Data Science Challenge

October 16, 2019

1 Context

Unbabel's translation service is a hybrid model where AI and Humans both play a well defined role. In this challenge we will focus on the human part of the pipeline and not on the Machine Translation (MT).

Our scenario is the following. A client sends their customer support tickets to Unbabel for translation. MT generates a translation for each ticket. After the MT step, Unbabel sends the content to post editors to make sure translations have good quality. In this process a translation job is broken into a number of tasks. These tasks are then randomly assigned to a set of post editors. After the editors work on the task, the content is regrouped and sent back to the client. In this process there are two very important variables: quality and price.

Price is a positive integer number that represents how much we pay an editor for a task. The price, $P(t)$, of a task t depends on the proficiency of the post-editor that proof-reads the MT text, relative to the language pair and domain of the task, LP_t and d_t respectively. It is defined as:

$$P(t) = \begin{cases} \alpha W_t S_e(d_t) & \text{if } LP_t \in \{LP_e\} \\ W_t \log(\gamma + S_e(d_t)) & \text{if } LP_t \notin \{LP_e\} \end{cases} \tag{1}$$

where

- α and γ are constants,
- W_t is the number of words of the task,
- $S_e(d_t)$ is an integer number representing the skill of the editor in the domain d_t, $S_e(d_t) \in [1, 5]$ (editors with a higher skill are more proficient than others),
- task domains d_t are one of five domains: travel, fintech, e-commerce, sports and gamming,
- $\{LP_e\}$ is the list of language pairs editor e is proficient in (e.g Portuguese to English would be pt_en).

The quality of a task, an integer between 0 and 5, is a metric used to evaluate how good a translation is. It is defined as:

$$Q(t) = \begin{cases} A_e(t) & \text{if } LP_t \in \{LP_e\} \\ 0 & \text{if } LP_t \notin \{LP_e\} \end{cases} \tag{2}$$

where

- $A_e(t)$ is the quality interval sampled from the distribution described below.

The selection of the quality interval should take in consideration the skill of the post editor. It is expected that post editors with higher skill produce tasks with higher quality, although they can sometimes produce some editions with lower quality. Hence, the probability that the work provided by a post editor e on a task t be assigned a quality interval $A_e(t)$ is given by:

$$P(A_e(t)) = \begin{cases} \frac{P(A)}{A} & \text{if } S_e(d_t) < \delta \\ P(A) \cdot \beta A & \text{if } S_e(d_t) \geq \delta \end{cases} \tag{3}$$

where

- A is a quality interval $\in \{1, \ldots, 5\}$,
- $P(A)$ reflects the a-priori probability that the performance of an editor is categorized as having quality A,
- β and δ are constants.

2 Problem

There have been some complaints lately from both customers and post editors.

Clients complain that quality is not stable and that they're having problems in trusting the service. This is critical and must be solved.

A large group of post editors claim that sometimes they're starving! In other words, they're not getting any tasks. This might cause them to leave and that is something that cannot happen.

Your mission is, based on the models defined above and a dataset (which we will provide shortly), to come up with a solution for this problem.

3 Notes

As a data scientist, you will have a vital role in understanding whether problems are well defined and counsel stakeholders as to the steps needed to address issues related to missing information or logical gaps. If you find any such issues in this

challenge, we encourage you to develop a solution. One of the most important things we'll evaluate is the consistency of your approach. It needs to tell a story, so make sure to document every step and decision you take, along with any advice you may have as to how the distribution of tasks could be improved in your view.

How you tell that story is entirely up to you but keep in mind that data scientists often need to communicate their findings with others that may not be as data-savy, so make sure to include some visualizations of your results if appropriate: an image can be worth a thousand words!

If you have questions please open an issue.

Good luck and have fun!

4 Datasets

Check the dataset.zip file.