Projeto de Formatura – 2023

PCS - Departamento de Engenharia de Computação e Sistemas Digitais

Engenharia Elétrica – Ênfase Computação

Tema:

Aprimoramento de Question Answering por Texto em Linguagem Natural para SQL

Objetivos

Modelos de linguagem enfrentam dificuldades para interpretar algumas entradas de usuários, por não capturarem corretamente as interações entre os termos da frase. Assim, por não terem compreensão referente à real intenção do usuário, geram saídas que não são correspondentes à frase de entrada. Esses problemas são vistos em tarefas particularmente sensíveis, por exemplo a tradução. No modelo de tradução de linguagem natural (Português) para SQL, por exemplo, se pedirmos “Quais carros foram fabricados no Brasil e pesam menos de 1500 kg?”, a saída gerada pelo modelo não traduz corretamente a relação aditiva expressa pela conjunção “e” para o operador SQL “AND”, mas interpreta como um OR. Esse trabalho investiga a apresentação das informações sobre as interações entre os termos da frase ao modelo de linguagem na entrada, para que tenha uma interpretação mais precisa sobre a intenção do usuário, na tarefa de tradução de linguagem natural para SQL.

Métodos e Procedimentos

As informações apresentadas ao modelo sobre a frase são de dois tipos: sintática e semântica. A informação sintática refere-se às funções sintáticas de cada termo e relações entre essas funções. Portanto, é o tipo de informação que revela o sujeito da frase, o verbo raiz, objeto direto, objeto indireto, por exemplo. Para informar tais relações, recorremos a um parser de dependências, o Spacy, que possui alta precisão. O Spacy gera as relações sintáticas para todos os termos na frase, num formato tabular, o que requer linearização para usar no refinamento do modelo na entrada. Isso é feito separando as linhas de cada termo, que têm as informações sintáticas relativas a ele, por um separador (“[row]”).

A informação semântica é a representação apenas dos termos que carregam o sentido da frase (responsáveis pelo núcleo semântico da frase), as relações entre eles e natureza delas. Assim, termos como determinantes são excluídos desse tipo de representação. A informação semântica é codificada por meio da *Abstract Meaning Representation* (AMR), por esta não ter vínculo com representações sintáticas; ou seja, é puramente semântica. Isso também nos permite avaliar qual informação é mais relevante para o modelo. A árvore AMR da frase é gerada por um parser implementado por uma biblioteca em Python, amrlib. Por essa implementação, uma rede neural Bart gera o correspondente AMR da frase do usuário na forma de uma árvore, assim também precisamos linearizar a saída para fornecer à nossa rede. Porém, nesse caso, a árvore tem um formato no qual os nós são separados por parênteses, o que nos facilita por servir de separador natural para a nossa rede neural, assim apenas removemos os espaços embutidos entre os termos da árvore de saída.

Assim que tenhamos as árvores sintática ou semântica linearizadas, concatenamos essas representações às correspondentes perguntas que as originaram. Essa estrutura é apresentada na entrada da rede neural no treinamento para a tradução de linguagem natural para SQL, como visto na figura 1.

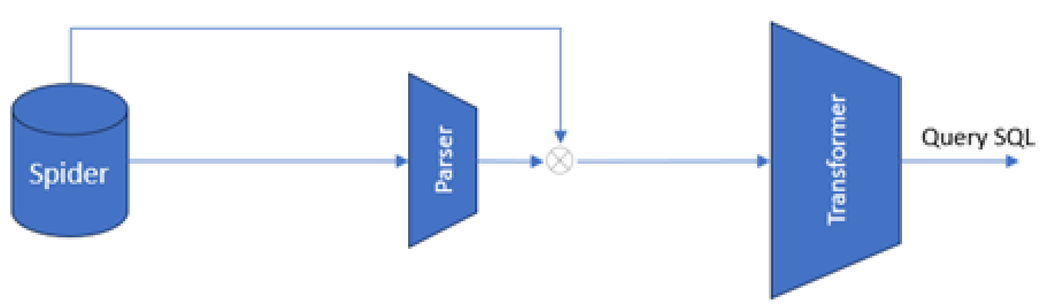

Fig.1: apresentação das informações sintáticas ou semânticas à rede no treinamento. O parser gera a representação, sintática ou semântica, sobre a pergunta da base de dados Spider, e concatenamos essa informação à pergunta. O resultado é apresentado à rede.

Na tarefa de tradução de linguagem natural para SQL, o benchmark para os modelos é a base de dados Spider. Nessa base de dados, temos perguntas referentes a diferentes níveis de consultas SQL e para várias tabelas. Treinamos um modelo de controle, usando apenas a base Spider com as perguntas traduzidas para Português, e dois modelos, um com as perguntas da base Spider traduzidas para Português e as informações sintáticas delas, e um com as perguntas da base Spider traduzidas para Português com as informações semânticas delas.

As métricas de avaliação são exact-set-match, que compara o quão próxima é a consulta gerada da consulta de referência (ouro), e execution accuray, que verifica se o resultado de execução da consulta gerada é o mesmo que o valor da execução da consulta de referência. Pela soma das duas, temos a pontuação.

Resultados

Os resultados apresentados na tabela são para o treinamento dos três modelos (sem informação, com sintaxe, com semântica), para modelo de linguagem T5. Os modelos foram treinados por 32 épocas. Na tabela, apresentamos as melhores pontuações para cada modelo treinado (independente do checkpoint a que pertence a pontuação):

Modelo	Exact-set-match	Execution accuracy
T5 sem informação	0.5502	0.5754
T5 com informação sintática	0.5909	0.6247
T5 com AMR	0.6392	0.6818

A tabela revela que os modelos treinados com informações sobre a frase tiveram pontuações importantes, principalmente quando comparamos ao controle.

Conclusão

Os resultados para tradução mostram que a inserção de informação, sintática ou semântica, é relevante para o modelo na tarefa de tradução. Além disso, a informação semântica é mais representativa para o modelo.

Integrantes: - Anton Bulle Labate

Professor(a) Orientador(a): Prof. Dr. Fabio G. Cozman

O trabalho foi conduzido no Centro de Inteligência Artificial USP/IBM/FAPESP (C4AI), com suporte da FAPESP e da IBM Corporation (2023).