Thesis Defense

Computer Science Master's Program

"Contrastive Filtering and Dual-Objective Supervised Learning for Novel Class Discovery in Document-Level Relation Extraction"

By **Nicholas Hansen**

Abstract:

Relation extraction (RE) is a task within natural language processing focused on the classification of relationships between entities in a given text. Primary applications of RE can be seen in various contexts such as knowledge graph construction and question answering systems. Traditional approaches to RE tend towards the prediction of relationships between exactly two entity mentions in small text snippets. However, with the introduction of datasets such as DocRED, research in this niche has progressed into examining RE at the document-level. Document-level relation extraction (DocRE) disrupts conventional approaches as it inherently introduces the possibility of multiple mentions of each unique entity throughout the document along with a significantly higher probability of multiple relationships between entity pairs.

There have been many effective approaches to document-level RE in recent years utilizing various architectures, such as transformers and graph neural networks. However, all of these approaches focus on the classification of a fixed number of known relationships. As a result of the large quantity of possible unique relationships in a given corpus, it is unlikely that all interesting and valuable relationship types are labeled before hand. Furthermore, traditional naive approaches to clustering on unlabeled data to discover novel classes are not effective as a result of the unique problem of large true negative presence. Therefore, in this work we propose a multi-step filter and train approach leveraging the notion of contrastive representation learning to discover novel relationships at the document level. Additionally, we propose the use of an alternative pretrained encoder in an existing DocRE solution architecture to improve performance in base multi-label classification on the DocRED dataset.

To the best of our knowledge, this is the first exploration of novel class discovery applied to the document-level RE task. Based upon our holdout evaluation method, we achieve significantly higher quality cluster propositions compared to the naive clustering approach. We then further enable the retrieval of both novel and known classes at test time provided human labeling of cluster propositions.

Date: Wednesday, June 12th, 2024
Time: 10:00 AM – 12:00 PM
Location: 14-238b
Committee: Dr. Khosmood, Dr. Ventura, Dr. Stanchev