Jichen Dai -- HW3

Adversarial Machine Translation Inputs

For this task, I translated sentences between Chinese, Japanese and English.

	Input	Machine Translation	Correct Translation
1	江山易改，本性难移	It's easy to change, but its nature is hard to change.	One's nature will never change.
2	未离海底千山暗	Qianshan is dark under the sea.	Before the moon rise, all the mountains are dark.
3	隔墙有耳	The wall has ears	Someone is tapping
4	随便	Casual	as one pleases
5	鬼知道	ゴーストは知っている	誰も知らない
6	太陽と月は同じ空にあります	太阳和月亮在同一天	日月同天
7	かまちょ	Kamacho	I am bored. Let's go out together.
8	空気読めない	Unable to read the mood.	Lack of interpersonal skills.
9	Irons in the fire	熨斗着火	进展神速
10	Lend me your ear	借给我你的耳朵	请认真听我说
11	Excuse my French.	不好意思，我法语讲的不太好	抱歉我说话不好听
12	It's all Greek to me	对我来说都是希腊文	完全看不懂
13	LOFT (Rolling On the Floor Laughing)	床で転がる笑い	笑いながら転がる
14	TFTI(Thanks For The Invite)	TFTI	招待してくれてありがとう

Error analyze:

I summarized **3** types of errors, and I will analyze them separately:

1. Sentences based on a specific cultural background or habit:

When we use slang in a certain language, it is hard for Google Translator to give the accurate translation. Such as case 3, 4, 5, 6, 8, 9, 10, 11, 12.

Possible Reason: The output of Google translator is trained mostly on normal articles, which means the translation model didn't get enough knowledge about these specific cases. This is reasonable, because trying to training the model to satisfy all cases may leads to a overfitting.

Additionally, there are some subtle concept that don't exist in another language. For example, in Chinese the word "你" and "您" both has meaning "you", but the second one is often used to show a kind of respect. However, in English, "you" don't has a form to show respect. So both "你" and "您" will be translated to "you".

2. The input text missing some grammatical elements:

When some grammatical elements is missing such as the subject, it might be easy for a man to understand what is the missing subject, but it is hard for a model to figure it out.

Possible reason: The translation model needs to analyze the component structure of the sentence. If the model want to guess what the missing component is, it needs to understand the background of this sentence. However, some certain forms of sentence is rare in teh database and it is easy for model to ignore these cases.

Moreover, sometimes, the sentences input by a user might be grammatical wrong and doesn't match any grammar or convention.

3. Abbreviation

It is very easy for Google Translator to make mistakes on abbreviations, especially those uncommon abbreviation or abbreviation of uncommon language.

Possible Reason: Similar to problem 1, the model is trained on normal sentences on the internet. Those uncommon cases is rare and trying to cover all of them may leads to overfitting.

Additionally, sometimes an abbreviation has multiple meanings based on different situation and only the user know which one he want. For example: the abbreviation ACD can be interpreted as Adelaide College of Divinity or Activity-centered design.

Ideas for Improvement

I have **2** ideas for improving this model:

1. Training multiple model and each model focuses on a specific type of situation. For example, one for normal text, one for classical literature, one for academic paper contents, one that specialize in handling missing components. When get the input, we first determine which class this sentence belongs to, then we throw this sentence to the model that is expertise in this kind of sentence. This gives the model a chance to handle those uncommon cases.

2. Updating the database to cover those rare case and subtle differences.