Introduction and Disclaimer

These mock examination questions span diverse disciplines and are designed for your practice in preparation for the International Research Olympiad (IRO) 2024. Endeavor to answer them to the best of your ability, utilizing this opportunity to enhance your skills and knowledge. For additional practice, it is advisable to engage in extensive reading of various papers; such efforts will contribute to a more comprehensive and nuanced understanding of the subject matter.

Try your best, and good luck! -International Research Olympiad 2024

Mock Examination Answer Key 1
Bolded answers are correct.

Paper 1: Computational Linguistics

Question 1

Question: In the study of natural language processing, different architectures are employed to develop high-quality vector representations of text. Which of the following methods is recognized in the paper for effectively learning these representations?

a.) Continuous Bag of Words (CBoW)

- This model predicts target words (e.g., 'milk') from context words ('I drank some ___ this morning')

b.) Neural Networks

- While these are a broad category of models, they are not specifically focused on text vector representation.

c.) **Skip-gram Model**

- **This approach is highlighted in the paper for its innovative way of predicting context words from a target word.**

d.) Bananagram

- This is a popular word game and is unrelated to the computational models for text representation.

Question 2

Question: According to the findings, the Word2Vec model demonstrates an ability to capture intricate relationships between words based on the contexts in which they appear. However, the model does have limitations. Based on the information provided, which of the following represents the largest limitation of the Word2Vec model as discussed in the abstract?

a.) Indifference to word order

- The model does not consider the sequence in which words appear, potentially overlooking syntactic structures.

b.) Inability to capture semantic relationships

- Despite contextual learning, the model fails to grasp the deeper meanings and associations between words.

c.) Lack of precision

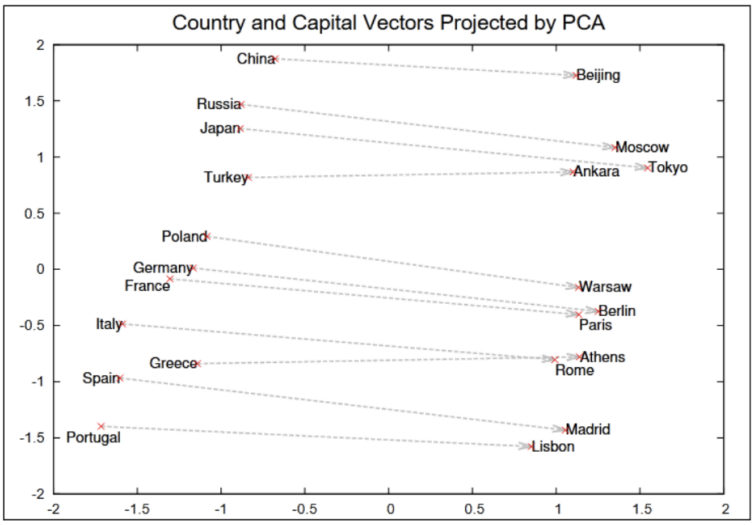
- The vectors generated by the model do not achieve the desired exactness in representing word contexts.

d.) **Inability to represent idiomatic phrases**

- **The model's approach to context and existence struggles with non-literal language and colloquial expressions.**

Question 3

Question: A researcher is using a Word2Vec model to understand the relationships between countries and their capitals as represented in a large corpus of text. The researcher uses Principal Component Analysis (PCA) to project the high-dimensional vectors onto a two-dimensional plane for visualization purposes. Upon examination of the resulting plot, it is observed that the vectors for countries and their corresponding capitals align in a manner that seems to reflect their latitudinal positions on a map. Which of the following hypotheses is best supported by the observation of the alignment of country and capital vectors along trajectories corresponding to latitude?



a.) The Word2Vec model was trained with a special emphasis on geographical data, which included explicit latitude and longitude coordinates.

- This answer is just untrue.

b.) The Word2Vec model inherently understands geographical concepts and can accurately map countries to their physical locations on Earth.

- Ignorant of function/workings of the model.

c.) The alignment along latitudinal lines is a coincidental outcome of the PCA reduction and does not necessarily reflect a true understanding of geographic locations by the Word2Vec model.

- Correct along lines but too strong to warrant this explanation.

d.) **The Word2Vec model has encoded a pattern where texts that mention countries and capitals also frequently include discussions of their geographical characteristics, such as relative latitudes.**

- **True, mentioned in the abstract.**

Question 4

Question: If the same associative logic is applied to the phrase "Apartment + Small," which of the following would be the most likely outcome?

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

a.) **Studio, One-Room, Bachelor Pad, Compact.**

- **Correct, closest match to small apartment.**

b.) Tiny, little, small, micro.

- Fails to consider apartment aspect.

c.) Apartment, residence, house, mini-home.

- Fails to consider small aspects.

d.) Division, section, sector, subset.

- Apartment confused with department.

Question 5

Question: Suppose a new neural network architecture is proposed for learning word representations that trains 10 times faster than previous models. Based on the paper, which of the following predictions is best supported?

a.) **The new model is using a better, faster approximation of the Softmax function.**

- **This is how the Skip-gram got so much faster.**

b.) The new model would still benefit from subsampling frequent words.

- Not a unique reason.

c.) The new model would not need as much training data to achieve good performance.

- It is possible, but not 100% true.

d.) The new model would struggle to capture precise relationships between rare words.

- Not necessarily implied training to be lower quality.

Question 6

Question: The Skip-gram model from the Word2Vec paper utilizes a specific formula to maximize the prediction accuracy of word occurrences in a corpus. The formula includes a double summation over the range of context words around a current word w_t. The notation for the range is given as $-c < j < c, j \neq 0$. What does this notation represent in the context of the Skip-gram model?

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\leq j\leq c,\\ j\neq 0}}\log p(w_{t+j}|w_t)$$

a.) The index j sums over all integers including zero, from -c to c, which represents the full range of words in the corpus.

- Fails to consider not including the current word.

b.) The index j represents the current word, and the summation is over all words in the corpus, excluding any context words.

- Nope, j is an index.

c.) The index j sums from -c to c and includes zero, which signifies that the current word is used twice in calculating the probability.

- Nope, this ignores the $j \neq 0$.

d.) **The index j sums over all context words in a specified range around the current word, excluding the current word itself.**

- **This explanation makes the most sense, correct.**