Adversarial attack to Semantic Parser

New Challenge to Attack Semantic Parser:
- The input is short, change of input is clearly distinguishable textually
- The input space is discrete, Semantic parser: “Natural Language” Logical Form

A new definition of adversarial example:
- Semantic(x) ≠ Semantic(x’)
- Semantic(Model(x)) ≠ Semantic(Model(x’))

Long input Continuous input Short discrete input Semantic task

What is the name of the loser when the winner was new england patriots…?

NLP2SQL Model

SELECT loser WHERE winner cond_op new england patriots...

Same NLP2SQL Model

SELECT winner WHERE winner cond_op new england patriots...

Generating Adversarial Examples

Basic Method: Fast Gradient Method

Algorithm:
1. \( f (\text{grad}) = \text{input} \times \text{grad} \)
2. \( \text{target} = \text{data} + \epsilon \times \text{grad} \)
3. \( \text{perturbed word} = \text{arg min}_{\text{word}} [f(\text{perturbed word}) - f(\text{original word})] \)

Experiment Result:

- The larger the \( k \), the higher diff ratio and lower valid ratio it will be when \( k \) is relatively small.
- Some pattern is shown in the successful perturbed examples:
  1. Among all the successful examples, 36% is done by changing a word in single form to plural form.

What is the air force cross when \( x \) is a small real number?
\[ \text{airforcecross} \]

What is height when \( x \) is a real number?
\[ \text{height} \]

Under fitting problem: some words are crowded in a small area, the word untrained is easily been misguided by the trained words around it.

New adversarial feature for NLP2SQL model: the header of SQL usually are the same type and sometimes very close to each other, the header can be vulnerable under adversarial attack.

Reason: Under-fitting problem in NLP2SQL task

The distribution of \( y \) (see below) on a plane in the high dimensional space

A more semantic consistency is shown after substitution

Improvement Using Bert

Cosine similarity is a more reasonable choice:
\[ \cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|} \]

loss(d2) > loss(d1)

cos_similarity describe the degree of following the gradient better since Bert ensured the small distance already

Unreasonable result occurs if using norm distance

Bert-FGM-norm

Bert-FGM-cos