College of
Computer Studies

DEPARTMENT OF
SOFTWARE TECHNOLOGY

NLP1000

Introduction to Natural Language Processing

Project 3: Machine Translation

Instructor : Nathaniel Oco
nathaniel.oco@dlsu.edu.ph

Course Site/Repository : AnimoSpace, Google Drive

Description

This project focuses applying strategies that make use of closely and distantly related languages to improve machine translation (MT). Students will first identify which languages are closely and distantly related. They will design and implement strategies to improve MT performance by leveraging resources from a related language. Approaches may include, but are not limited to:

- Data Augmentation: Add parallel or comparable text from a closely related language.
- Pivot MT: Translate via a high-resource or related language (e.g., Yami → Ivatan → Tagalog).
- Fine tuning: Fine-tune an MT model trained on a related language or multilingual model.

Deliverables

1. Report (PDF). A written report (3–5 pages) explaining your methodology, experiments, results, analysis, and declaration of AI usage (per person).
2. Demo Video. A short video demonstrating your MT system in action, explaining your approach, and summarizing results.
3. Source code (ZIP). A compressed folder containing: Source code, Instructions/README for reproducing your results.
4. Data sources (ZIP). A compressed folder containing: Training data, parallel corpora and related files.

Provide a declaration of AI usage (per person)

1. AI tool(s) used
2. Describe briefly how you used the AI tool(s) in the completion of this assignment. Examples include: idea generation, grammar checking, paraphrasing, summarizing references, formatting help, etc. Provide sample prompts and outputs.
3. Extent of Use:
   a. Minimal – used only for grammar/spell check.
   b. Moderate – used for inspiration, then reworded significantly.

In 2–3 sentences, reflect on how the use of AI contributed to your learning. If no AI was used, reflect on why you chose not to.

Rubric

Criteria	**Exemplary** Full points	**Satisfactory** ¾ points	**Developing** ½ points	**Beginning** ¼ points
Data selection (10 points)	The total parallel corpus size is 100,000 words and data sources are comparable	At least ¾ of the target corpus size	At least one half of the target corpus size	Less than 1,000 words
Computational Output & Methodology (60 points)	Output is reproducible, well-documented, and evaluated using standard metrics against a gold standard. Methodology is sound, justified, and effectively applied. Code and tools are appropriate, clearly explained, and properly referenced.	Output is mostly correct and reproducible. Methodology is appropriate but only partially justified.	Output is incomplete, inconsistent, or poorly explained. Methodology is vague or weakly justified.	Output is incorrect or missing; methodology absent or not credible.
Written report and demo video (30 points)	Clear, well-structured, and comprehensive. Strong introduction, methodology, results, and discussion with linguistic/sociocultural insights. Proper citations included.	Clear and organized with most required sections; some analysis but limited depth. Minor issues in clarity or citations.	Partial structure; limited analysis or weak discussion of results. Minimal sources.	Unclear, incomplete, or lacking analysis and discussion.

Bonus:

- The project includes the Yami language (romanized): +8 points
- The project includes human evaluation with at least 50 random sentences (with acceptable reliability scores): +8 points

Deadline

November 10, 11:30 am.