Improving the User Experience on MISP

Based on paper by Yao et al. (2019)

Presented by -
Noah LaPolt
Shlok Sah

Contents

- Problem

- Introduction

- UI

- Question Paraphraser

- Multilingual Translator

- Experiments & Results

- Summary

- Product in Action

- Future Works

- References

Problem

- The current question base is limited which makes it more like talking to a robot

- There is absence of a GUI and currently displays semantic information which may not be required by the user

- The business implication of current SQL Parser is lower as it only works for English speaking community.

- Building a multilingual semantic parser is a resource expensive task

Introduction

- Separate module from current framework

- Presented in web based gui

- Accessibility (vision)

- Utilizing T5 transformer for question parser

- Highly portable

- Support for more than 30 languages

- Separation of concern

  - Design does not depend on sql model, parser or translator

- Scaleable

- Theoretical simplicity

UI-Framework

- Flexx library for a web based interface, with python
- Serves web app to domain
- Supports multiple clients
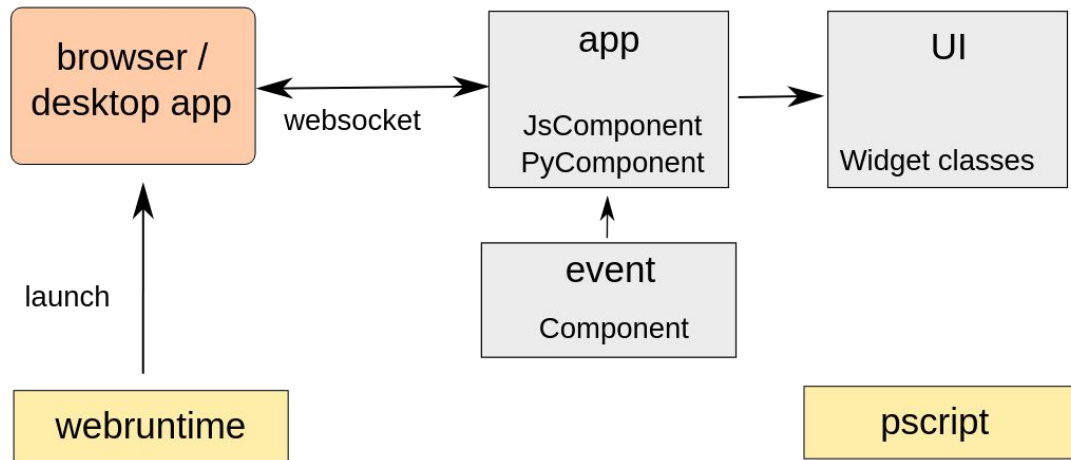
UI-design

- Chat box/Menu
- Google Icons
- Data display
- Accessibility
  - Color blindness
  - Language Support
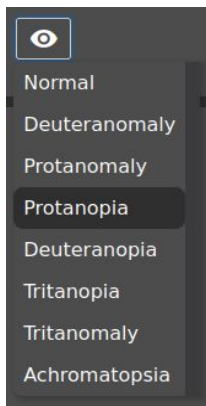
Deuteranomaly
92% SIMILAR
5% of men, 0.35% of women

Tritanopia
57% SIMILAR
0.001% of men, 0.03% of women

Question Paraphrasing using T5

- T5: Text-to-Text Transfer Transformer created by Google to develop general knowledge to "understand" text

- Reframes NLP tasks into a unified text-to-text-format

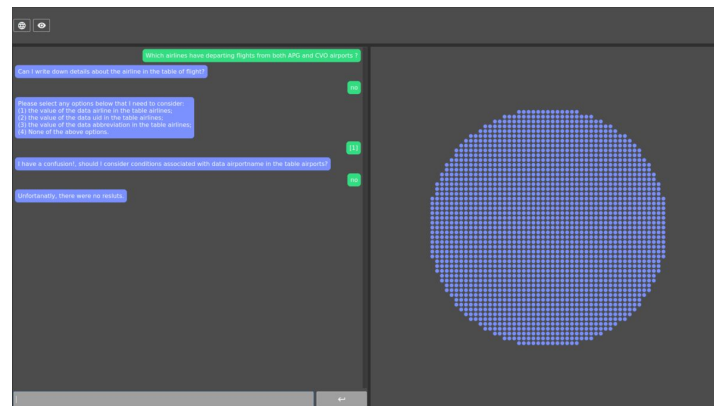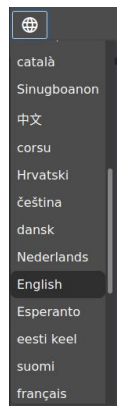- Pre-trained on unlabelled C4 (Colossal Clean Crawled Corpus). ~ 700 GB in size and is cleaned version of Common Crawl dataset.

- Pre-trained using fill-in-the blanks style denoising objective (recover missing words) but computationally expensive

- Difference with BERT is that T5 replaces multiple consecutive tokens with a single Mask keyword, unlike, BERT that uses Mask token for each word.

T5

Original text
Thank you for inviting me to your party last week.

Inputs
Thank you <X> me to your party <Y> week.

Targets
<X> for inviting <Y> last <Z>
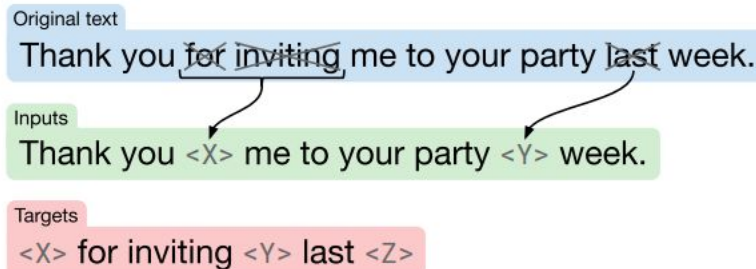
T5's mask language modeling (Raffel et al., 2019)

Quora Question Pairs (QQP)

- Released by Quora
- Originally developed to identify duplicates
- We used only the duplicate entries (paraphrases of original questions)

| id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

- T5 with QQP is implemented in hugging face transformer
  https://huggingface.co/ramsrigouthamg/t5_paraphraser

Training pairs = 134,333
Test pairs = 14,928

PAWS – Wiki

- Existing paraphrase identification datasets lack sentence pairs that have high lexical overlap without being paraphrases.
- Challenging pairs are generated by controlled word swapping, back translation and paraphrase judgments by human raters
- PAWS dataset gives improved generalisability
- This corpus contains pairs generated from Wikipedia pages

	Sentence 1	Sentence 2	Label
(1)	Although interchangeable, the body pieces on the 2 cars are not similar.	Although similar, the body parts are not interchangeable on the 2 cars.	0
(2)	Katz was born in Sweden in 1947 and moved to New York City at the age of 1.	Katz was born in 1947 in Sweden and moved to New York at the age of one.	1
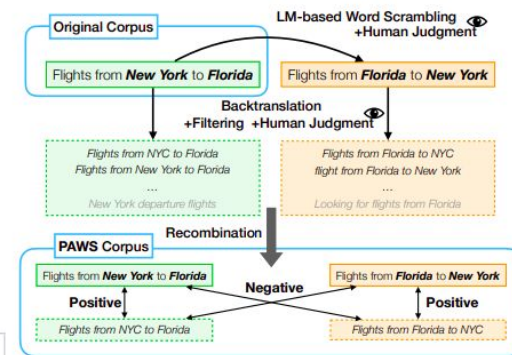
Training pairs = 49,822
Test pairs = 720

- We performed Cross database fine tuning using both QQP and PAWS-Wiki

Experiments

- T5 allows us to use same model, loss functions, hyperparameters across a diverse set of tasks
- Fine tuned a hugging face T5 transformer with different combination of datasets and optimizers
- Task-specific - "paraphrasing" prefix
- Batch Size: 6
- Length: 512
- Learning rate: 0.001
- Epochs = 2
- Training Time = 10 hours

Resource used = OSC Owens supercomputer with GPU acceleration

Usage for 30 days, ending: **02-DEC-21**

Core Hours per User

Results
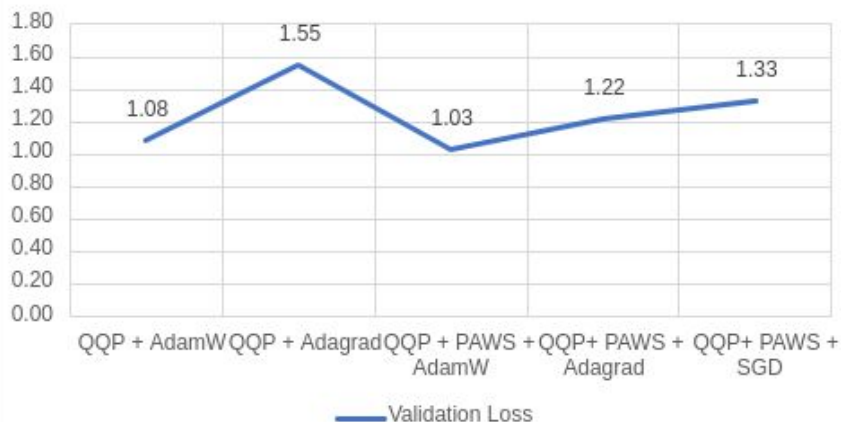
Original Question: Does the system need to return any information about city?

Paraphrased Questions:

Quora dataset + AdamW : Will our reservation system return any city information?

Quora dataset + PAWS+ AdamW: Do the system need to return any information about city?

Multilingual Translator

- **Deep_translator -** developed by Nidhal Baccouri
- Free tool to translate between different languages
- High level of abstraction
- Automatic language detection
- Easy to use and extend
- Supports many popular translators like Google translator with support upto 30+ languages

Usage

- Using it to allow users to interact with the Chat Bot
- Follow up questions are generated in User's language
- Easier to work with English parser like Edit SQL without having to train it on different languages.

Summary

MISP Agent

World Model

State

Error Detector

Action

Actuator

Environment

Percept

User

Intent
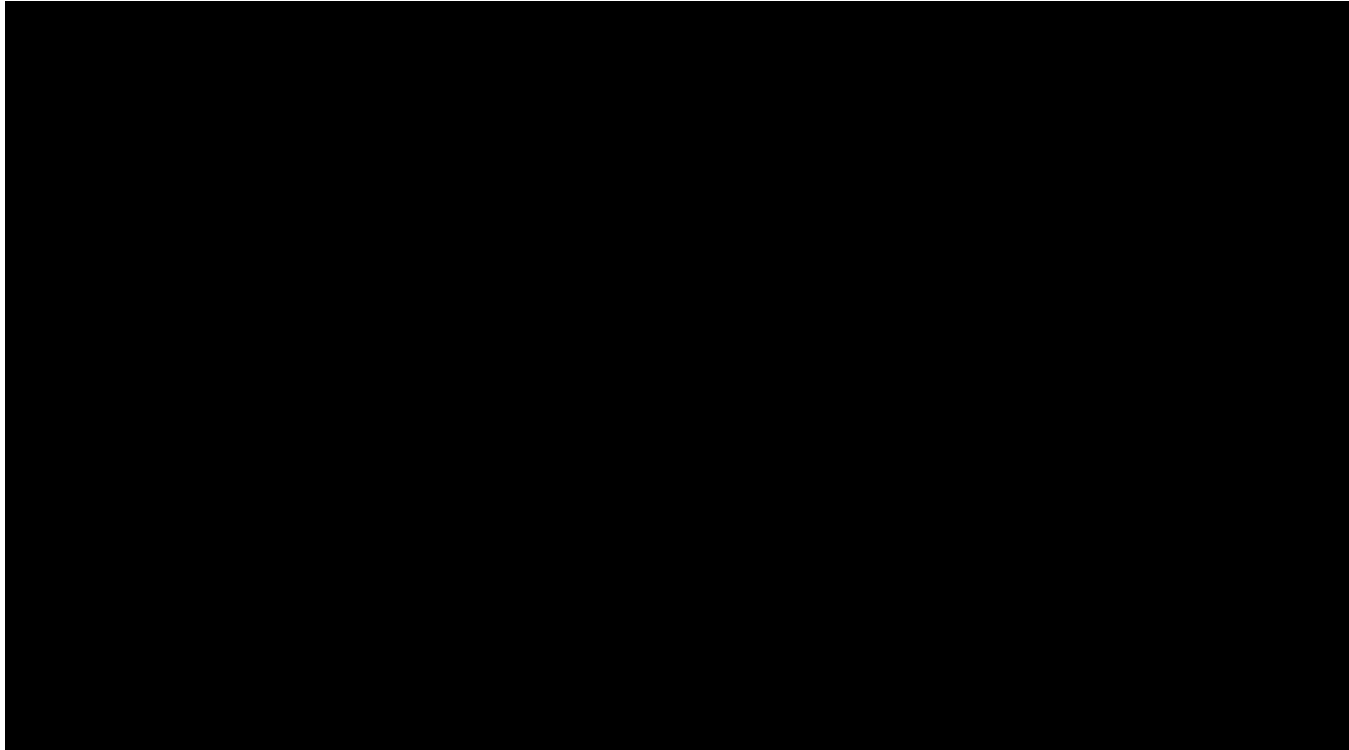
Question

Multilingual Translator

Question Paraphraser

ChatBot Interface

Real User
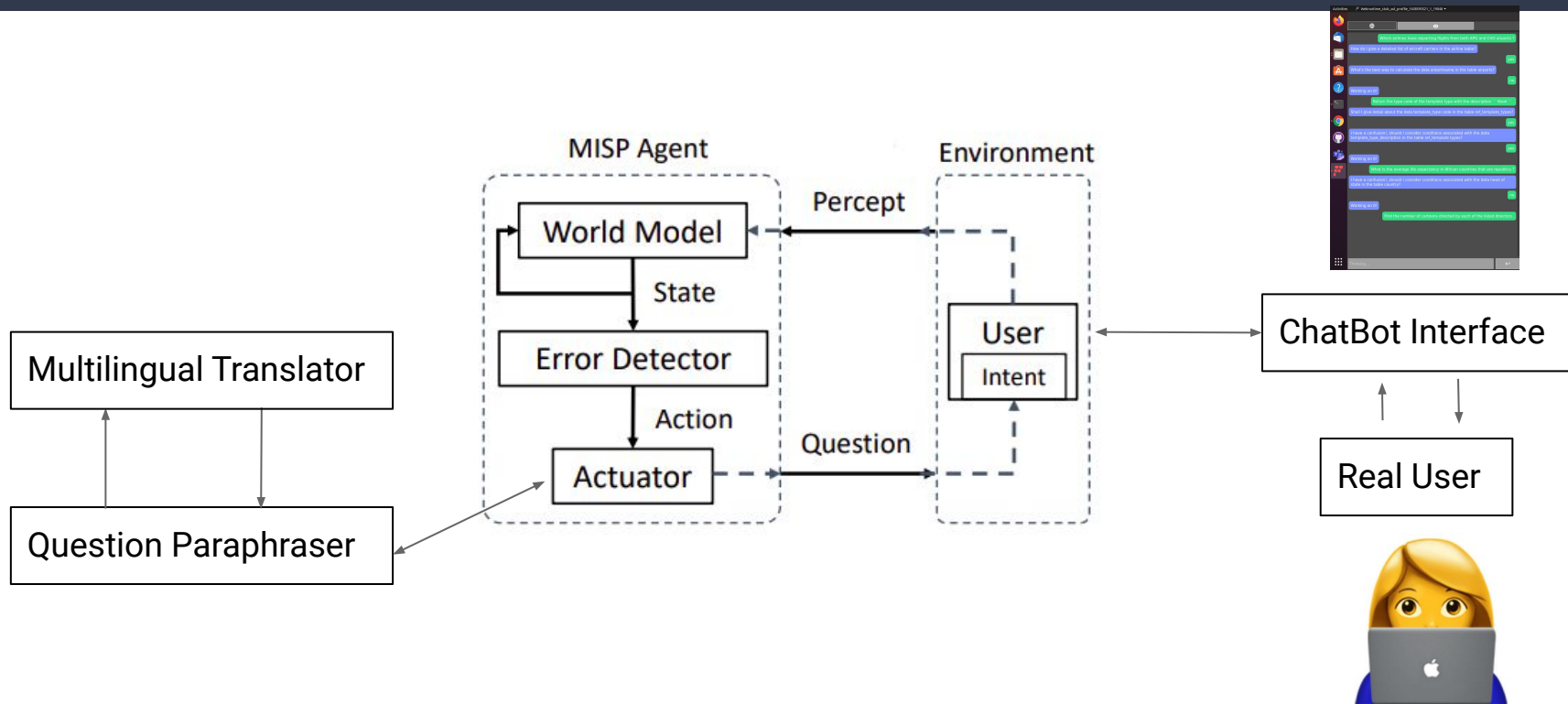
Product in Action

Future Work

- Using labels provided in the QQP and PAWS dataset

- Gradual unfreezing

- Implement our own translator using Fine Tuning task on T5 (but excessive resource consumption for multilingual translator)

- Using better transformers like Pegasus

- Same technique can be applied to better Semantic parsers like SQLova

References:

- Yao, Z., Tang, Y., Yih, W.-tau, Sun, H., & Su, Y. (2020). An imitation game for learning semantic parsers from user interaction. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.559
- *Exploring transfer learning with T5: The text-to-text transfer transformer*. Google AI Blog. (2020, February 24). Retrieved December 7, 2021, from https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html.
- *First Quora dataset release: Question pairs - data @ quora*. (n.d.). Retrieved December 7, 2021, from https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.
- Google-Research-Datasets. (n.d.). *Google-Research-datasets/PAWS*. GitHub. Retrieved December 7, 2021, from https://github.com/google-research-datasets/paws.
- Google. (n.d.). *Google colaboratory*. Google Colab. Retrieved December 7, 2021, from https://colab.research.google.com/drive/176NSaYjc2eeI-78oLH_F9-YV3po3qQQO?usp=sharing#scrollTo=SDVQ04fGRb1v.
- Nidhaloff. (n.d.). *Nidhaloff/deep-translator*. GitHub. Retrieved December 7, 2021, from https://github.com/nidhaloff/deep-translator.
- *RAMSRIGOUTHAMG/T5_paraphraser · hugging face*. ramsrigouthamg/t5_paraphraser · Hugging Face. (n.d.). Retrieved December 7, 2021, from https://huggingface.co/ramsrigouthamg/t5_paraphraser.
- Bianchi, F. (n.d.). *Coolors - the super fast color schemes generator!* Coolors.co. Retrieved December 5, 2021, from https://coolors.co/

Thank You!

Questions?