\%$<br>happen ing $0.91\%$<br>happen ing $0.91\%$<br>happen ing $0.26\%$<br>happen ingto g ether $0.03\%$<br>to g etherto g ether $1.35\%$<br>happ en ing $0.08\%$<br>happ en ing $0.08\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ingto g ether $0.03\%$<br>to g etherto g ether $0.35\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ing  | p er c ent                       | 0.18%  | per ce nt                      | 0.90%  | dif f ere nt                     | 0.13%  | diff ere n t           | 1.27%  |  |
| per cent $0.10\%$<br>per c entper c ent $0.37\%$<br>per c entd if fer ent $0.03\%$<br>differ entdiffer ent $0.03\%$<br>diff e rentdiffer ent $0.03\%$<br>diff e rentdiffer ent $0.01\%$<br>diff e rentdiffer ent $0.01\%$<br>diff e rent $0.41\%$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ together $88.88\%$<br>togethertogether $69.99\%$<br>happening $193\%$<br>happeninghappening<br>$1.93\%$<br>happen ing $70.05\%$<br>happen ingto ge ther $0.95\%$<br>to ge ther $3.12\%$<br>to ge ther $1.35\%$<br>happen ening $0.91\%$<br>happen ing $1.71\%$<br>happen ing $1.83\%$<br>happen ingto g ether $0.03\%$<br>to ge ther $1.33\%$<br>to ge ther $1.33\%$<br>happen ing $0.08\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ingto ge ther $0.03\%$<br>to ge ther $0.95\%$<br>to ge ther $0.95\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ing $0.03\%$<br>happen ing $0.03\%$<br>happen ing $0.94\%$<br>happen ing   | percent                          | 0.12%  | per ce n t                     | 0.39%  | dif fer ent                      | 0.07%  | dif fer ent            | 0.94%  |  |
| per cent $0.09\%$ per cent $0.26\%$ different $0.01\%$ different $0.41\%$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ together $88.88\%$ together $69.99\%$ happening $95.11\%$ happening $70.05\%$ to ge ther $9.94\%$ tog ether $14.68\%$ happening $1.93\%$ happen ing $14.80\%$ to ge ther $0.08\%$ to ge ther $3.05\%$ happe ning $0.91\%$ happen ing $4.26\%$ to ge ther $0.07\%$ to ge ther $1.35\%$ happen ing $0.08\%$ happen ing $4.26\%$ to ge ther $0.03\%$ tog ether $1.33\%$ happ en ing $0.08\%$ happen i ng $0.94\%$ to ge ther $0.03\%$ tog ether $0.95\%$ happ en ing $0.03\%$ happ en ing $0.89\%$ to ge th er $0.03\%$ tog ether $0.95\%$ happ en ing $0.03\%$ happ en ing $0.89\%$   | p er cent                        | 0.11%  | p er cent                      | 0.38%  | d iff ere nt                     | 0.03%  | diff er ent            | 0.94%  |  |
| BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ BPE-Dropout $p=0.1$ MaxMatch-Dropout $p=0.3$ together $88.88\%$ together $69.99\%$ happening $95.11\%$ happening $70.05\%$ to ge ther $9.94\%$ tog ether $14.68\%$ happen ing $1.93\%$ happen ing $14.80\%$ to ge ther $0.95\%$ tog ether $3.12\%$ happe ning $0.91\%$ happen ing $4.46\%$ to ge ther $0.08\%$ to ge ther $3.05\%$ happe ning $0.91\%$ happen ing $4.26\%$ to ge ther $0.07\%$ tog ether $1.35\%$ happe ning $0.16\%$ happen ing $0.94\%$ to ge ther $0.03\%$ tog ether $0.95\%$ happe ning $0.03\%$ happe ning $0.94\%$ to ge th er $0.03\%$ tog ether $0.95\%$ happe ning $0.03\%$ happe ning $0.94\%$  | pe r cent                        | 0.10%  | per c ent                      | 0.37%  | d if fer ent                     | 0.03%  | differ e n t           | 0.56%  |  |
| together $88.88\%$ together $69.99\%$ happening $95.11\%$ happening $70.05\%$ to ge ther $9.94\%$ tog ether $14.68\%$ happen ing $1.93\%$ happen ing $14.80\%$ to get her $0.95\%$ tog eth er $3.12\%$ ha pp ening $1.71\%$ happen ing $14.80\%$ to ge ther $0.08\%$ to get her $3.05\%$ happ ening $0.91\%$ happen ing $4.26\%$ to ge ther $0.07\%$ t og ether $1.35\%$ ha pp ening $0.16\%$ happen i n g $1.83\%$ to g ether $0.03\%$ tog eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ tog ether $0.95\%$ h app ening $0.03\%$ happ ening $0.94\%$  | per ce n t                       | 0.09%  | pe r ce nt                     | 0.26%  | di ff er ent                     | 0.01%  | diff e rent            | 0.41%  |  |
| to ge ther $9.94\%$ tog ether $14.68\%$ happen ing $1.93\%$ happen ing $14.80\%$ to get her $0.95\%$ tog eth er $3.12\%$ ha pp ening $1.71\%$ happen ing $4.46\%$ tog ether $0.08\%$ to get her $3.05\%$ happ ening $0.91\%$ happ ening $4.26\%$ to ge t her $0.07\%$ t og ether $1.35\%$ ha pp ening $0.16\%$ happen i n g $1.83\%$ to g ether $0.03\%$ tog eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ tog ether $0.95\%$ h app ening $0.03\%$ happ ening $0.89\%$   | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout p=0.3         |        | <b>BPE-Dropout</b> <i>p</i> =0.1 |        | MaxMatch-Dropout p=0.3 |        |  |
| to get her $0.95\%$ tog eth er $3.12\%$ ha pp ening $1.71\%$ happen in g $4.46\%$ tog ether $0.08\%$ to get her $3.05\%$ happ ening $0.91\%$ happ ening $4.26\%$ to ge t her $0.07\%$ t og ether $1.35\%$ ha pp en ing $0.16\%$ happen i n g $1.83\%$ to g ether $0.03\%$ t og eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ t og ether $0.95\%$ h app en ing $0.03\%$ happ en ing $0.89\%$  | together                         | 88.88% | together                       | 69.99% | happening                        | 95.11% | happening              | 70.05% |  |
| tog ether $0.08\%$ to get her $3.05\%$ happ ening $0.91\%$ happ ening $4.26\%$ to ge t her $0.07\%$ t og ether $1.35\%$ ha pp en ing $0.16\%$ happen i n g $1.83\%$ to g ether $0.03\%$ tog eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ tog ge ther $0.95\%$ h app ening $0.03\%$ happ ening $0.89\%$  | to ge ther                       | 9.94%  | tog ether                      | 14.68% | happen ing                       | 1.93%  | happen ing             | 14.80% |  |
| to ge t her $0.07\%$ t og ether $1.35\%$ ha pp en ing $0.16\%$ happen i n g $1.83\%$ to g ether $0.03\%$ tog eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ to ge ther $0.95\%$ h app en ing $0.03\%$ happ en ing $0.89\%$  | to get her                       | 0.95%  | tog eth er                     | 3.12%  | ha pp ening                      | 1.71%  | happen in g            | 4.46%  |  |
| to g ether $0.03\%$ tog eth e r $1.33\%$ happ en ing $0.08\%$ ha pp ening $0.94\%$ to ge th er $0.03\%$ to ge ther $0.95\%$ h app en ing $0.03\%$ happ en ing $0.89\%$   | tog ether                        | 0.08%  | to get her                     | 3.05%  | happ ening                       | 0.91%  | happ ening             | 4.26%  |  |
| to ge ther $0.03\%$ to ge ther $0.95\%$ h app en ing $0.03\%$ happ en ing $0.89\%$   | to ge t her                      | 0.07%  | t og ether                     | 1.35%  | ha pp en ing                     | 0.16%  | happen i n g           | 1.83%  |  |
|  | to g ether                       | 0.03%  | tog eth e r                    | 1.33%  | happ en ing                      | 0.08%  | ha pp ening            | 0.94%  |  |
| tog a ther $0.00\%$ tog at her $0.94\%$ happen pring $0.03\%$ happen pring $0.44\%$  | to ge th er                      | 0.03%  | to ge ther                     | 0.95%  | h app en ing                     | 0.03%  | happ en ing            | 0.89%  |  |
| 106 e then $0.0070$ tog et her $0.7770$ happ e hing $0.0570$ happ elling $0.4470$  | tog e ther                       | 0.00%  | tog et her                     | 0.94%  | happ e ning                      | 0.03%  | h app ening            | 0.44%  |  |
| t og ether $0.00\%$ to get he r $0.90\%$ ha pp e ning $0.02\%$ happ e ning $0.40\%$  | t og ether                       | 0.00%  | to get he r                    | 0.90%  | ha pp e ning                     | 0.02%  | happ e ning            | 0.40%  |  |
| tog eth er $0.00\%$ to get her $0.41\%$ h app ening $0.01\%$ ha p pe ning $0.27\%$   | tog eth er                       | 0.00%  | t o get her                    | 0.41%  | h app ening                      | 0.01%  | ha p pe ning           | 0.27%  |  |

## **D** Examples of Distributions from Data

Example 2: Frequencies of tokenizations of several words sampled from BPE-Dropout (with p = 0.1) and MaxMatch-Dropout (with p = 0.3). The top row in each is the canonical tokenization.