Revisiting Syllable-aware Language Modelling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language modelling (LM) is regularly analysed at word, sub-word or character level inputs, and this study reconsiders syllable units for the task. Rule-based syllabification typically requires less specialised knowledge than identifying morphemes, and the process can naturally work for low-resource cases, as we do not need an unsupervised model to extract sub-words. In this paper, we compare different granularities from characters to words in an open-vocabulary LM task, where syllables mostly outperform the rest of them for both English and Shipibo-Konibo languages. Thereafter, we obtain similarly positive results for syllable-level neural machine translation (NMT) with Spanish too. [All authors identify as Latinx]

1 Introduction

Previous work on syllable-aware LM in English failed to beat character-level models [2]; however, we propose to assess the task under two new settings. First, we could employ a plain-vanilla architecture, without additional composition functions, to analyse an open-vocabulary scenario with syllables [3]. Second, English has a weak correspondence between graphemes (written symbols) and phonemes (speech units), so we might include an study case with less-ambiguous splits. Therefore, we revisit syllable-aware LM by using simple recurrent neural networks [8] for open-vocabulary generation [15], and by also assessing a more phonetic language with a recent alphabetisation (Shipibo-Konibo [1]). We thereupon explore the syllables effect in another generation task such as NMT.

2 Methodology and Results

We evaluate syllables against words, Byte Pair Encoding [BPE, 14] sub-words, and characters, with a comparable perplexity [10] in LM; and character [18] and word level [13] metrics in NMT.

2.1 Languages and Datasets

For LM in English (*eng*), we use well-known datasets: Penn Treebank [PTB, 7] and WikiText-2 [9]. In the case of Shipibo-Konibo (*shp*), a low-resource and native language from Peru, we process the monolingual side of three parallel corpora aligned with Spanish (*spa*) [4]. For one of them, named Flashcards, we align *eng* sentences from the original *eng–spa* corpus used for its creation [16]. For comparison purposes, we also analyse the new *eng* monolingual text of Flashcards in LM. Afterwards, we study the NMT case only with the Flashcards dataset in both *shp–eng* and *shp–spa* language-pairs. We segment syllables in *shp* with rules [1] and with a dictionary-based method [5] for *eng* and *spa*.

Table 1-a-b describes the data for LM. We observe a vast amount of syllable types in the *eng* datasets, in contrast to *shp*, where syllables are closer to characters than to other granularities. Moreover, the Flashcards segmentation reveals the perplexing nature of *eng* syllables. For the LM task on *shp*, the significantly low amount of unique syllables could be interpreted as modelling a language with a larger alphabet (more characters types) and a smaller average length of token.

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

	Dataset	(a) Split size Train	Valid	Test	(b) # types Word	**Syl**	Char	BPE	(c) $ppl^c \downarrow$ Word	**Syl**	Char	BPE (*)
eng	PTB	887.0k	70.3k	78.6k	10.0k	6.1k	48	4.7k	2.36	**2.11**	2.52	2.42 (5k)
	WikiText-2	103.2M	217.6k	245.5k	33.2k	19.5k	274	1.3k	2.62	**2.15**	2.72	2.63 (1k)
	Flashcards	14.7k	1.4k	1.7k	2.5k	2.4k	63	2.3k	**2.12**	2.24	3.01	2.62 (3k)
shp	Flashcards	12.1k	2.1k	1.4k	2.6k	193	30	1.8k	2.70	**2.39**	2.64	3.30 (3k)
	Religious	82.4k	9.4k	10.2k	11.1k	331	26	1.0k	3.01	**2.37**	2.48	2.92 (1k)
	Educational	32.0k	3.6k	4.1k	4.0k	258	32	2.8k	2.65	**2.16**	2.29	2.77 (3k)

Table 1: (a) Split size in tokens; (b) Number of types per segmentation in Train; (c) ppl^c on Test for LM. For BPE, we show the best score given various merges between 1k–5k with a 1k-step.

	BLEU ↑ Word	**Syl**	Char	BPE 5k–10k–15k			characTER ↓ Word	**Syl**	Char	BPE 5k–10k–15k		
shp–eng	16.26	18.38	**19.60**	16.90	15.65	16.21	63.86	**53.57**	54.25	56.20	58.71	57.51
eng–shp	16.35	**19.70**	17.32	16.61	16.80	17.17	57.07	**51.76**	53.40	55.61	55.80	56.91
shp–spa	8.91	**13.20**	10.62	8.68	8.76	9.14	68.37	**55.33**	58.98	62.20	63.00	64.61
spa–shp	9.76	**14.78**	13.39	11.62	11.62	12.12	65.79	**55.05**	55.24	62.48	63.42	62.88

Table 2: NMT results at word (BLEU) and character level (CharacTER) on the Flashcards dataset.

2.2 Language Modelling with a Comparable Perplexity

For a fair comparison across all granularities, we evaluate all results with character-level perplexity: $ppl^c = \exp\left(L \cdot (s^{seg}+1)/(s^c+1)\right)$, where L is the cross-entropy loss of a string s computed by a neural LM, and s^{seg} and s^c refer to the length of s in the chosen segmentation and character level units, respectively [10]. Furthermore, we generate the same input unit as an open-vocabulary task, where there is no prediction of an "unknown" token [15], with an exception at word-level in PTB. We thereby differ from previous work [2], and refrain from composing the syllable representations into words to evaluate only word-level perplexity. Following other open-vocabulary LM studies [12, 11], we use a low-compute version of an LSTM neural network, named Average SGD Weight-Dropped [8], with a smaller embedding size (300 units) for faster training. Additionally, we use the SentencePiece [6] segmentation format in both characters and syllables, and the original one for BPE [14].

Table 1-c shows that syllables mostly result in better perplexities that the remaining granularities in LM, even for a low-phonetic language as *eng*, and with a very competitive score when they do not achieve the best one. Moreover, syllables outperform the rest in the open-vocabulary scope (excluding words). Beating characters implies a gain in time processing as well, given the shorter sequences of syllables. Other settings that could be further explored are working unsupervised morphemes or morphological-aided supervision [17], and constraining the BPE-vocabulary size to the number of syllable types [6].

2.3 Syllables for Neural Machine Translation (NMT)

To further explore the value of syllables, we build *eng–shp* and *spa–shp* NMT models with all granularities as inputs-outputs. Each model uses a two-layer LSTM encoder-decoder with a hidden layer of 512, an embedding size of 300, and joint BPE with various merges. Table 2 presents the BLEU [13] and CharacTER [18] scores, where syllables predominantly stand out again, with an exception against characters in *shp–eng* at the word level metric. This result reinforces the initial concern for the phonetic ambiguity of *eng*, as we infer an inherent difficulty to reconstruct a word with generated syllable sequences. We also hypothesise that the joint BPE models do not provide the best scores given the potentially small word and sub-word overlapping of *shp* with both *eng* and *spa*.

3 Conclusion

Our results suggest that syllables might be valuable for both open-vocabulary LM and NMT tasks, where they behave positively even for a poor phonemically-spelt language. Syllables do not have an embedded meaning; however, the required effort for their segmentation could be advantageous concerning other morphological-aware or unsupervised-driven methods. Finally, we are currently working on exploring a multilingual scope and the hybrid-LM scenario [12] with syllables.

References

[1] Carlo Alva and Arturo Oncevay. Spell-checking based on syllabification and character-level graphs for a Peruvian agglutinative language. In *Proc. of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116. Association for Comp. Linguistics, 2017.

[2] Zhenisbek Assylbekov, Rustem Takhanov, Bagdat Myrzakhmetov, and Jonathan N. Washington. Syllable-aware neural language models: A failure to beat character-aware ones. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1866–1872. Association for Computational Linguistics, 2017.

[3] Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the NAACL-HLT, Vol. 2 (Short Papers)*, pages 536–541. Association for Computational Linguistics, 2018.

[4] Héctor E. Gómez Montoya, Kervy D. Rivas Rojas, and Arturo Oncevay. A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23. EAMT, 2019.

[5] Kozea. Pyphen. Available in: `https://pyphen.org/`, 2013.

[6] Taku Kudo and John Richardson. 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. Association for Computational Linguistics, 2018.

[7] Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[8] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*, 2018.

[9] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[10] Sebastian J. Mielke. Can you compare perplexity across different segmentations? Available in: `http://sjmielke.com/comparing-perplexities.htm`, 2019.

[11] Sebastian J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. What kind of language is hard to language-model? In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989. Association for Comp. Linguistics, 2019.

[12] Sebastian J. Mielke and Jason Eisner. Spell once, summon anywhere: A two-level open-vocabulary language model. In *The Thirty-Third AAAI Conf. on Artificial Intelligence*, 2019.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics, 2016.

[15] Ilya Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1017–1024, 2011.

[16] Tatoeba. Tab-delimited bilingual sentence pairs. Available in: `http://www.manythings.org/anki/`, 2010. These are selected sentence pairs from the Tatoeba Project.

[17] Clara Vania and Adam Lopez. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 2016–2027. Association for Computational Linguistics, 2017.

[18] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Vol. 2, Shared Task Papers*, pages 505–510. Association for Computational Linguistics, 2016.