PhD proposal

Deep Learning for Data to Text and Text to Data Generation:

Contact: patrick.gallinari@sorbonne-universite.fr, Laure.soulier@sorbonne-universite.fr, vincent.guigue@sorbonne-universite.fr

When: starting in November 2021

Location: Sorbonne Université, Pierre et Marie Curie Campus, 4 Place Jussieu, Paris, Fr

How to apply: please send a cv, a motivation letter, grades obtained in master, recommendation letters when possible to the contacts above.

Skills: Master in computer science or applied mathematics, Engineering school. Strong background and experience in machine learning and/or natural language processing, and good technical skills in programming.

Context

Text generation has witnessed impressive progress with recent developments of Deep Learning (DL) methods. DL language models initially based on Recurrent Neural Networks and more recently on transformer architectures can be trained on huge quantities of data and then be used for generating high quality text, conditioned on some initial input. These systems currently achieve impressing results in different text generating tasks including free text generation (Radford et al. 2018, Raffel et al. 2020), abstractive summarization (Scialom et al 2020), or semi structured text generation (Bien et al. 2020).

Data to text and text to data

Knowledge sources are often encoded into structured format such as indexes, tables, triplets, ontologies, knowledge bases, or even raw numerical data. These data are easily readable by machines, but hardly interpretable by humans. On the opposite, textual information, easily accessible to humans is often complex to exploit by machines. A key challenge and an emerging field in machine learning and natural language processing, is the transcription of structured data to text and the inverse problem of transforming raw text into structured data.

The former problem is called data-to text generation and it occurs in several applications like journalism, medical diagnosis, financial reports, sport broadcasts, dialogue oriented tasks - generating responses, summarization. It may also be a component of explainable AI systems where by allowing humans to interact with machines. Data can come in different formats such as tables, graphs, etc. The RotoWire task for example consists in generating NBA game summaries from tables providing the game statistics (Figure 1).

TEAM	H/V	WINS	LOSSES	PTS	REB	AST	...
Hawks	H	46	12	95	42	27	...
Magic	V	19	41	88	40	22	...

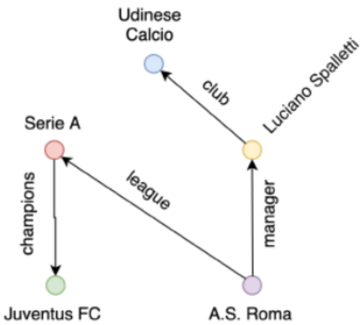
PLAYER	PTS	REB	AST	STL	BLK	CITY	...
Al Horford	17	13	4	2	0	Atlanta	...
Kyle Korver	8	3	2	1	2	Atlanta	...
Jeff Teague	17	0	7	2	0	Atlanta	...
N. Vucevic	21	15	3	1	1	Orlando	...
Tobias Harris	15	4	1	2	1	Orlando	...
...	...	...	...	...	...	...	...

H/V: home or visiting; PTS: points; REB: rebounds; AST: assists; STL: steals; BLK: blocks

The **Atlanta Hawks (46-12)** beat the **Orlando Magic (19-41) 95-88** on Friday. **Al Horford** had a good all-around game, putting up **17 points, 13 rebounds, four assists and two steals** in a tough matchup against **Nikola Vucevic**. **Kyle Korver** was the lone Atlanta starter not to reach double figures in points. **Jeff Teague** bounced back from an illness, he scored **17 points** to go along with **seven assists and two steals**. After a rough start to the month, the **Hawks** have won three straight and sit atop the Eastern Conference with a nine game lead on the second place Toronto Raptors. The **Magic** lost in devastating fashion to the Miami Heat in overtime Wednesday. They blew a seven point lead with 43 seconds remaining and they might have carried that with them into Friday's contest against the **Hawks**. **Vucevic** led the **Magic** with **21 points and 15 rebounds**. **Aaron Gordon** (ankle) and **Evan Fournier** (hip) were unable to play due to injury. The **Magic** have four teams between them and the eighth and final playoff spot in the Eastern Conference. The **Magic** will host the Charlotte Hornets on Sunday, and the **Hawks** with take on the Heat in Miami on Saturday.

Figure 1: example from the Rotowire Corpus (Wiseman et al. 2017). Given the game statistics (left), the objective is to generate a textual summary (right).

As a second example, the WebNLG challenge (Fereira et al. 2020) consists in transcribing knowledge graphs composed of RDF triplets to text and vice versa (Figure 2).



<S> Serie A <P> champions <O> Juventus F.C. <S> Luciano Spalletti <P> club <O> Udinese Calcio <S> A.S. Roma <P> manager <O> Luciano Spalletti <S> A.S. Roma <P> league <O> Serie A

AS Roma play in the Serie A league where Juventus FC are the champions. Their manager is Luciano Spalletti who has been associated with Udinese Calcio.

Figure 2: example from the WebNLG corpus (figure from (Kale 2020)). On the left a graph structure, on the right top its serialization as a set of RDF triples, on the right bottom, a possible associated text.

The latter problem is known as *semantic parsing* and comes in different instantiations like information retrieval, reasoning over the structured data (table or graph), generating symbolic queries (e.g. SQL) from text which may be used for example in dialog systems when the answer to a user requires querying a database, or generating abstract meaning representations (Belavicqua 2021). The 2020 edition of the WebNLG challenge (Figure 2) addressed the two tasks of data to text and semantic parsing.

PhD objective

Recent progress has been made for data to text tasks through the use of recent transformers (Agarwal et al. 2020, Kale et al. 2020, Guo et al. 2020) or recurrent sequence to sequence models (Rebuffel et al. 2020). Neural methods trained end to end achieve state of the art performance for different D2T tasks. They however suffer from different pitfalls, like data scarcity (small size annotated corpora), hallucinations (generation of linguistic but non-factual sentences), low coverage of table evidence. They are still restricted to tasks of low complexity. For the dual task of text to data generation, state of the art still relies on complex pipelines integrating different components and sequence-to-sequence models still lag behind the best models. Tasks are also of limited complexity.

The research will explore new paradigms for the dual tasks of data to text and text to data generation such as:

- Learning from unaligned corpora

Most current methods require learning from parallel corpora, where data and text are fully aligned and correspond closely one to the other. Besides limiting the amount of training data, this does not correspond to real world applications where strict alignment is seldom available. A first line of research will be the development of new unsupervised frameworks allowing training from unaligned data-text corpora. A related topic to be explored is transfer learning: how general knowledge from models trained in an unsupervised way on large corpora could be adapted to Data2Text and Text2Data.

- Learning from diverse sources

Current benchmarks focus on learning mappings from a unique structured data format to text. In practice data will be collected from different sources encoded through a diversity of structured formats and textual realizations. This requires rethinking the current data to text settings and a second direction will explore new formalisms for learning such multiple correspondences.

- Controlled text and data generation

Current research mainly focuses on the cases where there is a bijective correspondence between the data and text. A more general task as illustrated in (Wiseman et al. 2017), is to summarize information along different aspects of the data. This will allow for example to explain some specific information present in the data or to extract specific information from text. This requires acquiring the knowledge of the aspects and then performs controlled data or text generation. This has been for now explored in limited settings such as controlling the tone, tense, length. We will explore how to control generation according to different aspects and user needs.

References

Agarwal, O., Kale, M., Ge, H., Shakeri, S. and Al-Rfou, R. 2020. Machine Translation Aided Multilingual Data-to-Text Generation and Semantic Parsing. *WebNLG* (2020), 1–5.

Bevilacqua, M., Blloshmi, R. and Navigli, R. 2021. One SPRING to Rule Them Both : Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. *AAAI* (2021).

Bien, M., Gilski, M., Maciejewska, M. and Taisner, W. 2020. RecipeNLG : A Cooking Recipes Dataset for Semi-Structured Text Generation. *INLG* (2020), 22–28.

Deng, X., Awadallah, A.H., Meek, C., Polozov, O., Sun, H. and Richardson, M. 2020. Structure-Grounded Pretraining for Text-to-SQL. *http://arxiv.org/abs/2010.12773* (2020).

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT* (2019), 4171–4186.

Ferreira, T.C., Gardent, C., Ilinykh, N., Lee, C. Van Der, Mille, S., Moussallem, D., Shimorina, A., Loria, C. and Loria, D.L. 2020. The 2020 Bilingual , Bi-Directional WebNLG + Shared Task Overview and Evaluation Results (WebNLG + 2020). (2020).

Guo, Q., Jin, Z., Wang, Z., Qiu, X., Zhang, W., Zhu, J., Zhang, Z. and Wipf, D. 2020. Fork or Fail: Cycle-Consistent Training with Many-to-One Mappings. *http://arxiv.org/abs/2012.07412* (2020).

Guo, Q., Jin, Z., Qiu, X., Zhang, W., Wipf, D. and Zhang, Z. 2020. CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training. *WebNLG* (2020).

Kale, M. 2020. Text-to-Text Pre-Training for Data-to-Text Tasks. *INLG* (2020), 97–102.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *ACL* (2019), 7871–7880.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. 2018. Language Models are Unsupervised Multitask Learners. (2018).

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. *JMLR*. 21, (2020), 1–67.

Rebuffel, C., Soulier, L., Scoutheeten, G. and Gallinari, P. 2020. A Hierarchical Model for Data-to-Text Generation. *ECIR* (2020), 65–80.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B. and Staiano, J. 2020. Discriminative Adversarial Search for Abstractive Summarization. *ICML* (2020).

Wiseman, S., Shieber, S., Rush, A. 2017: Challenges in data-to-document genera- tion. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2253–2263. Association for Computational Linguistics, Copenhagen, Denmark, September 2017.