Data Science Challenge

March 25, 2019

1 Context

Unbabel's translation service is an hybrid model where AI and Humans 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) one.

Our scenario is the following:

A client will send Unbabel customer support tickets for translation. After the MT step, Unbabel will send the content 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.

A price, $P(t)$, of a task t is defined as:

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

Where:

- α, β and Γ are constants
- S_e is an integer number representing the skill of the editor. $S \in [1, 5]$. Editors with a higher skill are more proficient than others.
- $W(t)$ is the number of words of the task.
- LP is the language pair (e.g Portuguese to English would be pt_en)

Quality is an integer number between 0 and 100 and it's a metric used to evaluate how good a translation is.

For simplicity, the way we calculate the quality, $Q(t)$, of a task can be defined as:

$$Q(t) = \begin{cases} select\_element(Int_e) & \text{if } LP_e \in \{LP_t\} \\ 0 & \text{if } LP_e \notin \{LP_t\} \end{cases} \quad (2)$$

Where:

- Int_e is the quality interval.
- Quality interval - For simplicity we divide the space Q into $S-1$ buckets. In this example $S = 5$.

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. Hence, the probability of a post editor E be assigned to a quality interval A is given by:

$$P(A|E) = \begin{cases} \frac{P(A)}{A} & \text{if } S(E) \neq D \\ P(A).\beta A & \text{if } S(E) = D \end{cases} \quad (3)$$

Where:

- D is one of the five domains: travel, fintech, e-commerce, sports and gamming

2 Problem

There has been some complains lately from both customers and post editors.

Clients complaint 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's is something that cannot happen.

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

3 Notes

One of the most important things that we'll evaluate is the consistency of your solution. It needs to tell a story so make sure every step and decision you took.

How you tell that story is entirely up to you but make sure you include some visualizations (an image is worth a thousand words).

If you have questions please open an issue.

Good luck and have fun!

4 Data Models

```
from dataclasses import dataclass

domains = ['travel', 'fintech', 'ecommerce', 'sports', 'gamming']
category = ['Small', 'Medium', 'Enterprise']

@dataclass
class Ticket(object):
    id: str
    client_id: str
    tone: str
    topic: str
    number_words: int
    source_language: str
    target_language: str
    quality_score: float
    price: float

@dataclass
class Client(object):
    id: str
    domain: str
    category: str

@dataclass
class Task(object):
    id: str
    sequence_number: int
    number_words: int
    ticket_id: str
    cost: float

@dataclass
class Editor(object):
    id: str
    travel_skill: int
    fintech_skill: int
    ecommerce_skill: int
    sport_skill: int
    gamming_skill: int
```

5 Datasets

Check the dataset.zip file