Summary

Question Answering System for Regulations of University of Information Technology

Question Answering based Intelligent Assistants (or chatbots) have become popular, thanks to progressive achievement of NLP, AI and the blooming of text messaging application. With the fast-growing of the E-commerce market where experience and service quality are becoming important, the problem of traditional services are becoming obvious. The same problem can be seen at University of Information Technology (or UIT), that when a student has a question in mind, he/she must send an email to university staffs and wait for days/weeks for an answer. The student can also search for information on the UIT webpages, but they are hardly organized. What if we have a question answering system (core function of any chatbots) that can read regulations and infers answer for simple questions, leaves hard questions that require human participation for UIT staffs? Question answering systems, powered by deep learning approaches, have surpassed human performance on reading comprehension task, but Vietnamese researches are still underdeveloped (mainly use Information Retrieval approaches in search engines) since there is no large enough Vietnamese dataset for such deep learning model. Based on those observations**, *my hypothesis is that a question answering system using deep learning approaches & in Vietnamese can be built with a limited supply of hand-craft data*.** The work of this thesis is to test this hypothesis by building such a system for solving the UIT problem.

The proposed system consists of 2 main modules: (1) a search engine module is used to retrieved relevant documents from a UIT regulation corpus, and (2) a deep learning module that read each relevant document one by one, and infers the best answer from those documents, returns a short answer to the user.

The search engine module consists of 3 submodules: (1) a document indexing submodule that transform raw text corpus to a representation that is advantageous to search engine, (2) a question processing submodule that transform the question to logical query and (3) a relevance retrieval submodule that rank documents according to their importance based on the logical query. The module is built based on the Extended Boolean model with a modification of using the BM25F weighting scheme for term weight. Whoosh is used as a Python library for implementation.

The deep learning model use fine-tuned BERT, which achieves state-of-the-art accuracy on various NLP downstream tasks, including machine comprehension. BERT stand for Bidirectional Encoder Representation from Transformer, which is the first language model to achieve “deeply bidirectional”, unlike its related works (word2vec, glove, fasttext or even elmo). Such models can’t be trained normally, since word can indirectly see itself, lead to trivial representations, therefore the authors deploy 2 novel tasks (1) Masked Language

Model, that mask 15% of words by [MASK], and the model is asked to predict those words, which aim for word representations, and (2) Next Sentence Prediction, that 50% of the cases provide 2 sentences that are unrelated, and 50% of the cases provide 2 sentences that are related, and the model is asked to predict whether these 2 sentences are followed by each other or not, which aim for context understanding (important for various downstream tasks)

BERT consists of N = 12 layers of Transformer Encoders that are bidirectionally connected to each other (can be thought of a feed-forward network at word level where each cell is a Transformer Encoder). Transformer is a model architecture that relies entirely on attention mechanism to draw global dependencies between input and output.

BERT can be fine-tuned on SQuAD-like machine comprehension task by learning 2 additional vectors, a start vector S and an end vector E. From there, the start/end position of the answer is inferred through softmax.

To fine-tune BERT on Vietnamese task, a straightforward instance-based transfer learning approach is used, that is to translate the SQuAD dataset from English to Vietnamese. But bad translation could harm model performance, therefore the translated dataset is filtered using a weak model trained only on hand-craft dataset (Wikipedia + UIT regulations training sets), resulted in an additional 10% of F1 score, end up with an F1 score of 60% on the Wikipedia and the UIT regulations test sets.

Further analysis of the whole system shows that the system can easily answer factoid questions, and easy questions on lexical/syntactic variations, easy questions require multiple sentences reasoning, but struggle on hard questions on lexical/syntactic variations, questions require world knowledge, and questions with extremely long answer span, which pose a necessity for training data and word2vec substitute (for query expansion).