SatEYEre: Gaze Estimation for Computational Humor

Semester project DLAB – EPFL
Friday 7th June, 2019

Author: Nicolas Zimmermann
Under the supervision and collaboration of: P.h.d. Kristina Gligoric, Professor Robert West

***Abstract*— Humor is one of the core human reactions. There have been many efforts in the past to try to understand humor, what triggers it, in which context, what effects it has, etc. Here, we will be doing an eye-tracking study with the aim of understanding and comparing cognitive reactions by following the participants' gaze when confronted with either humoristic or serious headlines.**

I. Introduction

The novelties we introduce are two things, the eye-tracking and the comparison in pairs of nearly identical funny versus serious headlines. To get these pairs, we used data collected from Unfun.me, a website from a previous experiment where people can see satirical headlines from the satirical news website *The Onion* and have to do minimum changes in order to make them look non-satirical. For instance, going from the satirical sentence "City opens new art jail" to "City opens new art museum" so that it looks like any regular newspaper headline. Since the two sentences only differ by a few words, we can compare the participants' gaze patterns in both of them and check very precisely where they differ. It should allow us to see what role the modified word/group of words played (e.g. in the example above, "jail" replaced by "museum") and eventually if we get conclusive results, to build models to recognize these patterns.

II. Related work

This study is in the continuation of a previous work from R. West and E. Horvitz [1] where they created the website Unfun.me described in the introduction I. They collected many pairs of headlines and analyzed them to extract various information such as chunks of words representing a single concept (that we will be using later during the analysis part in this experience), the type of humor, if there was absurdity, violence, etc. This gave us a pool of pairs of headlines from which we picked the ones that we thought were the funniest and the easiest to understand.

There are not many papers related to this specific work, but there exist some studies about satire, such as a paper from S. Skalicky and S. A. Crossley [2] where they study what influences one's reactions/understanding of satire. They also took headlines from *The Onion* for their study, which is a good common basis. However, to our knowledge, no study tried to combine eye-tracking with humor analysis nor to compare pairs of satirical/serious headlines, hence this is mostly new work that we will be doing.

III. Research protocol

We will look for 30 participants and show to each of them 30 headlines (15 satirical, 15 serious), in random order with no two headlines from the same pair of satirical/serious with no indication about if the current headline is supposed to be satirical or not. After each headline comes a short questionnaire asking them to rate how realistic it was (as in could you see it in a real newspaper?) and how serious/funny. This will allow us to build a ground-truth to tell how different in terms of funniness, the two headlines from the pair were when we analyze the associated gaze pattern.

We will ask about half of the participants to come back one week after for a second part of the study where they will look at the complementary set of headlines: they will look at the serious-looking version of each satirical headlines they saw and vice-versa. We hope that within this week interval they will have forgotten most of the headlines and this will provide us with valuable information for each individual participant how the gaze and reaction changes when the same headline is satirical or not.

There are a total of 30 pairs, hence each participant will see exactly one headline from each pair and exactly the two headlines from each pair if they come for the follow-up experiment.

On top of the eye-tracker, we will record the participant's face and later perform emotional detection to find when they start smiling. This will allow us to know precisely when the participant understands the satire and starts smiling at it, even exactly on which word he was looking at. Although we certainly will not see that type of reaction on every satirical headline, this will give us good extra information when they do and ensure that they understood the sentence. We also ask in the questionnaire if they understood the sentence to remove undesired data points.

The experiment is done on a computer specifically equipped with an eye-tracker and the necessary software to use it. We will record every action from the user (mouse clicks and movements, keystrokes), as well as their gaze and faces.

When the participants register for the study online, they will have to fill their availabilities, accept the consent form and answer basic demographic questions (age, gender, education, GPA, etc.), as well as the TIPI (Ten Item Personality Measure).

The participants must meet the following criteria: having a minimum level of English of B2 (according to the Common European Framework of Reference for Languages) and not using glasses (contact lenses are fine). These requirements ensure that the participant will be able to understand the headlines and having no glasses improves the eye-tracker precision.

The online registration should take about 15 minutes and the actual study, as well as the follow-up, a maximum of 45 minutes. The participants will receive 20.- CHF after the first part of the study and an additional 10.- CHF if they participate in the follow-up. We chose to put the consent form and basic demographics questions in the consent form to save time during the on-the-spot study and have a higher chance people who registered would actually come because of the ”escalation of commitment”.

IV. Technical decisions

The eye-tracker we use are SMI fixed under the screen, they provide better precision than the glasses eye-tracker, in the Experiment Center software. We use BeGaze to export the raw data into a more practical format with aggregate events (such as eye saccade and fixations instead of continuous eye position) that are much easier to interpret. The participants will sit on a fixed chair and non-turning chair to avoid any movement that could affect the eye tracker's precision. The supervisor will help them to constantly stay within the right distance range away from the eye-tracker.

We will do two participants at the same time since we have two eye-trackers, to make things quicker. It will happen in two separate rooms, with one experiment supervisor in each, to avoid interference of reactions between participants.

We record the video using a regular webcam and OBS software to store the file. OBS also allows some compression into mp4 file right away and to record simultaneously the part(s) of the screen and the webcam, along with many other customizable parameters. The video output is sent to Microsoft Azure Cloud Services to be treated and we receive an output file with the different emotions at each frame of the video. Having a good emotion extractor is not easy to make or to find and we chose Azure because we had some existing credit on their platform. The synchronization of the video capture with the eye-tracker for the analysis is done with the sound of a “clap” when the participant presses the button to start the experiment (we might add extra precision by recording the button itself to look when it is pressed).

The actual sentences, questionnaires, etc. are displayed via a web browser. This option allows us good flexibility to change any parameters of the experiment, to be accessed from any computer easily and directly store the results in our server on MySQL tables. We could have used the Experiment Center integrated options but it would have been difficult to customize it to our needs. We synchronize the data with the eye tracker for the analysis by matching the press of the start button of the experiment. To match the eye tracker fixation positions with the position of the words, prior to the experiment we extract, for each sentence each, the words' position in pixel.

We decided to display the sentence on a single line because the precision of the eye-tracker is not very good on the Y-axis but rather fine on the X-axis. The font-size used is the same for every sentence and matches the longest sentence to display fully on the screen. For added precision, we put an extra space between the words so that the eye-tracker is less likely to have a fixation on the wrong word.

Before each headline, we shortly show a dot in the higher part of the screen that the participant has to look at for 4 seconds. It makes the gaze focuses on one particular part of the screen and avoid parasite remnant gaze data all over the screen when the sentence shows up. It allows us to have a cleaner analysis and understanding later on.

V. Analysis

Some of the analysis tools have already been made, based on the 3 people test tries we did, as we have not collected the real data yet. The actual analysis of the whole data set will be done at another time outside of the scope of this semester project. Let us describe the various metrics that will be used:

- **1.1** Time spent on each word for each sentence and for each participant, as well as the ratio of time spent a given word relative to the total time spent on that sentence.
- **1.2** Number of times the participant's gaze went on each word for each sentence and for each participant, as well as the ratio of each word relative to the sum on that sentence. We count +1 to a word when the participant's gaze fixates on this word and the previous fixation was on a different word/no word.
- **1.3** Time spent smiling while looking at a given word, for each sentence, and for each participant. We will use the recording of participants' faces to that end.
- **1.4** Same analysis as the three points above but doing prior to that a grouping of words that represent a single concept into a single chunk. For instance “new art jail” will be analyzed as a single chunk formed by the three words.
- **1.5** The time spent on each sentence and on answering questions for each sentence and for each participant.
- **1.6** Total time spent smiling on each sentence and the related questionnaire, for each participant.
- **1.7** Visual for each participant where the gaze is while looking at the screen in a video format. This will not bring quantitative data but can help in understanding how the participants acted.
- **1.8** Number of times the participant changed his mind while answering the questionnaire after a sentence, on each sentence, and for each participant. We can imagine that if a participant changed his mind many times, he may not have understood the sentence very well or only afterward, or at least that there is a bigger variance on a given funniness/realisticness rating.
- **1.9** Average funniness and realisticness score per participant, both in total and for all satirical headlines VS all serious headlines. This can show if the participant

was able to identify in general when the sentences were supposed to be funny & satirical or not.

Once we have all these pieces of information, we will perform aggregate analysis between the participants:

- **2.1** Comparison of the ratio results per word of **1.2** between the satirical and serious headline. We particularly look at the difference in the word that was changed. We expect to see that the satirical headline has a higher ratio than the serious on the word that was changed. This is one of the most important metrics we want to look at for this study.
- **2.2** For every possible pair of two users, use ratio results from **1.2** as a vector for each sentence and compute the average cosine similarity between the two participants on every shared sentence. This gives us a metric of how similar the two participant's gazes are. Using this data we can create 5 heatmaps, one for each personality type measured with the TIPI test, having on the axis the score on that personality type (1 to 7) and plot the average cosine similarities as heat values. If we observe the zone near/on the diagonal to have significantly higher values, it would mean this personality type is an indicator of how the gaze of the person will be like.
- **2.3** A histogram of average funniness/realisticness (one for both) per TIPI personality type, to see for each of these personality traits if they have an influence on the funniness/realisticness felt by the participants (hence a total of 5*2=10 plots, for each of the 5 personality traits with either funniness or realisticness values). Note that since we will only have 30 data points, it may not be very significant.
- **2.4** Average funniness and realisticness score per sentence. This is important to see if the participants really saw that one was supposed to be funny while the other one is more serious, the bigger the rating difference in a pair, the better it is for the sake of the study.

VI. Discussion and further work

This is really difficult to discuss the results as of now, since we have no data, except for the 3 test tries that we did but doing any aggregate analysis on so few data points would be meaningless but we can still observe a few interesting facts from this data.

	P1	P2	P3
Average funniness [0,4]	0.8	1.8	1.25
Avg. fun. on serious headlines	0.6	0.8	0.4
Avg. fun. on satirical headlines	1.0	2.8	2.1
Average realisticness [-100, 100]	-8.8	-19.75	12.9
Avg. real. on serious headlines	37.1	14.3	38.3
Avg. real. on satirical headlines	-54.7	-53.8	-12.5

TABLE I: Mean rating for the 3 test participants. 0 funniness means not funny, 4 means very funny. -100 realisticness means fake, 100 means real.

What we can observe already is a big difference in the funniness ratings between the participants, the scale ranges from 0 to 4 and the average for each of the test participants is between 0.8 and 1.8. It means correcting the ratings between the participants might be relevant, as a rating of 5 from one participant might mean the same as a 3 from another participant. The same could be applied to realisticness ratings.

We can also see that the average funniness for satirical sentences was lower than the serious ones for the three participants with a good margin except for P1. This is a good sign that the headlines were well chosen but a pair-wise comparison could give more detailed results. The same applies to the realisticness averages where the funny headlines were given a lower rating by a difference of at least 50 points.

These first results seem to show that while the satirical headlines are not necessarily much funnier than the serious ones, the participants where clearly able to tell which ones were more satirical.

Further work in continuation of this study would be a thorough analysis of the data collected as we did not have time to complete it. One could think of building a model for predicting if a sentence is funny and what precisely makes it funny given the gaze. It is difficult to think of real world applications but it might be useful for marketing and ads performance predictions. Eye trackers are already used in that area, hence it would be quite easily integrated and would provide useful information: being to tell if people get the message/punchline in your ad would be very useful.

VII. Conclusion

We were able to go up to the experience stage during this semester project and have a substantial part of the analysis already done while discovering completely new tools for this project. This is a good accomplishment considering the limited time we had for it.

References

[1] R. West and E. Horvitz, "Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances"," *AAAI — AAAI Conference on Artificial Intelligence, 2019.*, 2019. [Online]. Available: https://dlab.epfl.ch/people/west/pub/West-Horvitz_AAAI-19.pdf

[2] S. Skalicky and S. A. Crossley, "Examining the online processing of satirical newspaper headlines," *Discourse Processes*, vol. 56, no. 1, pp. 61–76, 2019. [Online]. Available: https://doi.org/10.1080/0163853X.2017.1368332