1 Multiple Choice Questions (4 Points)

For each of the following questions, multiple of the given answers can be correct. State whether each option is correct or wrong.

You receive points for correctly identified statements, while you lose points for wrongly identified statements. However, you cannot get negative points for a question. You can also "skip" an option. Then you will neither gain nor lose any points for that. For example, you can submit your answers like: "correct, wrong, skip, skip"

1. Why is it not the best idea to obtain the embeddings of a sentence by averaging the vectors of each word in the sentence? (2 points)

- ☐ Because this approach does not work even for simple text classification tasks.
- ☐ Because this approach cannot model the order of the words in the sentence.
- ☐ Because this approach cannot model the relations between words that occur in different parts of the sentence.
- ☐ Because this will largely increase the dimensions of the embeddings.

2. Which of the below is/are the advantage(s) of symbolic representations of meaning over continuous representations? (2 points)

- ☐ Symbolic representations can be easily interpreted and manipulated by developers of natural language understanding systems.
- ☐ It is straightforward to map words and sentences to symbolic representations.
- ☐ It is possible to combine symbolic meaning representations with other ontologies.
- ☐ Systems that map sentences to symbolic meaning representations usually do not require handcrafted linguistic resources or annotated datasets.

2 Neural language models (4 points)

1. (a) Word order plays an important role in natural language. E.g., "The cat chases the dog" has a different meaning than "The dog chases the cat". Which part of Transformer embeddings are used to model the position of each word in a text? (1 point)

   (b) Transformers have multi-head self-attention layers. What does “multi-head” mean? (1 point)

   (c) Why multi-head self-attention is necessary for Transformers to model natural language? Explain in terms of the characteristics of natural language.

   (2 points)

3 Machine Translation (12 points)

1. Consider the following word sequence:

   i. <start> Marie likes children <end>

   ii. <start> children like Marie <end>

   Given the following counts of unigrams and bigrams in a text corpus, calculate the probability of the above sentence predicted by a **bigram language model** trained on the corpus.

   (4 points)

unigram	frequency
<start>	1182
Marie	12
likes	105
like	89
children	33
<end>	1182

bigram	frequency
<start> Marie	2
<start> children	20
Marie likes	2
children like	13
Marie like	2
children likes	3
like children	27
likes children	54
like Marie	3
likes Marie	1
Marie <end>	2
children <end>	12

2. Consider the following German text, which is to be translated to English,

   Kinder mag Marie

   A statistical machine translation model is used to estimate if the above sentence should be translated to *Marie likes children* or *children like Marie*.

   (a) How is $P(\text{"}\textit{Marie likes children}\text{"}|\text{"}\textit{Kinder mag Marie}\text{"})$ estimated by statistical machine translation (SMT)?

   (1 point)

   (b) Explain which translation would the SMT model more likely predict by calculating:

   i. P("*Marie likes children*"|"*Kinder mag Marie*")

   ii. P("*Children like Marie*"|"*Kinder mag Marie*")

   given the following translation probabilities and the probabilities obtained in 1.
   (Note: assume that the translation model does not include reordering costs.) (5 points)

German (source)	English (target)	Probability
Kinder	children	0.85
mag	likes	0.5
mag	like	0.5
mögen	like	0.9
mögen	likes	0.001
Marie	Marie	0.99
English (source)	**German (target)**	**Probability**
children	Kinder	0.75
likes	mag	0.98
like	mag	0.001
like	mögen	0.6
likes	mögen	0.0008
Marie	Marie	0.99

   (c) Assume that reordering costs are not taken into account in the translation model.

   i. Which translation option should have higher reordering cost? Why? (1 point)

   ii. The SMT model still outputs the same translation as predicted in 2(b). Explain what could be the reason. (1 point)

Submission Details:

Upload your submission to our CMS in groups of two to three students until *January 15, 2025 at 17:59 am*. Late submissions will not be graded! The submission should be uploaded by exactly **one** team member. Make sure that your submission contains the name and matriculation number of each team member. Submit your solution as a ***pdf*** **file** with your answers.

1)

1. Why is it not the best idea to obtain the embeddings of a sentence by averaging the vectors of each word in the sentence? (2 points)

Wrong ☐ Because this approach does not work even for simple text classification tasks.

Correct ☐ Because this approach cannot model the order of the words in the sentence.

Correct ☐ Because this approach cannot model the relations between words that occur in different parts of the sentence.

Wrong ☐ Because this will largely increase the dimensions of the embeddings.

2. Which of the below is/are the advantage(s) of symbolic representations of meaning over continuous representations? (2 points)

Correct ☐ Symbolic representations can be easily interpreted and manipulated by developers of natural language understanding systems.

Wrong ☐ It is straightforward to map words and sentences to symbolic representations.

Correct ☐ It is possible to combine symbolic meaning representations with other ontologies.

Wrong ☐ Systems that map sentences to symbolic meaning representations usually do not require hand-crafted linguistic resources or annotated datasets.

2)

a) Positional encoding is used to model the position of each word in a text.

b) Multi-head means that it has multiple attention mechanisms which are used to capture different types of dependencies.

c) In natural language, words have different types of relationships which all have a part in the meaning. For instance, syntactic dependency like in the example of "The person sitting next to me is happy." or co-reference like in the example of "She likes herself." are two examples of these relationships. It's hard for single-head attention mechanism capturing all these dependencies. However, multi-head attention mechanism is able to capture all dependencies by enabling each head to focus on individual dependency. Hence, it is necessary for Transformers to make use of multi-head attention mechanism.