Multilingual Word Alignment with Optimal Transport

Joshua Hong, University of California, San Diego
Onizuka Lab, Associate Professor Yuki Arase
Graduate School of Information Sciences and Technology

The task of word alignment, where words in parallel sentences are identified to be translations of each other, is useful for many downstream tasks, including extending limited datasets for low resource languages and domains via annotation projection and general machine translation. While modern neural approaches for multilingual word alignment exist and outperform past statistical baselines, most of these methods take word embeddings generated by large language models and use simple probabilistic methods to extract word alignments. These methods are effective across different language pairs, but possibly struggle with asymmetricity, where a word has no alignment in the other sentence, and cases where multiple words align to the same word. In this study, we explore the viability of optimal transport as a method to extract word alignments from large language models for multilingual word alignment. Optimal transport is the problem of moving mass from one distribution to another while minimizing the total cost of moving the mass. We formulate the task of word alignment as an optimal transport problem by treating each sentence as a probability distribution and the cost of moving mass from one word to another as being some measure of similarity. Under this formulation, optimal transport can be used to find an optimal matching between two sentences, according to the cost between the word pairs. With this general definition, we assess the performance of different variations of optimal transport, including balanced optimal transport and unbalanced optimal transport, as well as different definitions of the cost between two words and the initial probability distributions of the two sentences. We evaluate these variations in both unsupervised and supervised settings and compare the results with state-of-the-art neural aligners.

Our experiments reveal that in the unsupervised setting, optimal transport is competitive against state of the art neural alignment methods, with some variations outperforming compared methods on a wide range of language pairs. However, optimal transport in a supervised setting achieves similar performance to current methods and doesn't provide significant improvements. Analysis on these models and past methods indicate that null alignment, where a word has no corresponding counterpart, and one-to-many alignment, where a word has multiple corresponding counterparts, continue to be challenging across language pairs. In spite of these issues, optimal transport continues to match state of the art benchmarks and contains multiple avenues for additional research, including alternative formulations for sentence distributions and cost functions. Therefore, we find optimal transport to be a promising avenue of improvement for multilingual word alignment. Building off of these experiments, we hope to further investigate why the success of optimal transport in an unsupervised setting doesn't carry over to a supervised setting for multilingual alignment. Additionally, we plan on analyzing current errors and addressing them with other techniques.