Homonyms Problem in Text

Table of Contents:

Introduction:

The motivation for the third project, focused on Homonyms, arises from the persistent challenges encountered in natural language processing (NLP), particularly with respect to nuanced homonym interpretation in English. Homonyms, words with shared spellings or pronunciations but distinct meanings, introduce ambiguity in language processing tasks. This complexity is compounded by the intricacies of the English language, posing obstacles to accurate homonym disambiguation.

Project Overview and Goals:

To initiate the exploration, I first assessed whether traditional sentiment analyzers, such as those from nltk, could effectively address the homonym challenge. Subsequently, the overarching goal of this project is to develop a homonym-aware model for English by leveraging transformer-based techniques.

The approach involves:

1. Exploring Homonym-Aware Transformers:

   - Investigate the effectiveness of pre-trained transformers in handling homonyms in English.

   - Assess the transformer's ability to capture context and distinguish between homonymous meanings.

2. Fine-Tuning on Sentiment Analysis Data:

   - Utilize sentiment analysis data to fine-tune the transformer model.

   - Aim to enhance the model's homonym-awareness by exposing it to sentiment-related nuances and context in English.

By embarking on this strategic exploration, the project endeavors to address the challenge of homonyms in English. The ultimate aim is to develop a transformer-based model that not only excels in homonym disambiguation but also demonstrates proficiency in sentiment analysis, thereby advancing the model's adaptability in diverse linguistic contexts.

Data Description:

The SST-2 dataset from the GLUE benchmark is organized into training, development (dev), and test sets, each with distinct sizes and compositions:

Columns:

- Sentence: Textual content of the sentences.

- Label: Sentiment label associated with each sentence.

- Idx: Index or identifier for each instance.

Dataset Characteristics:

- **Training Set**: 67,349 instances, 3 columns (sentence, label, idx)

- **Dev Set**: 872 instances, 3 columns (sentence, label, idx)

- **Test Set**: 1,821 instances, 3 columns (sentence, label, idx)

The training set constitutes the majority of the dataset, followed by the test set and the dev set. The **sentence** and **label** columns are of primary interest for sentiment analysis and model development, guiding our exploration and training efforts. The idx column provides unique identifiers for each instance in the dataset.

Baseline Experiment: Traditional Sentiment Analysis for Homonym Awareness

Goal:

Establish a baseline for Homonym-Aware Sentiment Analysis using a traditional sentiment analyzer, specifically the SentimentIntensityAnalyzer from the NLTK library. Evaluate its effectiveness in handling homonyms in English sentences.

Steps:

1. Sentiment Analyzer Initialization:

- Initialize Sentiment Analyzer: Implement the SentimentIntensityAnalyzer from NLTK for sentiment analysis.

2. Inference on Homonyms:

- Perform model inference on sample sentences containing homonyms in English.

3. Results:

- Conclude that the traditional sentiment analyzer, specifically the SentimentIntensityAnalyzer from NLTK, is not Homonym-Aware based.

```
Sentence: I hate the selfishness in you
Predicted Sentiment: negative
==============================
Sentence: I hate anyone who can hurt you
Predicted Sentiment: negative
==============================
Sentence: I will despise anyone despises you
Predicted Sentiment: negative
==============================
```

Conclusion:

The baseline experiment, utilizing the traditional SentimentIntensityAnalyzer from NLTK, aimed to assess its efficacy as a Homonym-Aware Sentiment Analyzer for English sentences. Upon evaluation of sentences containing homonyms, a notable observation emerged: the traditional sentiment analyzer lacks homonym-awareness. This finding serves as a pivotal baseline, clearly indicating that the analyzed model is not homonym-aware. The results underscore the necessity for exploring advanced techniques, particularly transformer-based models, to elevate the capabilities of sentiment analysis in the context of sentences with homonyms.

Experiment: Evaluating DistilBERT Base Uncased for Homonym-Aware Sentiment Analysis

Goal:

Investigate the capability of DistilBERT Base Uncased, a transformer-based model, in enhancing Homonym-Aware Sentiment Analysis for English sentences using the Hugging Face `pipeline` for simplicity and efficiency.

Steps:

1. Model Initialization:

- Load DistilBERT Base Uncased: Utilize the pre-trained DistilBERT model for English language understanding via the Hugging Face `pipeline`.

2. Sentiment Prediction on Homonyms:

- Use `pipeline` for Inference: Employ the Hugging Face `pipeline` functionality for tokenization and sentiment prediction on sentences with homonyms.

3. Results:

- The evaluation of DistilBERT Base Uncased, conducted using the Hugging Face pipeline for Homonym-Aware Sentiment Analysis, reveals that the model faces challenges in accurately capturing homonyms within the text. The model's performance indicates limitations in distinguishing between homonymous meanings, suggesting the need for further refinement.

```
Sentence: I hate the selfishness in you
Predicted Sentiment: negative
==============================
Sentence: I hate anyone who can hurt you
Predicted Sentiment: negative
==============================
Sentence: I will despise anyone despises you
Predicted Sentiment: negative
==============================
```

Comparison with Baseline:

Both the traditional SentimentIntensityAnalyzer from NLTK and DistilBERT Base Uncased exhibited limitations in capturing homonyms within the text. The comparison highlights the challenges faced by both approaches in accurately discerning between homonymous meanings.

Conclusion:

Summarizing the findings, the evaluation of DistilBERT Base Uncased, facilitated by the Hugging Face pipeline for Homonym-Aware Sentiment Analysis, indicates that the model encounters challenges in effectively handling homonyms within text. The observed limitations underscore the necessity for further refinement.

Experiment 3: Fine-Tuning DistilBERT Base Uncased on SST-2 Dataset

Goal:

Fine-tuning DistilBERT on the SST-2 sentiment analysis dataset was chosen for its efficiency, leveraging the model's lightweight nature to achieve faster training compared to the original BERT. This strategic choice aimed to capture contextual embeddings tailored to sentiment-related nuances while optimizing computational resources for expeditious experimentation and model refinement.Steps

Steps:

1. Tokenizer Modifications:

- Attempted POS Tagging and Dependency Parsing: Concatenated these features with each token before inputting them to DistilBERT tokenizer, but limited improvement observed, as DistilBERT relies on attention mechanisms for contextual embeddings

- Experimented with Word Sense Disambiguation: Investigated the impact of adding WSD information to tokens. Minimal impact, suggesting that the standard tokenizer proved to be the most effective choice for DistilBERT.

2. Training Configuration:

- Training Duration: 3 epochs

- Batch Size: 128

3. Optimization Strategy:

- Optimizer: AdaW

- Advantages of AdaW:

  - Adaptability to Parameters
  - Robustness to Sparse Gradients
  - Effective Exploration-Exploitation Balance

Results and Observations:

The fine-tuned DistilBERT model exhibited notable improvements in handling homonyms within sentences, demonstrating its enhanced homonym-aware sentiment analysis capabilities. The evaluation on a diverse set of sentences revealed nuanced outcomes:

1. Success on Specific Sentences:

   - The fine-tuned model demonstrated success in accurately capturing homonymous meanings in certain sentences. This success indicates the model's improved ability to discern context and disambiguate meanings effectively.

2. Challenges on Other Sentences:

- Despite the improvements, the fine-tuned model faced challenges in handling homonyms within certain sentences. This underscores the inherent complexities and nuances associated with homonym-aware sentiment analysis, highlighting areas for potential refinement.

```
Sentence: I hate the selfishness in you
Predicted Sentiment: negative
==============================
Sentence: I hate any one can hurt you
Predicted Sentiment: positive
==============================
Sentence: I will despise anyone despises you
Predicted Sentiment: positive
==============================
Sentence: I hate any one who can hurt you
Predicted Sentiment: negative
==============================
```

- Experimenting with Sentence Splitting:

In our efforts to enhance the fine-tuned DistilBERT model for homonym-aware sentiment analysis, we experimented with splitting sentences into two halves. This strategy intended to isolate potential homonyms by providing clearer contextual cues. We adapted our approach to address a specific case: if a split sentence exhibits two distinct sentiments, it likely contains homonyms. This insight allows us to focus on these instances, tackling homonym challenges by analyzing the divergent sentiments in split sentence parts. However, this method faced difficulties in preserving the overall coherence of meaning in sentences after splitting.

```
First Half of Sentence: I hate any one
Second Half of Sentence: who can hurt you
First Half - Predicted Sentiment: negative
First Half - Confidence Scores: 0.9981
==============================
Second Half - Predicted Sentiment: negative
Second Half - Confidence Scores: 0.9948
==============================
```

Comparison with Baseline:

A comparative analysis with the traditional SentimentIntensityAnalyzer revealed that the fine-tuned DistilBERT model surpassed traditional approaches, especially in scenarios involving homonyms. The nuanced contextual embeddings learned during fine-tuning contributed to the model's improved understanding of homonymous meanings.

Conclusion:

The results of the fine-tuning experiment signify a promising advancement in homonym-aware sentiment analysis using DistilBERT. While the model showcased success in capturing homonymous meanings in specific sentences, there are acknowledged challenges in handling certain nuanced contexts. This sets the stage for further iterations and refinements to enhance the model's adaptability and effectiveness in navigating the complexities of homonym-rich linguistic scenarios.

Implementation Tools and Challenges:

Language & Environment:

- Used: Python in Google Colab.

Challenges Faced:

- Resource Constraints: Managing limited GPU memory in Colab, requiring optimization and efficient data batching.

- Tokenization Compatibility: Ensuring alignment between tokenization strategies and transformer models for consistency.

- Fine-Tuning Complexity: Navigating intricacies in hyperparameters and the fine-tuning process.

Overall Conclusion:

The exploration into Homonym-Aware Sentiment Analysis witnessed a systematic progression through multiple experiments.

1. Baseline Experiment:

- The baseline experiment probed the effectiveness of traditional sentiment analyzers, revealing their lack of homonym-awareness. This critical insight became the benchmark for subsequent endeavors.

2. Transformer Assessment:

- Evaluating pre-trained transformers, specifically DistilBERT Base Uncased, demonstrated its incapacity to discern homonyms in English sentences. This revelation set the stage for the subsequent fine-tuning experiment.

3. Fine-Tuning DistilBERT:

- Fine-tuning DistilBERT on the SST-2 sentiment analysis dataset aimed to imbue homonym-awareness. The lightweight nature of DistilBERT was strategically chosen for efficiency, addressing computational constraints.

Key Findings:

- Fine-tuning DistilBERT showcased improvements, successfully capturing homonyms in certain sentences while facing challenges in others.

Implications and Future Directions:

The journey emphasized the need for advanced models, pushing the boundaries of homonym-aware sentiment analysis. While the fine-tuned model displayed promising results, ongoing efforts will focus on refining contextual embeddings and addressing nuanced challenges for a more comprehensive homonym-aware sentiment analysis model.