Milestone 1 - Literature review + Data

Akira Nair, Aneesh Edara, Prednya Ramesh, Ricky Raup

Literature Review:

CroCoSum: A Benchmark Dataset for Cross-Lingual Code-Switched Summarization[6]

In this paper, the authors address the lack of cross-lingual code-switched text in summarization research by introducing CroCoSum. CroCoSum, which we will use as part of our own dataset, contains English-to-Chinese summarization text where summaries mix both languages. The dataset also consists of over 24k English technology news articles paired with 18k human-written Chinese summaries collected from solidot.org. More than 92% of the summaries collected contain code-switched phrases and 55% of the sentences collected have some sort of language mixing.

The authors continue on to evaluate three different approaches, these being pipeline methods (translate-then-summarize or summarize-then-translate), end-to-end multilingual models, and zero-shot prompting with LLMs. Their main finding is that pretraining on existing cross-lingual summarization datasets does not improve performance on CroCoSum. This shows that existing cross-lingual summarization datasets do not transfer well to code-switched text. Also, through qualitative error analysis, they found that standard metrics like ROUGE fail when models produce correct content but are in different languages than the reference or omit code-switched terms. This shows that the evaluation of code-switched summaries requires human judgment on naturalness and relevance, mostly because current metrics cannot handle the complexity of the problem.

GLUECoS: An Evaluation Benchmark for Code-Switched NLP[3]

This paper introduces a very helpful comprehensive benchmark for NLP models trained on code-switched data. The authors compile datasets for six tasks across English-Hindi and English-Spanish settings: Language Identification (LID), Part-of-Speech (POS) tagging, Named Entity Recognition (NER), Sentiment Analysis, Question Answering (QA), and Natural Language Inference (NLI). These tasks establish a framework inspired by GLUE (for monolingual English). Along with the tasks, they introduce metrics to quantify the complexity of code-switching.

The authors evaluate multiple embedding strategies, including cross-lingual embeddings (MUSE, BiCVM, and BiSkip), and multilingual models such as mBERT. They show that while mBERT outperforms traditional cross-lingual embeddings, fine tuning on code-switched data improves its performance. Performance varies substantially across the chosen tasks. For example, while LID, POS tagging, and NER show strong results, NLI remains challenging for code-switched data. This suggests that there are fundamental difficulties in reasoning across mixed-language contexts. While GLUECoS established a benchmark for code-switched NLP across the above-listed six tasks including sentiment analysis and QA, it does not include summarization. Our work will address this gap by introducing code-switched thread summarization, attempt to build our own crude benchmarks for this task in line with what has

been introduced in the paper. We will do this so that we can properly evaluate how well our project does on the task.

GupShup: Summarizing Open-Domain Code-Switched Conversations[4]

GupShup introduces a large-scale dataset for Hindi-English code-switched conversations. It contains over 6,800 multi-party conversations and has human-annotated summaries. The paper provides linguistic analysis including token-level language tagging and metrics such as code-mixing index and switch-point frequency. The authors evaluate models for summarization, including GPT-2, BART, PEGASUS, T5, multitask T5, and mBART. They use metrics such as ROUGE, BLEU, METEOR, and BERTScore. For example, the paper discusses how, for generating English summaries from code-switched input, mBART gets the best ROUGE-1 score of 43.14, followed by multi-view seq2seq at 41.21. However, both of these models struggle significantly when generating code-switched summaries, with performance dropping to 19.98 for mBART and 36.28 for BART. Error analysis also reveals that models show 75.33% missing information rates, 77.83% incorrect inferences, and 45.50% wrong speaker attributions when processing code-switched conversations. Overall, this shows that multilingual pretraining alone is not good enough to provide effective code-switch awareness.

Dataset Description:

We train and evaluate on four normalised datasets and (optionally) a small scraped set for external validation:

- **CroCoSum[6]** (news, cross-lingual/code-mixed): use the official **train/val/test**; article text (title+body from src_docs.json) → summary (post_text).
- **CS-Sum[5]** (code-mixed dialogue): source has a single split; we parse #Speaker#: turns and create **80/10/10 by thread** (seed=42).
- **DialogSum[1]** (English dialogue): use the official **train/dev/test** JSONL; direct field mapping to our schema.
- **Kaggle Email Thread Summary[2]** (English email): merge message JSON and summary JSON **by thread_id**, sort by timestamp, strip quoted "Original Message" blocks/signatures, then split **80/10/10 by thread** (seed=42).

All data are converted to a unified **JSONL (one thread per line)** schema with messages[], a gold summary, and a lightweight domain tag (dialogue|news|email). This supports consistent training and per-domain reporting.

Optional external validation/fine-tuning: we may add a small set of **Reddit/YouTube** code-mixed threads later, normalised to the same schema. These will not be used for model selection; they are only for robustness checks. Collection will follow platform policies and avoid PII.

Works Cited:

[1] Chen, Yulong, Yang Liu, Liang Chen, and Yue Zhang. “DialogSum: A Real-Life Scenario Dialogue Summarization Dataset.” arXiv:2105.06762. Preprint, arXiv, June 16, 2021. https://doi.org/10.48550/arXiv.2105.06762.

[2] “Email Thread Summary Dataset.” Accessed November 6, 2025. https://www.kaggle.com/datasets/marawanxmamdouh/email-thread-summary-dataset.

[3] Khanuja, Simran, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. “GLUECoS: An Evaluation Benchmark for Code-Switched NLP.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.329.

[4] Mehnaz, Laiba, Debanjan Mahata, Rakesh Gosangi, et al. “GupShup: An Annotated Corpus for Abstractive Summarization of Open-Domain Code-Switched Conversations.” arXiv:2104.08578. Preprint, arXiv, April 17, 2021. https://doi.org/10.48550/arXiv.2104.08578.

[5] Suresh, Sathya Krishnan, Tanmay Surana, Lim Zhi Hao, and Eng Siong Chng. “CS-Sum: A Benchmark for Code-Switching Dialogue Summarization and the Limits of Large Language Models.” arXiv:2505.13559. Preprint, arXiv, May 19, 2025. https://doi.org/10.48550/arXiv.2505.13559.

[6] Zhang, Ruochen, and Carsten Eickhoff. “CroCoSum: A Benchmark Dataset for Cross-Lingual Code-Switched Summarization.” arXiv:2303.04092. Preprint, arXiv, May 23, 2024. https://doi.org/10.48550/arXiv.2303.04092.