TAPHSIR: Towards AnaPHoric ambiguity detection and reSolution In Requirements

About TAPHSIR

TAPHSIR is a tool associated with the ICSE'22 technical paper titled "Automated Handling of Anaphoric Ambiguity in Requirements: A Multi-solution Study". TAPHSIR provides a hybrid solution composed of two components. The first component employs machine learning (ML) for detecting anaphoric ambiguity. The second component uses a pre-trained language model, SpanBERT, for anaphora resolution.

The tool was developed at SnT / University of Luxembourg with funding from Luxembourg's National Research Fund (FNR).

What is released?

- **./artifact/**: is the folder that contains our ML and SpanBERT trained models and some python scripts that enable running TAPHSIR, e.g., text processing methods.
- **./example.txt**: is a txt file that contains an example from our technical paper. This example is used to illustrate how TAPHSIR can be used (see Usage Example below).
- **./taphsir.py**: is the main python file that contains the scripts to run TAPHSIR.
- **./output/**: is the folder that will contain the results produced by TAPHSIR. Output excel files will be generated and placed in this folder depending on whether one runs the detection compontent, resolution component, or both.
- **./requirements.txt**: is the file containing the required libraries needed to run TAPHSIR.

How to use TAPHSIR?

TAPHSIR is implemented in Python 3.8. Youn can run TAPHSIR using the following steps.

Getting the Project

- Clone the GitLab repository *TAPHSIR* to your local machine, for example using the following command:

```
git clone https://gitlab.uni.lu/sabualhaija/taphsir
```

- Navigate to the project main folder on your local machine

```
cd path/to/taphsir/
```

- Follow the instructions in the Installation File for setting up the environment for running TAPHSIR.

Usage Example

Once you set up your environment, you can run TAPHSIR using the following command:

```
python taphsir.py --doc path-to-doc --mode mode --detection model
```

where:

- *mode* is a parameter indicating what component from TAPHSIR you want to run. The value of this parameter is selected from three options: **1** (to run the detection component only), **2** (to run the resolution component only), and **3** (to run both components). Note that **3** is the default value.

- *model* is a parameter indicating which ML model one wants to apply for anaphoric ambiguity detection. The value of this parameter can be: **LF** (to use ML model pre-trained on language features only), **FE** (to use the ML model that is pre-trained on feature embeddings only), or **Ensemble** (to use ensemble ML pre-trained on both sets of features). More details on these models can be found in our technical paper.

For running TAPHSIR with the default parameters on **Example.txt**, use the following command (assuming that you have already navigated to the main folder of TAPHSIR on your local machine):

```
python taphsir.py --doc Example.txt
```

Output of TAPHSIR

Once the execution is completed, one or two excel files resulting from running the components of TAPHSIR according to the *mode* parameter will be generated and placed in the folder **./output/**. In the case of running the tool with the default parameters on **Example.txt**, two output files, **detection.xlsx** and **resolution.xlsx**, are generated.

The **detection.xlsx** file contains the following columns:

Column	Content
Id	A unique identifier for each pronoun in the context.
Context	The context where a pronoun is occurring.
Pronoun	The pronoun for which the analysis has been performed.
Detected As	The detection result of TAPHSIR for that pronoun ambiguous or unambiguous.

The **detection.xlsx** file contains the following columns:

Column	Content
Id	A unique identifier for each pronoun in the context.
Context	The context where a pronoun is occurring.
Pronoun	The pronoun for which the analysis has been performed.
Resolved As	The resolution result of TAPHSIR for that pronoun indicating to which antecdent this pronoun refers.

*Note: in case of multiple pronouns in the same context, the output files will contain one row per pronoun. *

How to cite?

If you wish to use or compare with TAPHSIR, please cite the following paper:
Ezzini, S., Abualhaija, S., Arora, C., Sabetzadeh, M. (2022, May). Automated Handling of Anaphoric Ambiguity in Requirements: A Multi-solution Study. In 44th International Conference on Software Engineering (ICSE).