**Documentation:**

Automatic Translation for Medical Terminologies with NLP Techniques

# 05-09-2021

****

# Toolkit

We use fairseq for developing our machine translation system. fairseq (Ott et al., 2019) is a sequence modelling toolkit that allows researchers and developers to train custom models for translation, among other tasks.

Our methodology is based on transfer learning. We use a pre-trained model (which is trained on a large general textual corpus) and then fine-tune (continue training) on a specialized dataset, in our case medical terminologies. For the pre-trained model, we select the CNN one offered by fairseq.

# Training sets

## 3rd round

The 3rd round includes the datasets mentioned in the LOUHI paper, with removing the ICPC dataset. ICPC has a problem with synonym labels ( different structure for English and French labels.)

## 4th round

The 4th round includes the dataset used in the 3rd round, plus the PatTR corpus.

## 5th round

Other datasets like EMEA or Medline, examine

The list of the final and best methods for biomedical machine translation is listed here:

- <https://github.com/biomedical-translation-corpora/corpora>

6th round

Transformer model

# Validation

We have gathered a sample of labels that already carry a human-verified translation. Nevertheless, this needs to be seen with caution as an expert’s opinion on a translation may not be necessarily considered as uncontested.

24k created from the previous paper

ATIH ground truth

# Truecasing

Trained a model with Pubmed French, and fixed errors with disease names.

# Metrics

Apart from BLEU we have evaluated the quality of the translation with BLEU2VEC.

|  |  |
| --- | --- |
| BLEU | Measuring how similar the translation is to reference by using n-grams (penalizes if words are not exactly the same and in a different order) |
| BLEU2VEC | Use word embeddings combined with BLEU metric to measure how far the translation from the reference is (thus not penalizing synonyms or similar words and a different order of words) |

# Translation service: web & API

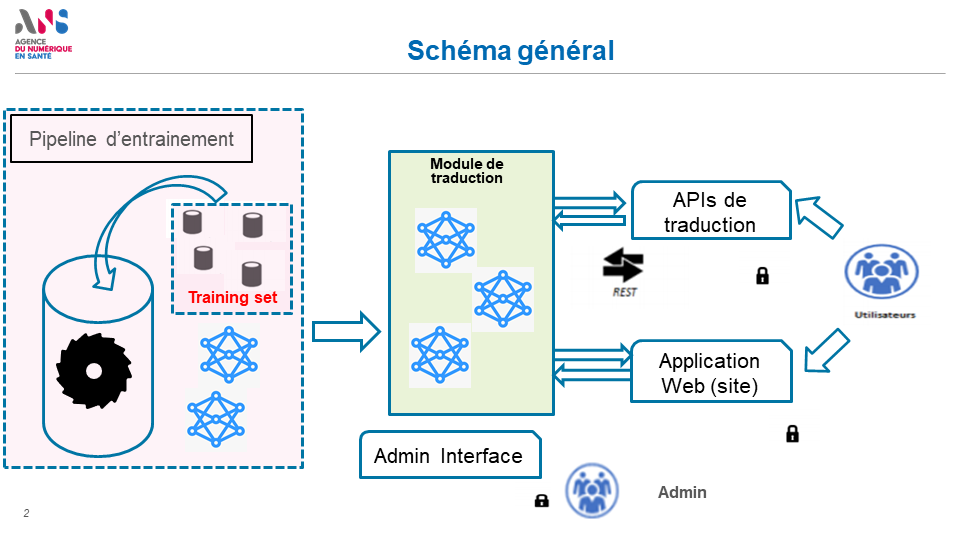
The idea is to divide the process in three (or two) releases:

* Release 1:
  1. Build the webApplication.
  2. Get the translation Models.
  3. Expose APIs.
  4. Admin Interface.

1. Release 2:
   1. Install and deploy on ANS premises.
   2. Deploy and run the Pipeline.

The web application will include the translations of our system for ICD-11 and SNOMED, and a user will be able to search them. If the user has a role of an expert, he could review and rate how good the translation is to an “acceptable” version.

# General schema of the pipeline



Regression

* + - Synonyms, multiple output ranking
    - Clear timelines

Improvement of the translation

* + - Gather all Yann proposals for QC and test the models and form it in a file (send the checklist by next week)

Examples: excision, resection

Compare with google translate, deepl

Propose roadmap of 2 months with the steps (send next week):

Send mail for progress

Create report for what we did (if we have 4 steps in the roadmap, we need 4 reports)

Create check list (which includes issues, bad translations, regressions) will be used for testing

Send:

* roadmap of 2 months
* checklist with qc data

A clear report for the use cases (

* Section 1: architecture
* Section 2: improvement and how to integrate it

For live web tool:

* section 2: usecase, experts users and what is implemented)

Schedule

|  |  |  |  |
| --- | --- | --- | --- |
| **Task** | **Description** | **Deliverables** | **Timeline** |
| 1 | * Gather all proposals for QC and form a checklist file * Test the existing models | * Checklist file * Results with BLEU and BLEU2VEC scores on models * Report | Last week of September 2021 (approx. Friday 30-09-2021) |
| 2 | * Improve models with new training sets (provided recently) * Test new models on the checklist file (and ground truth) | * New training datasets * Improved models * Report | Last week of October (27-10-2021) |
| 3 | * Transfer pipeline and live web tool (at current state) | * Environment setup, scripts * Report | Last week of November (26-11-2021) |
| 4 | * Complete live web tool | * Deliver final version * Report on live web tool | Mid December (15-12-2021) |
| 5 | * Develop unsupervised QC method | * Code and method * Brief report | Last week of December |
| 6 | * Delivery | * Deliver all updated files * Upload on Github | Last days of December |