Mining User-generated Bitext

— Constructing Parallel corpora from Movie Subtitles

Yizhan Wu, Bowen Chu, Lingbo Hu
Department of Computer Science and Engineering
State University of New York at Buffalo
{yizhanwu,bowenchu,lingbohu}@buffalo.edu

Fall 2018

Abstract

This project aims to introduce a methodology for constructing aligned English-Simplified Chinese corpora from movie subtitles. Subtitles that consist of two languages usually provide viewers with alignment of sentences manually done by the author. Since the common length-based algorithm for alignment is not desirable when provided with short spoken sentences, we present a simple methodology to use statistical lexical cues to align the subtitle. Along with the use of machine translation system, we will provide a viable solution for improving the quality of user-generated bitext.

1 Introduction

1.1 Bilingual Movie Subtitles

A parallel text is a text placed alongside its translation or translations, and in the field of linguistic studies, a bitext is a merged document composed of both source and target language versions of a given text. Bilingual movie subtitle is a common instance of bitext. A typical movie subtitle file format is based on the time. They are characterized by an identifier, a time frame and a sentence.

Usually, for a bilingual subtitle, text alignment is done manually by the author of the subtitle. However, since subtitle translations are not necessarily done by the same translator, directly extracting parallel corpora from movie subtitles can lead to noisy results, especially if we take rephrasing and different expression into consideration. (Wu, 1994) In addition, different translator has different culture background and different writing habits. When it comes to individual words, in English grammar we use prefix and suffix to define different usage of a word, but in Chinese even the same word can have different meaning in different situations and context. That's when we will introduce lexical cue extension as a solution to the issue. (Tiedemann, 2007)

1.2 Text Alignment

Text alignment is an important task in Natural Language Processing (NLP), it can be used to support many other NLP tasks. There are a lot of existing researches that have been done in the

area of bilingual sentence alignment using parallel corpora (Brown et al., 1991), and movie subtitle is one of which. Movie subtitles that consist of two languages are used as parallel corpora, and a large parallel corpus with good quality has great value in the field of statistical machine translation. In addition, not only is parallel data useful for training and testing machine translation systems, many other NLP tasks have also benefited from these kinds of resources.

1.3 Machine Translation

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. Both statistical and neural machine translation system rely on a large amount of parallel data that is aligned in sentence level for training the models. After the training of the translation system, it is capable of assist human translators and produce usable output. It is often the case that larger amounts of corpora lead to higher quality models, therefore parallel corpus that is of high quality is of great importance to machine translation system both in training and in improving the model, and bilingual movie subtitles are equally valuable in this field of studies.

1.4 Quality Control

Data quality plays an important role in machine translation models. Since the subtitles are from numerous internet resources, the results are prone to be noisy if quality control is not applied. One of the most common defects in parallel corpora is high dis-match between target and source language. It is also notable that translators tend to perform translation of certain sentences solely based on the deeper meaning of the context, which is not a bad thing in understanding context, however it introduces ambiguity and noise when used to train machine translation system. Therefore, quality control will be applied once all previous processing steps are finished.

2 Proposed Method

For this project, we will propose a simple method that consider both time frames and lexical content of the subtitles to automatically align sentence pairs. The purpose of our project is to improve the traditional method of alignment, and construct aligned parallel English-Simplified Chinese corpora from movie subtitles.

2.1 Existing Case Study

There are already a handful of researches that had been done in the area of constructing parallel corpora from movie subtitles, namely from English-Simplified Chinese subtitles. In our own research for this project, we reached out to several papers that were published online and utilized resources found towards our own project plan.

An article publish by Baobao Chang in 2004 discussed the application of constructing Chinese-English parallel corpus(Chang, 2004). The author acquired electronic Chinese-English bilingual text online to use as parallel corpora, while not necessarily the same as movie subtitles, it did provide reference value to our project, as online bilingual text are usually notably more noisy than movie subtitles. The author also provided us with the idea of searching for lexical cue in the context

to better align sentences.

Another article published in 2014 approached a similar topic of using dual subtitles as parallel corpora(Shikun Zhang, 2004). Though not considering machine translation as part of their research, they did provide a novel vision in extracting parallel corpora, and we benifit from their existing research find greatly when consider approach to solve our issue. Meanwhile, statistical machine translation is considered as a viable approach to constructing parallel corpus in another 2014 article(Liang Tian, 2014), where the researcher introduced a high quality parallel corpus designed for machine translation research. Case studies above have inspired our methodology in approaching and solving issues around the task given.

2.2 Method Analysis

Previous algorithms are solely based on the time frame, which is reliable at times, but it ignores the richer information in the context. To improve alignment accuracy, we need to consider the lexical content. Anchor points are needed to be found in the context of subtitle pairs, the most relevant word pairs will be decided afterwards, ultimately alignment for a whole sentence will be completed. After sentence alignment is completed, aligned English subtitles will be the input of the post-trained machine translation system to create a aligned corpus using machine translation. Quality check will be performed post translation to better the results by removing duplicate sentences and excess symbols.

3 Project Procedure

3.1 Pre-processing of Movie Subtitles

Before sentence alignment can be applied, the subtitle file needs to undergo a few pre-processing steps. Tokenization for the English subtitle and tokenization as well as segmentation for the Chinese subtitle will be performed for each movies.

3.2 Subtitle Alignment

After word tokenization and segmentation is done for both English and Chinese subtitle files, sentence alignment will be performed. We will use Hunalign as our main tool in assisting us in sentence alignment for this project. The input files of Hunalign is two different txt files of two languages which only contain the text of the subtitles. The output will be the alignment result of these two flies, which is stored in one file. Then the text in each language should be rematched with their timeline. Two timelines from subtitles can formed a new timeline. The next step is to do an alignment between English subtitle and the other subtitles. After that, these two alignment were combined in one file, which will be the final result of alignments.

3.3 Machine Translation

Moses Baseline System is used in our project as statistical machine translation system. Open Subtitles paraphrase corpora for English and Simplified Chinese were used as our training data, with over 10,000,000 lines in total. After building, training and tuning our machine translation

system, we will use our own English subtitle as testing data to acquire the translation result. Quality check will be done after machine translation to further enhance the result.

3.4 Quality Control

Quality control aims to remove low-quality sentences before training MT systems. Since we already used high-quality parallel corpora named Open Subtitles paraphrase corpora to train the MT system, we are now using quality control to clean our own source files before put them into MT system.

To implement quality control, we will use the toolkit named parallel-corpora-tools on Github.

Firstly, we will install all the prerequisite environment like Python with langid.py, PHP, Moses system and Subword NMT. As we only need to clean the source English subtitles, we will use Monolingual Corpora Tools.

Then we run the script 0-do-it-all.sh to process all source files to make it ready to be put into our trained MT system to translate target file containing Chinese to make our own parallel corpus.

4 Conclusion and Analysis

Alignment Result: The table shows the subtitle alignment result for Sky Scraper. The first column is the timeline, the second column is the English subtitle, the third and fourth column are different Chinese translation subtitle which is translated by different translators.

	00:01:25,458 --> 00:01:27,027	Suspect is wanted in connection with the murder	嫌犯因谋杀员警被通缉	犯罪嫌疑人涉嫌谋杀一名警务人员
	00:01:27,028 --> 00:01:28,028	of a police officer.	NaN	NaN
	00:01:28,029 --> 00:01:30,163	He should be considered armed and dangerous.	他是武装危险人物	应该持有武器 是极度危险人物
	00:01:30,164 --> 00:01:32,132	I'm not coming out, you hear me?	我是不会出去的，听见没?	我不会出去的
	00:01:32,133 --> 00:01:33,532	I'm not coming out.	我要所有人走开	明白吗 我不会出去的
	00:01:33,533 --> 00:01:37,403	I want everyone gone... you, your men, snipers, everyone.	你和你的人马都给我滚	你们全都撤走 还有狙击手 都滚开
	00:01:37,404 --> 00:01:39,873	If I don't see taillights in the next five minutes,	要是五分钟内没离开	五分钟之后你们还不走的话
	00:01:39,874 --> 00:01:42,142	you're not gonna like what happens next.	我不敢保证接下来会发生什么事	场面可就不太好看了
	00:01:42,143 --> 00:01:45,049	Ray?	雷伊?	雷
	00:01:46,580 --> 00:01:48,148	Can you get him back on the phone?	能让他再接电话吗?	继续和他谈谈吗
	00:01:48,149 --> 00:01:50,222	He's done talking.	他不肯谈了	他已经不想谈了
	00:01:51,518 --> 00:01:52,651	Do it.	上吧	NaN
	00:01:52,652 --> 00:01:54,959	Yes, sir.	遵命	行动 是 长官
	00:02:25,686 --> 00:02:27,520	Gold Unit, this is HR1.	黄金小队，这是HR1	救世主呼叫金骑士 准备行动
	00:02:27,521 --> 00:02:29,328	Move to green.	继续前进	NaN
	00:02:38,799 --> 00:02:41,473	Ma'am, come with me.	女士，过来	女士 请跟我来
	00:02:43,237 --> 00:02:46,106	FBI, show me your hands!	联邦探员，把手举起来!	联邦调查局 举起手来
	00:02:46,107 --> 00:02:48,475	Hands!	快	NaN
	00:02:48,476 --> 00:02:49,982	Turn around!	转身	转过来
	00:02:52,680 --> 00:02:54,353	Oh, shit.	该死	见鬼
	00:02:56,984 --> 00:03:00,187	- Got him clean, boss. - Ben, no.	我瞄准他了，老大	头儿 可以射击了 本 不行
	00:03:00,188 --> 00:03:01,894	He's not carrying.	班，不行，他没武器	他没有武器
	00:03:07,162 --> 00:03:09,263	Ray. Ray, look at me.	雷伊，看着我	雷 雷 看着我
	00:03:09,264 --> 00:03:13,134	It's all over.	结束了	都结束了
	00:03:13,135 --> 00:03:16,137	I didn't want it to end like this.	我也不想这样	我不想这样结束 不应该这样结束的
	00:03:16,138 --> 00:03:17,972	It's not supposed to end like this.	不应该是这样的	NaN
	00:03:17,973 --> 00:03:21,914	Put your son down and step away.	放下你的儿子，退开	把你儿子放下 跟我们走吧
	00:03:23,511 --> 00:03:25,545	I'm sorry.	对不起	NaN
	00:03:25,546 --> 00:03:27,487	It's okay.	没事的	我很抱歉 没关系
	00:03:33,155 --> 00:03:36,129	Good.	很好	很好
	00:03:37,159 --> 00:03:39,896	No!	不!	不
	00:03:43,198 --> 00:03:45,833	Command, this is Navy Alpha two-one-niner	NaN	NaN
	00:03:45,834 --> 00:03:48,573	en route to Beaufort Naval Hospital.	NaN	NaN
	00:03:51,106 --> 00:03:54,147	I need more saline over here!	海军阿尔发219正前往博福特海军医院	海军A队219号呼叫指挥部 正在前往博福特海军医院
	00:03:56,111 --> 00:03:58,779	Dr. McCullen, Surgery, please.	美国海军部	美国海军部
	00:03:58,780 --> 00:04:01,521	Three, two, one, lift.	三、二、一，抬	321 抬
	00:04:04,753 --> 00:04:07,358	He's ready for you, Lieutenant.	交给你了，上尉	都交给您了 上尉

Figure 1: Alignment Result

Quality Control Result: The table shows the subtitle quality control result for La La Land. All the

duplicate text are removed.

[' And when they let you down']	[]
[' The morning rolls around']	[' The morning rolls around']
[' It's another day of sun']	[' It's another day of sun']
[' It's another day of sun']	[]
['“ It's another day of sun Sun, sun, sun, sun”']	['“ It's another day of sun Sun, sun, sun, sun”']
[' It's another day of sun']	[]
[' Just one more day of sun']	[' Just one more day of sun']
[' It's another day of sun']	[]
[' Another day has just begun']	[' Another day has just begun']
[' It's another day of sun']	[]
[' It's another day of sun']	[]

Figure 2: Quality Control Result

Machine Translation result: below shows the result when we put text post-quality-control to our machine translaion system.

It's another hot and sunny Southern California day.	这是另一个热和阳光灿烂的南加州日子。
The temperature in downtown Los Angeles is 29°C...	洛杉矶市中心的气温是29°C .....
And at night will drop to...	并且晚上会降到.....
I think about that day	我想到那一天
I left him at the Greyhound station	我把他留在了灰狗站
West of Santa Fe	西圣达菲
We were 17 but he was sweet and it was true	我们17岁，但是他很甜蜜，这是真的
Still I did what I had to do	仍然我做了我一定做的事情
Cause I just knew	因为我才知道
Summer: Sunday nights	夏天：周日晚上
We'd sink into our seats right as they dimmed out all the lights	我们会沉入我们的座位，当他们暗淡所有的灯光
A Technicolor world made out of music and machine	一个Technicolor世界组成音乐和机器
It called me to be on that screen	它叫我在那个屏幕上
And live inside each scene	和住在每个场景里面

Figure 3: Machine Translation Result

References

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176, Berkeley, California.

Chang, B. (2004). Chinese-English Parallel Corpus Construction and its Application . In *The*

18th Pacific Asia Conference on Language, Information and Computation (PACLIC 18), pages 283–290, Waseda University, Tokyo.

Liang Tian, Derek F. Wong, L. S. C. P. Q. F. O. Y. L. S. L. Y. W. L. W. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In *Dual Subtitles as Parallel Corpora*, pages 1837–1842, University of Évora, Portugal.

Shikun Zhang, Wang Ling, C. D. (2004). Chinese-English Parallel Corpus Construction and its Application . In *Dual Subtitles as Parallel Corpora*, pages 1–6, Carnegie Mellon University, Pittsburgh, PA, USA.

Tiedemann, J. (2007). Improved Sentence Alignment for Movie Subtitles. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*, pages 582–588, Borovets, Bulgaria.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 80–87, Las Cruces, New Mexico.