Captain: Warren D. Child (only member)

UIUC ID: wchild2

Email: wchild2@illinois.edu

PLEASE NOTE: My original CMT link had an incorrect email address and is invalid. The email above is correct.

Topic: Japanese Patent Term Retrieval: Identifying the Immediate Context in Which Japanese Patent Terms Are Used for Linguistic Comparison

Language: Python

Other tools:

Google Programmable Search Engine

Kuromoji (Japanese Morphological Analyzer)

Stanza (Python Wrapper for Stanford Core NLP)

Abstract: Japanese patents use a large number of fairly esoteric terms and expressions that are not found in standard dictionaries, and that can therefore be tricky to translate. Such terms are a headache for patent translators, and typically one needs to look at a number of documents where the word is employed to understand its significance. This project seeks to build on top of the standard text retrieval capacities of a major search engine like Google and add functionality that will make it easy to quickly compare sentences containing such a term. Providing laser-focused contextual usage should prove beneficial for technical translators.

Discussion: Unlike normal text retrieval, where one is looking for topic relevance in documents, here one is looking for the correct usage of terms in context. For example, in normal text retrieval, the presence of a term within the title might suggest a higher

level of relevance. But from the standpoint of understanding how a word is used in context, it is actual sentences that one hopes to find. The topic and the term are not necessarily that closely related.

Difficulties: Obviously, one difficulty will be with accurately parsing words in Japanese, which does not employ spaces to parse separate words. Japanese kanji (*hiragana* and *katakana*) can sometimes help to indicate word boundaries, though as often as not, *hiragana* is used in place of *kanji.* Equally troublesome is the tendency to sometimes use hiragana (or even katakana) in place of characters, or to use variant characters for the same character. I will be researching ways to address this issue. It is unclear how successful Kuromoji will be with more esoteric patent terms. If it does not work or does not allow one to supplement their dictionaries, I may have to use another tool (Stanford's Stanza being a possible alternative).

Added Functionality: Taking the search results from the initial search, I will be parsing the retrieved documents to identify context sentences where the critical term occurs. Those sentences will be retrieved in a single document for comparison purposes. Search results that do not have the requisite vocabulary, or only have it in a non-sentence, will be eliminated from the results.

Success Benchmarks: If I am able to routinely prepare clean sets of sample sentences for problematic Japanese patent-specific terms, I will consider this successful. Often, search results return a potentially useful document, but because of document length, it is hard to see where the relevant portion or portions are. Having cleanly parsed relevant sample sentences would be a great time-saver for translators and other linguists.

Hours Expended: Between having to learn the Kuromoji API, possibly Stanza, features of the Google Programmable Search Engine, and writing scripts to parse the relevant parts of the returned documents should take more than 20 hours.

Note: I discussed this topic proposal with the professor, and the idea of adding functionality to Google was actually his suggestion. I am open to more suggestions on how to make this a useful tool for persons like myself who work as translators.