

Can a Transformer Assist in Scientific Writing?

Generating Semantic Web Paper Snippets with GPT-2

<http://swgpt2.amp.lod.labs.vu.nl/>

Albert Meroño-Peñuela, Dayana Spagnuolo, and GPT-2
{albert.merono, d.spagnuolo}@vu.nl

_MOTIVATION

>>> Current scientific crisis: unmanageable
pace at which new papers are being published

>>> 8-9% each year; only in bio-medicine
2 papers per minute are published in PubMed

>>> Humans have inherent limitations, such
as not being systematic, introducing errors,
having biases, and writing poor reports

>>> How can we use AI to address these
challenges?

_CLEANING ISWC PROCEEDINGS

>>> We clean the entire corpus of ISWC
proceedings (2002-2019)

>>> Only in 2019: 1,377 pages and 569,371
words

>>> The complete 2002-2019 series contains
over 51M tokens

>>> Preparation pipeline¹:

- > (a) Batch PDF processing with pdftotext²
- > (b) For each text file removal of: cover pages
and meta information about the book; running
headers with authors and paper titles; the list
of organisation committee and sponsors, and the
table of contents; copyright footnotes; list of
references; and author index.
- > (c) Cleaning of some instances of: tables;
extra spaces and indentation; and extra lines

_TRAINING GPT-2 WITH ISWC PROCEEDINGS

>>> Retrain GPT-2 using the 117M model,
32 core Intel(R) Xeon(R) CPU E5-2630 v3
@ 2.40GHz with 252GB of RAM

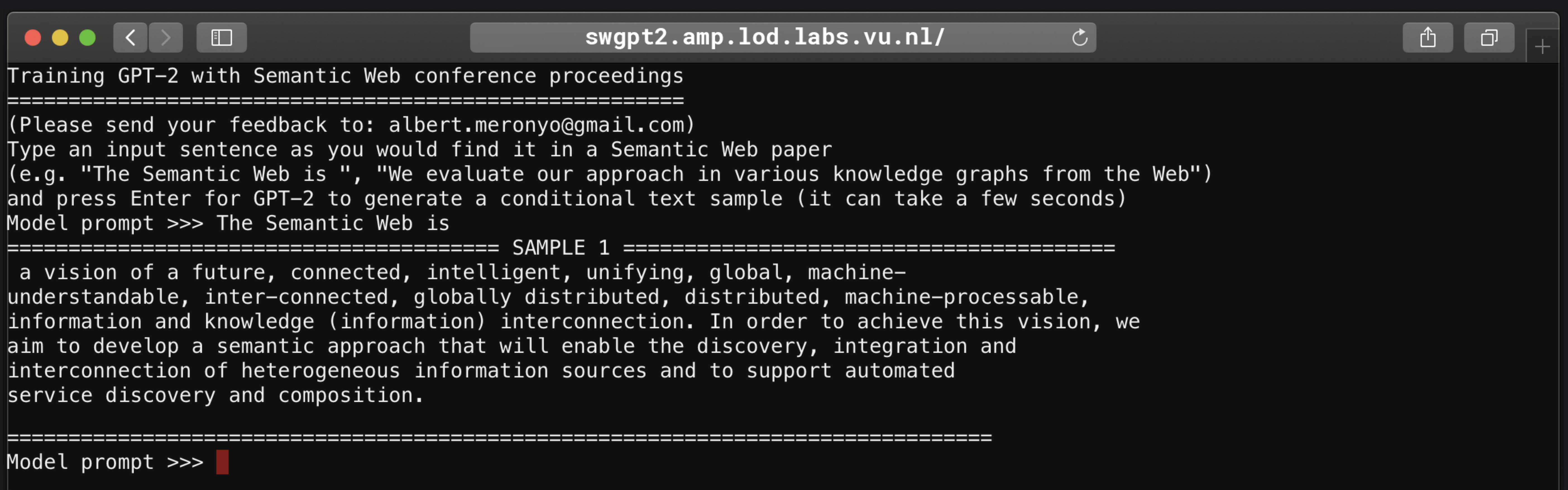
>>> We use OpenAI's framework³

>>> Dataset encoding (51M tokens)

>>> Training over 68,280 iterations and
332,533.83s (92.37h, average of 4.87s/it)

>>> Average loss is 2.19 (last loss 2.10)

_TEMPLATE-BASED CONDITIONAL TEXT SAMPLING



```
Training GPT-2 with Semantic Web conference proceedings
=====
(Please send your feedback to: albert.meronyo@gmail.com)
Type an input sentence as you would find it in a Semantic Web paper
(e.g. "The Semantic Web is ", "We evaluate our approach in various knowledge graphs from the Web")
and press Enter for GPT-2 to generate a conditional text sample (it can take a few seconds)
Model prompt >>> The Semantic Web is
===== SAMPLE 1 =====
a vision of a future, connected, intelligent, unifying, global, machine-
understandable, inter-connected, globally distributed, distributed, machine-processable,
information and knowledge (information) interconnection. In order to achieve this vision, we
aim to develop a semantic approach that will enable the discovery, integration and
interconnection of heterogeneous information sources and to support automated
service discovery and composition.
=====
Model prompt >>> █
```

_REFERENCES

¹ https://github.com/dayspagnuolo/lncs_template_cleaner
² <https://linux.die.net/man/1/pdftotext>
³ <https://github.com/openai/gpt-2>

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078 (2014)

Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), <http://arxiv.org/abs/1810.04805>

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P.E., Gil, Y.: Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. PLoS one 8(11) (2013)

Ghidini, C., Hartig, O., Maleshkova, M., Svatek, V., Cruz, I., Aidan, H., Song, J., Lefrançois, M., Gandon, F.: The Semantic Web-ISWC 2019. Springer (2019)

Gil, Y.: Thoughtful artificial intelligence: Forging a new partnership for data science and scientific discovery. Data Science 1(1-2), 119-129 (2017)

Landhuis, E.: Scientific literature: information overload. Nature 535(7612), 457-458 (2016)

Pearce, W., Niederer, S., Özkula, S.M., Sánchez Querubín, N.: The social media life of climate change: Platforms, publics, and future imaginaries. Wiley Interdisciplinary Reviews: Climate Change 10(2), e569 (2019)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. OpenAI Blog 1(8), 9 (2019)