Improved Machine Translation for CN-EN Novels through Targeted Dictionary Substitutions of Idioms and a Multi-Word-Phrase Synonym-Focused Automated Metric

Lisa Liu, Ryan Liu, Angela Tsai   Advisor: Jingbo Shang

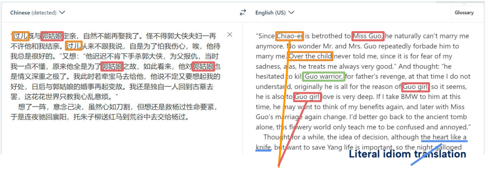
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

BACKGROUND

Machine translation is frequently used in the CN-EN webnovel translation scene, as human translators cannot keep up with an ever-growing body of texts. MT struggles with idioms, slang, and other culturally nuanced phrases found in webnovels. These phrases have meanings beyond the actual words used, but machines often incorrectly translate them literally. As such, our project aims to build a MT model that more accurately identifies and translates Chinese idioms could help reduce the amount of human post-editing needed to produce higher-quality translations, and increase the rate at which CN literature can be accessed by EN audiences.

SAMPLE TRANSLATION

Direct DeepL Translation:



"过儿既与郭姑娘定亲，自然不能再娶我了。怪不得郭大侠夫妇一再不许他和我结亲。过儿从来不跟我说，自是为了怕我伤心，唉，他待我总是很好的。"又想："他迟迟不肯下手杀郭大侠，为父报仇，当时我一点不懂，原来他全是为了郭姑娘之故，如此看来，他对郭姑娘也是情义深重之极了。我此时若牵宝马去给他，他说不定又要想起我的好处，日后与郭姑娘的婚事再起变故。我还是独自一人回到古墓去罢，这花花世界只教我心乱意烦。"

想了一阵，意念已决，虽然心如刀割，但想还是救杨过性命要紧，于是连夜驰回襄阳，托朱子柳送红马到荒谷中去交给杨过。

"Since Chiao-er is betrothed to Miss Guo, he naturally can't marry me anymore. No wonder Mr. and Mrs. Guo repeatedly forbade him to marry me. Over the child never told me, since it is for fear of my sadness, alas, he treats me always very good." And thought: "he hesitated to kill Guo warrior, for father's revenge, at that time I do not understand, originally he is all for the reason of Guo girl so it seems, he is also to Guo girl love is very deep. If I take BMW to him at this time, he may want to think of my benefits again, and later with Miss Guo's marriage again change. I'd better go back to the ancient tomb alone, this flowery world only teach me to be confused and annoyed."

Thought for a while, the idea of decision, although the heart like a knife, but want to save Yang life is important, so the night galloped

Literal idiom translation

Inconsistent names and titles

Pipelined DeepL Translation:

"**Guo'er** is betrothed to **Miss Guo**, naturally he can't marry me again. No wonder Mr. and Mrs. **Hero Guo** have repeatedly forbidden him to marry me. **Guo'er** never told me, because he was afraid that I would be sad, alas, he always treats me very well." And thought: "He hesitated to kill **Hero Guo**, for father's revenge, at that time I did not understand, the original he was all for **Miss Guo**'s sake, so it seems, he is also **Miss Guo** is also a very deep love. If I took the BMW to him, he might want to remember my favor again, and the marriage with Miss Guo would be changed again in the future. I'd better go back to the ancient tomb alone, this flowery world only teaches me to be disturbed."
After thinking for a while, I decided to **put my heartache aside** and focus on saving Yang's life...

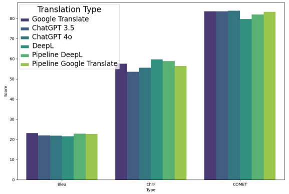
METHODOLOGY

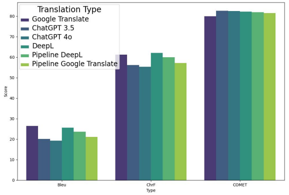
1. Use Jieba to identify names with its built-in dictionary
2. Select sample subset of sentences that do not contain any names identified by Jieba, and use LLM to find and translate any missed names
    a. Weight sentences earlier in the novel higher for selection, as characters are more likely to be introduced then
3. Combine newly identified names with Jieba's name dictionary; replace common suffixes and titles with standardized translation
4. Using names dictionary, replace all instances of names in original text with their translations
5. Using CC-CEDICT, replace all instances of idioms with their translations
6. Use low-cost translation model such as Google Translate to translate remaining text

CONCLUSIONS

1. Pipeline introduces much more consistency in resulting translations
    a. Consistent suffixes and names
    b. Prevent literal idiom translations
2. Compromise of cost/efficiency of NMTs and performance of LLMs
3. Currently only feeds 20% of the text into GPT for finding names
    a. 15% from first ¾, and 5% from last ¼
    b. Rest of text handled by NMTs
4. Can perform better than GPT in word-level evaluations (BLEU, ChrF)
5. Can perform better than NMTs in predicted human reception (COMET)

RESULTS





FUTURE WORK

In addition to improving the MT process with targeted idiom translations, one of our original goals was to also create our own automatic evaluation metric to more accurately judge the translation. Going forward, we aim to build an automated metric focused on multi-word synonyms, using another metric. METEOR, as a base METEOR is an existing metric that performs synonym matching using WordNet's synsets, which only contain unigrams, or single words. We will use a dictionary to form multi-word synsets for another stage of matching.

ACKNOWLEDGEMENTS

- WMT Conference on Machine Translation
- CC-CEDICT dictionary: https://cc-cedict.org/wiki/
- Jieba: https://github.com/fxsjy/jieba