Pathways to Semi-Unsupervised NLP Brain

M Saiful Bari

1 Background / Motivation / Introduction

The semanticness of the word and the flow of a sentence are the greatest mystery that NLP aims to solve. Complex NLP applications like Abstractive Summarization and Machine Translation need to model long-range structural dependencies in addition to complex interactions between input and output. Currently, most neural models for these tasks make local decisions and thereby suffer from exposure bias and lable bias problems. The pathway to generic systems and their real-life implication heavily depends on unsupervised and transfer learning. The design of traditional method and ideas try to formulate the brain with neurons with various architecture and learns from gradient based objective function optimization which leads to great success to task-oriented Artificial Intelligent system. The path to the generic representation of brain is yet to be discovered.

2 Research Objective

Among all other digital content, text is the highest contributor in terms of usability and impact. If we can understand this content properly then it will be possible to deploy real-life automated systems that can change the landscape of our daily lifestyle. The reliability of an artificial system depends on the function of proper decision making ability hence the term **brain** is used to denote the **algorithms** of it. The new genre of neuron motivated probabilistic methods and optimization techniques learn the representation of any abstract ideas, concepts or decisions. The representations are usually denoted with vectors and learned with gradient based optimization techniques. My research intends to design better model and techniques that lead to a better representation of various abstract concepts which will enable us to solve a set of downstream problems more accurately and elegantly.

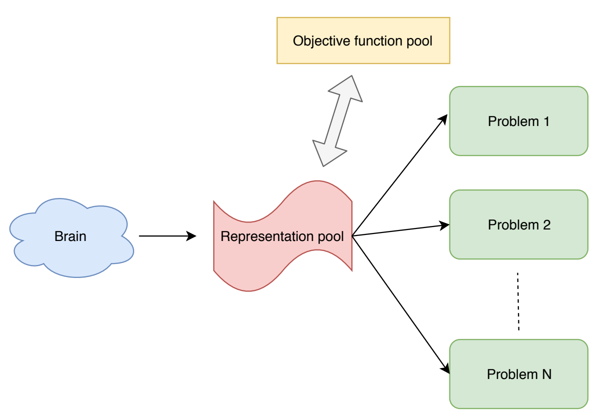

Figure 1: The big picture of NLP brain.

3 Research Scope

Having real-life impact in society as well as virtual space, concepts or proposal of better NLP systems should be applied in various tasks. Before going further let's define the **Semi-Unsupervised** wording. This term is proposed for the first time in this document.

> A system that is built purely in the intention of unsupervised learning but utilize surface level (not lable) knowledge of other resources and often do validation with task-oriented knowledge.

It is observed that as a human we learn most of our action/knowledge by unsupervised or one shot or few shot way. We are very good at observation but not memorization. In general case, it takes a significant amount of attention for us to memorize. But in current form of neural network, it just memorize/generalize the statistics of sample space. The ability to **adapt** and **evolve** is mostly ignored upto now. The recent trend of NLP is very task specific. Meaning various intelligent systems are developed with task-specific goals or objectives. I intend to work with a different approach where I want to build intelligent generic tool or approach that can capture various abstractive concepts and from there the downstream task can be learned more easily. Neural Named Entity Recognition, Machine Translation and Abstractive Summarization, Natural Language Generation etc. are some benefited downstream tasks for this usecase.

4 Research Plan / Methodology

In Natural Language Processing (NLP), our main objectives are to construct a structure of the text to explore information. For NLP, hand-crafted features are not always easy to find and needs individual interaction. However current NLP systems are incredibly fragile because of their atomic symbol representations. Deep learning algorithms attempt to learn multiple levels of representation of increasing complexity/abstraction. Having this advantage deep learning exploits Neural Machine Translation and Abstract Summarization.

Another form of the learning system is Reinforcement learning(RL). Reinforcement learning is a learning technique inspired by behaviorist psychology, concerned with how an agent may take actions in an Environment(ecosystem) so as to maximize the defined cumulative reward. The problem, due to its generality, is studied in many other disciplines, such as game theory, control theory, operations research, genetic algorithms etc. But RL is not yet been carefully applied to NLP problems like translation and summarization.

I plan to work on various downstream tasks while working with proposed methods. The description and analysis of some of the tasks are given below.

4.1 Contextual Representation

The vector representation of tokens and sentences are two very important aspects of generalization of representation of Natural Language contents. There's been a significant amount of contribution in this field in recent years [8] [10] [1] [11] [3] [12] [4] [14] [6]. This words or sentence representation gives a significant improvement boost on learning task-oriented job. Usually, we pretrain this vector representation and inject/finetune them to the model designed for the downstream tasks. The downstream tasks are mainly captured or learned by the encoder which encodes the task objectives.

4.2 Neural Machine Translation

Machine translation is the task of translating between languages(from one language to another). Unlike the traditional statistical MT, the NMT aims to build a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently[7] [18] for NMT often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. NMT uses LSTM-RNN as an encoder to encode a source sentence into a fixed-length vector. The decoder is also an LSTM-RNN that generates the translation.

The initial NMT model [2] was limited to handle long sentences as it used a single static vector to encode the whole sentence. This problem is resolved by attention mechanism [7] [18] which

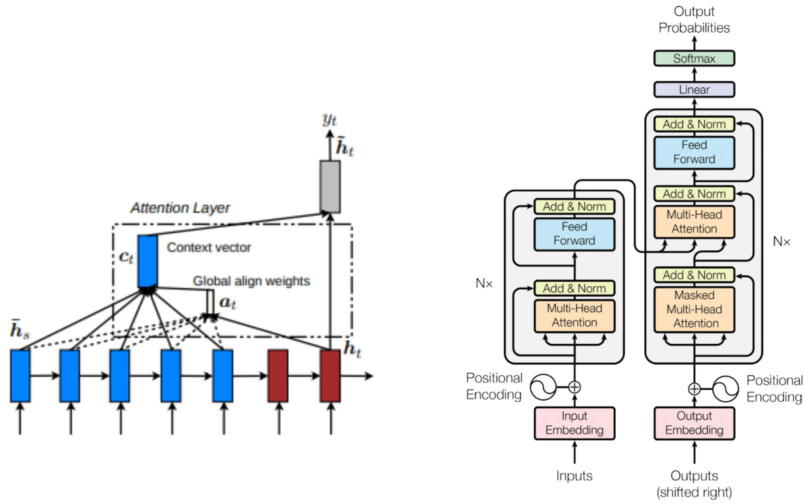

Figure 2: The sequence to sequence model. In the left the model proposed by luong et al [7] and in the right now the model proposed by Vaswani et al.[18]

selectively focuses on parts of the source sentence during translation. [9] sacle the training with reduced precision and large batch training and [19] improve the training with lightweight and dynamic convolution.

4.3 Abstractive Summarization

There are two types of text summarization, Extractive and Abstractive. Till now the majority of the summarization techniques are Extractive. However, because of the limitations of handling longer and complex text of copy-paste based Extractive Summarization, Abstractive summarization is necessary.

RNN based neural network is performed on a sequence of words and has become popular for summarization. The sequence-to-sequence model with attention is used in the early version of summarization where encode RNN first reads the source text word-by-word, producing a sequence of bidirectional encoder hidden states. From there attention distribution is calculated and with the help of learned vocabulary distribution decoder hidden states generates partial summery. Repetitiveness and factual inaccuracy are two problems of this mechanism solved by Pointer-Generator Networks [15].

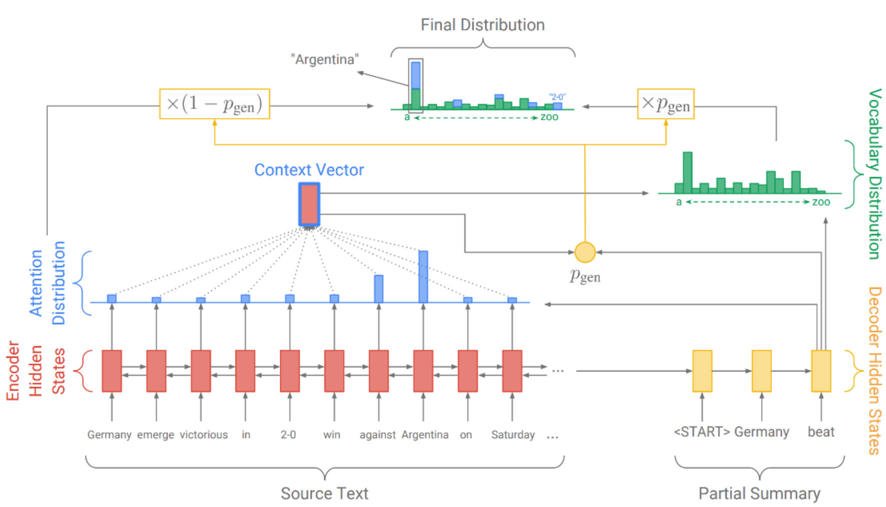

Figure 3: Pointer-Generator Networks proposed by Abigail et al [16]

4.4 Adversarial Training

The adversarial training pioneered by Goodfellow et al. [5] is a generative method where we project a distribution to another distribution with min-max game theory. The objective of this training is to select a saddle point where a discriminator become confused between a real/goal and fake distribution. The optimization function shapes the gradient in such a way the fake distribution transfer from noise to the real/goal distribution. This has become a great source of research scope and opportunity [13] [3] [20] in terms of transfer learning and generative models. Thus adversarial training can facilitate a brain where a lot of domain invariant generic representation will be captured.

4.5 Tree structure in hidden representation

Recently [17] proposes a rnn architecture which captures the tree paths in hidden representation and can retrieve Binary parse tree from the model. The proposed approach works tries to generate hidden representation for a tree which is "tree root to the current node". However a better approach may be possible to articulate the node or a subset of node systematically in the hidden representation.

5 Conclusion

An NLP task depends on the success of encoding-decoding the sequential contextual information along with a method to optimize the proper objective function. The intention of this proposal is to simulate an artificial brain that can facilitate the objectives of various task rather than learning task specific objective. The term semi-unsupervised introduced for validation (to measure best model in various task we need task level dataset) on downstream task. Hence the pathway to the semi-unsupervised NLP brain begins.

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.

[2] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[6] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015.

[7] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025, 2015.

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[9] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, 2018.

[10] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[11] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[12] Alec Radford. Improving language understanding by generative pre-training. 2018.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[14] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[15] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

[16] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368, 2017.

[17] Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron C. Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. *CoRR*, abs/1810.09536, 2018.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[19] Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*, 2019.

[20] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.