INFOMNLG *Natural Language Generation*

Individual coursework (coding)

General information

This worksheet accounts for 20% of your global grade for the course. Please do this coursework individually. A submission facility will be provided via Blackboard. Please submit this file with your answers.

Task Description

You will need the Python notebook accompanying this file (available from Blackboard). Task instructions are in the notebook.

Please answer all the questions below. Use this same document to insert your answers below the questions.

Question 1

1A [10 pts]

Reproduce below the prompt you used for the model to run it in zero-shot model (for Task 1). Only reproduce one example for a specific test instance.

Prompt:

You are a fluent English speaker who has been employed to generate natural language descriptions from structured data.

Your task is to generate a natural language description based on the following RDF triples:

Abilene_Regional_Airport | cityServed | Abilene,_Texas

Please generate a concise and accurate description of the data provided above.

Generated Text:

The Abilene Regional Airport serves the city of Abilene, Texas.

1B [5 pts]

Your prompt needs to include the subject/property/object triples. How did you format them? Explain your decision below.

I formatted the **subject-property-object** triples using a **pipe-separated format** (|), for example:

Abilene_Regional_Airport | cityServed | Abilene,_Texas

The pipe symbol ("|") provides a clear and visually distinct separator between the subject, property, and object. It is a common delimiter in data formats (e.g. RDFs, CSV, etc.), which makes it familiar and easy to parse. It is also not commonly used in natural language, and therefore reduces ambiguity.

In case of multiple triples, I am using a single line for each triple to avoid unnecessary complexity and keep the prompt easy to read.

I chose **not** to explicitly explain the format of the RDF triple ("subject | property | object") in the prompt because the structure is already **simple and intuitive** for both humans and the model to read and understand.

The goal was to provide the data in a simple and direct manner without adding excessive formatting or syntax, allowing the model to focus on the core information contained within the triples.

Question 2

2A [35 pts]

Complete the table below with the BLEU evaluation scores you obtained for the different experimental settings.

	BLEU SCORE
Zero-shot (Task 1)	55.80
One-shot with random example	60.43
One-shot with same category example	60.43

2B [10 pts]

Do you observe any differences in the scores between the settings reported in the above table? What are the reasons for the differences (or lack of differences) you observe?

Looking at the table, we observe that the **zero-shot setting** has a lower BLEU score (55.80) compared to the **one-shot settings** (60.43), while there is **no difference** in BLEU scores between the **one-shot with random example** and **one-shot with same category example** settings.

The increase in BLEU scores in the one-shot settings indicates that providing a single example significantly improves the model's ability to generate text that aligns with human-written references.

This is because the example provides the model with a concrete illustration of the desired output format and style, guiding the model towards producing more accurate and fluent descriptions. So, the model benefits from seeing how the input triples can be translated into natural language.

The lack of difference between the **one-shot with random example** and **one-shot with same category example** settings suggests that the model's performance is not significantly affected by whether the example is from the same category or a random one.

This could be because the model's pre-trained knowledge is already robust enough to generalize well from a single example, regardless of its category or the task (generating natural language descriptions from RDF triples) is relatively straightforward, so the additional context provided by a same-category example does not offer a significant advantage over a random example.

This implies that the model is primarily learning the format of the output from the example, rather than relying heavily on domain-specific knowledge. Because the format is the same across categories, the category of the example does not matter.

It is possible that the model is already very strong at understanding the general category of the data, so providing an example of the same category does not provide any additional benefit.

In conclusion, the **one-shot settings** outperform the **zero-shot setting** because the additional example provides useful context for the model and the **similarity in scores** between the two one-shot settings indicates that the model's performance is not highly sensitive to the category of the example provided, at least for this task.

Question 3

3A [20 pts]

Explain below the method you developed for measuring the variability in generated output. Your explanation should be clear and concise. Feel free to use an example, or to explain it using bullet points, formulas etc.

Since BLEU aims to assess how similar two sentences are, it can also be used to evaluate how one sentence resembles the rest generated sentences. By treating each generated sentence as a hypothesis and comparing it to all other generated sentences as references, we can compute BLEU scores for each and define their average as the Self-BLEU score as a measure of variability in the generated output overall.

A **lower Self-BLEU score** indicates higher diversity (less similarity between generated texts), while a **higher Self-BLEU score** indicates lower diversity (more similarity between generated texts) and the range is from 0 to 100.

Formula:

$$\text{Self-BLEU} = \frac{1}{N} \sum_{i=1}^{N} \text{BLEU}(\text{hypothesis}_i, \text{references}_i)$$

Where:

- N = Total number of generated texts.
- hypothesis_i = The i-th generated text.
- references_i = All generated texts except the i-th text.

In terms of limitations, self-BLEU measures only n-gram overlap and may not capture semantic diversity.

More specifically, two sentences with different meanings but similar wording could have a high Self-BLEU score. For example, the sentences *"The boy hit the ball."* and *"The boy was hit by the ball."* have very different meanings but have similar wording (high n-gram overlap), so the self-BLEU score would be high indicating low diversity.

Similarly, two sentences with similar meanings but different wording could have a low Self-BLEU score indicating high diversity. For example, the sentences "*Sophie went to the movies and enjoyed it a lot.*" and *"Sophie visited the cinema and really liked it"* have similar meanings but have very different wording (low n-gram overlap), so the self-BLEU score would be low indicating high diversity.

Ultimately, the limitations of Self-BLEU depend on the specific type of variability we aim to measure. If the goal is to assess lexical diversity, Self-BLEU can be useful, but if we are interested in semantic diversity, it may not always be a reliable metric.

3B [20 pts]

Report the results below for your variability metric. Does few-shot generation increase or reduce variation?

	VARIATION SCORE
Zero-shot (Task 1)	2.42
One-shot with random example	2.60
One-shot with same category example	2.59

Looking at the table, we observe that the **zero-shot setting** has the **lowest Self-BLEU score (2.42)**, indicating the **highest variability** in generated outputs, while the **few-shot settings** (both random and same-category examples) have **higher Self-BLEU scores (2.60 and 2.59)**, indicating **lower variability** compared to the zero-shot setting.

So, **few-shot generation reduces variation** compared to zero-shot generation. This can be explained by the fact that providing examples (even just one) guides the model toward generating outputs that are more consistent with the provided examples, reducing the randomness and diversity of the outputs. In contrast, the **zero-shot setting** relies entirely on the model's pre-trained knowledge without any guidance, leading to more diverse but potentially less consistent outputs.

The **one-shot with random example** and **one-shot with same category example** settings have nearly identical Self-BLEU scores (2.60 vs. 2.59). This suggests that the **category of the example** (random vs. same-category) does not significantly impact the variability of the generated outputs, which makes sense, since **one-shot with random example** and **one-shot with same category example** settings had the same BLEU score.

In conclusion, it seems that few-shot generation **reduces variation** in the outputs compared to zero-shot generation and the reduction in variability is likely due to the additional guidance provided by the examples, which makes the model's outputs more consistent but less diverse.