**Section 1: Linguistics**

**Question 1A (5 Marks)**

How many tokens are there in this sentence? Explain your reasoning.

*Boris Johnson has been the prime minister of the UK since last year.*

Answer: tokens are the number of all words in a sentence. Here, it is 13.

**Question 1B (10 Marks)**

Fill in the blanks in these statements:  
 government and govern are morphologically related through \_**derivation\_\_**

minister and ministers are morphologically related through **\_inflection\_\_**

road and roadmap are morphologically related through **compound**\_

**Question 1C (10 Marks)**

Describe in your own words the **difference between a parallel and comparable corpus**.

Give an **example of an NLP application that uses such corpora**.

**Answer:**

**Graphical user interface, text, application, email

Description automatically generated**

**Section 2: Parsing**

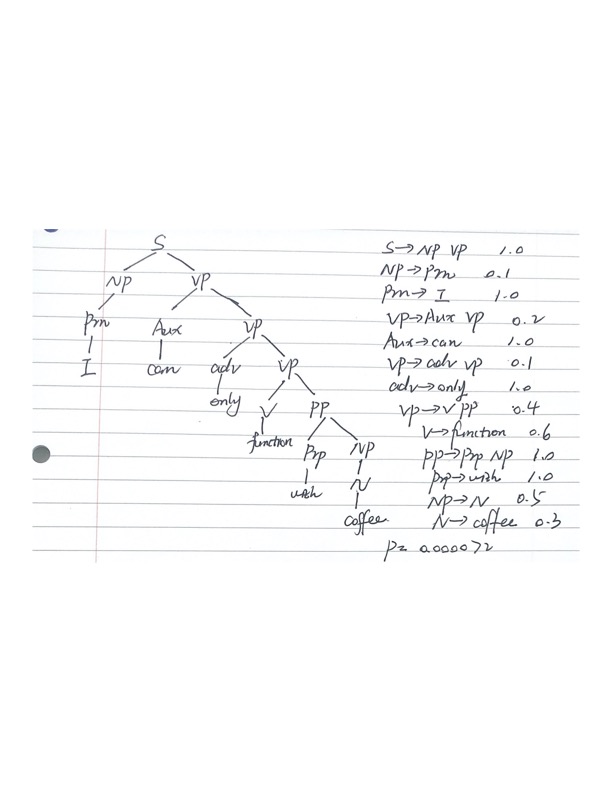
**Question 2A**  (10 Marks)

Consider the following grammar: What is the probability of the following sentence in this grammar?

“*I can only function with coffee*”

Show which rules in the grammar were used in the parse tree of this sentence.

**Answer:**

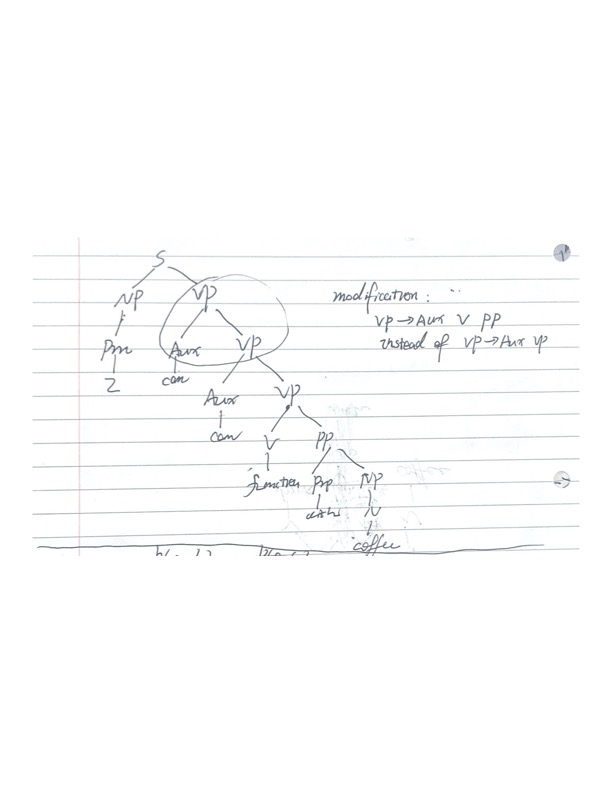


**Question 2B (10 Marks)**

Using the grammar of Question 2A, find a sentence that is accepted by the grammar but is not grammatical in English and suggest a modification to the grammar so that this sentence is not generated.

**Answer:**

I can can function with coffee.



Modification: use VP-AUX V PP instead of VP-AUX VP

**Question 2C (5** Marks)

Why do lexical dependencies cause an issue with a simple Probabilistic Context-Free

Grammar (PCFG) approach to parsing?

**Answer:**

In a PCFG the expansion of one non-terminal is independent of any other non-terminal.

However, subjects are 91% pronouns and 9 lexical noun phrases; direct objects are 34% pronouns and 66% lexical noun phrases. Thus, the assumption of independence is not valid.

Besides, lexical attachment is very word dependent, and PCFGs cannot model lexical dependencies. For example:

*Men in hats and women*

Chart, radar chart

Description automatically generated

Thus, we need a lexicalized PCFGs to further differentiate non-terminals by associating them with lexical items.

**Section 3: Semantics**

**Question 3A (5 Marks)**

Fill in the blanks in these statements on words that **are semantically related**:

government, cabinet, administration are **Synonym. (同义词)**

light and dark are **antonym.（反义词）**

**Question 3B (5 Marks)**

Explain in your own words how word senses are represented in WordNet. Give an

example.

**Answer:**

Word meaning is represented in “Wordnet” by use of “synsets”.

For example:

{board, plank} represents “board” in the meaning of “plank”.

{board, committee} represents “board” in the meaning of “committee”.

Text

Description automatically generated

**Question 3C (5 Marks)**

Explain in your own words how word senses are represented in FrameNet. Give an

example.

**Answer:**

With FrameNet, the most words can best be understood on the basis of a semantic frame: a description of a type of event, relation, or entity and the participants in it.

Frames express semantic roles-for example:A screenshot of text

Description automatically generated with low confidence

Text

Description automatically generated

Note: lexical units within a frame are synonyms.

**Question 3D (10 Marks)**

How can Wikipedia be used in word sense disambiguation?

**Answer:**

Through word sense disambiguation experiments, Wikipedia-based sense annotations are reliable and can be used to construct accurate and robust sense classifiers for WSD. The sense annotations for the corresponding concepts can be seen as a property particularly valuable for entities that are ambiguous.

Starting with a given ambiguous word, we derive a sense-tagged corpus following three main steps:

* First, we extract all the paragraphs in Wikipedia that contain an occurrence of the ambiguous word as part of a link or a piped link.
* Next, we collect all the possible labels for the given ambiguous word by extracting the leftmost component of the links.
* Finally, the labels are manually mapped to their corresponding WordNet sense, and a sense tagged corpus is created.

**Section 4: Applications**

**Question 4A (10 Marks)**

Explain in your own words **how a knowledge model** can **be used in information extraction**. Give an example.

**Answer:**

In IE system we can use knowledge nodes of a knowledge model to indicate the properties of object, the relationship between objects, or event. The tag set in the knowledge nodes is used to reflect objects, attributes, relationships and other information.

When users need the description information or the answers of questions about knowledge, the system will reply these queries with some sentences. The knowledge model use manual rules and fixed templates, coupled with knowledge tag, to generate the answer as output.

Take knowledge “AGE” as example. Knowledge feature in “AGE” is consisting of “<NAME>“, “<AGE>“. If a sentence represents the age property of object according to knowledge feature matching, the knowledge tag of “AGE” will be used to mark the sentence.

**Question 4B (10 Marks)**

Consider **Pointwise Mutual Information (PMI).** Given words a,b,c, explain how PMI(a,b)

for a given corpus can be higher than PMI(a,c).

**Answer:**

PMI can be used to rank words based on the correlative strength between the individual words in the NP.

PMI(a,b)=log2(P(a,b)/P(a)P(b))

PMI(a,c)=log2(P(a,c)/P(a)P(c))

For a given corpus, if PMI(a,b) > PMI(a,c), the probability of co-occurrence of (a,b) may much higher than (a,c); or P(b) is much lower than P(c).

**Question 4C 5 Marks**

Discuss **a limitation of a lexicon-based approac**h to sentiment analysis.

**Answer:**

1. Implicit sentiment

Neutral words used but POS sentiment implied. E.g. “go read the book!”

Neutral words used but NEG sentiment implied. E.g. “if you are reading this because it is your darling fragrance, please wear it at home exclusively and tape the windows shut.”

1. Ambiguity

Same words used in different contexts express NEG vs. POS sentiment.

e.g. “This car’s steering is unpredictable!”

e.g. “This film is unpredictable!”

3) Irony & Sarcasm (反讽与讽刺)

Positive words used but NEG sentiment implied.

e.g. “Great job Rogers! Raise rates but not service.”

4) negation

Positive words used but NEG sentiment implied.

“I don’t like this new Nokia model.”

“I didn’t enjoy it.”

5) Informal language

Social media content uses non-standard, informal language such as hashtags.

6) indirect sentiment

Sentiment expressed may not. Be that of the author.

“Although this product is disliked by many,..”