CS 410 Project: Spanish-English Machine Translation of AI Research

Team Members and Captain:

Name: Felipe Arias
NetID: felipea2
Role: Captain and sole team member

Free Topic Description:

Topic: Automated English-Spanish Translation of AI Research Papers
Description: The goal is to develop a system that can translate AI research papers (in PDF format) from English to Spanish, developing the basic framework for ensuring the accurate conveyance of complex technical terms and concepts.
Task: Design and implement a pipeline for PDF parsing, translation, and enforcement of specific translation of technical terms.
Importance: As AI advances, it may change intellectual work and technology as we know it. Ensuring researchers from non-English speaking countries can access and understand recent advancements is crucial. This project aims to bridge the language gap by using me, one of the few truly Spanish-English bilingual and multi-cultural ML experts, to verify that the translations are correct. In my preliminary evaluation of the translation of such technical content, the existing approaches make many mistakes due to the technical terminology and sound generally informal.
Approach:

- Extract text from 4 PDFs on machine translation using a programmatic approach (PyPDF2)
- Translate extracted text using an existing advanced translation API
- Use and develop a Spanish-English dictionary for technical terms (e.g., ensembles, self-attention, convolutions, backpropagation) and use it in the translations
- Save the translation in a readable format

Tools/Datasets:
Dataset: A small collection of AI research papers in English and an existing collection of parallel text of AI content (e.g., https://www.ibm.com/docs/es/spss-modeler/18.4.0?topic=networks-ensembles-neural and https://www.ibm.com/docs/en/spss-modeler/18.4.0?topic=networks-ensembles-neural)
Tools: PyPDF2 (for PDF parsing), googletrans or any advanced translation API, an existing Spanish-English AI dictionary (https://github.com/capitalone/AI_Dictionary_English_Spanish)

Expected Outcome: A system capable of translating AI research papers, emphasizing the correct translation of AI-specific terminologies.
Evaluation: Qualitative evaluation of the translations by me and, if time allows, a more thorough empirical evaluation with parallel text.
Programming Language: Python

Workload Justification:

Preprocessing & Cleanup of AI research papers: 5 hours
Dictionary Creation for AI terminologies: 5 hours
Translation System Development: 10 hours
Evaluation and Refinement: 5 hours
Documentation and Demo Preparation: 5 hours
Total: 30