On the Effectiveness of Quasi Character-Level Models for Machine Translation

Salvador Carrión salcarpo@prhlt.upv.es
Francisco Casacuberta fcn@prhlt.upv.es
PRHLT Research Center, Universitat Politècnica de València

Abstract

Neural Machine Translation (NMT) models often use subword-level vocabularies to deal with rare or unknown words. Although some studies have shown the effectiveness of purely character-based models, these approaches have resulted in highly expensive models in computational terms. In this work, we explore the benefits of quasi-character-level models for very low-resource languages and their ability to mitigate the effects of the catastrophic forgetting problem. First, we conduct an empirical study on the efficacy of these models, as a function of the vocabulary and training set size, for a range of languages, domains, and architectures. Next, we study the ability of these models to mitigate the effects of catastrophic forgetting in machine translation. Our work suggests that quasi-character-level models have practically the same generalization capabilities as character-based models but at lower computational costs. Furthermore, they appear to help achieve greater consistency between domains than standard subword-level models, although the catastrophic forgetting problem is not mitigated.

1 Introduction

Neural machine translation (NMT) has become the dominant paradigm in the field of machine translation due to the impressive results obtained with the encoder-decoder architectures (Sutskever et al., 2014; Cho et al., 2014; Wu et al., 2016; Vaswani et al., 2017) and the approaches proposed to tackle the open vocabulary problem such as subword-based models with byte-fallback or, more recently, token-free models.

However, despite these advances, low-resource languages are still problematic as there are many languages that are spoken but not written on the internet (e.g., Tigrinya, Sotho, Tsonga, etc.), and therefore, parallel text mining techniques are either not effective or not applicable at all. In these cases, it is common to use character-based models, since multiple authors have shown that these models usually perform better than (standard) subword- or word-based models in very low-resource settings.

Motivated by these ideas, we decided to study whether quasi-character-based vocabularies (defined as a subword-based vocabulary that is one or two orders of magnitude smaller than a standard subword-based vocabulary), had the same advantages as models with character-based vocabularies for low-resource languages, but with much lower computational costs, due to the exponential decrease in the average number of tokens per sentence when merging highly frequent character pairs.

Furthermore, given that the effects of the catastrophic forgetting problem are strongly related to the vocabulary of the model, we decided to study if these quasi-character-level vocabularies had the potential to mitigate them, since these vocabularies are closer to a universal-domain vocabulary (e.g., bytes or chars) than a word- or (standard) subword-based vocabulary.

The contributions of this paper are twofold:

- Quasi-character-level models appear to outperform character-based models in terms of performance, while offering practically the same generalization capabilities at much lower computational costs.

- Quasi-character-level models appear to achieve higher consistencies in performance between domains, but at the same time, they also seem to be more susceptible to the effects of the catastrophic forgetting problem.

2 Related work

Character-based models have been widely studied in the field of Natural Language Processing (NLP) to deal with the open vocabulary problem. Vilar et al. (2007) proposed one of the first character-based models, who treated source and target sentences as a string of letters. Similarly, Neubig et al. (2013) viewed translation as a single transduction between character strings in the source. However, these results were unsatisfactory, as their models generally performed worse than their word-based counterparts.

To overcome these problems, many authors have proposed strategies based on hybrid models (Luong and Manning, 2016), which mainly translated at the word level except when a rare or unknown word was encountered; subword-based model (Sennrich et al., 2016; Kudo, 2018; Kudo and Richardson, 2018), which allow to efficiently represent a word as a sequence of subwords; and more recently, token-free models (Xue et al., 2021; Clark et al., 2021), which operate directly on raw text.

Despite these improvements, character-based models are still interesting for low-resource languages since multiple authors have shown their benefits over other approaches. For example, Cherry et al. (2018) showed that character-level models have their greatest advantage when data sizes are small; Sennrich and Zhang (2019) showed that reducing the vocabulary size leads to improvements for low-resource NMT models. Similarly, by studying the Zipfian nature of languages in NMT, other authors have reached similar conclusions. Raunak et al. (2020) characterized the long-tailed phenomena in NMT, and Gowda and May (2020) proved that each dataset has an optimal vocabulary size. Although this optimal vocabulary size has been traditionally found by trial and error, very recently, Xu et al. (2020) has proposed a new technique to explore automatic vocabularization without trial and error. However, despite its impressive results, this method still requires a non-trivial amount of time[1]. Hence, heuristics will remain an effective solution to the vocabulary-size problem.

In this work, we focused our efforts on preserving the advantages of character-based models but at much lower computational costs. In this line, other authors have introduced new ideas, such as Lee et al. (2016), who used convolutional and max-pool layers to reduce the length of the character-level representations; Cherry et al. (2018) showed that alternative architectures for handling character input are better viewed as methods for reducing computation time than as improved ways of modeling longer sequences; Kreutzer and Sokolov (2018) proposed an approach to learning input and output segmentations for NMT, which favors character-level approaches; and more recently, Mielke et al. (2021) published a survey about tokenization and the open-vocabulary problem, where they concluded that it is likely that there will never be a silver bullet solution for all applications.

This work briefly studies the generalization capabilities of quasi-character-level models for different neural architectures. Many authors have extensively studied the limitations of existing tokenizations and neural architectures for text processing tasks. For instance, Conneau et al.

[1] 30 GPU hours on the WMT-14 English-German translation dataset

(2016) showed a state-of-the-art CNN architecture for text processing that operated directly at the character level; Araabi and Monz (2020) showed that the effectiveness of Transformer under low-resource conditions is highly dependent on the hyper-parameter settings; Banar et al. (2020) presented a fast character transformer via gradient-based subword tokenization.

Finally, this work ends with a brief discussion on the ability of quasi-character-level models to mitigate the effects of the catastrophic forgetting problem in NMT. As far as we know, this is the first work to address this problem from this perspective, since most of the works that we know of are based on regularization (Li and Hoiem, 2016; Kirkpatrick et al., 2016), dynamic architectures (Rusu et al., 2016; Draelos et al., 2016) or Complementary Learning Systems (CLS) (Kemker and Kanan, 2017).

3 Neural Machine Translation

3.1 Neural architectures for Machine Translation

The goal of any translation system is to transform an input sequence in a given language into an output sequence in a target language.

Nowadays, this is usually done using neural models based on the encoder-decoder architecture, also known as Seq-to-Seq models in the machine translation community (Sutskever et al., 2014). The encoder part transforms the input sequence into an internal representation, and then the decoder transforms this internal representation into the output sequence.

Recurrent architectures (RNNs) were the first to be successfully applied in an encoder-decoder setup for machine translation. Even though there are many RNNs, most chain a series of unit cells sequentially to process temporal sequences. We decided to use LSTMs (Hochreiter and Schmidhuber, 1997) because their unit cells are explicitly designed to deal with long-term dependencies.

Convolution-based architectures (CNN) do not contain any recurrent elements. They can do this because the idea behind this architecture is that the convolutional filters can slide through the sequence of tokens from beginning to end (Gehring et al., 2017).

Lastly, Vaswani et al. (2017) introduced the Transformer architecture, which is a state-of-the-art model based entirely on the concept of *attention* (Bahdanau et al., 2015; Luong et al., 2015) to draw global dependencies between the input and output. Unlike RNNs or CNNs, this architecture processes its temporal sequences all at once through masks that encode temporal information.

This work is focused on the Transformer as it is the current state-of-the-art model for machine translation. Nonetheless, RNNs and CNNs are briefly explored for completeness (See Section 5.4.3).

3.2 The open vocabulary problem

In the written language, it is common to find alternative spellings (i.e., *color-colour*) and typos (i.e., *acknowledge-acknowlege*) that slightly modify the spelling of a word but do not prevent us, the humans, from understanding its meaning. However, suppose a model is using a word-level representation. In that case, it will stop knowing a *known word* at the very first moment that it is slightly modified (and this modification is not in its vocabulary). Similarly, it has to be taken into account that many languages use agglutination and compounding mechanisms to form new words, making word-based vocabularies a very inefficient approach.

As a result, researchers have proposed multiple approaches to deal with the open vocabulary problem, such as bytes- or character-based models, hybrid models, subword-based models, or, more recently, token-free models.

Arguably, a character-based vocabulary[2] is the most straightway to solve the open-vocabulary problem, as it contains the minimum set of characters with which to form every possible word in a given language. Because of this, these types of models have the potential to translate every possible word, even rare or even unseen words, if enough information is present in the training set. However, despite the many innovations (Jaszczur et al., 2021; Banar et al., 2020; Chung et al., 2016; Kreutzer and Sokolov, 2018), these models tend to be much slower, resource-hungry, and harder to train than standard subword-based or word-based models, as they have to deal with longer long-term dependencies.

Given that subword-level vocabularies can degenerate to character- or word-based vocabularies, we decided to use this property to build vocabularies that are one or two orders of magnitude smaller than the standard subword-level vocabularies[3], with the goal of having virtually the same benefits of character-level models for low-resources languages, but at much lower computational costs.

4 Experimental setup

4.1 Datasets

The data used for this work comes mainly from the WMT tasks (see Table 1).

Dataset	Training set
Europarl (es/de/cs/sv/da/bg/zh/ru-en)	50K/100K/1-2M
CommonCrawl (es-en)	100K/1.8M
SciELO (es-en)	120K/575K
NewsCommentary (de-en)	35K/357K
IWLST'16 (de-en)	196K
Multi30K (de-en)	29K
Tatoeba (mr-en)	53K
CCAligned (or-en)	3K

Table 1: These datasets contain parallel sentences from different languages and domains (political, economic, health, biological, talks, etc.). All the values in this Table indicate the number of sentences.

4.2 Training details

For preprocessing and training we used AutoNMT (Carrión and Casacuberta, 2022), with *Unigram/SentencePiece* (Kudo, 2018; Kudo and Richardson, 2018) as the subword model, shared vocabularies, and Fairseq as the training framework (v1.0.0a0), on 2x NVIDIA GP102 (TITAN XP) - 12GB.

Initially, we started to experiment with the standard Transformer (45-93M parameters), but then we switched to a smaller version (4-25M parameters), as both performed quite similarly in terms of performance ($\pm 1 - 3$ BLEU), while the latter was notably faster. Similarly, other seq-to-seq neural architectures were used for completeness (Transformer, LSTMs, and CNNs).

In all cases, the set training hyper-parameters were pretty standard[4]. Despite using similar settings in most models, we noticed that as we used smaller vocabularies and training sets,

[2]Plus a byte-level fallback

[3]Vocabularies of 100-500 tokens vs. vocabularies of 30K-40K tokens)

[4]Hyper-parameters:`lr=[0.5e-4, 1e-3]; weight-decay=[1e-3, 1e-4]; criterion=[ce, label-ce(0.1)]; scheduler=[fixed, inverse-sqrt]; warmup-updates=[4000]; optimizer=[adam, sgd, nag]; clip-norm=[0.0, 0.1, 1.0]; beam-width=5]`

these models became more sensitive to the given hyper-parameters. The training time for most models was between a few hours to one or two days, and all models were evaluated with Sacrebleu (Post, 2018) and BERTScore (Zhang et al., 2019).

5 Experimentation

5.1 On Quasi-Character-Level Hypothesis

Given two different vocabularies, A and B, we could say that they are grammatically equivalent if both can represent any possible word of a given language. Because of this, the smaller a vocabulary is, the greater the generalization capabilities of the model used will have to be to end up with good translations, as the amount of information per token will be diluted by the number of tokens needed to encode each string.

Based on this premise, we can infer that the representation power of a given model will depend on the degree of generalization required by its vocabulary, the amount of data required to learn it, and if the complexity of the model can handle it. Hence, given a model with enough complexity, the advantages of character-based vocabularies will decrease with respect to subword-based or word-based vocabularies as the amount of data increases.

Based on these premises, supported by empirical evidence Sennrich and Zhang (2019), we hypothesized that quasi-character-based models should provide practically the same generalization capabilities as character-level models, but more efficiently, by exploiting highly frequent n-grams to decrease the sentence length exponentially.

5.2 Effects of the vocabulary and corpus size

In order to test the basis of our hypothesis, we chose a medium-sized corpus such as Europarl-2M (de-en). Then, two other versions were created, where the training set was artificially reduced from 2M sentences to 100k and 50k sentences. Similarly, we created two vocabularies:

- A standard subword-level vocabulary with 32k entries.
- A quasi-character-level vocabulary with 350 entries.

The aim of this experiment was twofold. First, we sought to confirm that smaller vocabularies tend to help in low-resource environments (Cherry et al., 2018), in addition to proving additional data points for smaller datasets (less than 2M sentences), languages, and domains. Second, we sought to establish baselines for our quasi-character-based models so that we could later study their computational advantage over purely character-level models.

As expected, in Figure 1 we see that when there is enough training data, standard subword-level models outperform quasi-character-level models (first column). In contrast, when the amount of training data was reduced (second and third columns), the quasi-character-level models outperformed the standard subword-level models.

In total, we performed this experiment for three different language pairs (Spanish-English, German-English, and Czech-English) to account for potential language biases and domains (political, economical, health, biological, transcripted talks, etc.). The low-resource settings were emulated in this experiment because high-quality datasets contain less noise. Consequently, we could generalize the findings of previous authors to much smaller corpora more confidently, and also, test the basis of our hypothesis for quasi-character-level vocabularies (See section 5.3 for actual low-resource languages).

5.3 On the Effectiveness of Quasi-Character-Level Models

As the results from our previous experiment could be compromised towards a sub-optimal vocabulary size, we repeated the previous experiment, but this time, we gradually increased the

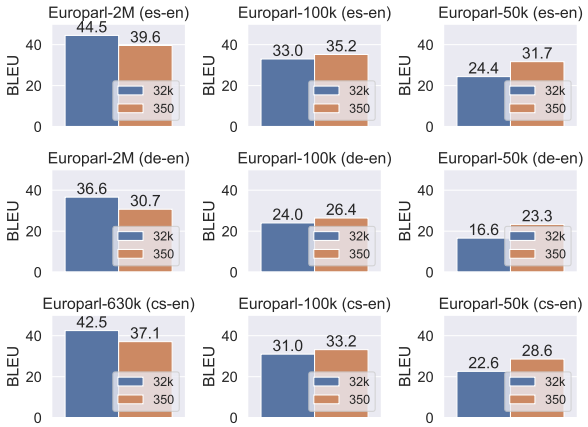
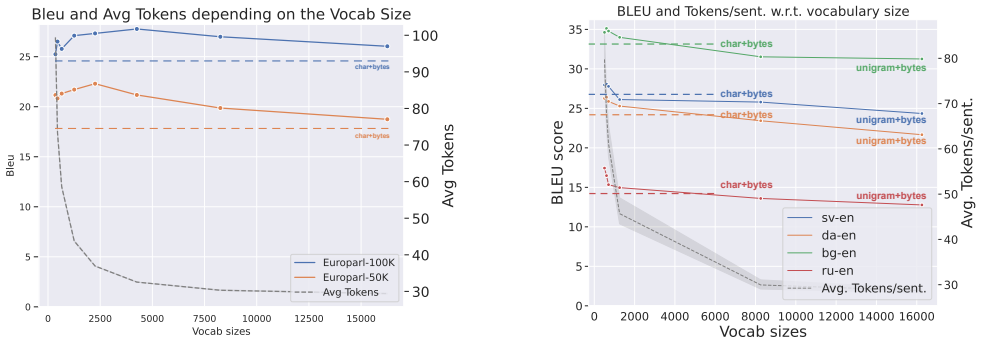


Figure 1: As we limited the training data (from left to right), quasi-character-level models perform better than standard subword-level models, regardless of language (top to bottom).

vocabulary size (at the subword-level) from 100 tokens to 16,000 tokens (plus 256 additional entries for the byte-fallback). Moreover, we added five actual low-resource languages (non-emulated) and three non-latin languages to the experiment in order to account for potential biases (See Figure 2).



(a) (Emulated) Low-Resource languages

(b) (Actual) Low-Resource languages

Figure 2: As we decrease the size of the vocabulary, the average number of tokens per sentence increases exponentially. Hence, more complex models and more training data are needed for exploiting the generalization capabilities of these vocabularies. In contrast, by merging a few highly frequent char-pairs into a single token, we can have models that practically generalize as character-based models but at much lower computational costs.

In Figure 2a we have the BLEU scores and the average tokens per sentence as a function of the vocabulary size, for two low-resource emulations of the Europarl (de-en) dataset, one with 50k sentences (orange line) and another with 100k sentences (blue line).

Firstly, we see that for both datasets, as the number of entries in the vocabulary decreases, the performance of our models increases. However, this phenomenon is much stronger on the smaller corpus (Europarl-50k), thing that might indicate that for high-quality corpus,

the advantages of character-level models could disappear much quicker than was previously thought (Cherry et al., 2018).

Secondly, we see that as vocabulary size approaches a character-level representation[5], the average number of tokens per sentence increases exponentially (dashed line). This phenomenon has a direct impact on the performance of the model due to: i) The additional complexity needed to handle the greater generalization capabilities of smaller vocabularies; ii) The problems imposed by having to deal with longer long-term dependencies; and iii) Higher computational costs at training and run-time.

Fortunately, the opposite is also true. As the vocabulary size increases, the average number of tokens per sentence decreases exponentially, and therefore, the models need less complexity. This can also be seen in Figure 2a, where the quasi-character-level models outperformed purely character-based models (dashed lines) by a significant margin without increasing the complexity of this model (or the training time).

Similarly, after repeating these experiments with actual low-resource language and non-latin languages, the results remained quite consistent for most languages (Swedish, Danish, Bulgarian, Russian) and scripts (latin and non-latin) (See Figure 2b).

However, there is no silver bullet as we noticed three language pairs where this phenomenon was not observed. The first was with Chinese-English, probably due to the large number of individual characters present. And then, with two very low-resource pairs (Marathe and Oriya), where the Bleu-Vocab curve remained flat.

5.4 On the Generalization of Quasi-Character-based approaches

In this section, we study whether the benefits of Quasi-Character-based approaches generalize to other domains, and neural architectures.

5.4.1 Domain generalization

To study whether the domain might be influencing the results from Section 5.2, we decided to repeat the same experiment but using parallel corpora from different domains, such as crawled data (CommonCrawl), political and economic news (NewsCommentary), health and biological sciences (SciELO), transcribed talks (IWLST'16) and multimodal transcriptions (Multi30k).

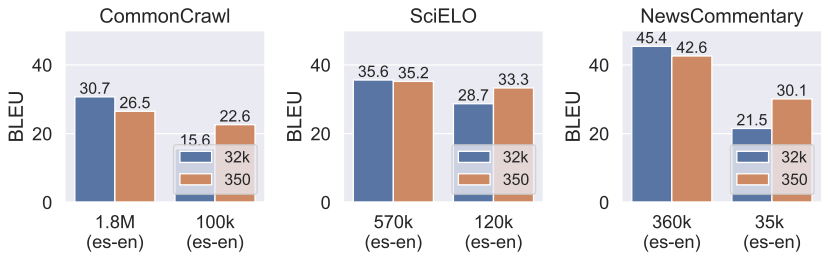


Figure 3: The benefits of quasi-character-level models for low-resource environments appear to be consistent regardless of domain.

The results from Figure 3 show the BLEU scores of the quasi-character-level and standard subword-level models trained on high- and low-resource settings[6], corresponding to different domains (Crawled data, Science and News). As in Section 5.2, the quasi-character-based models kept outperforming the standard subword-based models for the low-resource settings, re-

[5]Horizontal dashed lines indicate the character-level baselines

[6]Emulated

gardless of the training domain. These results seem to indicate that this phenomenon is not only language-agnostic, but also domain-agnostic.[7]

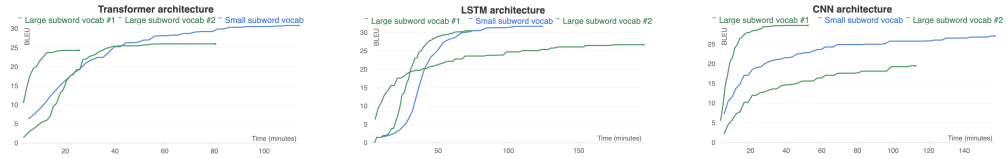
5.4.2 Performance comparison

In Figures 2a and 2b we see that the average amount of tokens (gray line) decreases exponentially (up to a point), when the vocabulary size is increased. As a result, our quasi-character-level models processed on average between 30% and 60% fewer tokens than the character-based models, depending on the language, the vocabulary size, and the dataset.

The exact speedup is highly dependent on the training setup, since it is not the same to limit the number of sentences per batch than to limit the number of tokens per batch. Nonetheless, in both cases, we obtained a non-negligible optimization. In the first case, the most significant improvement was in terms of memory consumption due to the quadratic complexity of the Transformer's self-attention. While in the second case, it was from reducing the number of batches needed to process a single epoch.

5.4.3 Neural architecture generalization

In this section, we study whether the above findings can be generalized to other architectures such as LSTMs or CNNs, or whether, on the contrary, the advantages of the quasi-character-level models are mainly due to the ability of Transformers to learn long-term dependencies.



(a) Transformer architecture (b) Bi-LSTM architecture (c) CNN architecture

Figure 4: The green lines refer to the best and worst performances of the standard subword-level models, while the blue lines refer to the best performance of the quasi-character-level models.

Specifically, we focused our study on bidirectional LSTMs with attention mechanisms, and fully convolutional architectures like the one described in (Gehring et al., 2017).

Although the comparison between different neural architectures is not a trivial task, we attempted to explore this topic by only comparing models that had a similar number of parameters for a given vocabulary (i.e., 25-30M parameters for 32k subword vocabularies).

From our experimentation, we observed that when the standard subword-level models were trained with sufficient data, they outperformed all the quasi-character-level models, regardless of their architecture. However, when this experiment was repeated in the low-resource regime, the quasi-character-based models performed better than their standard subword-level counterparts, regardless of their architecture[8] (See Figure 4).

In the left figure 4a, we see that quasi-character-level Transformers consistently outperform the standard subword-level models. This phenomenon is still present for LSTMs (middle Figure 4b), but it is not as evident as with the Transformer architecture due to the problems of RNNs with modeling long-term dependencies. Finally, in the right Figure 4c we see that CNNs

[7] The experiments done with IWLST'16 and Multi30K datasets yielded similar results. In these, the improvement for the quasi-character-based models was +6.2pts (BLEU) for the IWLST'16 dataset, and +2.3pts (BLEU) for the Multi30k dataset.

[8] In Figure 4c the quasi-character-level model did not outperform the standard subword-level models. This was due to stopping the training too soon. Nonetheless, we are confident that the quasi-character-level model would have caught the standard subword-level model.

do not benefit as easily from the quasi-character-level representations as they cannot model long-term dependencies so easily.

From these results, we conclude that the ability of a neural architecture to model long-term dependencies is critical to derive benefits from either character-based or quasi-character-based representations.

5.5 On the Catastrophic Forgetting Problem

In this section, we study whether quasi-character-level models could help to mitigate the effects of the catastrophic forgetting phenomenon, whereby neural networks forget previously learned information after learning new information.

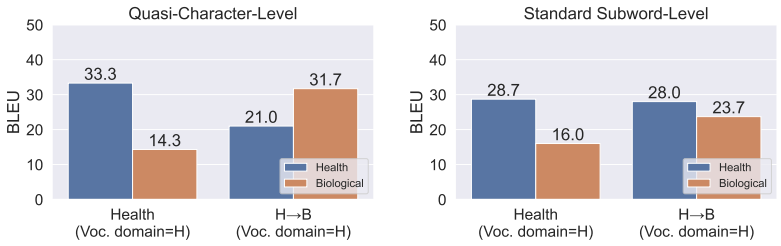


Figure 5: Vocabularies seem to have a strong impact on the catastrophic forgetting effects. While the quasi-character-level model lost 12.3pts, the large subword-level model only lost 0.7pts

To do this, we designed an experiment in which we first train a model in a domain *A* and it is evaluated in domains *A* and *B* to establish the baselines. Next, we fine-tuned the model trained in domain *A* with data from the new domain *B*, and then, it is evaluated it in domains *A* and *B*. In theory, the model trained in domain *A* should perform well in the domain *A*, and poorly in the unseen domain *B*. Similarly, after the fine-tuning on domain *B*, it should perform worse in *A* and better in domain *B* than the original model trained only on domain *A*.

In Figure 5a the quasi-character-level model trained on the health domain (SciELO) obtained a BLEU of 33.3pts on its domain (Health) and a BLEU of 14.3pts in the other domain (Biological). Then, when we fine-tuned it on the Biological domain (SciELO), the BLEU obtained on this domain increased from 14.3 to 31.7pts, while BLEU for the health domain fell from 33.3 to 21.0pts. Similarly, the standard subword-based model also suffered from the effects of the catastrophic forgetting problem. However, they were not as significant as in the quasi-character-based model, given that the BLEU score went from 28.7 to 28.0pts.

The vocabulary chosen seems to have a significant impact on the effects of catastrophic forgetting, given that the models with quasi-character vocabularies were more susceptible to the effects of catastrophic forgetting than those using standard subword-level vocabularies.

To explore this phenomenon in more detail, we repeated the previous experiment but taking into account the vocabulary domain. As a result, we found that the vocabulary domain has a more substantial impact on model performance than we thought. As shown in Figure 6, quasi-character-level models appear to be very consistent across domains, while standard subword-level models seem to be especially sensitive to their vocabulary's domain, to the point of obtaining opposite results across domains (see the right column of Figure 6).

Although the quasi-character-level models achieved better cross-domain consistencies, they also appear to suffer more severely from the effects of the catastrophic forgetting problem

Figure 6: Quasi-character-level models (left figures) appear to be more consistent between domains than models with standard subword-level vocabularies (right figures)

than standard subword-level models. Therefore, we expect that by using regularization-based techniques such as LwF (Li and Hoiem, 2016) or EWC (Kirkpatrick et al., 2016)), these effects could be mitigated to a great extent, leading to more robust and consistent models.

6 Conclusion

In this work, we have studied the effectiveness of quasi-character-level models in terms of performance and computational efficiency relative to purely character-based models and standard subword-level models. Furthermore, we have studied the generalization of quasi-character-level vocabularies and their ability to address the problem of catastrophic forgetting.

Our studies reveal that quasi-character-level models offer practically the same generalization capabilities as character-level models, but at much lower computational costs. Furthermore, these models outperformed both the standard subword-based and character-based models in low-resource environments, regardless of language, domain, and neural architecture.

Finally, we have shown that even though quasi-character-level models do not appear to mitigate the effects of the catastrophic forgetting problem, they achieved better cross-domain consistencies, which could lead to substantial improvements if specific regularization techniques are applied to deal with the catastrophic forgetting problem.

Acknowledgment

Work supported by the Horizon 2020 - European Commission (H2020) under the SELENE project (grant agreement no 871467) and the project Deep learning for adaptive and multimodal interaction in pattern recognition (DeepPattern) (grant agreement PROMETEO/2019/121). We gratefully acknowledge the support of NVIDIA Corporation with the donation of a GPU used for part of this research.

References

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. *CoRR*, abs/2011.02266.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Banar, N., Daelemans, W., and Kestemont, M. (2020). Character-level transformer-based neural machine translation. *CoRR*, abs/2005.11239.

Carrión, S. and Casacuberta, F. (2022). Autonmt: A framework to streamline the research of seq2seq models.

Cherry, C., Foster, G. F., Bapna, A., Firat, O., and Macherey, W. (2018). Revisiting character-based neural machine translation with capacity and compression. *CoRR*, abs/1808.09943.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111. Association for Computational Linguistics.

Chung, J., Cho, K., and Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *CoRR*, abs/1603.06147.

Clark, J. H., Garrette, D., Turc, I., and Wieting, J. (2021). CANINE: pre-training an efficient tokenization-free encoder for language representation. *CoRR*, abs/2103.06874.

Conneau, A., Schwenk, H., Barrault, L., and LeCun, Y. (2016). Very deep convolutional networks for natural language processing. *CoRR*, abs/1606.01781.

Draelos, T. J., Miner, N. E., Lamb, C. C., Vineyard, C. M., Carlson, K. D., James, C. D., and Aimone, J. B. (2016). Neurogenesis deep learning. *CoRR*, abs/1612.03770.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252.

Gowda, T. and May, J. (2020). Finding the optimal vocabulary size for neural machine translation. In *Findings of the ACL: EMNLP 2020*, pages 3955–3964.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Jaszczur, S., Chowdhery, A., Mohiuddin, A., Kaiser, L., Gajewski, W., Michalewski, H., and Kanerva, J. (2021). Sparse is enough in scaling transformers. *CoRR*, abs/2111.12763.

Kemker, R. and Kanan, C. (2017). Fearnet: Brain-inspired model for incremental learning. *CoRR*, abs/1711.10563.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2016). Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Kreutzer, J. and Sokolov, A. (2018). Learning to segment inputs for NMT favors character-level processing. *CoRR*, abs/1810.01480.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 66–75.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71.

Lee, J., Cho, K., and Hofmann, T. (2016). Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Li, Z. and Hoiem, D. (2016). Learning without forgetting. *CoRR*, abs/1606.09282.

Luong, M.-T. and Manning, C. D. (2016). Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1054–1063.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on EMNLP*, pages 1412–1421.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., and Tan, S. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508.

Neubig, G., Watanabe, T., Mori, S., and Kawahara, T. (2013). Substring-based machine translation. *Machine Translation*, 27(2):139–166.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Raunak, V., Dalmia, S., Gupta, V., and Metze, F. (2020). On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. *CoRR*, abs/1606.04671.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1715–1725.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *NIPS*, volume 27.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st NeurIPS*, NIPS'17, page 6000–6010.

Vilar, D., Peter, J.-T., and Ney, H. (2007). Can we translate letters? In *Proceedings of the Second WMT*, StatMT '07, page 33–39.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Xu, J., Zhou, H., Gan, C., Zheng, Z., and Li, L. (2020). VOLT: improving vocabularization via optimal transport for machine translation. *CoRR*, abs/2012.15671.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2021). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *CoRR*, abs/2105.13626.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.