NLP Assignment 1: Answers

Name: Alireza Dastmalchi Saei

Stu No.: 993613026

1 Question 1

The 4 stages of natural languages are:

- **Acoustic Ambiguity**: This type of ambiguity arises when words or phrases sound similar but have different meanings when put together. It occurs due to homophones or words with similar phonetic properties. Example:

"او بر روی چمن زار میزد" ← چمن + زار (گریه) / چمن‌زار

"ما در پیاله عکس رخ یار دیده‌ایم" ← ما + در / مادر

- **Syntactic Ambiguity**: Syntactic ambiguity occurs when the structure or arrangement of words in a sentence allows for multiple interpretations. This ambiguity arises from the grammar or syntax of the language. Different parsing or grouping of words within the sentence can lead to distinct meanings. Example:

"آن مرد با دوربین دیده شد" ← آن مرد توسط دوربین دیده شد / آن مرد که دوربین داشت دیده شد

"علی دوست های دخترش را صدا زد" ← دوست های علی که دختر هستند / دوست های فرزند (دختر) علی

- **Semantic Ambiguity**: Semantic ambiguity involves words or phrases that have multiple meanings or interpretations based on context. This ambiguity arises from the ambiguity of individual words or phrases themselves, rather than the structure of the sentence. The meaning of the word may change depending on the context in which it is used. Example:

"آیا میتوانیم در کلاس غذا بخوریم؟" ← آیا اجازه غذا خوردن در کلاس داریم؟ / آیا توانایی خوردن غذا را داریم؟

"او در باغ نیست" ← او در باغ حضور فیزیکی ندارد / او حضور ذهنی ندارد (در جریان اتفاقات قرار نگرفته)

- **Discourse Ambiguity**: Discourse ambiguity refers to ambiguity that arises at the level of larger units of language, such as conversations, paragraphs, or entire texts. This type of ambiguity often arises when pronouns or other referring expressions lack clear antecedents or when there are ambiguities in the overall context of the communication. Example:

"رضا، علی را در ماشین خودش به قتل رساند" ← ماشین متعلق به چه کسی است؟

"علی، رضا را در موزه دید. او همیشه آنجاست" ← او به چه کسی اشاره می‌کند؟

2 Question 2

Explanation: The Maximum Matching Algorithm is a method used in Natural Language Processing (NLP) and text processing to segment words or tokens from a sequence of characters. It trys to find the longest possible sequences of words or tokens from the input text based on a given dictionary or vocabulary.

This algorithm takes a sequence of characters (string) as input. Then, using a list of all valid words available (dictionary), the algorithm iterates over the input text and tries to match the longest possible sequence of characters with words or tokens from the dictionary. This greedy algorithm returns a list of words or tokens segmented from the input text.

Example: This algorithm acts as following:

Figure 1: Maximum Matching Example

Applications: Applications of the Maximum Matching Algorithm:

- Tokenization in Natural Language Processing.
- Segmenting words or tokens from raw text in Information Retrieval systems.
- Word segmentation in languages without clear word boundaries, such as Chinese or Thai.

3 Question 3

Lemmatization: Lemmatization is the process of reducing words to their base or dictionary form, known as the lemma. It involves identifying the root form of a word considering its meaning.

Stemming: Stemming is the process of removing suffixes or prefixes from words to extract their root form, known as the stem. It's a heuristic process that chops off ends of words based on common patterns. But this method is not always able to produce valid dictionary words.

```python
from hazm import Stemmer, Lemmatizer

def process_persian_sentence(sentence):
    stemmer = Stemmer()
    lemmatizer = Lemmatizer()

    tokens = sentence.split()
    lemmatized_tokens = [lemmatizer.lemmatize(token) for token in tokens]
    stemmed_tokens = [stemmer.stem(token) for token in tokens]

    return lemmatized_tokens, stemmed_tokens

persian_sentence = "من دارم به مدرسه می‌روم"
lemmatized, stemmed = process_persian_sentence(persian_sentence)

print("Original:", persian_sentence)
print("Lemmatized:", ' '.join(lemmatized))
print("Stemmed:", ' '.join(stemmed))
```

```
Original: من دارم به مدرسه می‌روم
Lemmatized: من داشت#دار به مدرسه رفت#رو
Stemmed: من دار به مدرسه می‌رو
```

Figure 2: Stemmed and Lemmatized