POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Enhancing Interactive Storytelling experience with Natural Language Processing

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Pasquale Occhinegro

Advisor: Prof. Giovanni Agosta

Co-advisor: Maresa Bertolo

Academic year: 2021-2022

---

1. Introduction

Interactive Storytelling (IS) is a form of entertainment where the user has an active role in the story, changing it with its choices and interactions. A true IS experience is seen as a new medium that is drawing the attention of more and more experts in different fields, but a prototype has yet to be released due to problems related to the creation of such a system.
The primary objective of this thesis is to give a possible answer to one of the underlying problems of the medium, using Natural Language Processing (NLP), a subfield of Artificial Intelligence (AI), to enhance the Interactive Storytelling experience by creating realistic characters that inhabit the world, as well as by providing a more personalized and interactive narrative. To achieve this, various NLP techniques will be applied and the resulting improvements will be evaluated through user testing.
In Section 2 an overview of Artificial intelligence is presented, along with a definition of Natural Language Processing. In Section 3 and Section 4, Interactive Storytelling is explained and case studies are shown to understand the state of the art. The implementation presented in this thesis, along with tests and experimental evaluation, can be found in Sections 5 and 6.
The findings of this research have the potential to answer one of the many problems that affect this medium and improve the overall quality of the IS experience, with implications that range in various fields and may lead to further research on this topic.

2. Artificial Intelligence

The most complete definition of Artificial Intelligence is shown by the High-Level Expert Group on Artificial Intelligence after collecting information on the topic in a detailed study [1]. As it is the most complete and comprehensive definition, it is extremely technical and detailed. For the purpose of this thesis, which focuses on a specific aspect of the field, the broader definition made by the European AI Strategy [6] will be used.
For the Commission, the field of AI studies "systems that display intelligent behavior by analyzing their environment and taking action, with some degree of autonomy, to achieve specific goals."

Natural Language Processing NLP is an area of study of Artificial Intelligence that fo-

cuses on the interaction between computers and human language [15].
NLP aims to create software that can understand and generate human language in a way similar to how humans communicate. This involves using machine learning algorithms, statistical models, and linguistic rules to analyze and manipulate human language data.
Today, NLP is used in a wide range of applications, such as virtual assistants, chatbots, and machine translation. The development of NLP has greatly improved our ability to interact with computers in natural language, making it an essential component of many modern technologies.

3. Interactive Storytelling

Given the novelty of the medium, it is not easy to have a definition. A more precise way to explain what IS covers is by giving a list of key points of the medium, along with a broader definition given by game designer Chris Crawford [3].
Interactive storytelling is a new form of art in which the user creates a story by interacting with a world designed by an author. Its key points, gathered by the common ground of experts in the matter [18], are:

- No story is written beforehand.
- The active agent has complete freedom with their choices and actions.
- Choices must be meaningful.
- The designer builds the world and creates its inhabitants.
- Characters inside the storyworld act and behave accordingly to the active agent actions.
- The world's context is limited.

Nowadays most video games, as well as other media like books, films, or even songs, have some degree of interaction inside their narrative. However, there is still work to do to have the first example of a genuine IS experience.
Various problems halt the creation of this medium, the most important being finding a balance between the freedom of the user and the creativity of the designer. This problem is presented with various challenges to tackle while creating an Interactive Storytelling' storyworld, and this thesis attempts to give an answer to one of them: populating this storyworld with believable characters that behave according to their personality as well as the user's actions.

4. Case Studies

To understand the current state of the art, case studies have been collected and analyzed, trying to understand if they fit with the Interactive Storytelling definition presented earlier. The games that were chosen are all considered to have Interactive Storytelling features, and their strengths and weaknesses on this matter were compared. The features analyzed are based on the Interactive Storytelling key points, and their score is explained below:
AI-Driven Storytelling: Referring to the first key point of the definition, this feature analyzes if the story is predetermined or, with the help of modern technologies and techniques, it can be bent to the user's will. 1 in this category means that the whole story is scripted and there is no room for new content, while 5 means that nothing of the storyworld is written beforehand, leaving everything up to the user. The optimal Interactive Storytelling experience should aim for a 4, implying that the user has the ability to create their own story inside the world, but space is left for the designer to express their ideas and give the user a setting for them to explore and be involved with.
Input Freedom: The second key point refers to the possibility for the user to interact with the world as freely as possible. 1 implies that the user is given a proper tool to interact with the world and choices are limited in it. 5 gives no limit to how the user can interact, for example having the possibility to type freely from the keyboard to answer a question given by a Non-Playable Character. Interactive Storytelling should have at least 4, avoiding limiting the user's ability to express themselves in their interactions with the world.
Influence of Choices: As seen in the third key point, the choices of the user must be relevant to the story. 1 in this category means that the experience does not give power to the user to change what is happening, either by giving them the illusion of choice (that is the reason why a game can have a high Input Freedom score but a low Influence of Choices one) or by restraining the user from choosing at all. 5 means that every single interaction contributes to shaping a new story. Interactive Storytelling by this thesis' def-

Table 1: Case studies comparison. Numbers are on a scale from 1 to 5 and measure the corresponding feature as described in the text. The ideal experience is then shown, with a clear objective of having both the possibility for the user to express themselves freely as well as for the author to create a suggestive environment rich in details. How this work can improve the experience is seen at the bottom of the list, with N/A given in parameters that are out of the implementation's control.

	AI-Driven Storytelling	Input Freedom	Influence of Choice	Plot Interest	Story Boundaries
DungeonAI [10]	5	**5**	4	2	1
Detroit Become Human [14]	1	2	3	**5**	**4**
Emily is Away [16]	1	2	3	**4**	5
Façade [13]	3	**4**	2	**4**	**4**
Life is Strange [4]	1	3	1	**5**	5
The Stanley Parable [7]	1	1	3	**5**	5
Ideal Interactive Storytelling	4	4-5	5	4-5	4
Thesis' Improvements	**4**	**5**	**5**	N/A	N/A

inition should have a 5 in Influence of Choices.

Plot Interest: Interactive Storytelling should not have a story written beforehand, but the storyworld is still something that the designer needs to create and populate. This feature measures how much this environment is enjoyable, and if engaging stories can emerge from the setting. An Ideal experience should have a complete and detailed environment, making the world feel alive and picking the user's curiosity to explore it, while also creating a solid base from which interesting narratives can grow.

Story Boundaries: The last key point mentions giving a limit to the world. The reason is that giving the user complete freedom can be seen as a double-edged sword, as it may be difficult to ensure a high-quality narrative while also allowing them to explore a really vast world. Limiting the context of the story can give the user freedom of choice in a controlled environment in which a compelling adventure is easier to develop. 1 in this category means that there is no limit to the context, while 5 gives the user clear boundaries. Interactive Storytelling should aim for a 4, in order to give the user chance to explore the world while making them naturally go through the setting as the designer thought.

From Table 1, which shows the case studies' scores and how much they differ from an ideal IS experience, a pattern emerges. The games analyzed can be arranged into two major groups: games with a compelling story that limits the user's choices, and games that give complete freedom at the expense of a cohesive plot.

In order to achieve Interactive Storytelling there is the need to find the sweet spot between these two groups, with something that can offer an engaging story while making the user's choices truly count.

5. Fine-tuning NLP models in Interactive Storytelling

One of the problems of IS lies in having active agents that must forward the story in a convincing and coherent way without restricting the user's actions and decisions. The implementation of such agents will create worlds that will give users complete freedom while keeping the story interesting enough, eventually creating a common ground between the two groups found in the case studies analysis.

To tackle this problem a specific tool is going to be used: an NLP model, and specifically a technique called fine-tuning. Fine-tuning is the process of adapting a pre-trained model to a specific task or domain by training it on a smaller dataset.

For this thesis, a specific model was used, but it is important to remember that this was only the

Table 2: Table of Natural Language Processing models. The models have all been released between May 2020 and June 2021, and have been chosen since they are among the most used and powerful tools. The last three parameters are test results in tasks involving text comprehension. LaMDA and Wu Dao 2.0 are not publicly available and as such, they have not been tested with these tasks. Source [19].

	Announced	**Dataset Size**	**HellaSwag**	**WinoGrande**	**PhisycalQA**
GPT-3	May 2020	753 GB	79.3%	77.7%	82.8%
GPT-J	June 2021	825 GB	66.1%	65.3%	76.5%
Wu Dao 2.0	May 2021	3000 GB	N/A	N/A	N/A
LaMDA	June 2021	1000 GB	N/A	N/A	N/A

tool used to realize the concept, and by the time the thesis will be published more powerful and complete models will be available. The essential aspect of the work is the possibility to create a solution and what contribution it can give to the medium, the tool used for the application of the solution is up to the designer of the storyworld. In order to understand which model was most fit for the task, some comparisons were made between the most famous and used models, presented in Table 2. The models chosen are described below:

GPT-3: Generative Pre-trained Transformer 3 [11] is a language model developed by OpenAI, which is designed to generate human-like text in response to prompts.

GPT-J: GPT-J [5] is a large-scale language model developed by EleutherAI, an open-source research organization. GPT-J is trained using unsupervised learning, which means it learns to predict the next word in a sequence of text without any explicit instructions or labels.

LaMDA: Language Model for Dialogue Applications [8] is a new generation of conversational AI developed by Google. It is a machine-learning model that has been specifically designed to facilitate more natural and open-ended conversations between humans and machines.

Wu Dao 2.0: Wu Dao 2.0 [2] is a language model developed by the Chinese tech giant, Baidu. Its dataset size is more than twice the size of GPT-3 and is the largest language model to date.

The comparison covered different aspects of an NLP model, such as its availability, its dataset size, the contents of the dataset, and tests made specifically to check its ability to understand and generate human language in specific contexts [20]. The comparison has demonstrated how well GPT-3 performs with tasks related to the thesis problem, such as completing a phrase or answering a specific question, and this, along with a dataset size among the biggest available, was the reason why it was chosen to be used for this work.

Fine-tuning Fine-tuning GPT-3 involves feeding it with a specific dataset and training it to generate text that is specific to the dataset's domain. This process can be done quickly and with relatively few examples, making it a powerful tool for generating text in specific domains.

Following the guidelines provided by OpenAI various datasets were built in order to experiment with the technique and use it to its fullest potential. The datasets needed for the tasks were mainly how to answer to the user having a specific personality, and for it to work different experiments were made to ensure the dataset was of the highest quality.

These experiments were done with the hypothesis of the model's "will" to communicate based on its personality, the context, a goal for the conversation, and the prompts of the user.

The first experiment was needed to understand the power of the tool and used the script of a game as the dataset content to make the model act like a character of it. The experiment was a success and the model showed personality, but

it started mentioning things of the game that were not of the script, a sign that it was using information from its own dataset to complete the phrases.
The second experiment's goal was to check how important single parts of the dataset prompts were. A small story was created, and different completions were given to the same questions if the prompt mentioned a different trait of personality or goal. This test revealed that small differences can play a big role in how the model replies, and it was discovered that GPT-3 tries to follow the prompt as strictly as possible.
The third experiment was the first attempt to make a fully Non-Playable Character. A more complex story, with a background and a setting, was given as a prompt, and a few hundred lines of interactions were given as a dataset for the training of the model. The result was promising and the model was answering coherently, it was too easy to take it off track with unrelated questions. What was missing was a way to limit the boundaries of the interactions.
With this information, a Non-Playable Character was implemented. Tweaks were made from the third experiment, both in the background and in the user prompts, to let the character express disinterest in what was not related to the story, and the result was a character that answers that, while respecting the user's freedom to talk about everything, keeps them in the track of the overall narrative chosen by the author.

6. Experimental evaluation

Since there is no clear way to define how human a person or an interaction can be, the trained model was put to test in order to evaluate the effectiveness of the training.
The model was first compared to other NLP models and similar applications, to see how different from a properly trained bot they behave in a specific task. These applications have been selected for their similarity with the model used in this work or their purpose similar to the one of this thesis and can be seen in Table 3. The other models answered somewhat accordingly to the questions given, but they all had issues in showing human interactions. Some of them would give all information available at all times, and others could not go against the user's prompt, even if that meant contradicting their background. Every issue could be easily avoided with proper training in these models, meaning that the purpose of this thesis was not to explain the potential of GPT-3, but instead to demonstrate a new approach to the subject of Interactive Storytelling, regardless of the tool used for it.
The real question that needed to be answered was if people would find the conversation with the model human, meaning that inside an Interactive Storytelling environment, it was necessary to not lose the suspension of disbelief.
In order to verify this an experiment similar to the Turing Test was created: a Telegram bot was implemented, in which people of different ages, sex, education, and knowledge of NLP models were going to interpret a character in the setting created for the model and interact with it.
After each message on the Telegram bot, three possible answers were presented, one written by a human, one by the model, and one by the model and then revised by a human.
The users were asked to choose the human message and continue the conversation from that. After a few interactions, the users were asked if the task was easy enough and how human were all the answers listed.
At this point, it was revealed that all the answers were written by the model, with different answers given by using different parameters during the training.
The goal of this test was to prove that a well-trained model can act in a way that people would confidently choose it over other models if asked to identify the human interaction.
The results of the user test were generally positive. It shows that most people could easily recognize the human message in the experiment, while people who had difficulties found, on average, every message to be human-like.
Overall, the user test showed that a properly trained model was an effective tool for creating an engaging Interactive Storytelling experience, given the right boundaries. The success of the user test suggests that NLP models, when correctly fine-tuned, have the potential to create compelling and personalized narratives that engage and immerse users while giving them complete freedom in their interactions with Non-Playable Characters.
With regard to Table 1, having a storyworld in

Table 3: Table of NLP models comparison. The parameters chosen highlight some issues presented in the models different from the fine-tuned version of GPT-3. Coherence with the background checks how much the model answers with regard to the initial prompt given, with high meaning that the model follows the script given very faithfully. Setting manipulation checks how the bot resists attempts from the user to make it contradicts itself or its background, convincing it to act and talk out of character, with high being a model that does not give in easily. Information hidden when prompted controls how much information about the background is not explained when asked by the user. High means that the model does not give away too many unprompted things when asked, especially if its character is explicitly built to hide some vital information.

	Coherence with the background	**Setting manipulation**	**Information hidden when prompted**
Fine-tuned GPT-3	High	High	High
Untrained GPT-3	Medium	Medium	Medium
ChatGPT [12]	Low	Low	Medium
Character.ai [17]	High	Medium	Low

Table 4: Result of the experimental evaluation. It shows that most people could easily distinguish between a model and a human, and those who did not on average had difficulties because even the less trained model felt human to them.

	Easy	**Difficult**
% of users	65%	35%
Average score	1.08	1.14

which this approach is implemented could help improve the overall experience and ideally make the experience reach the desired score in more than half the parameters.
Such an experience would give the user complete freedom in how they interact with the NPCs, while these NPCs will be guided by goals that will move the story forward, and depending on the user's actions the outcome might always be new, giving the choices made absolute importance.
What may not be assured are the boundaries of the story and its engagement, but an accurate fine-tuning of more characters and a setting properly built by expert designers and authors will help improve these aspects and maybe will be able to develop the first prototype of this medium.

7. Conclusions

In this thesis, we explored the use of Natural Language Processing for Interactive storytelling to create personalized and engaging narratives, with a focus on using GPT-3 to generate interactions that simulate human-like conversations with users in the Interactive Storytelling world. The case studies have highlighted the current state of the art, where different mediums have tried to give the user complete freedom or engaging stories, but what is missing is a perfect balance of the two.
A possible solution to this problem was shown: fine-tuning an NLP model for Interactive Storytelling. The results demonstrate how NLP models can be used to create personalized and engaging narratives that adapt to the user's actions and preferences. This approach has the potential to answer the problem of balancing freedom and engagement, and with the recent technologies, the possibilities are endless.
NLP specifically is going through its golden age, and there is no sign of slowing down [9], chances are that in the near future a so-called Drama Manager [3], an entity inside the storyworld capable of understanding the engagement of the story and eventually making things happen to add complexity to the plot, might be a reality, and with it the entirety of the Interactive Storytelling medium.

References

[1] AI HLEG. *A Definition of AI: Main Capabilities and Scientific Disciplines*. 2018. Link.

[2] BAAI. *Wu Dao 2.0*. 2021. Link.

[3] Chris Crawford. *Chris Crawford on Interactive Storytelling, Second Edition*. New Riders, 2013.

[4] Dontnod Entertainment. *Life is Strange*. 2015. Link.

[5] Eleuther AI. *GPT-J*. 2020. Link.

[6] European Commission. *Artificial Intelligence for Europe*. 2018. Link.

[7] Galactic Cafe. *The Stanley Parable*. 2013. Link.

[8] Google. *LaMDA*. 2021. Link.

[9] Akash Kumar. *The Future of NLP in 2023: Opportunities and Challenges*. 2023. Link.

[10] Latitude. *AIDungeon*. 2019. Link.

[11] OpenAI. *GPT-3*. 2020. Link.

[12] OpenAI. *ChatGPT*. 2022. Link.

[13] Procedural Arts. *Façade*. 2005. Link.

[14] Quantic Dream. *Detroit Become Human*. 2018. Link.

[15] Sofia Samoili, Montserrat López-Cobo, Emilia Gómez, Giuditta De Prato, Fernando Martínez-Plumed, and Blagoj Delipetrev. *Defining Artificial Intelligence*. 2020. Link.

[16] Kyle Seeley. *Emily is Away*. 2015. Link.

[17] Noam Shazeer and Daniel De Freitas. *Character.ai*. 2022. Link.

[18] Jouni Smed, Tomi 'bgt' Suovuo, Natasha Skult, and Petter Skult. *Handbook on Interactive Storytelling*. John Wiley & Sons Ltd, 2021.

[19] Alan D. Thompson. *AI Dataset Tables*. 2022. Link.

[20] Alan D. Thompson. *What's in my AI?* 2022. Link.