**Capstone Project** **Document**

# Neural Machine Translation (中英神经机器翻译)

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



# Problem statement

Although big players like Google Translate and Microsoft Translator offer near-accurate, real-time translations, some “domains” or industries call for highly-specific training data related to the particular domain in order to improve accuracy and relevancy. Here, generic translators would not be of much help as their machine-learning models are trained on generic data. These applications are used by small, medium and large enterprises. Some organizations offer multi-domain translation services–that is, customizable solutions across multiple domains–and other organizations offer translation solutions only for a specific domain. These solutions, although automated for the most part, still depend on human translators for pre- and post-editing processes. Some fields that warrant domain-specific machine translation solutions are:

* Government
* E-learning
* Software & technology
* Military & defence
* Healthcare
* Finance
* Legal
* E-discovery
* Ecommerce

As one of the education institution, during the Covid-19 break, our lectures are being delivered online to accommodate the change. For offshore teaching, currently using MS Teams platform replaces the face-to-face intensive workshop delivery locally conducted in China with a interpreter presented on site along with a lecturer. On the Ms Teams platform, the interpreter is often heavily loaded with the translation demand both from the lecturer and students for smooth communication to achieve efficient interaction. Thus, the project is going to investigate if we can reduce interpreters ‘s cognitive load/workload in the live streaming teaching without a drop in communication quality by exploring domain-specific machine translation model and see if the accuracy level is acceptable for online teaching scenario.

# Industry/ domain

This project lies in education industry, but there's no limit to any scenario where requires domain-specific translation e.g. conferences, healthcare, social media. In high education industry, it requires to provide equal opportunity for people with hearing impairment , potentially, the project will also enhance the learning experience for students with hearing impairment.

# Stakeholders

As this project was aiming to automate work task and improve efficiency and accuracy of the live Chinese-English translation by rapidly evolving machine translation technology, students, lecturers, translators and managers are our stakeholders. Students, lecturers can benefit from better communication experience. Translators’ workload will be reasonably reduced and potentially improve their translation quality, they can focus on prior or post machine translation editing and being more creative. For the managers, it will potentially reduce the labour cost of translators’ and provide better service for our students.

# Business question

* Can we reduce interpreters ‘s cognitive load/workload in the live streaming teaching without a drop in communication quality ?
* Is NMT a solution? Will the machine translation output make sense to human being (e.g. BLEU)>30?

# Data question

* How to prepare domain-specific data ?
* If NMT a solution, what’s the accuracy(BLEU score) can we achieve using available data?

# Data

The initial plan was to construct domain-specific English-Chinese pairs from textbooks (English version and Chinese version), lecture slides(English &Chinese), lecture recordings(English and Chinese) with subtitles, lecture recordings (English and Chinese) with both languages spoken.

However, due to time constraint, the data collection process will way beyond the project time allowed. Alternatively , as a compromise , the data used for this project comes from the link below:

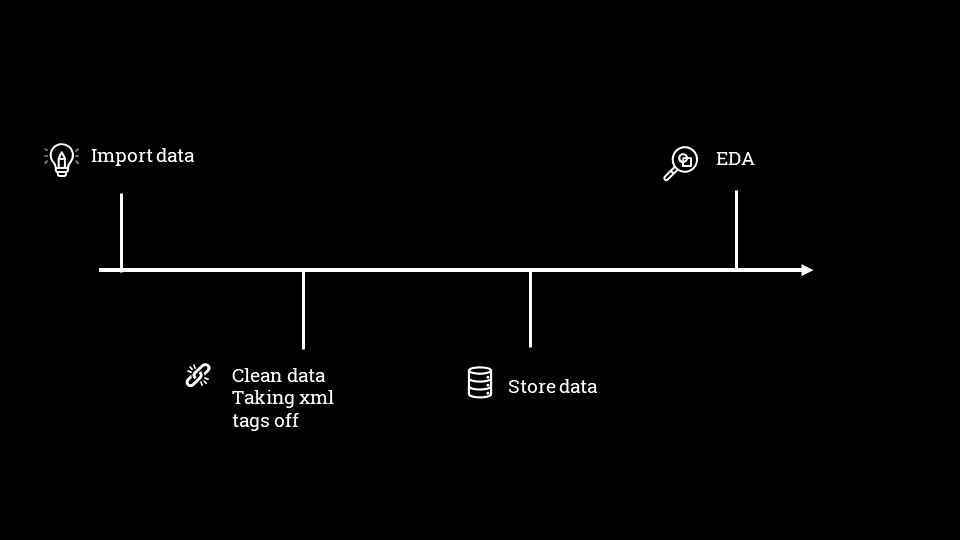
<https://wit3.fbk.eu/2017-01-c>

It is a well known dataset for machine translation, but as the dataset has little relevance to the domain-specific contents, therefore, this project will be just experimental for the real project process.

Data science process

## Data analysis

Data pipeline to wrangle the raw data as below:



Some highlights of the Exploratory Data Analysis (EDA):

|  |  |
| --- | --- |
|  | A picture containing qr code  Description automatically generated |
| Fig 1.Matched pair document wordcloud (English) | Fig 2. Matched pair document wordcloud (Chinese) |



Fig 3. Dataframe after data processing

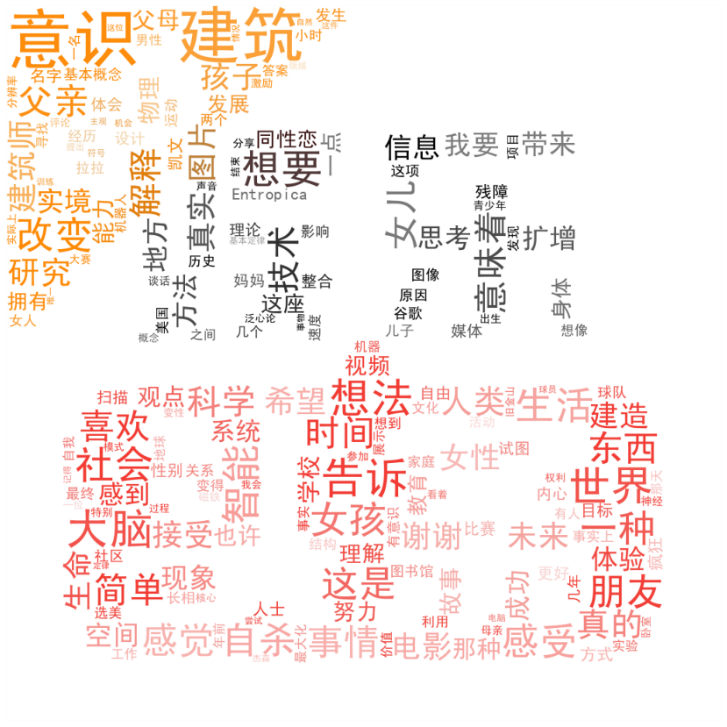


Fig 4. Matched pair document wordcloud (Chinese) in custumized shape

In a summery, the pipeline is reusable for taking xml tags off and then stored for tokenization and embedding of the paired texts.

## Modelling & Outcomes

Model generated & used along with their performances are like the table below:



RNN encoder without & with attention model trained on the training data, the accuracy score (BLEU) were 4.26, 2.15 respectively, which generated very poor output and were not able to produce any meaningful translation. I have also attempted on RNN with self attention model, but with no success due to error:

” RuntimeError: CUDA out of memory. Tried to allocate 2.00 MiB (GPU 0; 14.76 GiB total capacity; 13.72 GiB already allocated; 3.75 MiB free; 13.72 GiB reserved in total by PyTorch)”

Therefore I discarded this try and went on more robust pretrained model of ….

the result seemed very promising with accuracy score() of 30.42. Personally, I think in some case, it can translate better than some translators.

Working through this project, I have found that, for language translation models building , it requires high computation power due to its complexity and size of vocabulary. This project was run on Google colab pro with GPU and High RAM turning on.

Based on the rules that no need to reinvent the wheels, I can conclude that pretrained model from big player such as Google, Microsoft, Facebook should be used as base model and then we can fine tune the model with domain-specific data. Unfortunately, in this project, the domain-specific data are not in machine readable format and can’t convert them to machine readable format in short time.

## Implementation

In the future, once the model was fine-tuned with domain-specific data, it could link with other models such as speak to text to generate an Online Teaching Translation APP.

# Data answer

Due to time constraint, domain-specific data wasn’t available , comprised with simpler paired English-Chinese dataset (IWSLT 2017) well-known for machine translation.

Although, the accuracy level was poor for models built on training data, however, it seems being able to produce satisfied outcome if we use pretrained models(BLEU>30).

# Business answer

I am confident that NMT will be a big game change for translation tasks, once the models are fine-tuned with domains-specific data, it could potential outperform some experienced translators and no double it will reduce the existing translators’ workload , which might lead to less cost of online-delivery, higher learning experience for students and lectures.

# Response to stakeholders

Overall, NML will benefit all with potential lower cost and better user experience, eventually, it will contribute to high quality of teaching. I suggest that we are continuing to collect the teaching contents data, make a budget for a small project to convert existing data to machine readable data. It will be a step closer to our final product of Online Teaching Translation APP.

# End-to-end solution

Online Teaching Translation APP

# References

* Data files:

**Capstone Part1 #Folder**

2017-01-trnted #Zh-En text

Capstone\_Part1\_visualisation\_final\_Hongmei.ipynb

entertainment.jpeg

simhei.ttf

stopwords.txt #Chinese

**Capstone Part2** **#Folder**

Capstone\_Project\_Part2\_Hongmei\_final.ipynb

lstm\_w\_att\_dec\_1\_layer.pth #model

lstm\_w\_att\_enc\_1\_layer.pth #model

lstm\_wo\_att\_dec\_1\_layer.pth #model

lstm\_wo\_att\_enc\_1\_layer.pth #model

**Capstone Part3** **#Folder**

Capstone\_Part3\_Huggingface\_ZH\_EN\_final\_Hongmei.ipynb

* resources used in the project

<https://www.jianshu.com/p/37648e91bdc5>

<https://huggingface.co/transformers/>

<https://www.youtube.com/watch?v=jOf_AsBsrHA&list=WL&index=113>

<https://github.com/ZeweiChu/nmt-seq2seq/blob/master/pytorch/seq2seq.ipynb>

M. Cettolo, C. Girardi, and M. Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In Proc. of EAMT, pp. 261-268, Trento, Italy. pdf, bib.