Research Statement — Dani Yogatama

The ability to continuously learn and generalize to new problems quickly is a hallmark of general intelligence. In my research, I design **machine learning** models to advance **artificial intelligence** on this front. Within machine learning, my primary application area is **natural language processing**. I focus on language since it is a core component of intelligence and a primary medium through which humans acquire and communicate knowledge. I believe that solving language is an important step toward solving intelligence.

While deep learning has driven progress in many language understanding tasks, existing models have been shown to require a lot of in-domain training examples, rapidly overfit to particular datasets, and suffer from catastrophic forgetting (Yogatama *et al.*, 2019). In contrast, humans are able to learn incrementally and accumulate knowledge to facilitate faster learning of new skills. My long-term goal is to build an agent that is capable of performing multiple linguistic tasks (e.g., question answering, translation, summarization) in multiple languages and uses its experience to continuously improve over time. In order to make progress toward this goal (i.e., artificial general linguistic intelligence), we need machine learning models that can (i) deal with the full complexity of natural language, (ii) effectively store and reuse representations and combinatorial modules, and (iii) adapt to new tasks in new environments with little experience. My research seeks to answer the following questions:

- How do we find the best way to **represent language** for various tasks? (§1)
- How do we **store and reuse linguistic knowledge** to avoid **catastrophic forgetting**? (§2)
- How do we ensure **sample efficient generalization** to new problems? (§3)

I take inspirations from cognitive science and neuroscience and use techniques from deep learning, probabilistic graphical models, information theory, and many others to answer these questions. I have broad interests in many aspects of machine learning and natural language processing. In §4, I discuss future directions.

1 Representation Learning

The performance of a machine learning model heavily depends on how the data is represented in the model. For example, when working with text data, we can represent it as a sequence of words, subwords, or characters. Furthermore, each textual unit can be represented as strings, binary vectors, or real vectors. Recent advances have sought to automate the process of choosing the best representation in order to achieve better results.

My interest in representation learning started at CMU where I worked on sentence regularization (Yogatama and Smith, 2014), representation learning as hyperparameter selection (Yogatama *et al.*, 2015a), entity-type embeddings (Yogatama *et al.*, 2015b), and sparse word embeddings (Yogatama *et al.*, 2015c; Faruqui *et al.*, 2015). At DeepMind, I continue contributing theoretical foundations and analytical insights to improve representation learning methods.

In Kong *et al.* (2019a), we showed that state-of-the-art language representation learning methods maximize an objective function that is a lower bound on the mutual information between different parts of a sentence. Our formulation provides an information theoretic perspective that unifies classical word embedding models and modern contextual embeddings. It also leads to a principled framework that can be used to construct new self-supervised tasks. The resulting framework offers a holistic view of representation learning methods to transfer knowledge and translate progress across multiple domains (e.g., natural language processing, computer vision, audio processing).

In Artetxe *et al.* (2019), we presented a method to transfer a representation learning model for a particular language (e.g., English) to other languages. The project was motivated by our observation that human language learning is facilitated by abstractions that are independent of any particular language. We evaluated the model in a zero-shot cross-lingual setting (without labeled training data in the new languages) and demonstrated that a language model trained on English learns generalizable abstractions that are reusable in other languages. As a part of this project, we also created a new multilingual question answering dataset that was released to the research community.

In order to build effective language representations, we also need combinatorial modules which compose words into representations of phrases, sentences, and documents. In Yogatama *et al.* (2017b) and Maillard *et al.* (2019), we explored two methods based on reinforcement learning and differentiable parsers that compute representations of the meaning of sentences by composing representations of words and phrases. We showed that our automatic approaches yield better representations compared to methods that rely on explicit supervisions (e.g., syntactic parse trees of sentences).

2 Memory Models

Short-term and long-term memory systems is an integral part of human intelligence. Machine learning models work well on a single dataset given enough training examples, but they often fail to isolate and reuse previously acquired knowledge when the data distribution shifts (e.g., when presented with a new dataset)—a phenomenon known as catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990). I work on combining neural networks with memory modules to allow artificial agents to store knowledge and reuse it effectively.

My work in lifelong learning (i.e., where a model continuously learns from a stream of examples from multiple datasets) started at CMU where I designed a dynamic language model for streaming text (Yogatama *et al.*, 2014). At DeepMind, we presented a method that augments a neural network language model with an episodic memory module to mitigate catastrophic forgetting (de Masson d'Autume *et al.*, 2019). The memory module is used to store previously seen examples, which are then used for sparse experience replay and local adaptation (such a process bears some similarity to memory consolidation in human learning; McGaugh 2000). We showed that our model is able to accumulate knowledge throughout its lifetime in question answering and text classification experiments.

In Yogatama *et al.* (2018), we compared several working memory architectures (i.e., sequential, random access, and stack memory architectures) that are used to capture linguistic dependencies when learning a language model and proposed a new continuous stack memory model. We observed that stack-based architectures that encode a bias resembling hierarchical dependencies inherently found in natural language perform the best in explaining the distribution of words in natural language. This result is in line with linguistic theories that claim a context-free backbone for natural language (Chomsky, 1957).

3 Sample Efficient Learning

Humans are able to generalize to new problems quickly. Neural networks perform well at pattern recognition, but they still require a large number of in-domain training examples. One of my research goals is to build agents that exhibit sample-efficient generalization.

In Yogatama *et al.* (2019), we proposed a metric—online codelength—that quantifies how quickly an agent learns a new task. Existing metrics evaluate model performance on a held-out test set for a task of interest. These metrics capture an essential aspect of intelligence: being able to generalize to new inputs. However, none of them asses another defining attribute of intelligence: the ability to generalize *rapidly* to a new task. Online codelength rewards models that perform well with limited numbers of training examples, in addition to overall performance. It is rooted in information theory and is based on connections between generalization, compression, and comprehension (Wolff, 1982; Chaitin, 2007). It can be used across a number of tasks by any probabilistic model, allows seamless incorporation of other model and training properties (e.g., model complexity, training cost), and correlates well with standard evaluation metrics such as accuracy and F_1.

In Yogatama *et al.* (2017a), we characterized the performance of discriminative and generative recurrent neural networks for text classification. We found that generative models that have no task-specific modules approach their asymptotic error rate more rapidly than their discriminative counterparts (i.e., they are better in the small data regime). Our results empirically extended the theoretical results of Ng and Jordan (2001) from linear to nonlinear classification models. We then derived a more sample-efficient neural network classifier based on this finding.

My other projects in this area include a sample efficient hyperparameter tuning method (Yogatama and Mann, 2014) and variational language models (Kong *et al.*, 2019b).

4 Future Directions

Achieving general linguistic intelligence requires advances in many areas. I am especially interested in exploring the following directions in the next five years.

Semi-parameteric models. I think that limitations of existing approaches (e.g., data hungry, catastrophic forgetting, overfitting to a dataset instead of solving a task) are inherently caused by the way we train our agents as big parametric models. In my research, I have been exploring methods to combine non-parametric components such as memory modules with parametric neural networks. I am excited about this research direction, both in terms of how to combine these two modules effectively and how to consolidate (compress) past experiences (i.e., learning what to remember and forget) to manage the time and space complexity when using non-parametric components. I plan to continue taking inspirations from neuroscience and cognitive science to make progress in this area.

Hierarchical generative models. In addition to being more sample efficient, a perfect generative language model—in theory—should be able to do any linguistic task (e.g., by formulating the task as a question and querying the language model to generate answers). Generative modeling is also crucial for imagination and planning. I believe that investing in generative language models would lead to major advances. I am interested in designing hierarchical language models which meta learn from both task and example distributions. For example, a first step in this direction is to create a hierarchical model where each example for a task is drawn from a distribution that depends on a task variable (e.g., a task embedding which is a function of examples in the task). Such a model would be more robust to data distribution shifts and generalize to new tasks more efficiently.

Representation learning. Over the last few years, progress in this area has driven advances in many downstream tasks. The rapid pace of empirical progress created a gap between our theoretical understanding of state-of-the-art models and their practical applications. I think understanding these models is crucial to assess their limitations and provide a launchpad for future breakthroughs. I am eager to continue working on unsupervised and self-supervised representation learning. In language, a particular weakness that I plan to address is on how existing models fail to learn to encode persistent knowledge that is useful across sentences.

Summary. I think we should be moving toward one universal lifelong model that performs multiple tasks and uses its experience to continually improve over time. I have been fortunate to work with leaders in this area in my research career, and I am excited to continue working toward my long-term goal to build a general linguistic agent that is capable of understanding and generating natural language.

References

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv:1910.11856*.

Chaitin, G. J. (2007). On the intelligibility of the universe and the notions of simplicity, complexity and irreducibility. In *Thinking about Godel and Turing: Essays on Complexity, 1970–2007*. World Scientific.

Chomsky, N. (1957). *Syntactic Structures*. Mouton.

de Masson d'Autume, C., Ruder, S., Kong, L., and Yogatama, D. (2019). Episodic memory in lifelong language learning. In *Proc. of NeurIPS*.

Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., and Smith, N. A. (2015). Sparse binary word vector representations. In *Proc. of ACL*.

Kong, L., de Masson d'Autume, C., Yu, L., Ling, W., Dai, Z., and Yogatama, D. (2019a). A mutual information maximization perspective of language representation learning. *arXiv:1910.08350*.

Kong, L., Melis, G., Ling, W., Yu, L., and Yogatama, D. (2019b). Variational smoothing in recurrent neural network language models. In *Proc. of ICLR*.

Maillard, J., Clark, S., and Yogatama, D. (2019). Jointly learning sentence embeddings and syntax with unsupervised tree-LSTMs. *Journal of Natural Language Engineering*, **25**(4), 433–449.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier.

McGaugh, J. L. (2000). Memory–a century of consolidation. *Science*, **287**(5451), 248–251.

Ng, A. Y. and Jordan, M. I. (2001). On generative and discriminative classifiers: A comparison of logistic regression and naive bayes. In *Proc. of NeurIPS*.

Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological Review*, **97**(2), 285.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication*, **2**(1), 57–89.

Yogatama, D. and Mann, G. (2014). Efficient transfer learning method for automatic hyperparameter tuning. In *Proc. of AISTATS*.

Yogatama, D. and Smith, N. A. (2014). Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *Proc. of ICML*.

Yogatama, D., Wang, C., Routledge, B. R., Smith, N. A., and Xing, E. P. (2014). Dynamic language models for streaming text. *Transactions of the Association for Computational Linguistics*, **2**, 181–192.

Yogatama, D., Kong, L., and Smith, N. A. (2015a). Bayesian optimization of text representations. In *Proc. of EMNLP*.

Yogatama, D., Gillick, D., and Lazic, N. (2015b). Embedding methods for fine grained entity type classification. In *Proc. of ACL*.

Yogatama, D., Faruqui, M., Dyer, C., and Smith, N. A. (2015c). Learning word representations with hierarchical sparse coding. In *Proc. of ICML*.

Yogatama, D., Dyer, C., Ling, W., and Blunsom, P. (2017a). Generative and discriminative recurrent neural networks. *arXiv:1703.01898*.

Yogatama, D., Blunsom, P., Dyer, C., Grefenstette, E., and Ling, W. (2017b). Learning to compose words into sentences with reinforcement learning. In *Proc. of ICLR*.

Yogatama, D., Miao, Y., Melis, G., Ling, W., Kuncoro, A., Dyer, C., and Blunsom, P. (2018). Memory architectures in recurrent neural network language models. In *Proc. of ICLR*.

Yogatama, D., de Masson d'Autume, C., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., and Blunsom, P. (2019). Learning and evaluating general linguistic intelligence. *arXiv:1901.11373*.