Statement of Purpose, Computer Science PhD: William Chen wchen6255@knights.ucf.edu

My research interests are in NLP, focusing primarily on two research directions, machine translation and NLP for low-resource languages, and the intersection between them. These interests can be summarized in the following research question: “How can we better integrate linguistic knowledge to improve the performance of machine translation models given a low-resource scenario?”

I want to study low-resource languages in particular because they generally have smaller speaking populations and are thus more prone to endangerment. Developing machine translation models for these languages can help spread awareness and garner interest by exposing more people to their culture and content. Additionally, these languages have properties unique to the commonly-studied (mostly West European) languages, allowing us to evaluate the effectiveness of existing methods on a more diverse set of linguistic features.

My goal is to help expand machine translation to all the languages around the world, whether they be spoken, textual or signed. I arrived at this conclusion while independently researching subword segmentation for low-resource languages, and was only further affirmed of it when participating in the different translation shared tasks; I find that I particularly enjoy working with different languages and trying to understand them. A PhD would allow me to pursue this goal while helping me develop the skills necessary to achieve it.

I have been independently working on these topics since my final semester at the University of Central Florida (UCF) while completing my formal lab work under different professors in their areas of interest. This has especially helped me prepare for a PhD by learning how to apply the techniques and lessons I’ve gathered across different domains and projects. My experiences and their relevance to my research interests are detailed below.

Current state-of-the art machine translation systems are built upon on machine learning techniques for performance improvements over a base architecture, mostly relying on large models that can overfit to gigantic datasets. This is an issue for low-resource languages, which do not have the necessary amount of data to produce even remotely usable translation models, regardless of computing power and parameters. Furthermore, those techniques do not exploit the wealth of underlying linguistic structure that is available. Integrating structural knowledge, such as morphology and phonology, into language models may be especially useful for low-resource scenarios where there is plenty of existing linguistic knowledge but little amounts of labeled training data.

This idea drove my research project on applying a linguistically-driven subword segmentation algorithm to the task of machine translation. I lead a team of two in what began as a end-of-semester project for an introductory NLP class and ended up as Honorable Mention for the Best Paper Award at The 2021 Workshop for Low-Resource Machine Translation (LoResMT). We showed that exploiting morphological relationships during segmentation led to significant improvements in translation quality (in terms of BLEU score) for agglutinative languages, going as far as beating out similar models trained with larger amounts of data.

During our experiments, we observed that the combination of segmentation algorithm and language pair had a considerable effect on translation quality. However, most translation research work simply defaulted to an unsupervised algorithm without explanation. We also found that while Transformers were known to scale well to larger training corpora, LSTMs were far more effective in a low-resource scenario. We leveraged these findings in the 2021 LoResMT Shared Task, where we outperformed all other submissions in the constrained-data track for several language pairs. I attempted to further exploit this knowledge at the WMT 2021 FLORES 101 Multilingual Shared Task, which showed good initial results but was not ultimately able to make a submission due to computing constraints.

This latter project however aided in the work I did under Professor David Mohaisen at the UCF Security and Analytics Lab. I developed a method that improved models' ability to reword and summarize technical security data in laymen's terms. This idea was inspired by multilingual machine translation, where similar languages are trained together so that they (hopefully) improve each other's performance. Similarly, we separated out the identification of technical terms and vulnerabilities from the summarization as distinct subtasks, and trained the model on all of the tasks.

My experience with low-resource scenarios proved to be valuable in my work under Professor Wei Zhang at the University of Central Florida Computational Biology Lab. I worked with another student to develop multi-modal models for cancer sub-type prediction. Certain sub-types however, had fewer training samples than others. I proposed the use of transfer learning, which we utilized in the LoResMT shared task. We thus pre-trained on the higher- resource sub-types before fine-tuning on the lower-resource ones, improving the model's accuracy. I am currently working the problem of integrating structural knowledge from a computational biology perspective in a new project with Dr. Zhang, exploring how to integrate higher-level biological knowledge into our previous models. Our solution was to represent the higher-level knowledge as a graph structure embedded into the model architecture, while the lower-level inputs were used to create the features of the graph.

I believe that integrating linguistic structure into language modeling pipelines is a next step that will help improve all translation models. One direction I am particularly excited to explore is the effect of sub-wording on translation quality, such as if translation quality scales with the similarity between subword units and actual morpheme boundaries.

At John Hopkins University, I am especially interested in working with Professors Philip Koehn, Kevin Duh, Matt Post, and Kenton Murray due to their continued research in machine translation. I have enjoyed following their work, which has allowed me to see John Hopkins University as an excellent fit for my research interests and skills.