A Comparative Study of English-Chinese Translations of Court Texts by Machine and Human Translators and the Word2Vec Based Similarity Measure's Ability To Gauge Human Evaluation Biases

Ming Qian
Pathfinders Translation,
Interpretation & Research
513 Elan Hall Rd
Cary, NC 27519, USA
qianmi@pathfinders-transinterp.com

Jessie Liu
California Court Certified Interpreter
15421 Hoover Ln
Fontana, CA 92336, USA
jessiel@middlebury.edu

Chaofeng (Joseph) Li
California Court Certified Interpreter
1300 E Main Street. #209G
Alhambra, CA 91801, USA
jl.interpreting@gmail.com

Liming Pals
ATA Certified Translator
Ivy Tower International LLC
3114 Whitetail Ln, Ames, IA 50014
Limingpals@gmail.com

Abstract

In this comparative study, a jury instruction scenario was used to test the translating capabilities of multiple machine translation tools and a human translator with extensive court experience. Three certified translators/interpreters subjectively evaluated the target texts generated using adequacy and fluency as the evaluation metrics. This subjective evaluation found that the machine generated results had much poorer adequacy and fluency compared with results produced by their human counterpart. Human translators can use strategic omission and explicitation strategies such as addition, paraphrasing, substitution, and repetition to remove ambiguity, and achieve a natural flow in the target language. We also investigate instances where human evaluators have major disagreements and found that human experts could have very biased views. On the other hand, a word2vec based algorithm, if given a good reference translation, can serve as a robust and reliable similarity reference to quantify human evalutors' biases beacuse it was trained on a large corpus using neural network models. Even though the machine generated versions had better fluency performance compared to their adequacy performance, the human translator's fluency performance was still far superior. The lack of understanding by machine translators led to inaccurate and improper word/phrase selections, which led to bad fluency.

1 Objective

The purpose of this study is to evaluate the quality of machine translation by comparing the target texts generated by multiple machine translation tools with texts translated by an expert human translator/interpreter.

Three expert human translators/interpreters evaluated the target texts. We also evaluate the word2vec as an algorithm tool to measure the similarity between machine generated sentences and human generated sentences. In addition, we analyzed various quality problems of the machine translation results, their severity levels, and possible causes.

2 Methods

We used a video clip of a judge giving a jury instruction (Pastor 2011) as the test script.

A certified court interpreter with many years of experience interpreted what the judge said in

English into Chinese in real-time. In addition, the same interpreter got the chance to take as much time as she wanted to translate the same content from English to Chinese.

Three machine translation tools (Google Translator, Microsoft-Bing Translator, and Mr. Translation by Tencent) were used to translate the same content from English to Chinese.

Two certified court interpreters and an ATA-certified translator were asked to evaluate the five versions of targeted text generated (three generated by machines, and two generated by a human expert). They were asked to fill out a questionnaire with a 5-level Likert Scale regarding adequacy and fluency (relying on an intuitive understanding of these notions by the evaluators).

Human experts also discussed the various quality issues of machine translated results, their severity levels, and possible causes.

3 Translation and interpretation Results

Due to the limitation on the number of pages allowed, we list five versions of targeted text generated by machines and humans for ten text segments on the following website:

https://sites.google.com/Pathfinders-transinterp.com/mainsite/machine-translation-summit-tables?authuser=0

This is the raw data for the analyses below.

4 Results of Quality Evaluation

4.1 Questionnaires on Translation Adequacy and Fluency

Two California court certified interpreters and one ATA (American Translator Association) certified translator evaluated the adequacy and fluency of the results (Koehn 2017). Adequacy answers the question of whether the translation output conveys the same meaning as the input. Is part of the message lost, added, or distorted? Fluency answers the question on whether the translation output can be considered fluent Chinese or not? This involves both grammatical correctness and idiomatic word choices. Evaluators relied on an intuitive understanding of these notions to make judgments and were asked to provide reasons for sentences on which evaluators had major opinion differences.

The Likert Scale was used for the questionnaires. Evaluators were offered a choice of five pre-coded responses with the neutral point being neither agree nor disagree.

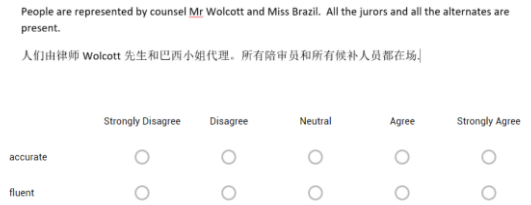

Figure 1: Likert scale evaluation questionnaires are used to evaluate adequacy and fluency of the target texts.

4.2 Questionnaire Results

The results based on feedback from the three evaluators are listed in Table 1. Annotations S1-S10 represent text segments 1 to 10. Check marks (√) represent the votes given by the interpreters and translators. For example, in the "Strongly Agree" cell, √√√ (S1) means that all three evaluators chose to "Strongly Agree" that the translation using the specific translation version has good adequacy or fluency.

The evaluation results showed that compared with the human translator/interpreter, the machine generated versions were of poorer quality (in terms of both adequacy and fluency). In addition, the adequacy in these cases was on average worse than the fluency.

4.3 Adequacy Analyses

For text segment 2, the word "exhibits" was translated as "展览" (the items in an exhibition) by the google translator, while a better Chinese word would be "证物" (forensic evidence).

For text segment 3, all three machine generated versions translated the word "feverishly" as "狂热", while the human translator chose to forgo direct translation and intentionally left out the word. This is because the jury in this case did not really exhibit fanatic enthusiasm.

For text segment 5, all machine generated versions seemed to have a hard time figuring out who the bailiff (court room policeman) was. Google version chose to leave the word in its English form, The Microsoft-Bing version used the transliteration approach (translating the sound of pronunciation), and Mr. Translator generated the Chinese word "联谊会", which means "friendship association", which is obviously a mistranslation.

For text segment 7, the Chinese texts produced by all machine translation tools said that "可能会有一点不同" (there might be some variability), but no explanation was given. The human translator, in this case, mentioned that the (reading)

time might be a little different. Therefore, a reason was given to explain the variability.

For text segment 8, the human translator chose to omit "Page 1 is always the best page", since it was just a small quip by the judge that was not elaborated on. Leaving this phrase in could potentially confuse readers, and for this reason the human translator chose to omit it.

For text segment 10, Google Translator translated the phrase "master set" as "主人套装". Here "主人" means master relative to slave. Obviously, this is not the right word. Microsoft-Bing and Mr. Translator translated the phrase as "主集" which is a good translation. The human translator provided an even better solution "公用文件" (a document set shared by the group).

source language because it is apparent from either the context or the situation. For example, in sentence 7, all machine generated versions used the literal translation "会有一点不同(变化)" (there could be varieties). The human translator added the implicit reason "时间或长或短" (the reading time could a bit longer or shorter).

4.4 Human Adequacy Evaluations and word2vec Similarity Results

In this section, we select instances of adequacy evaluation in which three evaluators had very different opinions. We asked the evaluators to provide the reasons for their choices.

In addition, we compare the machine generated version with the human translation version which

Evaluation Categories	Likert Scale Choices	Google Translator	Microsoft-Bing Translator	Mr. Translator by Tencent	Human Interpretation (real-time)	Human Translation
Adequacy	Strongly Disagree	√√(s1) √√(s2) √√(s3) √√√(s4) √√√(s5) √√(s6) √(s7) √√√(s8) √√(s10)	√√√(s1) √(s2) √(s3) √(s4) √√√(s5) √(s8) √(s10)	√(s1) √√(s2) √(s3) √(s4) √√(s5) √(s7) √(s8)	√√(s10)	√(s8) √(s10)
	Disagree	√(s1) √(s2) √(s3) √(s6) √(s7) √(s9) √(s10)	√√(s3) √√(s4) √(s6) √√(s7) √√(s8) √(s9) √√(s10)	√√(s1) √√(s3) √√(s4) √(s5) √(s6) √(s8) √√√(s9) √(s10)	√(s1) √(s3) √(s5) √√ (s6) √√√(s7) √(s9) √(s10)	
	Neutral	√√(s9)	√√(s2)	√(s2) √(s7) √√(s6)	√(s3) √√(s4) √(s5) √(s9)	√(s8)
	Agree	√(s7)	√√(s6) √(s7) √√(s9)	√(s7) √(s8) √√(s10)	√(s1) √√√(s2) √(s4) √(s5) √(s6) √√√(s8) √(s9)	√(s1) √(s3) √(s4) √(s5) √(s6) √(s7) √(s9) √(s10)
	Strongly Agree				√(s1) √(s3)	√√(s1) √√ (s2) √√(s3) √√(s4) √√(s5) √√(s6) √√(s7) √(s8) √√(s9) √(s10)
Fluency	Strongly Disagree	√(s3) √(s5) √(s6) √(s8) √(s9) √(s10)	√(s2) √(s3) √(s6) √(s8) √(s9) √(s10)	√(s1) √(s2) √√(s5) √(s6) √(s8) √(s9)	√(s10)	
	Disagree	√√(s1) √(s2) √(s3) √(s4) √(s5) √(s6) √√(s7) √(s8) √(s9) √(s10)	√√(s1) √(s3) √(s4) √√(s5) √(s6) √√(s7) √(s8) √√(s9) √(s10)	√(s1) √(s3) √(s6) √√√(s7) √(s8) √(s9)	√(s2) √(s5) √(s6) √(s10)	
	Neutral	√(s2) √√(s4) √(s5) √(s8) √(s10)	√ (s2) √(s4) √(s5) √(s10)	√ (s2) √(s3) √(s4) √(s6) √(s8) √(s9) √(s10)	√√(s4) √(s5) √(s6) √(s7)	√(s10)
	Agree	√(s2) √(s3) √(s6) √(s9)	√(s3) √(s4) √(s6) √(s7) √(s8)	√√(s4) √(s5) √√(s10)	√ (s1) √(s2) √(s3) √(s4) √(s6) √(s7) √(s8) √(s9)	√(s4) √(s8) √(s9)
	Strongly Agree	√(s1) √(s7)	√(s1) √(s2)	√ (s1) √ (s2) √(s3)	√ (s1) √(s2) √√(s3) √(s5) √(s7) √√(s8) √√(s9) √(s10)	√√√ (s1) √√√(s2)√√√(s3) √√(s4) √√√(s5)√√√(s6) √√√(s7) √√(s8) √√(s9) √√(s10)

Table 1: Evaluation Results on Adequacy and Fluency of the Ten Text Segments.

As described in the ATA's (ATA BOD 2019) position paper on machine learning, machines understand neither the source nor the target text. The problems we found in the examples show that the adequacy suffered significantly due to the lack of understanding of context.

On the other hand, the human translator applied the tactics of strategic omission to reduce distraction. For example, the human translator chose to omit the phrase "The first page is always the best" in Sentence 8 because it was a distractor deviating from the main message.

The tactic of explicitation (Vinay et al., 1958/1995 and Gumul, 2006) was used by the human translator as well. This tactic made explicit in the target language what remains implicit in the

has the best adequacy. The comparison is done using word2vec2. Word2vec is a two-layer neural net that processes text (Artificial Intelligence Wiki). Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus. The objective of Word2vec is to group the vectors of similar words together in a vector space. Therefore it can be used to detect the similarity of two sentences. The Chinese word2vec model used can be found on the website below: https://pan.baidu.com/s/1TZ8GII0CEX32ydjsfMc0zw, and the 64-dimension model was trained using the news, Baidu Encyclopedia, and Chinese novels. The python code used to calculate the word2vec similarity between two sentences is listed in table

2. The similarity score is between 0 and 1 and a higher score indicates higher similarity.

```
import gensim
import jieba
import numpy as np
from scipy.linalg import norm

model_file = 'word2vec/news_12g_baidubaike_20g_novel_90g_embedding_64.bin'

model = gensim.models.KeyedVectors.load_word2vec_format(model_file, binary=True)

def vector_similarity(s1, s2):
    def sentence_vector(s):
        words = jieba.lcut(s)
        v = np.zeros(64)
        for word in words:
            v += model[word]
        v /= len(words)
        return v

    v1, v2 = sentence_vector(s1), sentence_vector(s2)
    print(v1, v2)
    return np.dot(v1, v2) / (norm(v1) * norm(v2))

strings0 = [
    '随后我们会做简短的休息',
    '我们会做简短的休息'
  ]

print("Sentence 0 Vector Similarity Results Below:");
print(vector_similarity(strings0[0], strings0[1]));
```

Table 2: Python code to calculate word2vec based sentence similarity.

Table 3 shows an example in which the three evaluators exhibit disagreement. Evaluator 1 believed that the Google version is literal and accurate, while evaluator 2 and 3 observed some mistranslation and grammar/syntax errors. We found that some evaluators can have very biased view. For example, evaluator 2 believed that "才能阅读完" (finish reading in 28 minutes and 14 seconds) is very different from "需要28分14秒才能阅读" (needs 28 minutes and 14 seconds to read). Realistically, those two expressions are not that different from each other.

Since word2vec measures cannot be performed on empty spaces and punctuation, we measured the clause similarities between the google result and the result generated by the human translator (as the reference). Two clauses had similarity scores of around 0.88 and 0.78, while one clause "这会有所不同" (that vary a little bit can) had a very low similarity score (0.1058) compared to the human version "时间或长或短" (the reading time could be longer or shorter). This is in line with the comment by evaluator 3. The major difference is that the human version mentioned "时间" (time), while the Google version did not specify what varies. If we add the time ("时间") to the Google generated clause and change it to "时间会有所不同", then the word2vec based similarity changes from 0.1058 to 0.6617. That shows that the word2vec provides a very reliable similarity measure, and a good indicator of human bias. In this case, only the evaluator 3's opinion was supported by the word2vec results.

Original English: It should take about 28 minutes and 14 seconds for me to read these instructions to you. That's going to vary a little bit but it'll take just about a half an hour for me to read the instructions to you.
Google translator result: 我需要大约28分14秒才能阅读这些说明（word2vec similarity calculated against human generated translation = 0.8806）。这会有所不同（word2vec similarity calculated against human 0generated translation = 0.1058），但我需要大约半小时的时间（word2vec similarity calculated against human generated translation = 0.7865）才能阅读说明书。
Evaluator 1 -> Agree: The translation is accurate and literal, but without taking into consideration the English conversation style and properly converting that to the target language, an accurate translation doesn't necessarily convey a message accurately. Evaluator 2 -> Disagree: Here "to read" is translated to "才能阅读", which is a literal translation, but in Chinese, a more accurate translation should be "才能阅读完", meaning ""to finish reading". Evaluator 3 ->Strongly Disagree: 1. Missing "to you". 2. Literal translation of "vary a little bit" which can cause confusion.

Table 3: Google translator result for sentence No.7, evaluators' Likert scale evaluation and comments, and word2vec similarity measures.

Table 4 shows another example in which the three evaluators disagreed. Evaluators 1 and 2 believed that the translation is not adequate because it should be a polite request instead of a conditional statement. On the other hand, Evaluator 3 believed that the adequacy is acceptable. Using the word2vec measure, we measured the

clause similarity between the Tencent result and the result generated by the human translator "请大家翻到指示文件的第一页" (as the reference), and the similarity score is 0.5166. If we change the Tencent result based on what Evaluators 1 and 2 suggested, the new similarity score is 0.6507. This shows that Evaluators 1 and 2 had a valid point, but while the improvement is significant, it is limited. Again, the word2vec based similarity measure

Original English: So if you would turn to page 1 which is always the best page. Page 1 of the instructions. Post-introductory series.
Tencent Mr Translator: 因此，如果你想翻到第一页(0.5166)，这始终是最好的一页。说明第1页：审判后介绍性系列(0.6646)。
Evaluator 1 -> Strongly Disagree: The translation is too literal to keep the intended meaning intact. "If you would turn to page 1" in the sentence isn't a conditional statement, although grammatically incorrect, it's a polite way to give a command, to tell the jurors to do something, and such command shall be reflected in the translation, instead of a conditional statement. Evaluator 2 -> Disagree "so if you would turn to page 1" is translated to "如果你想翻到第一页", which means "if you want to" But the original meaning is basically a polite way ofrequesting jury member to "please turn to page 1". A more accurate translation is "请翻到" Evaluator 3 -> Agree: The overall quality is OK. However, there are a few places that can be improved. Literal translation is an issue.

Table 4: Tencent Mr. translator result for text segment No.8, evaluators' Likert scale evaluation and comments, and word2vec similarity measures.

These examples showed that a word2vec based algorithm, if given a good reference translation, can serve as a robust and reliable similarity reference to quantify human evaluators' biases because it was trained on a large corpus using neural network models.

Human Evaluator's opinion	Word2Vec similarity measure given a good reference translation	Human evaluator's bias
Table 3 example:	Similarity score changes from	The evaluator's opinion is confirmed
"这会有所不同" (that vary a little bit can) is not accurate given the context	0.1058 to 0.6617 after adding the word "time" ("时间")	by the Word2Vec result (the similarity score changed by more than 0.5)
Table 4 example: The sentence "if you would turn to page 1" should be translated as a request instead of a conditional statement.	Similarity score changes from 0.5166 to 0.6507 after the sentence was translated as a request	The evaluator's opinion has some merits. But the improvement is as significant based on the Word2Vec result (the similarity score only changed by less than 0.15))

Table 5: Word2vec based similarity measures serve as a robust and reliable similarity reference to quantify human evaluators' biases.

4.5 Fluency Analyses

The subjective evaluation results (Table 1) show that the fluency performances of machine translators were not as bad compared to their adequacy performances. Nevertheless, their fluency performances were still inferior to the human translator's.

The machine translated results for Sentences 6, 8, and 9 by all three machine translators were categorized as the worst by human evaluators, since because majority of evaluators answered "strongly disagree" or "disagree" in regards to these sentences being fluent and adequate translations.

For Sentence 6, the human translator chose to omit the phrase "As I'm going to be telling you in just a few moments". This is an intelligent move because obviously the judge is telling them right at that moment, not a few moments later. Also, for the sentence "It was provided to each of you", the human translator used "刚才就给你们提供过的" to represent the past tense while all three machine translators failed to reflect the past tense. This is important because unlike English, the form of a Chinese verb never changes, regardless of whether it is present, past, or future tense. The past tense has to be represented using a timing word such as "刚才".

For Sentence 8, by leaving out the phrase "Page 1 is always the best page", the human translator avoided confusing the readers, and made the sentence as a whole flow much better. In addition, all machine translators used "做完那以后" (after I am done), which is not a conventional Chinese expression. "读完以后" (After finishing the reading) is a better Chinese expression in this case.

For Sentence 9, the human translator used "你们可能需要休息一下了" while all three machine translators used "你们可能需要它". In Chinese, "它"(with the meaning "it") is usually not used in this context. "休息一下" (with the meaning "take a break") better matches Chinese convention.

Again, our observation is that machine translators lack understanding of the source and target texts, and the lack of understanding context and background led to inaccurate and improper word/phrase selection, which led to unnatural flow (bad fluency).

5 Conclusion

In this comparative study, we used a jury instruction scenario to test multiple machine translation tools and a human translator with extensive court experience. Three certified translators/interpreters evaluated the target texts generated using adequacy and fluency as the evaluation metrics.

We found that machine generated results had much worse adequacy performance compared with their human counterparts. Since machine translation tools understand neither the source nor the target text, unlike the human translator, they cannot minimize the misunderstanding across language and culture. Human translators can use strategic omission and explicitation strategies such as addition, paraphrasing, substitution, and repetition to remove ambiguity.

We also evaluate the word2vec as a tool to evaluate the similarity between machine generated results and human generated results. Word2vec trained neural network models on a large corpus to map words onto a vector space. Therefore it can be used to detect the similarity of two sentences. We use multiple examples to show that the word2vec serves as a robust and reliable similarity reference to quantify human evalutors' biases.

Even though the machine generated versions had better fluency performance relative to their adequacy performance, the human translator's fluency performance was still far superior. The lack of understanding by machine translators led to inaccurate and improper word/phrase selections, which led to unnatural flow (bad fluency).

References

The ATA Board of Directors (BOD). 2019. *ATA Position Paper on Machine Translation: A Clear Approach to a Complex Topic,* https://www.atanet.org/chronicle-online/extra/ata-position-paper-on-machine-translation/

Artificial Intelligence Wiki, *A Beginner's Guide to Word2Vec and Neural Word Embeddings,* https://skymind.ai/wiki/word2vec.

Gumul, Ewa, 2006. Explicitation in Simultaneous Interpreting: A Strategy or A By-product of Language Mediation. *Across Languages and Cultures* 7 (2), pp. 171-190.

Pastor, Michael (Judge), Conrad Murray Trial: Judge Instructions to Jury. 2011. https://www.youtube.com/watch?v=GRf9bZkE-mE

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. *In Workshop on Neural Machine Translation. Vancouver, BC. ArXiv*: 1706.03872. 171-190.

Vinay, J-P. and Darbelnet, J., 1958/1995. Comparative Stylistics of French and English: A Methodology for Translation, Amsterdam/Philadelphia: John Benjamins.