به نام خدا

تمرین سری سوم

درس مبانی پردازش گفتار و زبان

سینا علی‌نژاد

سوالات تئوری:

الف)

- **Ambiguity:** Words can have multiple meanings. For instance, "Apple" could refer to the tech company or the fruit, depending on the context. NER systems need to be able to understand these subtleties.
- **Context Dependence:** Meaning often comes from surrounding text. "Washington" might be a city or a state, and "Java" could be the island or the programming language. NER models need to consider the context to make the right call.
- **Misspellings and Variations:** Text data can be messy, with typos and informal variations. NER systems need to be robust enough to handle these imperfections.
- **New and Out-of-Vocabulary Entities:** NER models are trained on specific data sets. They might struggle to recognize entities they haven't seen before, especially if they are completely new or not spelled conventionally.
- **Language-Specific Issues:** Languages have different grammatical structures and naming conventions. An NER model that works well in English might not perform as well in Spanish or Chinese.
- **Data Quality and Scalability:** NER systems rely on high-quality training data that is properly labeled. This can be expensive and time-consuming to create, especially for specialized domains. Additionally, as the amount of text data grows, NER systems can struggle to keep up with the processing demands.

ب)

Positive effects:

- **Disambiguation:** Context can help disambiguate between different meanings of the same word. For example, "Apple" could refer to the fruit, the company, or a verb depending on the surrounding words. By analyzing the context, an NER system can choose the most likely interpretation.
- **Identifying Relationships:** Context can help identify relationships between named entities. For example, if a sentence mentions "Paris" and "France" together, the NER system can recognize them as a location (city) and its corresponding country.
- **Identifying Implicit Entities:** Context can help identify entities that are not explicitly mentioned but implied. For example, if a sentence talks about "the White House," the system might recognize it as a location (government building) even though "White House" isn't a proper noun.

Negative effects:

- **Ambiguity:** Complex or ambiguous sentences can confuse NER systems. For example, a sentence like "The ring was made of gold" could be interpreted in two ways:

"ring" as a piece of jewelry or "ring" as a group of people. Context might not always be sufficient to resolve such ambiguity.

- **Rare Entities:** If an NER system hasn't been trained on a specific type of entity appearing in the context, it might miss it. For example, a system trained on general news articles might not recognize the names of fictional characters in a fantasy novel.
- **Negation and Modality:** Context can include negation ("not a doctor") or modality ("might be a location"). These can affect the NER system's confidence in identifying an entity or its type.

Here are some additional points to consider:

- **Window Size:** The size of the context window considered by the NER system can affect its accuracy. A larger window might provide more information but also increase computational cost.
- **Preprocessing:** Techniques like text normalization, part-of-speech tagging, and coreference resolution can help improve context understanding for NER systems.
- **Deep Learning Models:** Newer deep learning-based NER models can leverage larger contextual windows and learn complex relationships between words, leading to better context utilization.

Overall, context plays a crucial role in NER. By understanding the surrounding words and their relationships, NER systems can achieve more accurate identification and classification of named entities in text.

ج)

Hidden Markov Models (HMMs) are a popular approach for sequential data modeling, but they have limitations that Conditional Random Fields (CRFs) address. Here's how CRFs improve upon HMMs:

Limitations of HMMs:

1. **Independence Assumption:** HMMs assume that the probability of a label at a given position depends only on the previous label, not on the entire sequence or features of the current observation. This can be unrealistic, as current labels can be influenced by more than just the immediate predecessor.
2. **Label Bias Problem:** HMMs favor labels that appear more frequently at the beginning of sequences. This is because the initial state distribution in an HMM can bias the prediction towards frequent starting labels.
3. **Limited Feature Representation:** Traditional HMMs primarily rely on the previous label for prediction, making it difficult to incorporate rich contextual information like word embeddings, part-of-speech tags, or other features.

How CRFs address these limitations:

1. **Conditional Probability:** CRFs model the conditional probability distribution of label sequences given an observation sequence. This allows them to consider the entire sequence and potentially all features during prediction, leading to more accurate labeling.
2. **No Label Bias:** CRFs don't have a separate initial state distribution, eliminating the label bias issue. All label transitions are modeled jointly, taking into account the entire sequence.
3. **Flexible Feature Integration:** CRFs can incorporate a wide range of features beyond just the previous label. They can leverage various feature functions based on words, parts of speech, dictionary information, or any other relevant data source. This allows them to capture richer context and make more informed predictions.

Additional Advantages of CRFs:

- **Efficient Inference:** Despite their increased modeling power, CRFs can still be efficiently trained and decoded using algorithms like Viterbi decoding, making them practical for real-world applications.
- **Discriminative Model:** CRFs are discriminative models, meaning they directly model the conditional probability of labels given observations. This can be advantageous over generative models like HMMs when the focus is solely on prediction accuracy.

In summary, CRFs overcome the limitations of HMMs by:

- Relaxing the independence assumption for labels.
- Eliminating label bias.
- Enabling the use of rich feature representations.

These improvements make CRFs a more powerful and flexible approach for various sequence labeling tasks, including Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and protein structure prediction.

(د

- I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN

The word ‘Atlanta’ is a NNP

- Does/VBZ this/DT flight/NN serve/VB dinner/NNS

The word ‘dinner’ is a NN, it is not plural

- I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP

The word ‘have’ is a VBP

- Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

The word ‘Can’ is a MD (Modal)

(۵

In BIO labeling technique, beginning token of a named entity is shown by B and other parts of it are shown by I. Tokens that are not part of a named entity, are shown by O. Below picture is an example of this method.

Winnie	B-PERSON
the	I-PERSON
Pooh	I-PERSON
is	O
naïve	O
and	O
slow-witted	O

In BIOES, beginning token of a named entity is shown by B, inner parts of it are shown by I and the last token represented by E. If a named entity consists of just one token, then it is represented by S. Tokens that are not part of a named entity, are shown by O. below picture is an example:

In IO, tokens that are part of a named entity are labeled I and other are labeled O.

Method	Description	Advantages	Disadvantages
IO	**Inside-Outside**	- Simple and easy to understand. - Efficient for memory usage, as only two tags are needed (I - Inside, O - Outside).	- Doesn't differentiate between the beginning and inside of an entity. - Can be ambiguous for nested entities.
BIO	**Beginning-Inside-Outside**	- More informative than IO, as it distinguishes between the beginning (B) and inside (I) of an entity. - Suitable for handling nested entities.	- Requires more tags compared to IO (B, I, O).
BIOES	**Beginning-Inside-Outside-Single-ton**	- Most detailed tagging scheme, differentiating beginning (B), single-word entities (S), inside (I), and outside (O). - Useful for identifying single-word entities.	- Requires the most tags (B, I, O, S, E). - Can be more complex to learn for NER models.

سوال ۳-

الف) یکی از مشکلات، این است که برخی از این اسامی هم نام یک کتاب یا رمان و هم نام فیلم هستند و با توجه به context باید مشخص شود که منظور کدام است. برای مثال The Godfather را میتوان از این دسته نام برد.

مشکل دوم، کلمات اضافه ای مثل of,the,and,with یا حتی period sign هستند که مدل اینها را به عنوان I-MOV در نظر گرفته که درست نیست. عکس زیر نشان دهنده این موضوع است:

```
('the', 'I-MOV') biggest film ('of', 'I-MOV') ('the', 'I-MOV') pandemic ('.', 'I-MOV') ('A', 'B-MOV') ('hit', 'I-MOV') ('with', 'I-MOV') critics ('and', 'I-MOV')
```