Nameling: Creative Neologism Generation with Transfer Learning

Gabriel R Lencione
University of Campinas
São Paulo - Brazil
gabriellencione@gmail.com

Rodrigo F Nogueira
University of Campinas
Campinas - Brazil
rfn@unicamp.br

Paula Y Pasqualini
University of Sao Paulo
São Paulo - Brazil
pypasqualini@gmail.com

Abstract

This work proposes the adoption of deep pre-trained models to generate neologisms, approaching the word generation problem as a supervised summarization task in which we provide the definition of the topic as input and expect a new summarized word as output. We explore subword (T5) and character-level (ByT5) models for this task, fine-tuning them with two different datasets and assessing the quality of the outcomes. We demonstrate the success of the proposals on learning the basic rules of word formation and generating neologisms. A demo of our method is available at https://nameling.org

Introduction

Creating a new word capable of summarizing an idea or concept can be a challenging task. Needed in a variety of domains and used for different purposes, its importance is present in business branding, product naming, art, popular culture etc.

A neologism can be produced, for instance, through the composition of stems and affixes or through the blend of existing words, adapting phonemes and morphemes. A fan of Harry Potter novels (Rowling 1997) may be called a 'pottermaniac', a neologism formed by the concatenation of 'Potter' and 'maniac', which may seem simple and intuitive, but requires a profound knowledge about the meaning of the terms and the rules of word formation. These complexities are treated by humans as a mixture of art and science (Özbal and Strapparava 2012), making it more difficult for computer-based methods to achieve the supposed outcome.

To the best of our knowledge, the literature of neologism generation is limited to a few works, especially when approached by Deep Learning (DL) techniques. Ozbal and Strapparav (2012) developed a handcraft-based method, gathering words related to the input, a fixed scheme consisting of a category and the desired properties, blending them to form the respective neologism. In the work of Deri and Knight (2015), a data-driven Finite State Machine (FSM) cascade is proposed to combine input words by their sounds and meaning into a new original word, a portmanteau. They demonstrated the success of their approach, achieving 45% of exact match on test set. Malkin et al. (2021) proposed an approach similar to ours with an inverse application. They employed language models to generate definitions for neologisms.

In this work, we propose a novel and extended manner to treat the neologism generation problem, approaching it as a summarization task in which we present a concept definition (one or more sentences defining the subject) and produce a new word as its representation. It may also be seen as an extreme form of sentence compression (Cohn and Lapata 2008). We adopted deep pre-trained encoder-decoder models for this task, whose main purpose is to process input sequences and return output sequences (seq2seq), profiting from their previous knowledge to fine-tune them on our customized dataset.

The paper is organized as follows: Proposed Method Section covers a detailed description of the proposed method and the experimental setup. The results and their quantitative and qualitative analysis are presented in the Results and Discussion Section. Conclusion Section is devoted to the concluding remarks and further steps of the research.

Proposed Method

Since the proposed method is essentially based on fine-tuning deep seq2seq models, we present, in the following subsections, the elements employed in our word generation task: the baseline DL models and the dataset.

Baseline Models

One family of encoder-decoder models that has gained attention for its high performance on a variety of Natural Language Processing (NLP) tasks, including sentiment analysis, translation and summarization, is the T5 family (Raffel et al. 2020; Xue et al. 2022). The first T5 paper (Raffel et al. 2020) introduces a transformer-based seq2seq model capable of performing several tasks by converting each of them to a text-to-text format. The authors demonstrated the benefits of their multi-task learning approach, by achieving the state-of-the-art on many benchmarks. As a fair compromise between performance and feasibility (according to our limited resources), we selected the T5-base model, which has about 220 million parameters and rarely achieved 10% worse than the best results.

Another important member of the T5 family is the ByT5 (Xue et al. 2022). Instead of tokenizing whole sentences into words and subwords, like the SentencePiece (Kudo and

Richardson 2018) in the T5 model, the authors proposed a byte-level (UTF-8) encoder-decoder based on T5 architecture, which was shown to be more robust to noise and easier to preprocess. They also proposed different-size models, leading us to adopt, by the same reasons, the ByT5-small, which has 300 million parameters, being the most comparable to the T5-base.

Dataset

In order to have our models generating new and original words, they should first learn the basic patterns of word formation in an inductive manner. Therefore, we should present a significant amount of examples linking existing words to their meaning, which is basically the definition of a dictionary. Then, we adopted, as a first step of our research, the English version of the free online Kaikki dictionary (Kaikki 2022).

This dictionary is composed of more than 1 million distinct words. It contains words from all major morphological categories, from inflected forms, compound words and a great diversity of semantic fields, including slangs, places, famous personalities etc. It was preprocessed keeping only single or hyphenated words and removing words with less than 5 characters, with digits and with punctuation other than hyphen. This was done in order to avoid potential noise present in our dataset and retain the most meaningful terms. Hereinafter we shall refer to this dataset as Dic. We proceeded with a random train/validation/test (70%/15%/15%) split for both datasets, coming up with 663k/132k/151k examples.

Evaluation

The assessment of neologism generation quality is not a trivial task. There are not any available metrics, to the best of our knowledge, that could contemplate all the possible implicit and explicit manners to generate a new original word. Here, we proceeded with a qualitative analysis, evaluating with limited human judgment the quality of a random sample of new words generated by our proposals.

It is also fundamental that our models learn the basic rules of word formation, as they should use this knowledge to build neologisms from concept definitions. It can be addressed by a quantitative analysis, in which we assess the distance from the predicted words to their respective targets employing the Character Error Rate (CER) and the Word Error Rate (WER) (Morris, Maier, and Green 2004).

Experimental Setup

For both T5-base and ByT5-small, we let the maximum input and output lengths be 300 and 32 respectively, as it presented a good compromise to the choice of batch size on our varying training environment and the mean length of our samples (41.6 characters). We adopted the *Trainer* framework from Hugging Face to perform the training stage, which took place on different GPUs, mostly Nvidia Tesla T4 and P40, according to their availability in our environment. The batch size was set to 40 and the maximum number of epochs was set to 15.

Metrics (Test Set)	T5-base	ByT5-small
CER (Dic)	**0.3028**	0.3032
WER (Dic)	0.5024	**0.4770**

Table 1: Performance on test set for T5-base and ByT5-small.

Results and Discussion

Quantitative Analysis

In Table 1, we present the previously discussed performance metrics on test set. We shall note that both models presented very similar CER, with ByT5-small achieving a 5% lower WER. We should remark that, although low errors are important because of the previously discussed reasons, extremely low ones may imply a massive reproduction of existing words instead of creative neologism generation.

Qualitative Analysis

Table 2 presents some examples sampled from Dic test set and some elaborated by us. The output words were produced using beam-search with 15 beams, from which we selected the top-3 candidates. The main goal is to verify if our proposals are able not to match the actual words, but to generate new creative and meaningful words that could serve as neologisms to the proposed concept definitions.

The first example evidences the versatility of the four proposals in applying common suffixation to derive adjectives from the noun 'plant'. It is a perfect illustration of our previous discussion about how low similarity between outcome words (such as 'planty', 'plantish' etc) and target words ('phytoid') does not imply poor quality neologisms.

The second one represents very well the type of neologism generation we aimed. The input 'A fan of entertainer Nora Aunor' was summarized by new words such as 'Noraholic' and 'Noramaniac', which contain humor.

The fourth example clearly demonstrates the ability of the proposals to derive the present participle tense of 'befal'. It is a Dic example with a deviation from the verb 'befall'. Both T5 and ByT5 models generated words with and without the addition of 'l' before adding the suffix '-ing'.

The last three examples were proposed by us and provide interesting results on creative generation. The input 'mixture of romeo and juliet' tries to assess the ability of our models to directly blend two words and resulted in successful neologisms such as 'romeoliet', 'romiet' and 'rojuliet'. The input 'to love a book' returned high quality neologisms such as 'booklove', 'bibliolove' and 'bibliophile'. For 'a modern internet shopping company', our proposals generated words such as 'cybershop', 'cyberstore', 'neoshop' and 'neoshopper' which are creative branding suggestions for the respective business description.

Conclusion and Future Steps

We can conclude, therefore, that our proposals were capable of significantly learning the rules of word formation and applying them to generate neologisms from concept definitions. This endorses the success of our proposed methodology, which treats neologism generation as a summarization

Input (concept definition)	Actual Word	T5-base (Dic)	ByT5-small (Dic)
Resembling a plant, plantlike	phytoid	'planty', 'plantish', 'plantlike',	'plantish', 'plantly', 'plantiform'
A fan of entertainer Nora Aunor	Noranian	'Norahead', 'Noraholic', 'Noramaniac'	'Noraholic', 'Noraphile', 'Norahead'
Of or pertaining to Africa entirely	transafrican	'african', 'panafrican', 'Panafrican'	'Afroafrican', 'African', 'Afrocentric'
present participle of befal	befaling	'befalling', 'befaling', 'befallling'	'befalling', 'befaling', 'befallling'
is a derivative of the amino acids arginine and alanine. it was the first member of the class of chemical compounds known as opines to be discovered. gets its name from octopus octopodia from which it was first isolated in 1927... is formed by reductive condensation of pyruvic acid...	octopine	'octopidine', 'octadine', 'octanoid'	'opine', 'octopine', 'pyruvine'
mixture of romeo and juliet		'romeojuliet', 'romeoliet', 'romeojuliette'	'romejuliet', 'rojuliette', 'romiet'
to love a book		'booklove', 'bibliophile', 'bibliolove'	'bibliophilize', 'bibliophile', 'booklove'
a modern internet shopping company		'cybershop', 'cybershopping', 'cyberstore'	'cybershop', 'neoshop', 'neoshopper'

Table 2: Examples of words generated by our proposals. The first four examples were sampled from Dic test set, while the last three are proposed by us.

task. Both of the baseline models presented similar CER and WER and the qualitative analysis demonstrated that they performed closely well.

The limitations of our qualitative analysis shall guide us through the next steps of our research, employing a wider and more systematic human-based evaluation, specially with concept definitions that do not correspond to any existing English word. We shall also explore the expansion of this methodology to larger seq2seq models, the study of customized datasets to specific neologism generation domains and the comparison of our baseline proposals with zero-shot and fine-tuned large language models.

References

Cohn, T., and Lapata, M. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 137–144. Manchester, UK: Coling 2008 Organizing Committee.

Deri, A., and Knight, K. 2015. How to make a frenemy: Multitape FSTs for portmanteau generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 206–210. Denver, Colorado: Association for Computational Linguistics.

Kaikki. 2022. Kaikki english dictionary. [Online; accessed 30-July-2021].

Kudo, T., and Richardson, J. 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*, 66–71. Brussels, Belgium: Association for Computational Linguistics.

Malkin, N.; Lanka, S.; Goel, P.; Rao, S.; and Jojic, N. 2021. GPT perdetry test: Generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5542–5553. Online: Association for Computational Linguistics.

Morris, A. C.; Maier, V.; and Green, P. D. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

Özbal, G., and Strapparava, C. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 703–711. Jeju Island, Korea: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140):1–67.

Rowling, J. K. 1997. *Harry Potter and the Philosopher's Stone*, volume 1. London: Bloomsbury Publishing, 1 edition.

Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; and Raffel, C. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics* 10:291–306.