AI and Life Science

Molecule Retrosynthesis

Muhammad Ammaduddin Qazi
K12343657

JOHANNES KEPLER UNIVERSITY LINZ

Predicting Reactants from Products using UALIGN Model

Objective:

To predict reactants from given product molecules using retrosynthesis analysis.

Challenge:

Utilizing machine learning models for accurate prediction in chemical synthesis.

Approach:

Implementing a pre-trained model from the UALIGN GitHub repository.

Link: UALIGN GitHub Repository

Dataset and Model

Dataset: USPTO-FULL

- Comprehensive database for chemical reactions.
- Used for training and testing the model.

Model: UALIGN pre-trained model

- Tokenization: Utilized tokenizer specific to USPTO-FULL dataset.
- Graph-based Neural Network: GATBase with TransformerDecoder.
- Modifications: Custom changes in graphs.util and inference_one.py

Modifications to Graph Utility

Objective:

- Convert SMILES strings to graph objects for model processing.
- Used by the model to predict the next character.

Key Modifications:

- Handling atom mapping and feature extraction.
- Adjusting edge and node features for better graph representation.

Inference Script Adjustments

Objective:

- Fine-tune the model inference process for retrosynthesis prediction.
- Optimizer: Adam with a learning rate of 0.001 for efficient gradient descent.

Key Modifications

- Set parameters for model dimensions, layers, and device handling.
- Manual adjustments for beam search and data processing.

```
args = argparse.Namespace(
    dim=768,
    n_layer=8,
    heads=12,
    negative_slope=0.2,
    seed=2023,
    device=0,
    checkpoint='model.pth',
    token_ckpt='token.pkl',
    use_class=False,
    max_len=100,
    beams=1,
    product_smiles='product_smiles_test.csv',
    input_class=-1,
    org_output=False,
    output_file='results.csv'
)
```

Results and Conclusion

Execution

- Ran the inference on a test set using Jupyter Notebook.
- Parameters manually set in the inference_one.py script.

Results

- Successfully predicted reactants for the provided product molecules.
- Results saved in results.csv

Score

23	StartedFromTheBottomNowWeHere	0.691	7	4 days ago

Experimentation

Custom Seq2Seq LSTM Model

- Developed a custom sequence-to-sequence model using LSTM
- Result:
  The model did not achieve satisfactory accuracy for retrosynthesis prediction.

Pre-trained BART Model from Hugging Face

- Utilized a pre-trained BART (Bidirectional and Auto-Regressive Transformers) model.
- Result:
  Similar to the custom LSTM model, the BART model did not yield good results in predicting reactants from product molecules.