1. Semantic Parsing and NLP/ML for Automated Programming

Structure-Grounded Pretraining for Text-to-SQL

Accomplishment

In this paper, a novel Structure-Grounded pretraining framework (STRUG) has been proposed for more efficient text-to-SQL by leveraging a large parallel text-table corpus and learning words in both unstructured and structured tabular data. 3 new prediction tasks are identified to obtain text-table alignment knowledge, which are Column Grounding, Value Grounding, and Column-Value Mapping. Moreover, the authors have also built a realistic evaluation set based on Spider to both maintain features of multi-domain datasets and reduce the new query structures and dataset conventions introduced. Results have indicated that STRUG outperforms BERT-large in all the evaluation sets utilized.

Key Approach

For Column Grounding, the key approach used is to treat it as a binary classification task: whether the column header is mentioned in the sentence or not.
Regarding Value Grounding, the key approach used is also to classify the token into 2 categories: whether it is a part of a grounded value.
Concerning Column-Value Mapping, the key approach is to calculate the probability of each token matching each column.

Potential Future Work

The total loss is calculated as the sum of all the 3 losses of the 3 prediction tasks, since losses with different weights did not bring significant improvement. However, these 3 tasks are not independent of each other, it is reasonable to assume that there exists an obscure mathematical relationship among them and their weights can be derived accordingly. Apart from the increased size of pretraining corpus and incorporation with MLM and synthetic data, future works may also focus on the potential relationship among the weights of losses of Column Grounding, Value Grounding, and Column-Value Mapping.

2. Pre-training and Semi-structured Table Representation and Understanding

Leveraging 2-hop Distant Supervision fromTable Entity Pairs for Relation Extraction

Accomplishment

In this paper, a novel REDS2 model for relation extraction which combines the information from 1-hop DS and 2-hop DS. 2-hop is newly introduced here to utilize entity pairs sharing the same relation, defined as anchors, since current RE models tend to depend on the entity pairs with a number of supporting sentences. The framework consists of 4 components: Table-aided Instance Expansion, Sentence Encoding, Hierarchical Bag Aggregation, and Optimization. The results have indicated that this method outperforms many other baseline models. Moreover, 2-hop DS is shown as helpful in 2 aspects: when the information from 1-hop DS is insufficient since it is quite stable, and when entity pairs do not have direct supporting sentences while having anchor entities.

Key Approach

For Table-aided Instance Expansion, the key approach used here is to categorize KB entities into topic entities, subject entities, and body entities, find anchor entity pairs, and then build 2-hop DS bags.
Regarding Sentence Encoding, the key approach used is to concatenate the word embedding and the position embedding of each word, and then feed it into the encoding layer, such as CNN and PCNN to obtain the fixed-size sentence representation.
In Hierarchical Bag Aggregation, the key approach is to assign different sentence weights given a certain relation with a selective attention, and use calculated attention and the query vector of the relation to generate the final representation to solve the problem of noise and imbalanced size.
In Optimization, the key approach used is the cross entropy loss and the stochastic gradient descent(SGD).

Potential Future Work

The information obtained by 2-hop DS is redundant, as shown in Table 5, which suggests that a more representative set can be selected to reduce time complexity or improve the model performance. This may be achieved by better sampling or bag-level aggregation methods. This paper also advises that headers and column names in the tables can also be incorporated.

3. Knowledge Representation and Reasoning in (textual) graphs, with Emphasis on Interpretability

Rationalizing Medical Relation Prediction from Corpus-level Statistics

Accomplishment

In this paper, a new framework simulating human memory's cognitive process, which is based on the corpus-level statistics, a global co-occurrence graph of a clinical text corpus able to preserve patients' privacy well, is proposed for relation prediction between entities. This framework consists of 3 stages, which are Global Association Recall, mimicking the recall process, Assumption Formation and Representation, mimicking the recognition process, and Prediction Decision Making, determining the final prediction. This model takes an advantage over other neural networks in that it both achieves satisfactory performance and rationalizes its prediction, which is quite helpful to acquire experts' trust in this framework.

Key Approach

For Global Association Recall, the key approach used is to calculate the conditional probability of an entity associated with another entity and to optimize this conditional probability to be close to the defined empirical distribution.
Regarding Assumption Formation and Representation, the key approach used is to measure the probability of relations holding for a pair of head and tail entities, formulate OWA rationales by calculating the conditional probability of a relation with a pair of associations, and combine entity vectors and relation vector to obtain the representation for assumptions for a association pair.
Concerning Prediction Decision Making, the key approach used here is to aggregate all assumption representations and estimate their accountability for the final prediction. The assumption with the highest probability is selected as the rationale. A negative log likelihood loss function is adopted to improve efficiency rather than using the margin based one.

Potential Future Work

According to Table 2, which shows the comparison results of model predictive performance, all other neural baseline models have the lowest F1 scores on the task of MAY_PREVENT, and in most of these models, the second lowest F1 scores come from the task of CAUSES. It is then logical to assume that MAY_PREVENT and CAUSES are 2 special relations in the clinical corpus, which requires extra handling in dataset preparation or training process. For example, in this framework, since the CAUSE relation is sometimes difficult to derive even in reality for experts, maybe more associations or stronger associations should be requested for predicting this relation.

4. Question Answering and Reading Comprehension, with applications to the clinical domain

CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering

Accomplishment

In this paper, a simple while effective framework CliniQG4QA is proposed to improve question answering training in terms of generalization, with the use of question generation to construct new QA pairs. In the process of QG, a new phase Question Phrase Prediction(QPP) is introduced to generate diverse questions, which consist of more types of questions. This framework consists of 3 stages, which are Answer Evidence Extractor(AEE), Question Phrase Prediction, and Training. Experimental results have shown that the QPP module is able to better both the generation relevance and diversity of base QG models, and CliniQG4QA does perform well on new context different from the training set.

Key Approach

For Answer Evidence Extractor, the key approach used here is to extract potential evidence sentences (long text spans) with BIO sequence labeling. Moreover, some heuristic rules are developed and then applied to the classification results to enhance the quality of the extracted spans.
Regarding Question Phrase Prediction, the key approach used is to build a vocabulary of all available question phrases of a fixed length in the training corpus, and then predict whether a phrase in the vocabulary can be a evidence through sequence prediction following a predefined order with an attention based seq2seq model, rather than through multi-label classification. During this stage, the number of question phrases for each answer evidence can be decided dynamically, increasing the credibility of diverse generation.
Concerning Training, the key approach used is to train AEE, QPP module, and QG model by minimizing the negative log-likelihood loss respectively following this order. Based on AEE, QPP is able to be used in QG, and then the QA model can be trained.

Potential Future Work

The heuristic rule *merge-and-drop* applied in AEE, may have some potential issues since it is not always the case that 2 spans close to each other should be recognized and merged as one. A more suitable rule can be developed. In addition, more advanced QG and QA models can be experimented to further test the performance of CliniQG4QA.