Model-based Evaluation of Multilinguality

Jannis Vamvas
Department of Computational Linguistics
University of Zurich
vamvas@cl.uzh.ch

The aim of this thesis was to extend the methodological toolbox for evaluating the ability of natural language processing systems to handle multiple languages. Neural machine translation (NMT) took the central role in this endeavor: NMT is inherently cross-lingual, and multilingual NMT systems, which translate from many source languages into many target languages, embody the concept of multilinguality in a very tangible way. In addition, NMT and specifically the perplexity of NMT systems can themselves be used as a tool for evaluating multilinguality.

Limitations of targeted evaluation methods for machine translation

In (Vamvas and Sennrich, 2021a), we identified a limitation of an existing targeted evaluation method, **contrastive evaluation using minimal pairs**. We discussed this limitation from a theoretical perspective by drawing a comparison between the conditions of contrastive evaluation and the concept of exposure bias.

We then performed experiments with English–German machine translation and demonstrated that testing implausible hypotheses using contrastive evaluation could lead to incorrect conclusions about the errors actually made by a system in practice. Finally, we proposed an effective mitigation approach, deriving minimal pairs from NMT-generated translations instead of human-written reference translations.

Contrastive conditioning: A novel approach to targeted evaluation

In (Vamvas and Sennrich, 2021b), we proposed **contrastive conditioning**, a novel targeted evaluation method for machine translation. Our idea is to analyze machine translations by measuring the perplexity of an "expert" NMT system that we provide with privileged information via a modified source sequence. Unlike some previous methods, contrastive conditioning can be used for a targeted evaluation of **black-box systems** such as commercial translation APIs. Another advantage of contrastive conditioning is that it requires few assumptions about the specific target language used, which allows for the scaling of automatic evaluation to many languages.

Two applications of contrastive conditioning

- In (Vamvas and Sennrich, 2021b), we used the method to quantify **overgeneralization bias** when translating ambiguous source expressions, which is a major challenge for machine translation. We hypothesized that lexical overgeneralization is more pronounced in NMT systems trained with **knowledge distillation**. Through the use of contrastive conditioning, we showed that distilled models are indeed more biased than non-distilled models, even if their overall quality is equal.

- In (Vamvas and Sennrich, 2022a), we demonstrated how contrastive conditioning can be applied to the automatic recognition of **erroneous omission and addition of content**. We performed a human evaluation study to validate our simple approach and found that the accuracy in detecting omission errors is comparable to that of a specialized quality estima-

tion model that was trained on a large amount of synthetic data.

Translation cross-likelihood for semantic similarity

In the final publication included in the thesis (Vamvas and Sennrich, 2022b) we proposed a novel and robust way of using NMT perplexity for judging the similarity of sentence pairs, called **translation cross-likelihood**. We evaluated our approach on paraphrase identification and found that cross-likelihood tends to have a higher accuracy than previous approaches. We also found that translation-based similarity measures strongly outperform embedding-based measures in distinguishing between paraphrases and adversarial non-paraphrases. Finally, we highlighted the potential of evaluation based on NMT perplexity on the example of multilingual data-to-text generation.

Dissemination and Impact

A focus of this thesis has been the open sharing of research artifacts. All research code has been released on GitHub[1], including the `NMTScore` library[2] for computing translation perplexity. Whenever possible, open-source models and open datasets were used. Every paper was accompanied by a lay summary on the candidate's research blog.[3]

Acknowledgments

The author would like to thank his Ph.D. supervisors, Rico Sennrich and Lena A. Jäger, and the doctoral committee members, Lena A. Jäger and Bill Byrne. The work presented in this thesis was funded by the Swiss National Science Foundation (project MUTAMUR; no. 176727).

Relevant Publications

Vamvas, Jannis and Rico Sennrich. 2021a. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vamvas, Jannis and Rico Sennrich. 2021b. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vamvas, Jannis and Rico Sennrich. 2022a. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Dublin, Ireland, May. Association for Computational Linguistics.

Vamvas, Jannis and Rico Sennrich. 2022b. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

[1] `https://github.com/ZurichNLP`
[2] `https://github.com/ZurichNLP/nmtscore`
[3] `https://vamvas.ch`