Dimitrios Roussis
Nikos Episkopos

Learning Neural Templates for Text Generation - S. Wiseman, S. M. Shieber, A. M. Rush (2018)

Data Science & Information Technologies
Database Systems
Spring 2020

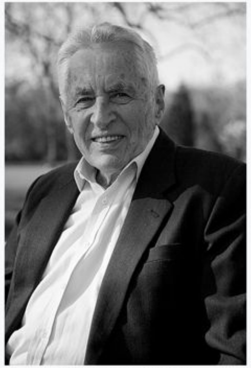
Instructor:
Georgia Koutrika

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

Table-to-text generation

Core task of Natural Language Generation (NLG) which involves the generation of textual descriptions $y_{1:T} = y_1, \ldots, y_T$ from knowledge bases which contain a collection of records $x = \{r_1, \ldots, r_j\}$.

George Mikell



Mikell in March 2017

Born	Jurgis Mikelaitis 4 April 1929 (age 88) Bildeniai, Lithuania
Nationality	Lithuanian, Australian
Occupation	Actor, writer
Years active	1957–present
Known for	*The Guns of Navarone* *The Great Escape*
Height	6' 0" (1.83m)

WIKIPEDIA
The Free Encyclopedia

George Mikell (born **Jurgis Mikelaitis**; 4 April 1929) is a Lithuanian-Australian actor and writer best known for his performances in *The Guns of Navarone* (1961) and *The Great Escape* (1963).

Contributions of the paper

- **Encoder-decoder** architectures are the SOTA in table-to-text generation. However, they lack: **(1) interpretability** and **(2) controllability**.

- Proposition: interpretable and controllable neural generation systems.

- The paper focuses on learning template-like structures for conditional text generation.

- Utilizes a **neural hidden semi-markov model** (HSMM) as a decoder.

Hidden Markov and semi-Markov models

A **hidden Markov model** (HMM) assumes that the hidden states of an unobservable process X can be learned by observing a dependent process Y. The emission probability is used in a single time step *t*.

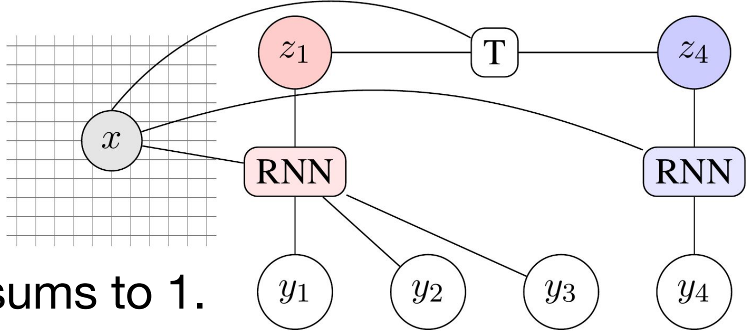
A **hidden semi-Markov model** (HSMM) is a HMM in which emission probabilities may last multiple time-steps.

The HSMM utilized in this research:

➢ Takes into account the length of a sequence and whether it should continue generating text or not.

➢ Makes use of a Recurrent Neural Network (RNN) for continuous emissions.

Neural HSMM decoder



Transition distribution $P(z_{t+1} \mid z_t , x)$

- K × K matrix of probabilities, where each row sums to 1.
- Apply row-wise softmax to get the desired probabilities.

Training

1. Maximize the log marginal-likelihood of the observed Y given X
2. Use the backward-algorithm of HMM to parameterize the emission distributions of RNNs
3. Associate text-segments with the discrete labels that frequently generate them, for guided generation
4. Collect the most common sequences of hidden states to use them for targeted generation

WikiBio dataset, benchmarks and metrics

WikiBio dataset [Lebret et al., 2016]:

➢ 728,321 biographies from the English Wikipedia, with ~ 400k word types

Models compared in benchmarks:

- Autoregressive and non-autoregressive HSMM
- **Kneser-Ney** (KN) language model [Heafield et al., 2013]
- 2 variants of a **feed-forward neural language** model [Lebret et al., 2016]
- **Seq2seq** [Liu et al., 2018]; SOTA at the time (encoder-decoder architecture)

3 comparison **metrics**:

- BLEU [Papineni et al., 2002]
- NIST [Belz & Reiter, 2006]
- ROUGE-4 [Lin & Och, 2004]

E2E dataset, benchmarks and metrics

E2E dataset [Novikova et al., 2017]:

➢ ~ 50K records with 945 distinct word types; knowledge base of restaurants

Models compared in benchmarks:

- Autoregressive and non-autoregressive HSMM
- SUB, a non-parametric template-like baseline
- The system from [Dusek and Jurcicek, 2016] (an **encoder-decoder followed by a reranker**)

5 comparison **metrics**:

- BLEU, NIST, ROUGE (like before)
- CIDEr [Vedantam et al., 2015]
- METEOR [Banerjee & Lavie, 2005]

Flat MR	NL reference
name[Loch Fyne], eatType[restaurant], food[French], priceRange[less than £20], familyFriendly[yes]	Loch Fyne is a family-friendly restaurant providing wine and cheese at a low cost.
	Loch Fyne is a French family friendly restaurant catering to a budget of below £20.
	Loch Fyne is a French restaurant with a family setting and perfect on the wallet.

Results on WikiBio dataset

Test set	BLEU	NIST	ROUGE-4
Template KN*	19.8	5.19	10.7
NNLM (field)*	33.4	7.52	23.9
NNLM (field & word)*	34.7	7.98	25.8
NTemp**	34.2	7.94	35.9
NTemp+AR**	34.8	7.59	38.6
Seq2seq***	43.65	-	40.32

* from [Lebret et al., 2016]

** from [Wiseman et al., 2018], our paper

*** from [Liu et al., 2018]

Results on E2E dataset

Test set	**BLEU**	**NIST**	**ROUGE**	**CIDEr**	**METEOR**
D&J*	**65.93**	**8.59**	**68.50**	**2.23**	**44.83**
SUB	43.78	6.88	54.64	1.39	37.35
NTemp	55.17	7.14	65.70	1.70	41.91
NTemp+AR	59.80	7.56	65.01	1.95	38.75

* from [Dusek and Jurcicek, 2016]

Qualitative evaluation

- **Controllable diversity** (manipulate template while leaving database constant)
  Choice of template: affects word-ordering and the mentioned fields.

- **Interpretable states** (capture the variation in information within the dataset)
  A discrete state corresponds to a particular piece of information.
  For example, nationalities, occupations, etc. share the same hidden state.

Average purity: Percentage of state's words that come from the most frequent record type that the state represents.

Average Purity	NTemp	NTemp+AR
E2E	89.2	85.4
WikiBio	43.2	39.9

Research and Implementation Plan

- Set up and train the HSMM model (2 variants)
  - WikiBio dataset
    - Preprocessing needed in order to test
    - Set up or implement & train the benchmarks:
      - Kneser-Ney
      - NLM (2 variants) by Lebret et al.
    - Compare the:
      - BLEU
      - NIST
      - ROUGE-4
  - E2E dataset
    - Minimal Preprocessing
    - Set up or implement & train the benchmarks:
      - Dursek & Jurcicek
      - Seq2seq
    - Compare the:
      - BLEU
      - NIST
      - ROUGE
      - CIDEr
      - METEOR

(a) Implement models (b) reproduce results
(c) perform qualitative evaluation

Implementation details and resources

<u>Tools</u>: Python, Bash, Anaconda

<u>Source code of implemented models from</u>:

HSMM: https://github.com/harvardnlp/neural-template-gen

Dursek & Jurcicek: https://github.com/UFAL-DSG/tgen

Kneser-Ney: https://kheafield.com/code/kenlm/

Seq2seq: https://github.com/tyliupku/wiki2bio

<u>Datasets from:</u>

WikiBio: https://github.com/DavidGrangier/wikipedia-biography-dataset

E2E: http://www.macs.hw.ac.uk/InteractionLab/E2E/

Possible research paths and alternatives

HSMM parameters:

- The authors determine the number of K (hidden states) based on BLEU performance on held-out validation data.
- How can we intuitively determine the most appropriate metric for the dataset/task? Can we choose the parameters based on it?

Implementation:

- Model implemented using Python 2.7 (deprecated) and PyTorch 0.3.1 (released in Feb 2018).
- Will porting the model to Python 3.8 and TensorFlow 2.2.0 yield any performance improvements and/or other advantages?

Datasets:

- Can we use the new and cleaned E2E dataset?

References

- Wiseman, S., Shieber, S.M. and Rush, A.M., 2018. Learning neural templates for text generation. *arXiv preprint arXiv:1808.10122*.
- Lebret, R., Grangier, D. and Auli, M., 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.
- Novikova, J., Dušek, O. and Rieser, V., 2017. The E2E dataset: New challenges for end-to-end generation. *arXiv preprint arXiv:1706.09254*.
- Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.
- Lin, C.Y. and Och, F.J., 2004, June. Looking for a few good metrics: ROUGE and its evaluation. In *Ntcir Workshop*.
- Belz, A. and Reiter, E., 2006, April. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Vedantam, R., Lawrence Zitnick, C. and Parikh, D., 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575).
- Banerjee, S. and Lavie, A., 2005, June. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization* (pp. 65-72).
- Dušek, O. and Jurčíček, F., 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. *arXiv preprint arXiv:1606.05491*.
- Liu, T., Wang, K., Sha, L., Chang, B. and Sui, Z., 2018, April. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Heafield, K., Pouzyrevsky, I., Clark, J.H. and Koehn, P., 2013, August. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 690-696).