Word Sense Disambiguation using Word Embeddings and WordNet

Om Khare
112003066
Division 1

This code performs Word Sense Disambiguation (WSD) using a combination of Word Embeddings (Word2Vec) and WordNet. Word Sense Disambiguation is the task of determining the correct sense of a word in context, particularly when a word has multiple meanings or senses. The code uses the Semcor corpus for training and evaluation, which is a corpus annotated with WordNet senses.

Let's break down the code and understand its key components:

Imports and Setup:

- The code imports necessary libraries, including NLTK, NumPy, TQDM, and scikit-learn.
- It defines custom stopwords and initializes a WordNet lemmatizer.

Helper Functions:

- cosineSimilarity(a, b): Calculates the cosine similarity between two vectors.
- isNumber(s): Checks if a string represents a number.
- n2w(w): Converts numeric word representations using the num2words library.
- lemmatize(w, tag): Lemmatizes a word based on its POS (Part of Speech) tag.
- clean(tokens): Tags and lemmatizes tokens, removes stopwords, and returns cleaned tokens.
- getVec(w): Returns the Word2Vec vector for a word.
- syn2sense(syn): Converts a WordNet Synset to its sense representation.
- treebank2wn(ttag): Maps Treebank POS tags to WordNet POS tags.
- sent2vec(tokens): Generates a sentence vector by averaging Word2Vec vectors of words.
- parse(d): Parses a Semcor sentence, extracting tokens and associated senses.
- getCandidates(w, tag): Retrieves candidate sense vectors and labels for a word.

WSD Evaluation on Semcor Corpus:

- The code uses the Semcor corpus (semcor.tagged_sents) for training and evaluation.
- It iterates through each sentence in the Semcor corpus, performs WSD, and evaluates the predictions.

Evaluation Metrics and Analysis:

- The code calculates accuracy, precision, recall, and F1-score for the WSD task.
- It prints and displays statistics about the model's performance.

Prediction Function:

- The code defines a predict function to perform WSD on new sentences.

Example Predictions:

- The code provides example sentences (sents) and uses the predict function to predict WordNet senses for each word.