ABSTRACT

This thesis argues that domain-specific language understanding tasks need more effective domain adaptation techniques than simple fine-tuning. Domain adaptation techniques are crucial for achieving superior performance in natural language processing (NLP) tasks. The 'domain shift' issue is addressed, as the training data distribution of open-domain models differs significantly from the downstream datasets of a specific domain. While standard domain adaptation techniques are designed to address such domain shift issues, they prove inadequate when applied to domains such as Customer Support, Scientific Papers, and Law due to limited availability of annotated data and the need to incorporate structured domain knowledge into deep learning models. The thesis addresses this research gap by developing novel in-domain pre-training techniques optimized for domain-specific language understanding across several domain-specific tasks. The first work focuses on developing an automated question answering system for E-Manuals - Technical documents that provide device instructions and specifications but are often lengthy and challenging to navigate. It addresses unique challenges such as answers dispersed across multiple, non-contiguous sections and long-range dependencies arising from hierarchical instructions. Due to limited domain-specific labeled data and the formal language style of manuals, the approach involves pre-training RoBERTa on a large, newly curated corpus of over 300,000 E-Manuals . The second work focuses on enhancing procedural text understanding by introducing novel pre-training methods that exploit the sequential order of steps as a supervision signal, while also proposing a contrastive learning cum masked step modeling framework, **CLMSM**. It addresses the entity tracking task, requiring models to capture how entities' states change across procedural steps, as well as the action alignment task, requiring fine-grained alignment of actions across different procedures. The third and final work focuses on the ***FastDoc*** framework, which replaces traditional local context-based pre-training objectives (MLM, NSP) with document-level supervision using document similarity learning and hierarchical classification tasks, leveraging domain-specific metadata and taxonomy.

Keywords: Domain Adaptation, Natural Language Processing, Incorporation of Domain Knowledge, Efficient Pre-training, Procedural Text, Question Answering

Abhilash Nandy
IIT Kharagpur, India