My research is primarily centered on the development of *simpler*, more *resource-efficient* alternatives to address current NLP problems, and my work is dedicated to exploring methodologies that aim to minimize computational requirements while boosting performance.

By investigating innovative algorithmic designs and distribution strategies, I aim to democratize access to NLP technologies, making them more accessible across diverse applications and user communities. I have a particular interest in *information extraction*, with a specific focus on structured prediction-based methods, which I find to be both intriguing and instrumental in this process.

Structured prediction in machine learning focuses on mapping a sequence of inputs to a sequence of outputs within a vast output space, each prediction interconnected with others. My dissertation focused on harnessing the power of structured prediction to enhance and simplify intricate solutions in areas such as machine translation, entity linking, and question answering. The following sections will provide a more detailed explanation of my published contributions.

Structured Prediction for Machine Translation

Introduction of BERT [2] was a revolutionary breakthrough in NLP and Jawahar et al. [6], among others, showed that BERT captures syntax and semantics very well. We demonstrated that a lightweight structured prediction-based linguistic information extraction module from a non-finetuned BERT is effective in enhancing translation quality in different settings considering a spectrum of training set sizes for Machine Translation [14]. Importantly, this enhancement was achieved without imposing supplementary inference time or introducing undue complexity to the model.

Structured Prediction for Entity Linking

Knowledge bases such as Wikipedia serve as extensive repositories of information, and Entity Linking (EL) sifts through unstructured text to identify spans of text (mentions) that correspond to entries in these repositories. Given that knowledge bases often encompass millions of entries, EL faces a formidable challenge due to the vast pool of potential candidate entries it must navigate even when correctly identifying a mention. We proposed SpEL [15], a novel efficient structured prediction-based approach with a single affine transformation atop a RoBERTa [10] encoder. We devised several mechanisms including context sensitive prediction aggregation and careful entity vocabulary construction to ensure state-of-the-art performance of SpEL on a common entity linking benchmarking dataset. In addition to exceptional performance, SpEL can smoothly operate on CPU for inference and can be fine-tuned with a GPU memory of less than 6GBs.

Unified Examination of Entity Linking

In a subsequent study [11], we comprehensively examined recent neural entity linking methods within a unified black-box evaluation framework. Our experiments focused on how these models responded to an entirely unseen test set of recent news articles (AIDA/testc), aiming to understand the impact of adaptive overfitting [9] on recently published entity linking contributions. Additionally, we benchmarked these models in scenarios where candidate sets (shortlists of knowledge base entities for each mention span) were completely absent. We found that while

adaptive overfitting is not a significant issue in entity linking, only a handful of off-the-shelf entity linking models, including SpEL, maintain effectiveness in the absence of handcrafted candidate sets which are not typically available in specific domains such as medical NLP or languages other than English.

Entity Retrieval for Answering Entity-Centric Questions

Following the emergence of Large Language Models [LLMs; 12, 7, *inter alia*], retrieval-augmentation [8, 5] has become a common approach to enhance the knowledgability and factual reliability of LLMs [19, 13, 21]. We studied the application of Entity Linking, focusing on our proposed SpEL framework, as an alternative to the widely used dense retrieval methods in retrieval-augmented question answering, especially for entity-centric questions about the real world. Our findings indicated that our proposed *Entity Retrieval* strategy [16] surpasses other retrieval methods in performance, while requiring less time and resources.

Other Work: Multi-class Multilingual Classification of Wikipedia

Wikipedia stands as a valuable resource of world knowledge for humans, yet its lack of structured data poses challenges for NLP models in comprehending contextual cues when making predictions, particularly evident across non-English languages. To address this gap, we created SHINRA-5LDS [17], a multilingual hierarchical classification dataset featuring fine-grained annotations across three levels of hierarchy for Wikipedia articles in Japanese, English, French, German, and Farsi. Our evaluation with contemporary classification models underscored the formidable hurdles text classifiers encounter when tasked with large datasets and fine-grained tag sets.

Other Work: Improving the Efficiency and Quality of Machine Translation

In a separate line of work, I have contributed to improving the efficiency and quality of language translation through various innovative methods, as briefly discussed below:

- A data annotation algorithm to identify optimal segmentation boundaries considering both latency and translation quality in spoken language translation [18].
- Integration of the segmentation model of [18] and an incremental decoding algorithm to create an automatic simultaneous translation framework, which outperformed other comparable systems available at the time [20].
- A supervised learning approach for training an agent in simultaneous translation, which minimizes the average lagging in producing target tokens while maintaining high translation quality [1].
- A hierarchical tree-based decoding approach incorporating syntactic dependencies to enhance translation fluency and improve reordering accuracy in machine translation [3].
- A constrained decoding algorithm integrating bilingual lexicons into the copy mechanism for machine translation, effectively reducing model parameter size and training costs without compromising performance [4].

References

[1] Ashkan Alinejad, **Hassan S. Shavarani**, and Anoop Sarkar. Translation-based supervision for policy generation in simultaneous neural machine translation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1734–1744, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.130. URL https://aclanthology.org/2021.emnlp-main.130.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[3] Jetic Gū, **Hassan S. Shavarani**, and Anoop Sarkar. Top-down tree structured decoding with syntactic connections for neural machine translation and parsing. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 401–413, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1037. URL https://aclanthology.org/D18-1037.

[4] Jetic Gū, **Hassan S. Shavarani**, and Anoop Sarkar. Pointer-based fusion of bilingual lexicons into neural machine translation. *arXiv preprint arXiv:1909.07907*, 2019. URL https://arxiv.org/pdf/1909.07907.

[5] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74.

[6] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL https://aclanthology.org/P19-1356.

[7] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/pdf/2401.04088.pdf.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-

augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[9] Shuheng Liu and Alan Ritter. Do CoNLL-2003 named entity taggers still work well in 2023? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8254–8271, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.459. URL https://aclanthology.org/2023.acl-long.459.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. URL https://arxiv.org/pdf/1907.11692.pdf.

[11] Nicolas Ong, **Hassan S. Shavarani**, and Anoop Sarkar. Unified examination of entity linking in absence of candidate sets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico, 2024. Association for Computational Linguistics. URL https://openreview.net/forum?id=p8U0sZWOrt.

[12] OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/pdf/2303.08774.pdf.

[13] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023. URL https://arxiv.org/pdf/2302.12813.pdf.

[14] **Hassan S. Shavarani** and Anoop Sarkar. Better neural machine translation by extracting linguistic information from BERT. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2772–2783, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.241. URL https://aclanthology.org/2021.eacl-main.241.

[15] **Hassan S. Shavarani** and Anoop Sarkar. SpEL: Structured prediction for entity linking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11123–11137, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.686.

[16] **Hassan S. Shavarani** and Anoop Sarkar. Entity retrieval for answering entity-centric questions. In *review for the 2024 Conference on Empirical Methods in Natural Language Processing*, Florida, United States, November 2024. Association for Computational Linguistics.

[17] **Hassan S. Shavarani** and Satoshi Sekine. Multi-class multilingual classification of Wikipedia articles using extended named entity tag set. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and*

Evaluation Conference, pages 1197–1201, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.150.

[18] **Hassan S. Shavarani**, Maryam Siahbani, Ramtin Mehdizadeh Seraj, and Anoop Sarkar. Learning segmentations that balance latency versus quality in spoken language translation. In Marcello Federico, Sebastian Stüker, and Jan Niehues, editors, *Proceedings of the 12th International Workshop on Spoken Language Translation: Papers*, pages 217–224, Da Nang, Vietnam, December 3-4 2015. URL https://aclanthology.org/2015.iwslt-papers.14.

[19] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023. URL https://arxiv.org/pdf/2301.12652.pdf.

[20] Maryam Siahbani, **Hassan S. Shavarani**, Ashkan Alinejad, and Anoop Sarkar. Simultaneous translation using optimized segmentation. In Colin Cherry and Graham Neubig, editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 154–167, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1815.

[21] Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*, 2023. URL https://arxiv.org/pdf/2305.14002.pdf.