Punalogy: a generator of puns based on analogical templates

SARAN, Camille
LETTIFI, Abir
SILLAIRE, Maeva
BERRANE, Louis

January 27, 2023

Abstract

We introduce Punalogy, a generator of puns based on analogical templates. Punalogy takes a word as input and generates a pun by substituting a part of the word with a homophone or a similar-sounding word. For example, given the input "mango," Punalogy might generate the pun "man is to mango as woman is to womango." In addition to generating the pun, Punalogy also creates a meme using a template and existing image generation models.

We discuss humor, meme generation and address several technical challenges. We also present a protocol for evaluating the quality of the puns generated by our system. Our experiments show that Punalogy is able to generate a variety of puns, but the funniness is not always obtained. But as it is functional, it is a valuable tool for language enthusiasts and researchers on developing generated humor.

Introduction

Punalogy is a project that generates puns using an analogy template: *A is to B as C is to D*. The project combines semantic and morphological aspects to create a pun, *i.e.* a type of humorous wordplay, from a given input word. Additionally, our system is able to generate not just the pun itself, but also a corresponding meme image (using Stable diffusion and Dall-e Mini models), which adds an additional layer of humor. This paper reviews state-of-the-art aspect related to humor creation, and then develops on our approach and our results. This project showcases the creative potential of Natural Language Processing and its ability to generate playful and humorous content.

1 Background

1.1 Analogy

As stated by Kedar-Cabelli, word *Analogy* comes from Greek, meaning "mathematical proportion". Analogy is used to describe the process of comparing two or more phenomena to draw connections by identifying similarities or differences.

An analogy can be proportional, adopting the form "A : B :: C : D", which should be read as: "A is to B as C is to D". But others forms of analogy can be found. From a logical point of view, analogy is a type of argument in which one thing is said to be similar to another, based on the assumption that if two things are similar in some respects, then they are probably similar in other respects as well. From a cognitive point of view, analogy is a way of thinking that allows us to understand new or complex ideas by breaking them down into smaller, more familiar parts.

Recent research has proven that analogy is really important in cognition and that analogical reasoning has many applications, mainly in Artificial Intelligence (AI) and Natural Language Processing (NLP) (Turney, 2008). Indeed it is a type of reasoning or thought process that allows us to understand new concepts or ideas by comparing them to something familiar.

In our project, we use a pun template of the form of analogy. We think this template should be able to produce laugh.

1.1.1 Semantic & Morphology

Semantic studies meaning of words. In our project, we create a pun analogy by combining semantic and morphological properties.
We want to access to one word antonym's or synonym's. This can be done using WordNet (Fellbaum, 1998). WordNet is a lexical database, accessible with python (Van Rossum and Drake, 2009) using its API. It allows the one that uses it to retrieve some lexical information about a word: its lemma, meanings, synonyms, antonyms, hyperonyms... We can notice that these set of words (antonyms, synonyms) are restricted to strict ones. For example, there are not any antonyms for *Man*. Even if *Woman* could considered as a type of antonyms, as WordNet only retrieve the strict one, it will not retrieve *Woman*.
Morphology studies how words are formed in a given language. It is a key element in our pun. Indeed, the pun: *"Man is to Mango as Woman is to Womango"* relies on morphological similarities between words. By drawing a morphological closeness between words, that exists (but that is just a coincidence, and not driven by meaning) we expect the pun to be funny, as explained in General Theory of Verbal Humor (see *Incongruity theories*).

Analogies can be used in puns. Indeed, its structure allows easily to create humor by doing a cratylistic logical mechanism (Attardo, 2020), we will explain more about humor theories in the coming section.

1.2 Humor Theory

There exists three types of humor theories: (1) hostility theories, (2) release theories and (3) incongruity theories. Hostility theories state that humor come from the fact that one laughs by feeling superior to someone via humor. Release theories stipulate that humor is used to maintain a balance in us, *i.e.* to deal with homeostatic reasons. We will not discuss about hostility and release theories as the majority of literature in computational humor focuses more on incongruities theories. There exist a variety of incongruity theories, and we will review some of them, and show they can be used in our project.

1.2.1 Incongruity theories

In Incongruity-Resolution (IR) theory (Ritchie, 1999), humor comes from incongruity . From the Oxford dictionary (oxf, 2012), incongruity is defined as *"the state of being incongruous; incompatibility."*. Thus, the incongruity-resolution theory states that humor comes from the fact that an initial incongruity is then resolved. This is why IR can also be described as Surprise disambiguation (SD). Ritchie writes that there exist three steps in a joke: (1) The first (more obvious) interpretation of the set-up, (2) the second-hidden interpretation of the set-up; and (3) the meaning of the punchline, *i.e.* the IR or SD. He also explains that 5 features appear in a joke : obviousness, conflict, compatibility, comparison and inappropriateness.
Another interesting theory is General Theory of Verbal Humor (GTVH) by Attardo. In this theoritical framework, humor is linked to six knowledge resources that should be taken into account when analyzing humor : (1) Script Opposition, (2) Logical Mechanism, (3) The Situation, (4) The target, (5) The narrative strategy, (6) the language. It is an incongruity theory since Script Opposition is composed by Script Oppositness (SOp) and the Script Overlap (SOv). For example, a joke J should denote two different scripts (say *S1 and S2*). These two scripts should show SOv, so that they can be understood at the same time, but they should also present SOp so they could be interpreted in different ways.
In puns, humor come from the resolution of these two scripts by a cratylistic Logical Mechanism (LM). The cratylistic LM, is a false logic LM. It is based on the belief that there exists a link between a word's form and it's meaning. And thus, similar words by form should also show similarities by meaning. Obviously, it is false since words form are not iconic (*i.e.* they represent what they mean) but are mostly symbolic in oral languages (*i.e.* their form is not motivated by iconicity, but are learned by locutors (Atkin, 2010)). The question of iconicity of words in languages and the extent to which language is representational is of course more complex than that, but it is a subject of study in itself, and for our purpose, we will conclude that words are not iconic. Thus, it is false to assume a link of meaning between words that show similarities of form.
In GTVH, humor comes from the discovery that the cratylistic LM does not work. A joke could be

explained like that:

- Joke J contains two opposite scripts (SO) that present SOp and Sov: S1 and S2
- By using false cratylistic LM, locutor try to resolve the SO.
- The resolution of SO create humor.

As one could see, this theory is totally an incongruity one because humor comes from a SO, that is resolved by using a false logic mechanism, that lead to laugh.
These theories of humor always show that humor comes from the script opposition. Script opposition is expressed via semantic opposition. Thus, semantic opposition is really important. As Ritchie stated it, obviousness and conflict are necessary in humor. The more opposition there exist between the two scripts, the more it will be considered funny.
But what is the opposition's nature exactly ? This is where cultural and personal differences come. Indeed, what is considered opposite or inappropriateness is very subjective, leading to subjectivity of humor. Concerning topics of humor the article *Characterizing humor: An Exploration of Features in Humorous Texts* (Mihalcea and Pulman, 2009) shows that humorous and non-humorous texts can be differentiated. Indeed, there are a few characteristics of verbal humor: human-centric vocabulary (*i.e.* words are related to human), negation (*i.e.* majority of jokes contains word negation), negative orientation (*i.e.* negative polarity appears often), professional communities (*i.e.* often jokes contains job related, that can be also linked to human-centric vocabulary), and human weakness (*i.e.* nouns and verbs that are related to "poor" moral value are mostly used).

1.2.2 Visual humor

As we generate memes, the visual modality is important. Work on computational visual humor is very recent and an understudied area. In the article *We are humor beings: Understanding and predicting visual humor*, Chandrasekaran et al. consider publishing the first work on computational visual humor. In this article, they use the GTVH theoretical framework to study it. Thus, it is the opposition between two scripts that would always bring humor. Their model allows to alter an image to make it more or less funny. To do this, they try to introduce or remove what might constitute a break in visual normality.

Tsakona's research focuses on humor in comics (mixing text and image) and is also based on the GTVH. According to this article, the difficulty to analyse humor is greater when several modalities are mixed (visual and language). The results of his study show that the image does not in itself return a comic effect, but that it serves as a support for the text. Thus, the image creates the necessary context to accentuate the joke of the text. The author also finds that exaggeration, contradiction and metaphor are constitutive elements of visual and verbal humor. Finally, humor is largely based on our knowledge, so this characteristic is naturally found in visual humor. Visual literacy', i.e. the ability to recognise images and interpret them, is important in finding an image funny or not.

1.3 Humor generation

Humor generation is the process of creating humorous text using NLP techniques. It can be a challenging task. Indeed, as discussed earlier, humor is highly subjective and can be difficult to define.

1.3.1 Humor generation approaches

Neural Networks approaches Neural networks approaches in joke generation were used from 2017, beginning with Ren and Yang's LSTM RNN. They used GloVe as vector representation. For training, they used around 8.000 jokes. The user chooses a word, and their model tries to perform incongruity by choosing a word with a probability chosen by the output layer.
Another try with not neural networks method was done by Yu et al. (2018). They used a seq2seq model that takes a polysemic word, and two of its meanings as input. Then, they used a encoder-decoder network to create a sentence that combines both meanings. Then they use a second model that enhances the generated sentence.
From our knowledge, both models did not totally succeed in creating jokes, as results showed that jokes were not considered very funny by humans.

Template based approaches A large part of humor generation systems are template-based. Template-based systems are the ones that use templates, *i.e.* a text with slots that have to be filled. A schema is then associated with a given template, and specifies the variable relationships. There exists a few types of variable selection: ontologies (*i.e.* system uses a specific lexicon containing lexical relationships. For example: homonymy, antonymy...), quantitative measures (*i.e.* relationships between words are not explicit but expressed in quantitative measure as n-gram co-occurence) and hybrid systems(*i.e.* ontological and quantitative approaches are mixed). We will review some of these project in the upcoming section *Related Works*.

1.3.2 Humor Evaluation

The question of evaluating humor is a difficult one. Indeed, as we have seen, humor comes from an opposition between two scripts which are both conveyed by the message. However, scripts, although denoted by the statement (both verbal and visual), only make sense in the mind of the person receiving them. Thus, humor is in reality a co-construction between the one who produces it and the one who receives it. A joke cannot be intrinsically funny: it may tick all the boxes of what is considered necessary for a comic effect by the GTVH, but in reality what makes it funny is that someone finds it funny. This is why assessing humor is a complicated task. Indeed, it is not possible to find a joke that is funny, per se, outside of any audience. Humorists are proof of this: even if some are considered funny by more people than others, it seems impossible to establish a ranking from least to funniest that is not a personal judgment. In NLP, evaluating humor in then a very challenging task. In literature, attempts to evaluate humor have to deal with this issue.

Humor's subjectivity explains why humor evaluation was mainly evaluated by humans. Human evaluation is the one that is conducted the most in NLP. For example, by rating jokes on a likert Scale. but this method is not optimal, because it truly depends on the participants background.

Another type of evaluation was proposed by **?**, who uses HF (humorous frequency) *i.e.* the number of funny items in a set. However this method is not optimal. First because it assumes that items can be funny per se, and second because it is not suitable for our project. Indeed, in Punalogy humor should come from the combination of the words together, forming a false logic mechanism and not from the word taken individually.

Binsted and Ritchie (1997) proposed a soft Turing test to evaluate humor. This test consists of saying if the joke was generated by computer or by human. This method do not really evaluate humor in itself, but more on the faculty of machine to imitate humans.

A last type of humor evaluation uses user engagement on social media. User engagement is the set of all actions that user have to interact with a post on internet. Aldous et al. consider that user engagement can be seen on four different levels : (1) views, *i.e.* "private engagement", (2) likes *i.e.* "exposing user preferences", (3) comments*i.e.* "express opinion or feeling" and (4) shares and repost *i.e.* "public sharing". This four levels reflect engagement along the continuum of private to public actions.

In reality, testing humor on user engagement evaluates a post reception. As humor is born from interaction with people, in a logic of co-construction. This method would therefore make it possible to assess people's interest in a post. However, this method has its limits since it actually measures engagement and not the fact that people find the post funny. Engagement is in fact subject to many other variables:

- Timing of the post,
- Subject of the post,
- Platform on which the post is (which can change both engagement in general, and the type of engagement (level 1, 2, 3 or 4)

Thus, variables must be discarded because engagement is not only due to the post itself, but also to many other contextual factors. Moreover, it is the "virality" of the post that is studied rather than its funny effect. What is viral is not necessarily viral for good reasons. This method also restricts the type of audience that will see the post. However, in view of all the methods presented, this one allows access to reactions in a simpler, more effective way. Many works use this measure, such as the work of Kurochkin, which we will discuss in more detail in the *Related Works* section.

All these methods are not optimal, and Alessandro believe that an objective criteria should be to evaluate if a joke is considered funny or not. For our project, we considered that a practical and efficient way of assessing humor was to use a combination of various methods.

1.4 Meme

A meme is generally understood as a image accompanied with a text, forming a funny media object that is shared online. Figures 1 and 2 show examples of variations of a meme. A meme is usually composed of what is called a 'template', in this case it is the *Distracted boyfriend* template. This is the text that is subject to change, as in figure 2.
In this section, we will explain what is an internet meme. First by defining what the concept of meme means, then by explaining more on internet memes.

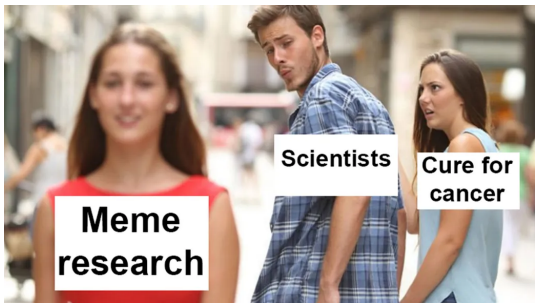

Figure 1: Meme "Distracted Boyfriend", example 1

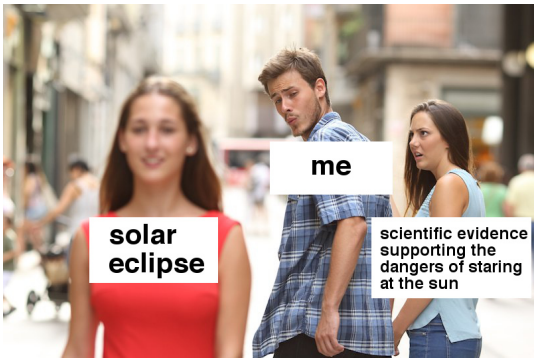

Figure 2: Meme "Distracted Boyfriend", example 2

Meme. One could not discuss about Internet Meme (IM) without reviewing what a meme is. In this section, we will first develop what is a meme, then we will analyze the rise of internet meme.

The term meme was first introduced by Richard Dawkins in 1976 in his book *The Selfish Gene*. In this book, he describe what he called a "meme". A meme etymologically comes from ancient greek *"mimema"*, (*something imitated*). Dawkins used term meme to define all objects (behavior or ideas) that are spread through individuals. This definition is quite large, and define all type of "spreading-like" object in human which is not genetic. His theory was inspired by genes or virus behavior. In fact, at first memes were described as virus (horizontal transmission trough individual with potential mutation) or genes (vertical transmission trough individuals with potential mutation)Castano Diaz (2013). Indeed, this comparison was allowed because of the followings facts :

- A virus is spread trough people, a meme is transmitted through users.
- A virus can mutate, a meme can be transformed by users.
- A virus source is not always well-known, meme's paternity is not known (because meme are often not attributed to someone in particular, thus making it easier to transmit (Miltner, 2018)).

Even if this view of what a meme is criticized (see Cannizzaro (2016) for a more in depth explanation), authors tend to agree that a meme have some particularities : (1) it is an object that is transmitted by people, (2) this object can be prone to mutation.

This very succinct description of a meme is clear enough to continue our description of internet memes. As you can see, humor notion is not present in this description. We will discuss further about it in the following sections.

Internet Meme. IM is a type of meme whose support is internet. Even if pre-digital meme existed (like the demotivational poster, parody of motivational US Army poster. See Miltner (2018)), first memes on internet are believed to be emoticons Davison (2012). An emoticon (blending of *emotion* and *icon*) mimics human facial expression to give informations about how the text it corresponds should

Form properties	Use	IM general properties
Can be an image, a GIF or a video.	Vehicule for a joke	Simplicity (to make creation possible)
Form of the image is not really important (it can be bad quality), but the message conveyed is.	Express certain emotions, not expressable by written language	Transferability (to induce replicability)
Multimodality is a key component as texts and images are mixed together.		Majority in english (lingua franca)
Source: usually taken from anywhere in culture (from politician, to movies, videos games ..).		

Table 1: Properties of IM

be interpreted. Around 2000, the Hamster Dance website (still accessible here) was a website created to view an animated image of an hamster who is dancing. This website only goal was to entertain, and reached 15 000 views per day Davison (2012). As one could see, IM were constant on internet: they were used from the beginning. Davison says that emoticons and hamster website are truly precursor of what is a meme today because they have respectively these function: to be better understood, and to entertain. Nowadays, IM do these function, even if others functions can be observed (as a political critique, as stated in *Internet memes* (Miltner, 2018).)

Origins of modern IM are not clear. Indeed, a big component of meme is the lack of attribution. Even if personal attribution can not be drawn to meme creation, Miltner states that IM were first populated of a said *"meme/troll space* on 4chan[1] and Reddit [2]. For him, this space is deeply rooted into hacker culture and thus could explain the lack of attribution for meme.

Research on IM is recent, and often not consensual, but we will align our self with Shifman's definition of meme: *"[an internet meme] (a) share common characteristics of content, form, and/or stance; (b) are created with awareness of each other; and (c) are circulated, imitated, and transformed via the internet by multiple users "* (from (Shifman, 2014)). Thus we do not view meme as unit, but as part of a wider systems that share properties. For Shifman, IM are *"operative signs" i.e.* they have an autonomous function apart from the semantical point of view.
To sum this part, we reviewed characteristics of IM in the table 1.

In conclusion, IM are present since Internet ever existed. They are a subject of research in various fields (semiotics, cybernetics, communication...). IM is a media object composed by a form, which is multimodal, to convey a meaning. Meanings vary from political views to puns. IM can be considered as a relational entity. Via internet, users transmit (share without modifying) or transform (share and modify) IM, an act that is facilitated by the fact IM's property is not own by anyone.

2 Related Works

In this section, we discuss related work of our project. We will firstly focus on works in the field of humor generation, with a specific focus on puns and jokes.

In *Automatic Joke Generation: Learning Humor from Examples* , Winters et al. implemented a Generalized Analogy Generator (GAG) that generates jokes of the form: *"I like my X like I like my Y, Z"*. For evaluation, they used JokeJudger.com. It is a platform they created to allow user to create and rate jokes already existing in the platform. To evaluate a joke, users had to answer on a Likert Scale (from 0 (not funny) to 5 (Very funny)). Just like in our project, they used a template of an analogical form. They did not generate jokes from explicit models or schemas, but their model learns the template from rated jokes examples. Jokes generated by Gag are perceived as funny half as often as human-created jokes.

In *An implemented model of punning riddles* , Binsted and Ritchie created one of the first joke generators. They used what they called a "schemata". A schema *"stipulates a set of relationships which must hold between the lexemes used to build a joke"* Binsted and Ritchie (1994). They also used a template that produces the general form of the pun that specifies relationships between words of the schema. They created JAPE, which generates jokes. They used lexicon, schematas, templates and a post-production checker. Their evaluation procedure consisted of three steps: data acquisition, common knowledge and joke judging. 14 human evaluators had to rate the joke from 0 to 5, and were

[1] 4chan is an anonymous imageboard website, available here
[2] Reddit is a social news and discussion website where users can submit, vote, and comment on content. See here

asked a few questions (how the joke can be improved ? Have you heard jokes like this before ?). Also, they did not create a joke control group.

Concerning meme generation, a few project had been conducted. MEMEGERA 2.0 is a project where, given a headline is generated a meme template selection and a text to go in it in portuguese (Oliveira et al., 2016). However, this project mainly focuses on the combination between a prompt and its template, which we do not have interest in.
In *Meme Generation for Social Media Audience Engagement* (Kurochkin, 2020), author use deep learning techniques and GPT-2 fine-tuned model to generate memes and post titles. They select a meme template, the text to go inside it and the post title that will go with it. For evaluation, they conducted a human-evaluation using Amazon Mechaniical Tuck - participants had to say if they liked the meme (yes or no)-. Then, they compared the engagment of their memes to the one of human-created memes. Their results is that their memes showed less engagement than random human-created memes.

From our knowledge there exist no project similar to our. Indeed we want to create a pun (based on analogical template) and then to generate an associated meme. GAG (Winters et al., 2018) uses a template to create a joke pun, Kurochkin use neural approach to create a great combination of text and image, but non of them create a pun with its associated meme.

3 Our approach

We implemented Punalogy, a program that generate a pun and a meme associated. The github project is accessible online here [3]. The pun is of form *"A is to B as C is to D"*, where A, B and C are existing english words and D is a neologism created by the generator. The general architecture of Punalogy is given by Figure 3. We will explain more in detail about all components of our program in the following section.

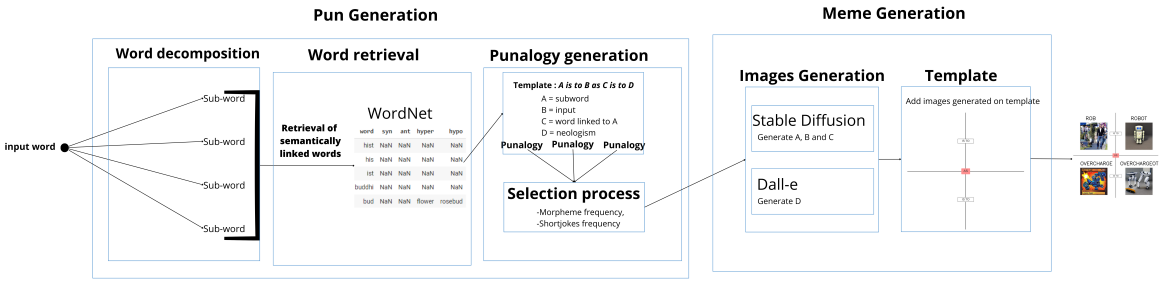

Figure 3: Punalogy architecture

3.1 Pun generation

The first step of Punalogy is to create a pun with an input word. Our approach is rule-based. This step is decomposed into four: word decomposition, word retrieval, fill the template blank and selection process.

3.1.1 Word Decomposition

This step is the first of our program, it is represented as the first rectangle from the left in figure 3. It's task is to decompose the input word into subwords. Our approach is the following: (1) the input word is splited into all possible set of caracters. Only word whose lenght is superior or equal to three and inferior to input word lenght are kept.
For example, the word *Buddhist* is splited into : *bud, budd, buddh, buddhi, buddhis, udd, uddh, uddhi....* Then, only subwords that are english words are kept, by checking its existence with **nltk**. For *buddhist*, only 5 subwords corresponds to english words: *hist, his, ist, buddhi and bud.*

3.1.2 Word Retrieval

The second step is the word retrieval (see second rectangle in figure 3). All subwords that were validated are now thrown into WordNet (Fellbaum, 1998) to retrieve its related words.
WordNet is a lexical database that organizes English words into sets of synonyms (synsets) and defines the relationships between them. WordNet also includes definitions and examples of usage for each synset, as well as relationships between words, such as synonymy, antonymy, and hyponymy.
This step allow to create a dataframe that store all existing subwords and their synonyms, antonyms, hyperonyms and hyponyms. For example, with all subwords from *buddhist*, only *bud* got other words related: *flower (hyperonym), rosebud (hyponym).*
Another word retrieval was also conducted with similar words. Indeed, subwords were added one letter, and checked if existed, and then retrieve related words via WordNet.

3.1.3 Filling Pun Template

This step take all result from now to fill the analogical template of the form *A is to B as C is to D*, where:

- A is a subword,

[3]For the complete url: https://github.com/Camille-saran/puns_analogies

- B is the input word,
- C is the word retrieved with WordNet,
- D is a neologism

For A,B and C, we just have to fill the blank. For D, we have to generate it. To do that, we take word B, delete the A part and replace it with C.

In our example, A is *bud*, B is *Buddhist*, C is *flower, or rosebud* and D should be created. To create it, we just substitute *bud* in *Buddhist* by *flower*. This create the following pun : *Bud is to buddhist as flower is to flowerdhist*.

At this step, all possible pun are generated and the following step will elect one.

3.1.4 Selection process

This step is the one that chooses one pun among the various generated. In order to do that, our program check frequencies for each word. Indeed, by looking at our first results, we discovered that the main problem was that words were not really known, and very domain-specific. This is mainly because of WordNet that propose very strict antonyms, synonyms, hyperonyms and hyponym. To deal with that, we have two steps:

1. Retrieval of frequency by each word, using NLTK (Loper and Bird, 2002),
2. Retrieval of each word frequency in ShortJokes dataset, available on Kaggle. This dataset contains around 0.2 million jokes, collected from the web.

By getting the frequencies for each words in the analogy, we compute a "selection score" which the mean of median of frequencies in english and median of frequencies in jokes dataset. The punalogy who obtain the higher selection get selected. This step marks the end of pun generation by selecting one punalogy.

3.2 Meme Generation

The second task of our project is to generate a meme. As for the pun, we use a template. Our goal is then to generate the four images (corresponding to the four words of the pun), and then we add it on a template.

3.2.1 Image Generation

We generate the images for A, B and C with Stable Diffusion (Rombach et al., 2021), accessed via Hugging face. It is a latent text-to-image diffusion model.
We generate image D with Dalle-E mini (Dayma et al., 2021), it is a transformer-based text-to-image generation model. It is an open-source version of Dall-E (OpenAi).
We use Stable diffusion model because images A, B and C because images can be generated in a very detailed, photo-like way. On the contrary, we use Dall-e mini model to generate D since it generates "funny picture", that is to say images generated can be absurd. We wanted to mix text-to-images models, because we want to create a SD and create conflictness, which (according to GTVH) will provoke humor.
Prompts used to generate images where the words itself. We tested various prompts on it, and it appears that better results come from using only the word as prompts.

3.2.2 Template

We created a template shown in figure 60. This is a template we created. It has four spaces that corresponds to all elements of punalogy. The word template *"...is to ... as... is to ..."* is written directly on the image template. Having our four images generated, we now have to fill this template. We use PILLOW library (Clark, 2015) to paste all generated image on it.

First analysis of our results suggested that our template is sometimes not really funny, and we were suggested to use different kind of templates. This is why we used well-known meme templates :

- Pam Beesby's Interview template, see Figure 5.
  The image is from the US version of The Office, a comic TV-show.
  On the 26 of january 2023, it was considered the sixth most used on Reddit by the website Meming Wiki.

- Gru's plan template, see Figure 6.
  The image is from the movie *Despicable Me*.
  On 25 of january 2023, this template is considered the tenth most used on Reddit by the website Meming Wiki.

- Winnie's template, also known as *"Tuxedo Winnie the Pooh"*, see Figure 7.
  The image is from the movie *"Winnie the Poo and Tigger Too"*.
  On 25 of january 2023, this template is considered the seventh most used on Reddit by the website Meming Wiki.

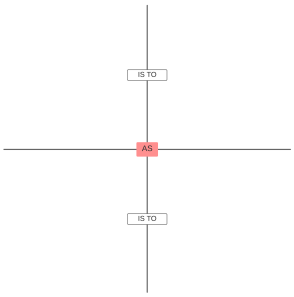

Figure 4: Original Square Template

Figure 5: Gru's plan template

Figure 6: Pam Beesby's interview template

Figure 7: Winnie template

4 Evaluation

To evaluate our results, we conducted two experiments: the first one is a human evaluation assessed via an online questionnaire, the second one is the use of a metric that compute engagement score on Reddit, here you can find our user account link `https://www.reddit.com/user/meme_pun/`.

4.1 Human evaluation

In order to conduct our **Online Survey**, we created three online questionnaires, each filled with 15 randomly selected memes. We privileged these ways of working because it is too heavy to ask our respondents to note more than 50 memes. We came to the conclusion that 15 memes is a good compromise. After asking some personal information like the age, the gender and the familiarity with memes, we demand to rate our memes on a scale of 0 to 5, where 0 means not funny at all and 5 means

really funny. We also put a comment's section after each meme to permit our respondents to give us more precise feedback.

4.1.1 Quantitative results

We managed to collect more than 70 responses. Majority of our respondents are under 32 years old, 58% born between 1990 and 1999 and 27% between 2000 and 2009. It would have been preferable for us to interrogate a wider range of people in order to evaluate specific aspects of our memes like comprehension and accessibility. But in an other hand, people born between 1990 and 2009 are usually the target audience for memes.
The survey permits us to create our **HE Score** (Human Evaluation Score), which is the average score given by respondents for a meme. This score, between 0 and 5, is one of the tools we use to evaluate our analogies. As we expected, most of the respondents rate the memes quite low with an average of 1,47. We do not notice significant difference between the different templates, but it can be interesting to look further into this direction in future human evaluations. Also we have collected a lot of interesting comments that will greatly help us to improve our results.

4.1.2 Comments

First, we had a lot of comments of people who did not understand the memes because they did not have the reference: *"I don't have much knowledge about english expressions"*, *"Idk a lot of word puns"...* Or sometimes because of the absurdity of the meme: *"Makes absolutely no sense"*, *"what does this even mean im so confused"*, *"Cannot comprehend the meaning"*, *"very absurd"....* But as some comments mentioned it, absurdity can produce a very good humorous effect: *"It doesn't make sense! But that kind of makes it funny"*, *"I don't get it but the randomness is funny lol"...* Also we had a lot of comments about the templates. In general, people found the templates good, but we had some responses who raise that some memes don't match with their templates: *"This doesn't really fit in this meme format"*, *"Most of them would be funnier without the template"*, *"Funny but not with this template"...*
Some comments also raise some very good points:

- *"feels like it could be racist lol"*: This is a danger that we need to keep in mind, we need to control our results in order not to risk to offend anyone with our memes.
- *"it makes me think but it doesn't make me laugh"*: These comments show us that even if the meme is not funny, it can have an interest. So it can be interesting to try to evaluate our memes on others criteria than humor.

Finally, we also had some very positive comments that encouraged us to pursue our efforts: *"Analogy generation not seen, which is good"*, *"very adapted to current events"*, *"Made me laugh, good meme"...* Globally, those results are in line with our expectations. We already knew, before sending our surveys, that our memes can be improved. But these results gave us some really good improvement axis.

4.2 RAS (Reddit Appreciation Score)

Reddit is a social media platform that enables users to create and manage their own communities known as "subreddits." These communities are dedicated to specific topics or themes, allowing users to post content such as text, images, and links related to those topics. Other users can then interact with the content by commenting, upvoting or downvoting. The front page of the website displays a selection of popular posts from various subreddits, and users can also browse and search for specific subreddits based on their interests.

Our goal was to utilize this platform to gather direct feedback on the output of our program. After creating an account and we started to put meme to evaluate . We used the upvotes (equivalent to a "like") and number of view as a measure of the appreciation of the puns generated by our program along with corresponding images.

Given that humor is subjective and can vary, we considered Reddit to be a valuable resource for studying and understanding different types of humor and how they are received by different audiences. Its large and diverse user base allows for multiple communities and subcultures with their own sense

of humor. Additionally, the platform's voting system and active community facilitate discussions and feedback, providing a rich source of data for researchers to understand how their content is received by their target audiences.

To evaluate our project, we created a metric by selecting important information from the platform. For each meme, we assigned an evaluation score "RAS" (Reddit Appreciation Score) that aligns with the French phrase "Rien A Signaler." RAS is calculated using the number of shares, the number of upvotes, and the total number of views of the meme. To make this metric useful and make comparison, we created the following equation:

$$RAS = ((share * uptove)/nbviews) * 100 \tag{1}$$

We use uptoves and views on Reddit as a way to evaluate the success of memes within our community. The number of uptoves indicates the level of appreciation from the community, while the number of views shows how widely the meme has been seen. As discussed in *Background* section, the impact of a meme can vary depending on factors such as the reference and the community it is being shared in.

The results presented in 4.2 show that while there are some interesting starting results, the scores, such as RAS/HE, are relatively low. But we are thinking that there is always room for improvement.

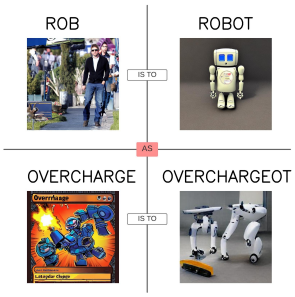

Figure 8: Score RAS : 2

Figure 9: Score RAS : 7.2 HE: 2.9

Figure 10: Score RAS: 3.5 HE: 1.20

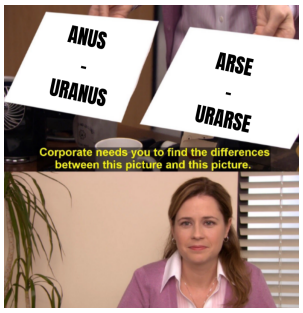

Figure 11: Score RAS : 3 HE: 2.2

The results presented in section index A.1 show that the metric RAS may not be fully optimized. To improve visibility and variation in the results, we are planning to work on the metric to take into account factors such as downvotes. We also want to ensure that the number of shares is not overvalued. As a next step, we will look for ways to increase the score by modifying the generation process.

5 Discussion

Our evaluation results are deceiving. Indeed, overall our memes were not considered as funny by participants.

Concerning the generation of puns, we find that it can be improved. Indeed, the results of our evaluations show that the generated puns are not very funny. We think that the word retrieval function with WordNet does not work for our project. For example, very few antonyms are found, and it is from semantic opposition that humor is born.

Concerning the choice of template, our results indicate that the basic template (the one we created) obtains worse results than the existing templates. This finding seems normal to us, since according to memes research it is precisely the fact of taking a template and transforming it that is funny. Our results also indicate that while some puns become funnier when combined with certain templates, they become less fun when combined with others. Thus, the combination between the template and the pun should be studied more in order to generate a funnier pun.

Concerning the generation of images, we think that the absurd effect of the generation could make people laugh, since this visual opposition between the word and the generated image could introduce a notion of opposition, and thus become funny. However, the link between the image and the word needs to be studied in more detail in order to know exactly which link would be funnier.

One of the main disadvantages we faced during this project was the evaluation of humor, as it can be quite subjective. To address this issue, we decided to use social network Reddit as our evaluation platform. Additionally, we conducted human evaluations with several quizzes sent to people to gather their subjective opinions on the generated analogies. We also used a metric to measure the performance of our program, but we realize that it has its own limitations. However, we found that the metric has some limitations and the number of views and shares can be biased, so we are likely to rework on it.

In future implementations, we plan to incorporate machine learning techniques to improve the performance of the program. Currently, we are only using rule-based program to generate analogies, but by incorporating machine learning, we hope to achieve more accurate and diverse results. Machine learning models can learn patterns in the data and generalize to new examples, which can be useful in the context of analogies generation. We expect that the integration of machine learning will enhance the performance and robustness of the program.

Our next current focus is on testing new ideas for generation that involve restricting output to specific universes and combining them (index A.2). We plan to incorporate a defined theme to enhance humor, blend cartoon and other elements, and explore the use of machine learning and humorous metrics. Additionally, we are considering creating a metric to objectively evaluate humor, which would be beneficial for the future of humor assessment. Our goal is to continue refining the model and its results to produce interesting and useful outcomes.

6 Conclusion

Overall, this project should be understood as the prototype of a larger project. Indeed, although functional, the comic effect is not often achieved. For this, Punalogy should be further developed. Similar works with better results are those based on neural network approaches. Thus, in the following work the focus should be on:

- Obtaining a finer representation of words (to allow a better semantic opposition, and thus a greater comic effect). This representation should be done by words embeddings. However, as our project also needs morphological features for the words, the embeddings should also contain morphological information, as in Alsaidi et al. (2021).

- The selection of the pun should be improved by directly selecting the funniest pun. This step could be learned by a model, using a dataset of pun and it's associated scores. Thus, given enough data, the model could learn what is funny and select it. For this, a platform like Jokejudger (Winters et al., 2018) for retrieving data should be used to build the dataset.

- The combination of template and pun should be further studied. As in Kurochkin's work, we should find a way to combine template and text in the best way.

- Finally, we should also think about other types of puns. Indeed, it has been suggested to us that we should only extend to a few specific domains (like a video game, a TV show, or a theme). As we have studied, the script opposition is intrinsically linked to what an individual considers to be 'normal'. But normality is based on what we know. Thus, limiting ourselves to a few areas would allow us to effectively target a specific population, and be sure that they have the necessary references to understand the meme.

Final findings of the project are somewhat disappointing. However, we believe that this project is still valuable. Indeed, although results are not the ones expected, the structure of the program is functional. Thus, changes should be made to improve tasks, but not to the structure of the project. Therefore, we believe that Punalogy provides a basis for memes generation, although it needs to be improved.

References

Dictionary (6th Edition). Oxford University Press, 2012. Compton Effect.

Kholoud Khalil Aldous, Jisun An, and Bernard J. Jansen. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):47–57, Jul. 2019. doi: 10.1609/icwsm.v13i01.3208. URL https://ojs.aaai.org/index.php/ICWSM/article/view/3208.

Valitutti Alessandro. How many jokes are really funny? towards a new approach to the evaluation of computational humour generators. In *Proceedings of International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2011)*, pages 189–200, Denmark, 2011. Samfundslitteratur. International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2011) ; Conference date: 20-08-2011 Through 21-08-2011.

Safa Alsaidi, Amandine Decker, Puthineath Lay, Esteban Marquer, Pierre-Alexandre Murena, and Miguel Couceiro. A neural approach for detecting morphological analogies, 2021. URL https://arxiv.org/abs/2108.03945.

Albert Atkin. *Peirce's Theory of Signs*. 01 2010.

Salvatore Attardo. The General Theory of Verbal Humor. In *The Linguistics of Humor: An Introduction*. Oxford University Press, 06 2020. ISBN 9780198791270. doi: 10.1093/oso/9780198791270.003.0007. URL https://doi.org/10.1093/oso/9780198791270.003.0007.

Kim Binsted and Graeme Ritchie. An implemented model of punning riddles. 07 1994.

Kim Binsted and Graeme D. Ritchie. Computational rules for generating punning riddles. 1997. doi: 10.1515/humr.1997.10.1.25.

Sara Cannizzaro. Internet memes as internet signs: A semiotic view of digital culture. *Σημειωτκή-Sign Systems Studies*, 44(4):562–586, 2016.

Carlos Mauricio Castano Diaz. Defining and characterizing the concept of Internet Meme. *CES PsicologÃa*, 6:82 – 104, 12 2013. ISSN 2011-3080. URL http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S2011-30802013000200007&nrm=iso.

Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4603–4612, 2016.

Alex Clark. Pillow (pil fork) documentation, 2015. URL https://buildmedia.readthedocs.org/media/pdf/pillow/latest/pillow.pdf.

Patrick Davison. The language of internet memes. *The social media reader*, pages 120–134, 2012.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khc, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021. URL https://github.com/borisdayma/dalle-mini.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Smadar T. Kedar-Cabelli. Analogy — from a unified perspective. 1988.

Andrew Kurochkin. Meme generation for social media audience engagement. 2020.

Edward Loper and Steven Bird. Nltk: The natural language toolkit, 2002. URL https://arxiv.org/abs/cs/0205028.

Rada Mihalcea and Stephen Pulman. Characterizing humour: An exploration of features in humorous texts. Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 9783540709381. doi: 10.1007/978-3-540-70939-8_30. URL https://doi.org/10.1007/978-3-540-70939-8_30.

Kate M Miltner. Internet memes. *The SAGE handbook of social media*, 55:412–428, 2018.

Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes!–explorations on the automatic generation of internet humor. In *Proceedings of the Seventh International Conference on Computational Creativity (ICCC 2016). Paris, France*, 2016.

He Ren and Quan Yang. Neural joke generation. 2017.

Graeme D. Ritchie. Developing the incongruity-resolution theory. 1999.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Limor Shifman. The cultural logic of photo-based meme genres. *Journal of visual culture*, 13(3): 340–358, 2014.

Villy Tsakona. Language and image interaction in cartoons: Towards a multimodal theory of humor. *Journal of Pragmatics*, 41(6):1171–1188, 2009.

Peter D. Turney. A uniform approach to analogies, synonyms, antonyms, and associations. 2008. doi: 10.48550/ARXIV.0809.0124. URL https://arxiv.org/abs/0809.0124.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Thomas Winters, Vincent Nys, and Daniel De Schreye. Automatic joke generation: Learning humor from examples. In Norbert Streitz and Shin'ichi Konomi, editors, *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*, pages 360–377, Cham, 2018. Springer International Publishing. ISBN 978-3-319-91131-1.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1153. URL https://aclanthology.org/P18-1153.

A Index

A.1 Generated puns:

Figure 12: Score RAS:0.2

Figure 13: Score HE: 2.2 / Score RAS: 1.9

Figure 14: Score HE: 0.8

Figure 15: Score HE: 1.2 / Score RAS: 3.47

Figure 16: Score HE : 1.2

Figure 17: Score HE: 1.57 / Score RAS: 0.14

Figure 18: Score HE: 1.79

Figure 19: Score HE: 0.8

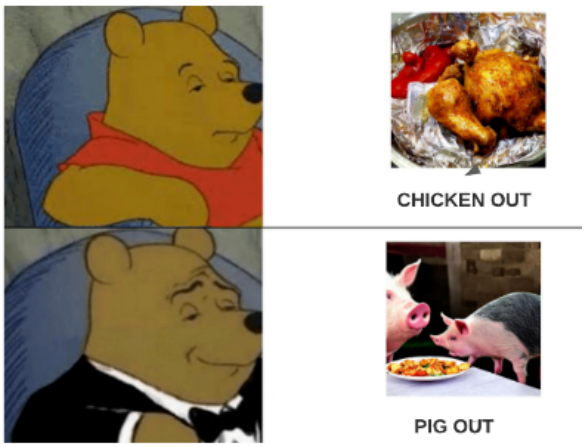

Figure 20: Score RAS: 0.35

Figure 21: Score HE : 1.6

Figure 22: Score HE: 1.23

Figure 23: Score RAS: 0.6

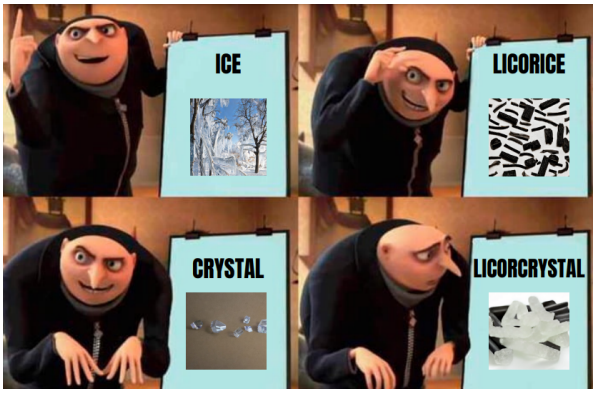

Figure 24: Score HE : 1.01

Figure 25: Score RAS : 0.1

Figure 26: Score HE: 1.23

Figure 27: Score HE : 2.06

Figure 28: Score HE: 3.85 / Score RAS: 1.7

Figure 29: Score HE: 1.15

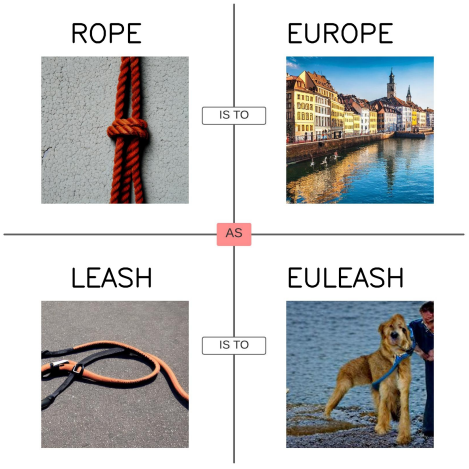

Figure 30: Score HE: 1.06 / Score RAS: 0.1

Figure 31: Score HE: 1.72

Figure 32: Score HE: 1.36

Figure 33: Score HE: 1.33

Figure 34: Score HE: 1

Figure 35: Score HE:1.06

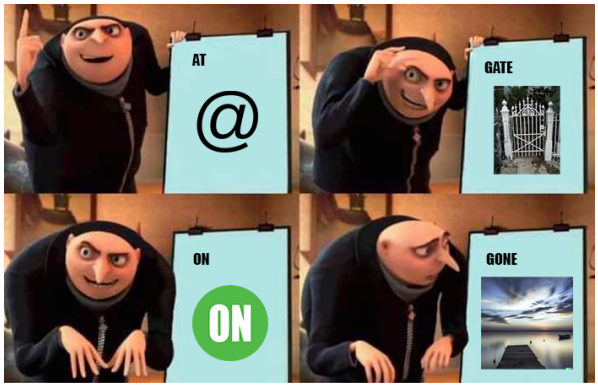

Figure 36: Score HE:0.66

Figure 37: Score HE: 2.7

Figure 38: Score HE: 0.66

Figure 39: Score HE:2.9 / Score RAS: 7

Figure 40: Score HE: 1.13

Figure 41: Score HE: 1.9

Figure 42: Score HE: 0.8

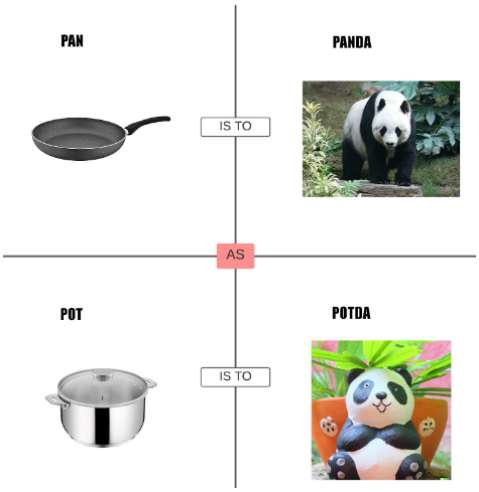

Figure 43: Score HE:1.23

Figure 44: Score RAS: 2

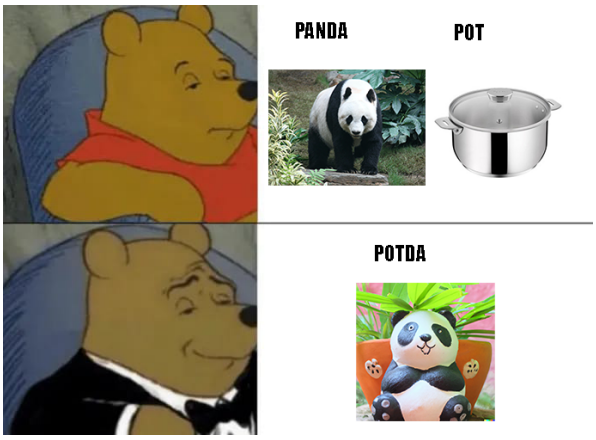

Figure 45: Score HE: 1.33

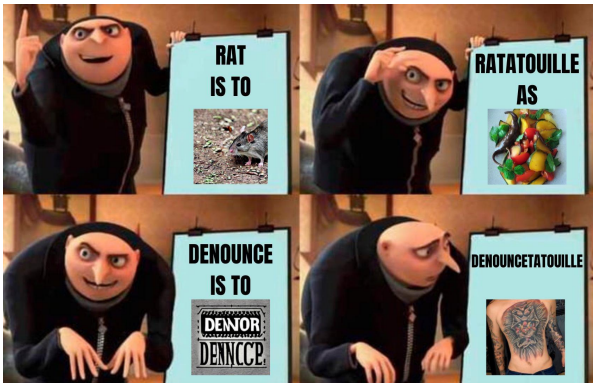

Figure 46: Score RAS : 2.93

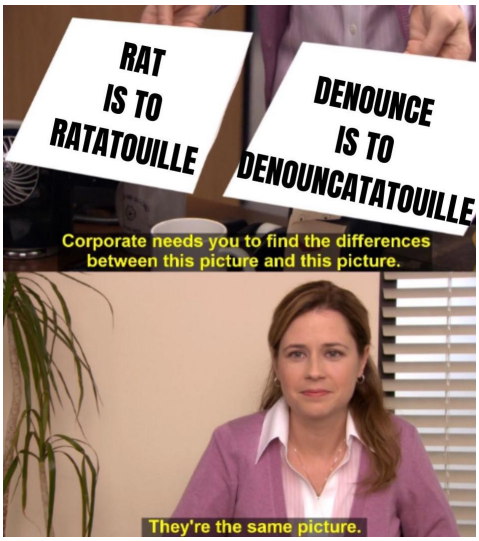

Figure 47: Score RAS :0.375

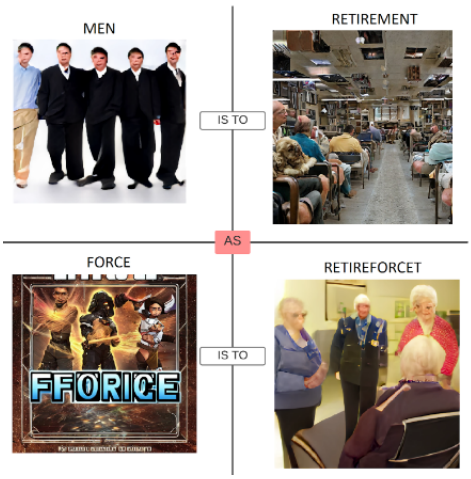

Figure 48: Score HE: 1.18

Figure 49: Score HE: 2.65 / Score RAS: 0.57

Figure 50: Score HE: 0.33

Figure 51: Score HE: 0.77

Figure 52: Score HE: 2.45 / Score RAS: 2

Figure 53: Score HE: 1.06

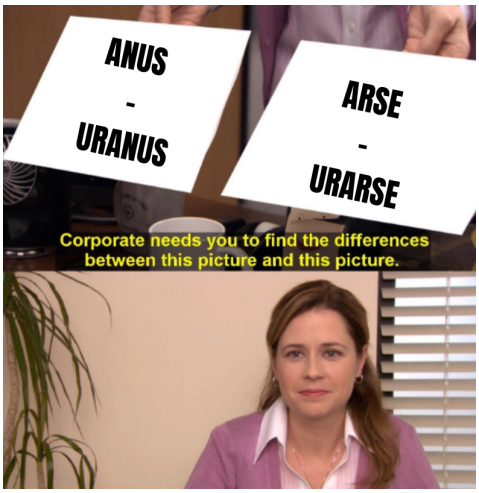

Figure 54: Score HE : 2.63 / Score RAS : 2.9

Figure 55: Score HE: 0.81

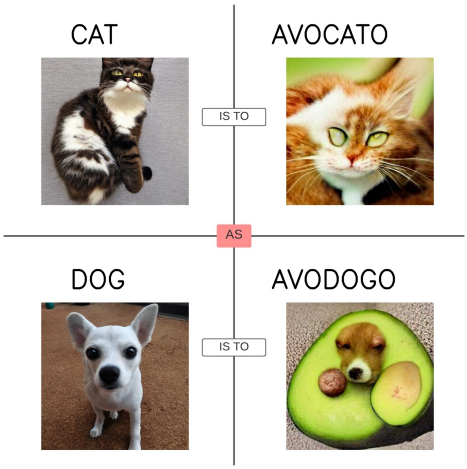

Figure 56: Score RAS : 1

Figure 57: Score RAS : 0.2

Figure 58: Score RAS: 2.8

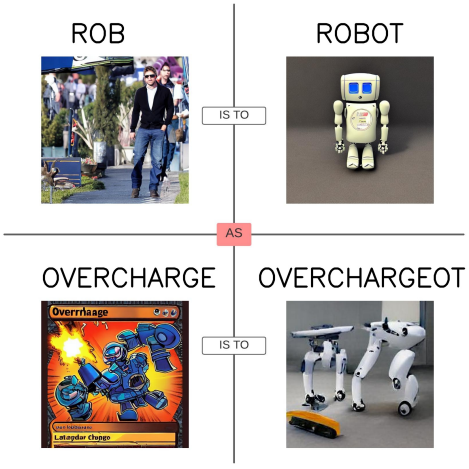

Figure 59: Score RAS : 2

A.2 Test

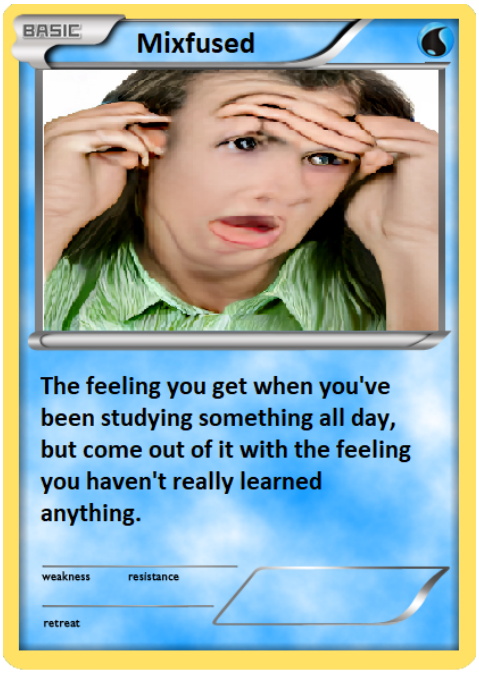

Figure 60: Score RAS :1.3

Figure 61: Score RAS: 102

Figure 62: Score RAS: 10.1

Figure 63: Score RAS: 28.87