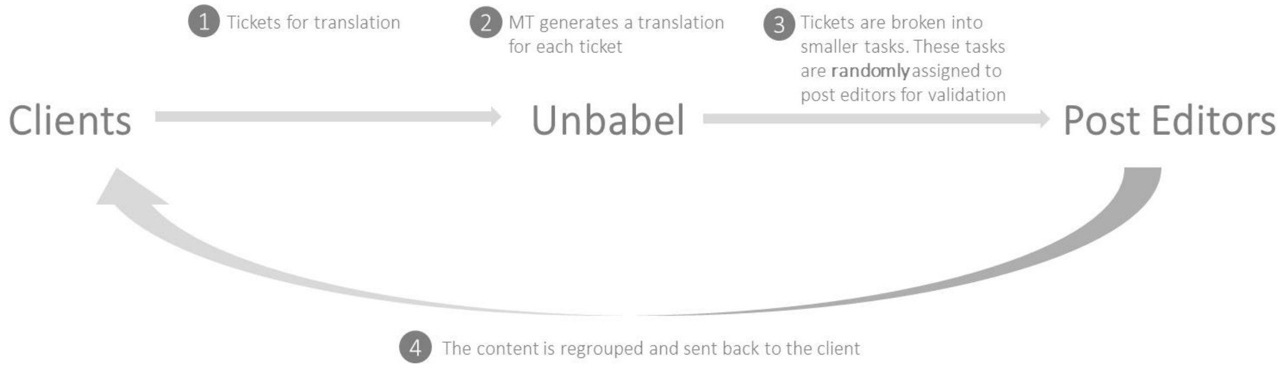
Unbabel | Data Science Challenge *solution proposed by Catarina Pinto*

1. Context and Approach

Unbabel's translation service follows the flow described in the schema bellow:

Image 1 – Unbabel Context



In this process the variables quality and price must be considered.

<u>Price</u>

The price represents how much Unbabel pays an editor for a task. Due to missing information regarding the language pairs in which an editor is proficient, ***I assume that editors are proficient in all language pairs considered in the dataset provided [1].***

The price of a task is defined as: $P(t) = W_{t.}S_e(d_t).0.01$,

Where: W_t is the number of words of the task and $S_e(d_t)$ represents the skill of the editor in the domain

Constant is assumed to be 0.01, this assumption is based on my guess of what a fair price for a translation is [2].

<u>Quality</u>

The quality of a task is defined as: $Q(t) = A_e(t)$

Where $A_e(t)$ is the quality interval sampled from the distribution:

$$P\big(Ae(t)\big) = \begin{cases} \dfrac{P(A)}{A} & if\ \ Se(dt) < 3 \\ P(A).A & if\ \ Se(dt) \geq 3 \end{cases}$$

Where A is a quality interval and P(A) reflects the a-priori probability that the performance of an editor is categorized as having quality A.

The constants were assumed to be β = 1 and δ = 3 [2] [3].

The priori probability that the performance of an editor is categorized as having quality A is not given, I assume that it follows the same distribution of the editors' skills [4].

Problems

Unbabel needs to solve 2 problems:

- Clients complain that quality is not stable
- Post editors claim that they're not getting any tasks.

Approach

Firstly, the datasets were analyzed to get to know the data. After that, the editors were distributed randomly by the tasks and the main variables: quality and price of a task were calculated. In this process, the result of the current method was obtained. The third step is the development of a task assignation method that increases the quality of the tickets and avoids large gaps between the numbers of tasks assigned among post editors. Finally, the results of the new and old methods were compared and evaluated. The comparison includes aspects directly related with the identified problems: quality and gaps between the numbers of tasks assigned among post editors but it also takes in consideration price variation.

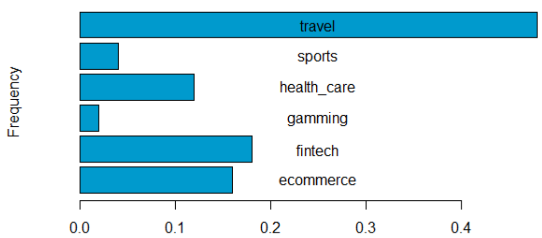
2. Datasets

Clients

Clients dataset includes information of the id, domain and category of 50 clients. The category variable was eliminated since it is not relevant for the current analysis. ***I assumed that the clients can only have a domain, which means that all tickets/tasks of a certain client have the same domain [5].***

The distribution of the clients is given by the graph bellow. 48% of clients have travel as domain, followed by fintech (18%), ecommerce (16%), health care (12%), sports (4%) and gamming (2%).

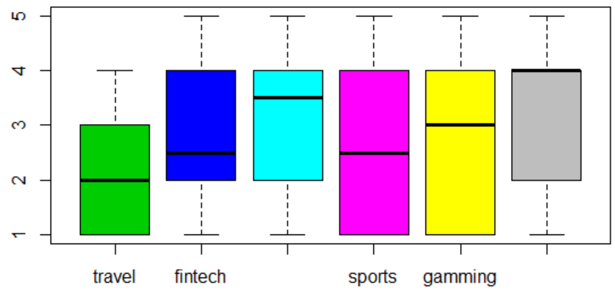
Image 2 - %Clients by Domain



Editors

Editors dataset shows the skill according with the different domains for each of the 418 editors working with Unbabel. The variation of the skills is shown in graph below:

Image 3 - Variation of Editors Skills by Domain



The editors have on average the lowest skill in the domain travel and there isn't any editor with the maximum skill (5) in this domain. More information below:

Image 4 - Variation of Editors Skills by Domain (cont.)

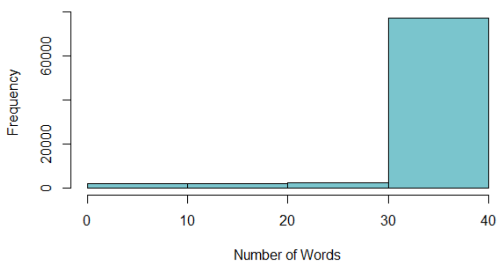
```
    travel          fintech         ecommerce        sports           gamming        health_care
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.00
 Median :2.000   Median :2.500   Median :3.500   Median :2.500   Median :3.000   Median :4.00
 Mean   :2.005   Mean   :2.835   Mean   :3.167   Mean   :2.665   Mean   :2.675   Mean   :3.17
 3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00
 Max.   :4.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00
```

Even though 48% of the clients have the domain travel, the skill of the editors in the travel domain is the lowest. It is possible to conclude that Unbabel should either hire editors with a higher skill in the travel domain or invest in post editors training in this domain.

Tasks

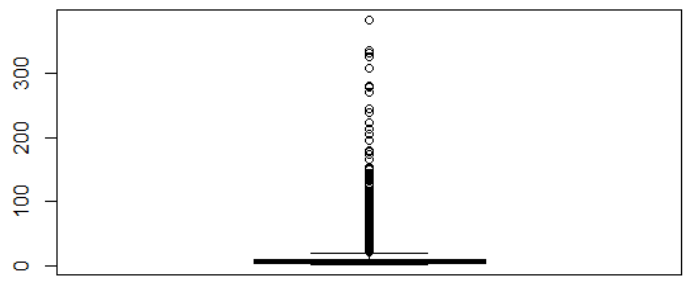
Tasks dataset contains information regarding the number of words of each task and the id of the ticket they belong. 83 153 tasks are considered. Most tasks have about 40 words.

Image 5 - Distribution of Tasks by Number of Words



The tickets usually are divided into 10 or less tasks. However, there are some outliers, the maximum is a ticket divided into 384 tasks.

Image 6 - Variation of Tickets by Number of Words



<u>Tickets</u>

Tickets dataset comprises information regarding 7 788 tickets. The information includes the number of words of each ticket and the client id associated with the ticket. The information regarding language pair and the tone were eliminated since they are not used in the current analysis.

3. Current Method

To analyze the current method, I merged the data frames. Firstly, I associated the client and respective domain to each ticket using the client id field. Once I had information regarding the relation between ticket and domain, I merged this info with tasks dataset by the ticket id.

To reproduce the current method, the tasks were randomly assigned to the editors. The skills of the editors in the domain of the assigned tasks were attributed to task itself and Q(t) and P(t) were obtained through the formulas presented previously.

<u>Quality</u>

As referred in ***[4]***, the probabilities distribution of editors' skills was used as a proxy for the a-priori quality probabilities distribution. This distribution follows the table below, where A is a quality interval and P(A) reflects the a-priori probability that the performance of an editor is categorized as having quality A.

Image 7- A-priori Quality Probabilities Distribution

A	P(A)
1	28%
2	19%
3	17%
4	22%
5	14%

Since there are two distributions - one for the editors with higher skills (3,4,5) and other for editors with lower skills (1,2), the dataset was divided in 2, and both distributions were applied. The distributions of qualities for both datasets are the following:

Image 8 - Quality Probabilities Distribution

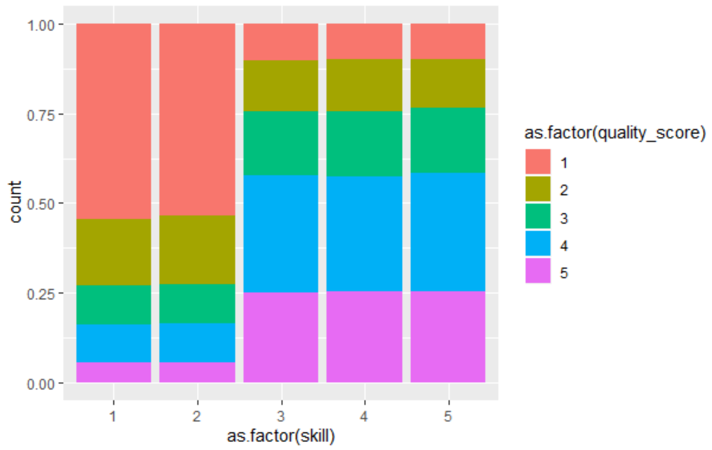
	1	2	3	4	5
Lower Skill	54%	19%	11%	11%	5%
Higher Skill	10%	14%	18%	32%	25%

Results

After applying the description above, the results obtained indicate that: 15% of tasks - quality 5, 21% of tasks - quality 4, 14% tasks quality 3, 17% quality 2 and 34% quality 1.

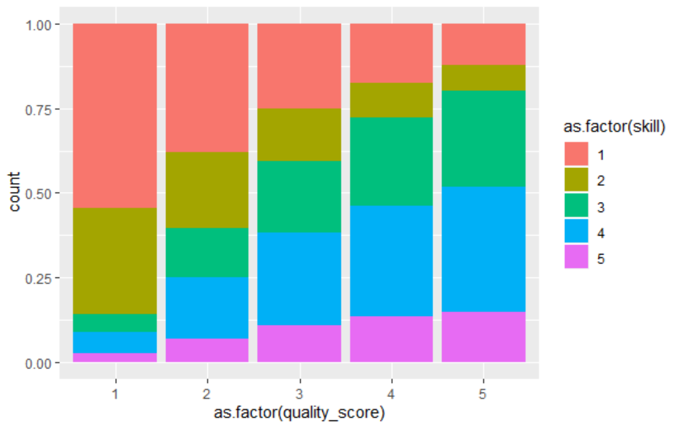
By observing the graph below, we observe that the group of higher skill editors produce more tasks with a superior quality. However, there is no difference between an editor with skill 5 and an editor with skill 3. Since the editors with skill 3 produce cheaper tasks than the ones with skill 5 and with the same quality level, it is a smart move for Unbabel to choose first the ones with skill 3.

Image 9 - Quality Frequency by Skill (Current Method)



Another interesting analysis is to observe that editors with higher skill can sometimes produce some editions with lower quality.

Image 10- Skill Frequency by Quality (Current Method)



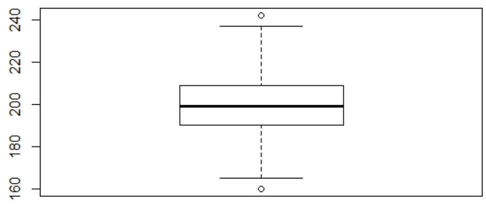
Price

Regarding the price, by applying the random method, a total price of 79 849 was obtained.

Editors Gap

The editor with the maximum number of tasks get 242 tasks while the one with less tasks assigned get 160. Distribution: 1st quartil: 190 - median: 199 - 3th quartil:209.

Image 11- Editors Gap (Current Method)



4. New Method

In the selection of a certain editor *e* to a task *t*, 3 aspects are considered:

- The skill of *e* in the domain of *t*
- The skills of the *e* in the domains different from the one in *t*
- Number of tasks *e* already has

Firstly, a skill order preference was set: 3, 4, 5,1, 2 [6]. This order means that an editor with skill 3 is the most preferred and an editor with skill 2 is the less favored. As explained above, the preferred skill is 3 since an editor with skill 3 gets the same results in terms of quality as the editor 4 and 5 and it is cheaper. Since one of the main goals is to increase the quality, the skill order preference starts with 3, 4 and 5.

Secondly, the maximum number of tasks an editor can get was assumed to be 5% higher than the division of the number of tasks by the number of editors [7]. This assumption enables to reduce the gap between the numbers of tasks assigned among post editors. The maximum number of tasks is:

(number of tasks/ number of editors) *1.05 = 208

In the third step, all editors with the preferred skill (3) in the domain of *t* are selected.

After the selection of the editors (4th step) the group of editors selected in the previous step are analyzed. All editors with number of tasks equal to 208 are eliminated. If the subset is empty in the end of this selection, the process goes back to the third step and select all editors with the second preferred skill (2). The process continues interactively until a filled data frame is obtained.

In the fifth step, the skills of the editors in the different domains from the one in the task are examined. Following the logic of our skill order preference, we attribute points to editors according with the skills. The points were attributed using the following logic: 50 points – skill 3, 40 points – skill 4, 30 points – skill 5, 20 points – skill 1, 10 points – skill 2.

Example: In the assignation of travel task, an editor with the following skills: health care (1), ecommerce (3), sports (4), travel (3), fintech (1) and gamming (5) receives 20+50+40+20+30 = 160 points. The skill in the domain of the task (travel) is not relevant for points attribution.

After the point attribution, the editor with less points is selected. In case of a tie, one of them is selected randomly.

Summary

The new method enables the selection of the editor that is simultaneously the best (according with the preference assumed) in the domain of a certain task and the worst in the remaining domains without exceeding the limit of tasks per editor.

Results

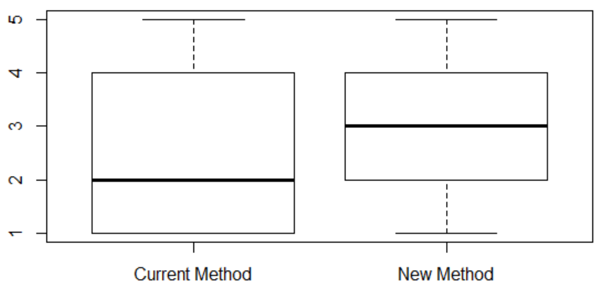
The new method used only editors with skills 3, 4 and 1. And it obtained the following results in terms of the quality: 21% tasks with quality 5, 28% tasks with quality 4, 16% with quality 3, 15% with quality 2 and 20% with quality 1. The total price of the new method is 88574.56. And the editors distribution in terms of tasks follows the distribution: min: 146 - 1st quartil: 205 - median: 208 - 3th quartil: 208 –

Compare Methods

Quality

The quality increases if Unbabel decided to apply the new method as it possible to observe in the image below. The quality median increases from 2 to 3 and the means increase from 2.666 to 3.154.

Image 11- Quality Distribution (New vs Current Method)



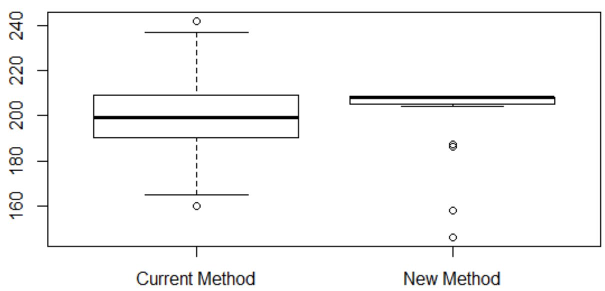
Price

It was verified an increase of 11% (8725.51) of the price with the new method.

Editors Gap

Editors gap reduces considerably with the implementation of the new method as shown in the image below.

Image 12- Editors Gap (New vs Current Method)



5. Notes

- Along the current report, all the assumptions are identified with *[ ]*.
- This solution was solved by using R.