Research Highlights

1. **I worked on evaluating and understanding model performance**. This allows insights into performance bottlenecks in otherwise opaque neural models, providing more concrete way forward, instead of single metrics, such as BLEU for MT, which conflate various performance issues into an often non-informative number.

    - Word embedding usually capture only the most common sense of ambiguous words [9].
    - Machine translation is gender biased (for example, assumes doctors are always men, nurses are always women) [11] **(Nominated for Best Paper at ACL 2019)**.
    - BLEU and F1 are quite poor evaluation metrics for text generation in QA [1] **(Best Paper at MRQA workshop 2019 at EMNLP)**.
    - Semantic representations prove difficult to learn within the active learning paradigm [4]. **(Received an Honorable Mention at CoNLL 2019)**.

2. **I created large-scale datasets for various NLP tasks.** These enable progress in the field by providing benchmarks on which state-of-the-art models can be tested and improved against a myriad of challenging semantic phenomena.

    - **I co-organized a SemEval workshop and shared task on Math QA**. As part of this workshop we published a standard corpus for evaluation from different Math textbooks (SAT exams, practice books, etc.). The shared task had over dozens of system submissions [3].
    - We recently published DROP: a 96K QA questions over long paragraphs requiring discrete mathematical operations over dates and quantities [2].
    - I created a unified version of the factuality corpus [12], which we showed achieves state-of-the-art results, and is being used as the standard ML setting for the task.

3. I worked extensively within the Open IE paradigm, showing that **QA-SRL subsumes the information in Open IE**. This insight opened the door for a much-needed first standardization of the popular Open IE task, by creating the first large-scale train and test corpus for Open IE. Follow up work have used this dataset to develop new and improved methods for the task.

    - We created a large corpus by conversion from the QA-SRL corpus [7]. This is being used as the conventional benchmark for the task in academia and industry. Several startup companies were in contact with me, saying that they use this benchmark to show progress in their technologies (e.g., plasticity.ai, diffbot).
    - Using this corpus, we trained the first neural model for the task [10]. This model is now hosted on AllenNLP library and the online demo gets over a thousands requests per month.[1]
    - I showed that Open IE is a useful representation for downstream tasks, such reading comprehension and lexical similarity [8].

4. **I advanced the QA-SRL paradigm to make it more expressive and useful in downstream applications.** These advancements enable annotation of a complete and intuitive meaning representation which does not require expensive annotator training or the development of predefined annotation schema, thus lending itself for annotation in low resource domains and new languages.

    - We extended the verbal QA-SRL templates to QAMR: a more expressive formalism between any two words in the sentence, using open-form questions. Also showing that given dense enough annotations, we can derive full sentence graph representations. These can be used to annotate relations for low-resource domains, without requiring expert annotation [5].
    - Besides the Open IE use-case, I also used QA-SRL to identify restrictive vs. non-restrictive clauses (which can be omitted without changing the meaning of the sentence [6].

5. In all of these projects, **I collaborate extensively** – I coauthored 24 papers with 40 NLP researchers in different research groups in both academia and industry. and was invited for talks at various seminars (e.g., Nuance, IBM, Darmstadt, UW). During my PhD, I was invited to two research internships, at AI2 and IBM. The work I did at IBM was published and used in a product which detects adverse drug reactions in social media [13].

[1] https://demo.allennlp.org/open-information-extraction

References

[1] Anthony Chen, **Gabriel Stanovsky**, Sameer Singh, and Matt Gardner. Evaluating question answering evaluation. In *Machine Reading for Question Answering (MRQA) Workshop at EMNLP 2019*, Hong Kong, 2019 **(Received the Best Paper award)**.

[2] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, **Gabriel Stanovsky**, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*, 2019.

[3] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, **Gabriel Stanovsky**, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[4] Omri Koshorek, **Gabriel Stanovsky**, Yichu Zhou, Vivek Srikumar, and Jonathan Berant. On the limits of learning to actively learn semantic representations. In *Computational Natural Language Learning (CoNLL)*, 2019 **(Received an Honorable Mention award)**,.

[5] Julian Michael, **Gabriel Stanovsky**, Luheng He, Ido Dagan, and Luke S. Zettlemoyer. Crowdsourcing question-answer meaning representations. In *NAACL-HLT*, 2018.

[6] **Gabriel Stanovsky** and Ido Dagan. Annotating and predicting non-restrictive noun phrase modifications. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016.

[7] **Gabriel Stanovsky** and Ido Dagan. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, November 2016. Association for Computational Linguistics.

[8] **Gabriel Stanovsky**, Ido Dagan, and Mausam. Open ie as an intermediate structure for semantic tasks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, 2015.

[9] **Gabriel Stanovsky** and Mark Hopkins. Spot the odd man out: Exploring the associative power of lexical resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[10] **Gabriel Stanovsky**, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In *Proceedings of The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[11] **Gabriel Stanovsky**, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Association for Computational Linguistics (ACL)*, Florence, Italy **(Nominated for Best Paper award, top 1% of submissions)**, June 2019.

[12] **Gabriel Stanovsky**, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, Vancouver, Canada, August 2017.

[13] **Gabriel Stanovsky**, Daniel Gruhl, and Pablo N. Mendes. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics*, Valencia, Spain, April 2017. Association for Computational Linguistics.