

Multiscale sequence modeling with a learned dictionary

Bart van Merriënboer¹, Amartya Sanyal², Hugo Larochelle³, Yoshua Bengio^{1,4}

¹MILA, ²IIT Kanpur, ³Google Brain, ⁴CIFAR Fellow

Abstract

We propose a generalization of neural network sequence models. Instead of predicting one symbol at a time, our multi-scale model makes predictions over multiple, potentially overlapping multi-symbol tokens. A variation of the byte-pair encoding (BPE) compression algorithm is used to learn the dictionary of tokens that the model is trained with. When applied to language modelling, our model has the flexibility of character-level models while maintaining many of the performance benefits of word-level models. Our experiments show that this model performs better than a regular LSTM on language modeling tasks, especially for smaller models.

1. Introduction

Sequence modeling is the task of learning a probability distribution over a set of finite sequences. We consider sequences of elements drawn from a finite set of symbols, $s_t \in \Sigma$. In the context of language modeling this is the probability mass function of a set of strings given some alphabet of characters.

$$p(s_1 \dots s_n), \quad s_t \in \Sigma \quad (1)$$

Most approaches in language modeling follow [1] and model sequences as acyclic Markov chains, exploiting the fact that natural languages have strong temporal dependencies (see figure 1).

$$p(s_1 \dots s_n) \approx \prod_{t=1}^n p(s_t | s_1 \dots s_{t-1}), \quad s_t \in \Sigma \quad (2)$$

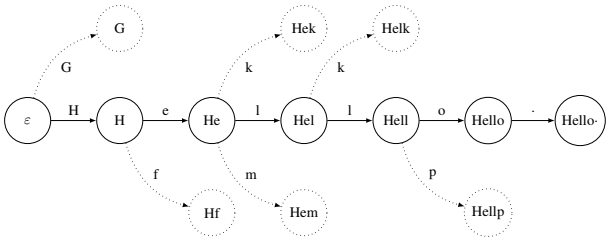


Figure 1: Diagrammatic representation of a character-level language model as a Markov chain. The probability of the string “Hello” is the probability of reaching the absorbing state “Hello.” starting from the empty string (ϵ), where \cdot is a special end-of-string (EOS) token. Each state transition is analogous to concatenating a token from the dictionary Σ to the state.

Recurrent neural networks (RNNs) can be used to efficiently model these Markov chains [2, 3]. The hidden state of the network can encode the subsequence that is conditioned on ($s_1 \dots s_{t-1}$) using constant memory. Let x_t be an embedding

of symbol s_t , then an RNN model is of the form

$$h_t = f(x_t, h_{t-1}) \quad (3)$$

$$y_t = g(h_t) \quad (4)$$

Typically the function f is a long-short term memory (LSTM) unit or gated recurrent unit (GRU), and g is a linear transformation followed by a softmax activation.

1.1. Tokenization

Natural language is naturally represented as a sequence of characters [4, 2, 5]. However, in practice text is usually ‘tokenized’ and modeled as a sequence of words instead of characters (see figure 2). Word-level models display superior performance to character-level models, which we argue can be explained by several factors.

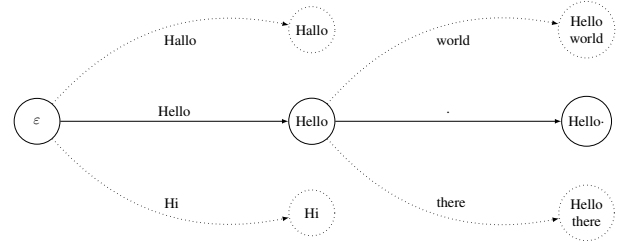


Figure 2: A word-level language model, requiring fewer transitions in order to reach the state “Hello.”. The state space is significantly reduced, which means that many strings cannot be modeled.

1.1.1. Training difficulties

Tokenization reduces the length of the sequences to model. Learning long-term dependencies with RNNs can be difficult [6, 7, 8], and in natural language dependencies such as agreement in number or case can span tens or even hundreds of characters.

Furthermore, the softmax generally used in neural networks can never assign a probability of exactly 1 to the correct token. The product of many probabilities less than 1 in equation 2 causes the probability of a sequence to decay quickly as it grows longer. To counteract this behaviour a network will quickly learn to fully saturate the softmax. However, this slows down learning [9].

1.1.2. Compositionality

In the context of natural language it can be argued that character-level and word-level language modeling are fundamentally different.

Word-level models rely on the principle of compositionality [10] and can learn to represent the meaning of a sentence

by combining the semantic representations of its constituent words [11]. This mapping is arguably ‘smooth’: If words have similar semantics, the sentences they form are likely to have a similar meaning and representation as well.

A character-level model performs a second, different task: Mapping a sequence of characters to the semantic representation of a morpheme. The principle of compositionality does not apply in this case. It is a lookup operation which is entirely non-linear: ‘the’, ‘then’ and ‘they’ are entirely unrelated. It is possible that character-level RNN models perform worse than word-level models because RNNs are ill-suited to perform this lookup operation. It is feasible that other application domains have a similar hierarchical structure in their sequential data e.g. sequence motifs in genetics.

2. Multi-scale sequence modeling

In typical sequence models, we model the likelihood of the next symbol individually. A single symbol (e.g. a character) is selected from a dictionary of mutually exclusive options. In this paper we propose a more general setting in which at each step we make predictions over multi-symbol tokens, potentially multiple of which are correct if they share a prefix (see figure 3).

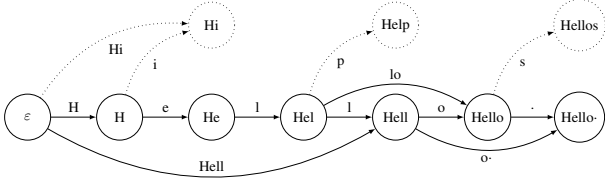


Figure 3: The multi-scale model allows multiple outgoing transitions, maintaining the flexibility of a character-level model while incorporating many of the benefits of word-level models. Any path through the Markov chain from ε to Hello is a segmentation of the string Hello using the tokens in the dictionary. The probability of the state Hello is the sum of the likelihood of each segmentation. When modeled using an RNN, each state corresponds to a hidden state \mathbf{h}_t , and each arrow corresponds to the application of the transition function f which takes inputs \mathbf{h}_t and token embedding \mathbf{x}_i .

Formally, given a set of symbols Σ , consider a dictionary of multi-symbol tokens T , where $\Sigma \subset T$. (This condition guarantees that the space of sequences we can model is the same as for typical symbol-level models.) Let $|t_i|$ denote the number of symbols in token t_i . The Markov chain (see figure 3) for a sequence $s_1 \dots s_n$, $s_t \in \Sigma$, can be modeled using an RNN as follows:

$$\mathbf{h}_t = \frac{1}{|T_t|} \sum_{T_t} f(\mathbf{x}_i, \mathbf{h}_{t-|t_i|}), \quad (5)$$

$$T_t = \{t_i : t_i \in T, t_i = s_{t-|t_i|+1} \dots s_t\} \quad (6)$$

$$\mathbf{y}_t = g(\mathbf{h}_t) \quad (7)$$

where \mathbf{x}_i is an embedding of token t_i . Note that a typical RNN model (e.g. a character-level language model) is a special case of this model where $T = \Sigma$.

The likelihood of this model is tractable and can be easily calculated using dynamic programming. We can optimize this likelihood directly using gradient descent. This is similar to the forward-backward algorithm used in hidden Markov models and connectionist temporal classification (CTC) [12].

$$p(s_1 \dots s_t) = \sum_{T_t} p(t_i | s_1 \dots s_{t-|t_i|}) p(s_1 \dots s_{t-|t_i|}) \quad (8)$$

This approach can be used in general for the modeling of Markov chains without cycles in the case of a finite set of transitions (even if the state space is infinite). The recurrent neural network predicts the transition probabilities over this finite set of transitions for each state using a representation of the state and a learned representation of each transition. We believe this is a novel approach to modelling acyclical Markov chains using RNNs.

In this work we consider a multiscale generalization of LSTM networks. For transition functions, f , with multiple operations we can perform the averaging at any point. We choose to average the cell states and output gates. Note that performing the averaging earlier on reduces the amount of computation.

$$\begin{pmatrix} \mathbf{f}_i \\ \mathbf{i}_i \\ \mathbf{o}_i \\ \mathbf{g}_i \end{pmatrix} = \mathbf{W}_h \mathbf{h}_{t-|t_i|} + \mathbf{W}_x \mathbf{x}_i + \mathbf{b}$$

$$\mathbf{c}_t = \frac{1}{N} \sum_{i=1}^N (\sigma(\mathbf{f}_i) \odot \mathbf{c}_{t-|t_i|} + \sigma(\mathbf{i}_i) \odot \tanh(\mathbf{g}_i))$$

$$\mathbf{h}_t = \sigma\left(\frac{1}{N} \sum_{i=1}^N \mathbf{o}_i\right) \odot \tanh(\mathbf{c}_t)$$

2.1. Model characteristics

The computational complexity of a regular RNN model grows as a function of the sequence length, $O(T)$. The multiscale model’s complexity instead grows as a function of the number of arcs. The number of arcs in a sequence is theoretically bounded by $\frac{T(T+1)}{2}$, but in practice it grows sublinearly with the size of the dictionary. For example, for a dictionary with 16384 tokens we find an average of 2.7 arcs per time step for the text8 dataset.

It should be noted that the computation of arcs can be entirely parallelized, so on a parallel computer (e.g. a GPU) the span (depth) of the computation is equivalent to that of a normal RNN, $O(T)$.

During training time the memory usage of an RNN model grows as $O(T)$ because of the need to store the hidden states for the backward propagation. The multiscale model’s memory usage grows the same and does not depend on the number of arcs, since the averages (see formula 5) can be calculated by accumulating values in-place. The need to keep token embeddings in memory means that the memory usage grows as $O(T + D)$ where D is the dictionary size.

In conclusion, the multiscale model is both computationally and memory efficient. On a parallel architecture it has the same computational complexity as a regular RNN and only requires a small amount of extra memory in order to store the token embeddings.

2.2. Dictionary learning

The formulation of our multi-scale model requires the construction of a dictionary of multi-symbol tokens. Heuristically speaking, we would simplify our modeling problem if we construct a dictionary which allows each sequence to be segmented

into a short sequence of tokens, minimizing the shortest path length through the graph (see figure 3).

In natural language processing, word-level models usually construct a dictionary by splitting strings on whitespace and punctuation. The dictionary then consists of the N most frequent tokens, with the rest of the words replaced with a special out-of-vocabulary (OOV) token. Note that many other application domains (e.g. modeling DNA sequences) don't have any straightforward heuristics to tokenize the data.

Even in language modeling this type of tokenization is problematic for a variety of reasons. The number of words in natural language is effectively infinite for synthetic languages, which means there will always be OOV tokens. Furthermore, it is arguably arbitrary from a linguistic perspective. Whereas English is a rather isolating language, with ~ 1.67 morphemes per word [13] on average, synthetic languages such as Turkish or Eskimo have ~ 2.33 and ~ 3.70 morphemes per word respectively. For example, the Dutch word *meervoudigepersoonlijkheidsstoornis* (multiple personality disorder) contains 10 morphemes. For these types of languages, we might want to consider a tokenization that contains subword units. On the other hand, for highly isolating languages we might want to model several words as a single token e.g. *chng ti*, Vietnamese for 'we'.

2.3. Dictionary coders

Instead of arbitrarily splitting on whitespace, a more principled approach is to 'learn' the tokens to be modeled. Here we propose an approach which is grounded in text compression and inspired by the byte-pair encoding (BPE) algorithm. BPE has been used in the domain of neural machine translation to learn subword units, reducing the number of OOV tokens [14].

Dictionary coder algorithms like BPE learn dictionaries of tokens with the purpose of representing strings with as few tokens as possible, increasing the level of compression. This reduces the effective depth of the unrolled RNN network (i.e. the shortest path through the graph in figure 3), which is a reasonable learning objective for our dictionary.

Algorithm 1 Adapted byte-pair encoding algorithm

Require: $T_{max}, T = \{s_1, \dots, s_m\}, \mathbf{s} = s_{i_1}, \dots, s_{i_n}$ {Initial dictionary $T = \Sigma$, string \mathbf{s} }

while true **do**

$j, k \leftarrow \arg \max_{j,k} \text{paircount}(s_j, s_k)$

$s_{new} \leftarrow [s_j | s_k], T \leftarrow T \cup \{s_{new}\}$

if $|T| = T_{max}$ **then**

break

end if

Substitute each occurrence of s_j, s_k in \mathbf{s} with s_{new}

for all $l \in \{l : \text{count}(s_l) < \text{count}(s_{new})\}$ **do**

$T \leftarrow T \setminus \{s_l\}$

Substitute each occurrence of s_l in \mathbf{s} with s_o, s_p s.t. $[s_o | s_p] = s_l$

end for

end while

return T

Regular BPE starts with a dictionary of characters and consecutively replaces the most frequent pairs of tokens with a single new token, until a given dictionary size T_{max} is reached. We extend the algorithm by reversing the merger of two tokens whenever a token becomes too rare. As a motivating example consider the string *abcabcabc...*. In this case a and b are

#	Token
1	and·
2	a·
3	the·
7	s·
11	in·the·
17	ing·
65	ation·
239	people·
296	ed·by·the·
540	at·the·end·of·the·
608	united·states·
1468	in·the·united·states·
2727	one·of·the·most·

Table 1: A sample from the tokens in the dictionary of size 8192 constructed using the text8 dataset by our adapted BPE algorithm. Spaces are visualized with the · character.

merged into ab , followed by a merger between ab and c into abc . Our extension makes sure that the token ab , which now occurs zero times, is removed from the dictionary. This removal prevents us from wasting space in the dictionary on rare tokens.

Our implementation of this algorithm uses a bit array of the size of the input data, where each element signifies whether the corresponding character is merged with the subsequent character. We maintain two d -ary heaps of tokens and token pairs sorted by their frequency. The algorithm proceeds by repeatedly popping the most common pair from the heap and searching the text and bit array for occurrences. If an occurrence is found, the bit array is updated to represent the merge, and the d -ary heaps are updated to reflect the new token and pair counts. This requires a minimum of D passes over the data to construct a dictionary of size D but uses a relatively small amount of memory.

3. Experiments

3.1. Penn Treebank

We evaluate the multiscale model on the widely used Penn Treebank dataset using the training, validation and test split proposed by Mikolov [15]. Our baseline is an LSTM with 1024 hidden units and embeddings of dimension 512 trained with truncated backpropagation-through-time (TBPTT) on sequences of length 400. These optimal values were found using a grid search. We train using the Adam [16] optimizer (learning rate of 0.001) and to increase convergence speed we use layer normalization [17].

Our baseline model achieves a score of 1.43 bits per character. The multiscale model is trained using the exact same configuration but using a dictionary of 2048 tokens. It achieves a test score of 1.42 bits per character. Note that the multiscale models improvements are orthogonal to what can be achieved by straightforwardly increasing the capacity of the network. The regular LSTM networks with more than 1024 units showed decreased performance in our experiments due to overfitting.

Moreover, our network is able to achieve better performance with far fewer parameters. The multiscale model with 512 hidden units, embeddings of size 256, and 2048 tokens has 51% fewer parameters compared to our baseline, but achieves a score of 1.41 bpc, compared to 1.48 bpc for a regular LSTM with the same embedding and hidden state size.

Samples

the ·independ-ence ·in the ·third quarter ·the ·chief ·ex-port ·stock ·prices ·for the ·year ·and ·into the ·disa-ster-l-and
gains ·so ·on the ·economy ·because ·in addition ·to a ·compl-ex ·closed ·higher ·comm-ut-e ·pres-sure ·of ·his ·company
meeting ·in ·the ·trust ·is ·expected to ·be ·an-ticip-ated ·\$ ·offic-es ·during the ·past
actu-ally ·have ·spok-es-man ·with ·hous-ing ·their ·junk ·bond ·due ·\$ N billion from ·most ·important ·next ·day ·at N

Table 2: Samples from the multiscale model trained on Penn Treebank. Token boundaries are marked with the · symbol. The samples show the model’s ability to model a sentence by predicting entire words or phrases (‘in addition’, ‘into the’) at a time, while also being able to exploit subword structure (‘comm-ut-e’) and maintaining the flexibility of character language models to output unseen words (‘disa-ster-l-and’).

3.2. Text8

Text8 [18] is a text dataset of 100 million characters built from the English Wikipedia. The characters are limited to the 26-letter alphabet and spaces. We use the traditional split of 90, 5 and 5 million for the training, validation and test set respectively.

We compare the performance of our multi-scale model with a single-layer character-level language model with 2048 units. The same training procedure as described in the previous subsection is used. The baseline achieves a score of 1.45 bits per character. The multiscale model improves on this performance using a dictionary of 16,384 tokens, achieving a test score of 1.41 bits per character.

4. Related work

In language modeling a variety of approaches have attempted to bridge the gap between character and word-level models. The approach in [19] is to apply a convolutional neural network (CNN) with temporal pooling over the constituent characters of a word. The CNN filters of this network can be interpreted as character n -gram detectors. The output of this network is used as the input to an LSTM network which models the word-level dynamics. Note that the resulting model still requires information about word-boundaries.

Other approaches use multi-scale RNN architectures. The model in [20] uses both a word-level and character-level RNN, the latter being conditioned on the former. This model too still requires knowledge of word boundaries. The approach in [21] does not require word boundaries, and instead uses the straight-through estimator to learn the latent hierarchical structure directly. Their model does not learn separate embeddings for the segments however, and can only output a single character at a time.

The latent sequence decomposition (LSD) model introduced in [22] is related to our multiscale model, and was shown to improve performance on a speech recognition task. Instead of using compression algorithms the LSD model uses a dictionary of all possible n -grams. Since the number of n -grams grows exponentially, this limits the the dictionary to very short tokens only. The LSD model uses a regular RNN which is trained on a set of sampled segmentations instead of averaging the hidden states using dynamic programming. This complicates training and makes the likelihood of the model intractable. The recent Gram-CTC model [23] is also related and does use dynamic programming but still uses a dictionary of character n -grams.

Although our model is competitive with recent methods such as MI-LSTM [24] and td-LSTM [25], which achieve 1.44 and 1.63 bits per character on the text8 dataset respectively, other recent models such as HM-LSTM [21] have achieved lower scores (1.29 bpc). Since many of the LSTM variations

in the literature can be extended to the multiscale model, we believe it is possible to improve the performance of multiscale models further in the future. Similarly, deeper multi-layer extensions to our model are feasible.

5. Discussion

Through arithmetic encoding it can be shown that modeling data is equivalent to compressing it [26]. Using neural networks to improve upon text compression algorithms is a common technique [27], but as far as we are aware the reverse has not been researched. One can see our model as a mix between non-parametric and parametric approaches: As discussed in section 1.1.2, character-level models learn a parametric mapping from constituent characters to semantic representations of morphemes. Word-level models avoid learning this highly non-linear function by constructing a dictionary and learning a representation for each word, which is non-parametric. Our multiscale model generalizes this approach, combining non-parametric dictionary coders and parametric RNN models. The size of the dictionary allows us to choose the balance between the two approaches.

A rough parallel can be drawn between our multiscale approach for sequences and superpixels in the computer vision domain [28], where pixels are clustered in order to improve computational and representational efficiency

The multiscale model can also be related to work on text segmentation. The hierarchical Pitman-Yor language model in [29] learns how to segment a string of characters into words, while simultaneously learning a word-level n -gram model. Each path through the graph of the multiscale model (figure 3) can be considered a single segmentation of the text, with the likelihood of the string being the marginalization over all possible segmentations.

A large number of combinations of data compression and neural network sequence modeling are still open to investigation. Besides BPE, there are many other dictionary coder algorithms out there. Another consideration would be to learn the dictionary and the sequence model jointly. Subsequently, a variety of neural network models can conceivably be adapted to work with the multi-scale representation of text used in this paper e.g. bag-of-words (BOW) models could be replaced with bag-of-token models instead, similar to the approach in [30] which uses character n -grams.

6. Acknowledgements

The authors would like to acknowledge the support of the following agencies for research funding and computing support: NSERC, Calcul Québec, Compute Canada, the Canada Research Chairs and CIFAR.

7. References

- [1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [2] T. Mikolov, "Statistical language models based on neural networks," Ph.D. dissertation, Brno University of Technology, 2012.
- [3] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTER-SPEECH*, vol. 2, 2010, p. 3.
- [4] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *ICML'2011*, 2011, pp. 1017–1024.
- [5] A. Graves, "Generating sequences with recurrent neural networks," arXiv:1308.0850, Tech. Rep., 2013.
- [6] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*. ACM, 2013. [Online]. Available: <http://icml.cc/2013/>
- [7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Nets*, pp. 157–166, 1994.
- [8] S. Hochreiter, "Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München," 1991. [Online]. Available: <http://www7.informatik.tu-muenchen.de/~Ehochreit>
- [9] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural Networks, Tricks of the Trade*, ser. Lecture Notes in Computer Science LNCS 1524. Springer Verlag, 1998.
- [10] Z. G. Szab, "Compositionality," in *The Stanford Encyclopedia of Philosophy*, fall 2013 ed., E. N. Zalta, Ed., 2013.
- [11] J. Mitchell and M. Lapata, "Vector-based models of semantic composition," in *ACL*, 2008, pp. 236–244.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML'2006*, Pittsburgh, USA, 2006, pp. 369–376.
- [13] J. H. Greenberg, "A quantitative approach to the morphological typology of language," *International journal of American linguistics*, vol. 26, no. 3, pp. 178–194, 1960.
- [14] R. Senrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.
- [15] T. Mikolov, I. Sutskever, A. Deoras, H.-S. Le, S. Kombrink, and J. Cernocký, "Subword language modeling with neural networks," preprint (<http://www.fit.vutbr.cz/~imikolov/rnnlm/char.pdf>), 2012.
- [16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [18] M. Mahoney, "Large text compression benchmark," URL: <http://www.matmahoney.net/text/text.html>, 2009.
- [19] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models," *arXiv preprint arXiv:1508.06615*, 2015.
- [20] P. Bojanowski, A. Joulin, and T. Mikolov, "Alternative structures for character-level rnns," *arXiv preprint arXiv:1511.06303*, 2015.
- [21] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *arXiv preprint arXiv:1609.01704*, 2016.
- [22] W. Chan, Y. Zhang, Q. Le, and N. Jaitly, "Latent sequence decompositions," *arXiv preprint arXiv:1610.03035*, 2016.
- [23] H. Liu, Z. Zhu, X. Li, and S. Satheesh, "Gram-ctc: Automatic unit selection and target decomposition for sequence labelling," *arXiv preprint arXiv:1703.00096*, 2017.
- [24] Y. Wu, S. Zhang, Y. Zhang, Y. Bengio, and R. R. Salakhutdinov, "On multiplicative integration with recurrent neural networks," in *Advances In Neural Information Processing Systems*, 2016, pp. 2856–2864.
- [25] X. Zhang, L. Lu, and M. Lapata, "Top-down tree long short-term memory networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 310–320. [Online]. Available: <http://www.aclweb.org/anthology/N16-1035>
- [26] M. V. Mahoney, "Text compression as a test for artificial intelligence," in *AAAI/IAAI*, 1999, p. 970.
- [27] —, "Fast text compression with neural networks," in *FLAIRS Conference*, 2000, pp. 230–234.
- [28] X. Ren and J. Malik, "Learning a classification model for segmentation," in *ICCV*, vol. 1, 2003, pp. 10–17.
- [29] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested pitman-yor language modeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 100–108.
- [30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.