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 in-domain training examples that are often costly to annotate, rapidly overfit to the idiosyncrasies of 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 artificial agent with human-level linguistic ability; that is capable of performing multiple tasks (e.g., question answering, translation, summarization) in multiple languages and uses its experience to continuously improve over time. I endeavor to provide a better understanding of natural intelligence in the process of advancing artificially intelligent systems.

In order to make progress toward this goal (i.e., artificial general linguistic intelligence), we need models that can (i) deal with the full complexity of natural language, (ii) 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:

- What **memory mechanisms** are needed to store and reuse linguistic knowledge? (§1)
- How do we **represent language**? (§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 Memory Models

Memory in humans consists of specialized systems, which forms a basis for intelligent behaviors (Tulving, 1985; Rolls, 2000; Eichenbaum, 2012). Machine learning models work well on a dataset given enough training examples, but they often fail when the data distribution shifts (e.g., when presented with very long context or a new dataset)—a phenomenon known as catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990). I work on memory modules to allow agents to store knowledge and reuse it effectively throughout its lifetime.

My work in lifelong learning (i.e., where a model continuously learns from a stream of examples drawn from evolving distributions) started at CMU, where we designed a dynamic language model for streaming text (Yogatama *et al.*, 2014). At DeepMind, we presented a method that augments a 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 experience replay and local adaptation. Such a process bears some similarity to human memory consolidation (McGaugh, 2000).

In another project (Yogatama *et al.*, 2018), we analyzed several working memory architectures that are used to capture short-term linguistic dependencies 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).

As a continuation of the above two projects, inspired by the modular design of human memory systems, we presented a language model that combines a large neural network with both short-term working memory and long-term episodic memory components in an integrated architecture (Yogatama *et al.*, 2020). Our model is able to use short-term or long-term memory (or a combination of them) on an ad hoc basis depending on the context. We showed the efficacy of our model for word-based and character-based language modeling.

2 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. Distributed representations—which has been argued to have strengths and weaknesses that match those of the human mind (Hinton *et al.*, 1986)—have emerged as the leading approach to represent objects in artificial systems.

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.* (2020), 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.* (2020b), 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 and argued for a more rigorous comparison of cross-lingual learning methods (Artetxe *et al.*, 2020a).

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).

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 generalize efficiently.

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: the

ability 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 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 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 used this finding to derive a more sample-efficient neural network classifier.

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.

Semiparameteric 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 parametric neural networks with non-parametric components such as memory modules. I am excited about this research direction, both in terms of (i) how to compress past experiences to manage the time and space complexity when using non-parametric components (i.e., learning what to remember and forget); and (ii) how to incorporate data of different modalities (e.g., text, images, structured knowledge bases) to improve a language model. I plan to continue taking inspirations from neuroscience and cognitive science to make progress in this area (Nematzadeh *et al.*, 2020).

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 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 a lifelong model that learns from multimodal data and uses language to interact with humans. 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 linguistically intelligent agent and provide a better understanding of natural intelligence in the process.

References

Artetxe, M., Ruder, S., Yogatama, D., Labaka, G., and Agirre, E. (2020a). A call for more rigor in unsupervised cross-lingual learning. In *Proc. of ACL*.

Artetxe, M., Ruder, S., and Yogatama, D. (2020b). On the cross-lingual transferability of monolingual representations. In *Proc. of ACL*.

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*.

Eichenbaum, H. (2012). Memory systems. *Handbook of Psychology, Second Edition*, **3**.

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

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986). Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume I, chapter 3. MIT Press, Cambridge, MA.

Kong, L., de Masson d'Autume, C., Yu, L., Ling, W., Dai, Z., and Yogatama, D. (2020). A mutual information maximization perspective of language representation learning. 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.

Nematzadeh, A., Ruder, S., and Yogatama, D. (2020). On memory in human and artificial language processing systems. In *Proc. of ICLR Workshop on Bridging AI and Cognitive Science*.

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.

Rolls, E. T. (2000). Memory systems in the brain. *Annual Review of Psychology*, **51**(1), 599–630.

Tulving, E. (1985). How many memory systems are there? *American Psychologist*, **40**, 385–398.

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

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*.

Yogatama, D., de Masson d'Autume, C., and Kong, L. (2020). Adaptive semiparametric language models. *In review*.