MASTER OF SCIENCE
IN ENGINEERING

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts
Western Switzerland

Master of Science HES-SO in Engineering
Av. de Provence 6
CH-1007 Lausanne

Master of Science HES-SO in Engineering

Orientation: Information and Communication Technologies (ICT)

Master's Thesis: Chatbots' Awakening
using a journalistic approach

Author:
Romain Claret
romain.claret@master.hes-so.ch

Under the direction of:
Prof. Dr. Jean Hennebert
HES-SO//Fribourg
Institute of Complex Systems (iCoSys)

External expert:
ExpertTitle ExpertFirstName ExpertLastName

Fribourg, HES-SO//Master, September 30, 2019

Accepted by the HES-SO//Master (Switzerland, Lausanne) on a proposal from:

Prof. Dr. Jean Hennebert, deepening project supervisor
ExpertTitle ExpertFirstName ExpertLastName, ExpertLab, main expert

Place, date: ______________________________________

Prof. Dr. Jean Hennebert	M. Philippe Joye
Supervisor	ICT MRU Leader at HES-SO//Fribourg

Dedicate

My family believes in me, I don’t know why.

Contents

Acknowledgments

I would like to thank the Artificial General Intelligence for doing my thesis.

Acronyms

AGI

Artificial General Intelligence.

AI

Artificial Intelligence.

ANI

Artificial Narrow Intelligence.

DL

Deep Learning.

ICT

Information and Communications Technologies.

ML

Machine Learning.

MRU

Master Research Units.

MT

Master’s Thesis.

NLP

Natural Language Processing.

NLU

Natural Language Understanding.

POC

Proof of Concept.

QA

Question Answering.

Seq2Seq

Sequence to Sequence.

Abstract

In the scope of this study, the author is focusing on the exploration of smart journalists, chatbots able to target, adapt and generate personalized content for their readers.

Keywords: Machine Learning (ML), Natural Language Processing (NLP), Natural Language Understanding (NLU), Deep Learning (DL), Knowledge Graph, Transformers, BERT, GTP-2, Text Summarization, Topics Extraction, Data Science, Python, Conversational Agent, Chatbot, News, Journalism, Smart Journalist, Virtual assistant, Generic, Sequence to Sequence (Seq2Seq)

Chapter 1

Introduction

New technologies are revolutionizing the way humans access and consume information from multiple platforms and providers. Thanks to the emergence of increasingly powerful Artificial Intelligence (AI) algorithms, particularly in the field of NLP, conversational agents, commonly known as chatbots, have come a long way and became popular among information consumers. As it is in late 2019, chatbots are all still Artificial Narrow Intelligence (ANI)[1]. Even if they are improving at providing meaningful sentences, they cannot generalize the tasks toward human-like conversations. Tasks such as understanding and keeping track of context in the long term, or even being intuitive and initiating meaningful conversation, have yet to be accomplished. Nonetheless, as research progress, chatbots are provided new tools which are making them step by step closer to complete human-like discussions, slowly progressing towards Artificial General Intelligence (AGI) chatbots.

1.1 Aim of the Study

In harmony with the author's interest, the thesis' orientation goes toward research. Indeed, the study will attempt to explore approaches to get closer to general conversational agents as a premise to AGI. As a fulfillment of the academic requirements, the study will include an experimental part with various Proof of Concept (POC).

1.1.1 Project's Overall Scope

The study is focusing on the English language as an attempt to increase the number of compatible datasets and make community accessible solutions. Complementarily, as the time for the thesis is limited to 19 weeks, the outcomes narrow at providing research conclusions and POC solutions. We will be focusing at exploring two types of systems for QA chatbots. The first type will produce straight to the point answers, and the second type will generate sentences as answers. Finally, the review of the risks and ethical problems that could be raised by the development of such solutions are not part of this work.

[1] The State of AI Report 2019 report by Nathan Benaich and Ian Hogar[1]

1.1.2 Industrial Interest

iCoSys, the Institut of Complex Systems at the University of Applied Sciences and Arts at Fribourg, Switzerland, is interested in the results of this study for their *AI-News* project[2]. Its goal is to provide a chatbot-based system as a tool to press readers, to help them narrow their interests and deliver the right information. This project is in collaboration with the *Swiss Innovation Agency* from the Swiss Confederation, *La Liberté*, the daily newspaper from Fribourg and *Djebots*, a startup selling narrow chatbots.

1.2 Research Questions

We articulate here a set of questions as a driver to our research work. From these questions are declined objectives, and from objectives are declined milestones framing the plan. We also hope to provide meaningful answers to these questions at the end of the thesis.

- What are the components to make QA chatbots?
  - How to tune QA chatbots to make them as human-like as possible
  - How to tune such systems for the field of journalism?
- What is the state of the art for generative QA chatbots?
  - What are the components to make make generative QA chatbots?
  - Are generative chatbots only as good as the data they consume?
  - Could generative chatbots be a step toward AGI?

1.3 Objectives

1.3.1 Intrinsic

This subsection presents the general objectives related to the master's thesis.

Primaries

- Suggest project specification and planning.
- Analyze the state of the art of existing technologies and technics of QA systems and generative AI.
- Overview digital transformation in journalism and review the current status of the AI-News project.
- Document the study and write the thesis.

1.3.2 Fact-based QA Chatbot

The first objective is to make a state of the art software that takes a question as input and outputs a response.

[2] AINews.ch

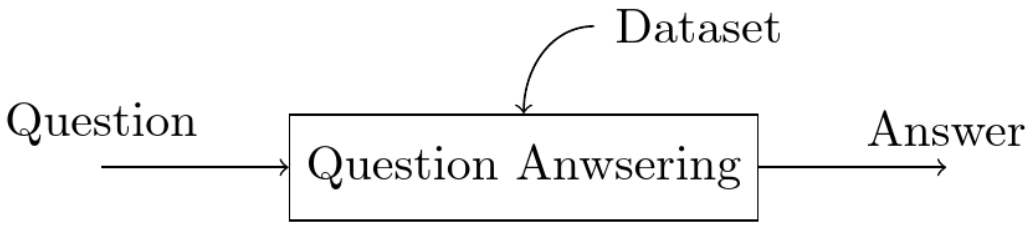

Figure 1.1: Suggested QA diagram

Primaries

- Select existing papers and projects treating the subject as a starting point.
- Identify relevant datasets.
- Develop one or more POC.
- Test and evaluate solutions.
- Suggest improvements, possible continuation, and future outcomes.

Secondaries

- Extended the QA chatbot using tailored knowledge.

1.3.3 Generative QA Chatbot

The second objective is to improve the output from the prior objective into enhanced answers.

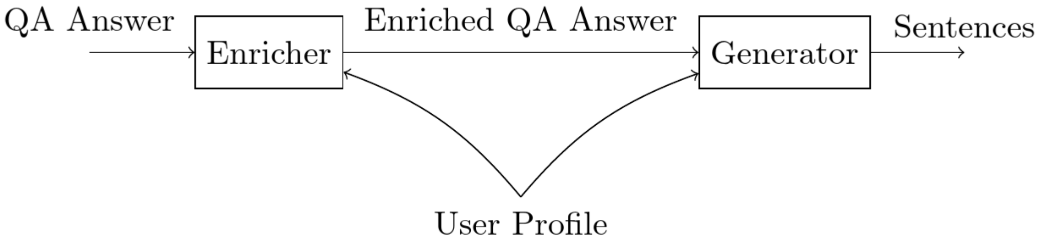

Figure 1.2: Suggested Generative QA diagram

Primaries

- Investigate a rule-based system for keyword enrichment.
- Generate sentences with keywords.
- Identify relevant datasets.
- Develop one or more POC.
- Test and evaluate solutions.
- Suggest improvements, possible continuation, and future outcomes.

Secondaries

- Use advanced strategies to enrich keywords.
- Use advanced text generation technics such as GTP-2[3].
- Use user profiles to customize the outputs.

[3] OpenAI's GTP-2 Algorithm[2]

1.4 Plan

1.4.1 Contraints

Timeframe: 19 weeks
Starting date: 16.09.2019
Ending date: 07.02.2020

1.4.2 Methodologies

For consistency, the project is split into two methodological parts. The first third, as the project's orientation is going toward information gathering and self-study, uses a standard sequential project management methodology. For the next two-thirds of the project, the author is using an agile methodology intending to reach incremental progress while exploring.

Back to level Milestones

(6 weeks) First third of the study, from **16.09.19 to 25.10.19**.

M1. Initial Master's Thesis (MT) plan and project specification

M2. Review the state of the art of the NLP and NLU technologies and refine the plan if needed.

Diving into the subject Milestones

(13 weeks) From 28.10.19 to 07.02.20, the following two-third of the thesis is composed 6 sprints of two weeks and one week to finalise the thesis.

M3. Basic QA Chatbot

M4. Evaluation of basic QA Chatbot

M5. Basic generative QA Chatbot

M6. Evaluation of basic generative QA Chatbot

1.4.3 Gantt

The Figure 1.3 represents the chart for the initial plan.

1.5 How to read this document

Describing the structure of the document with a redline and its reasoning.

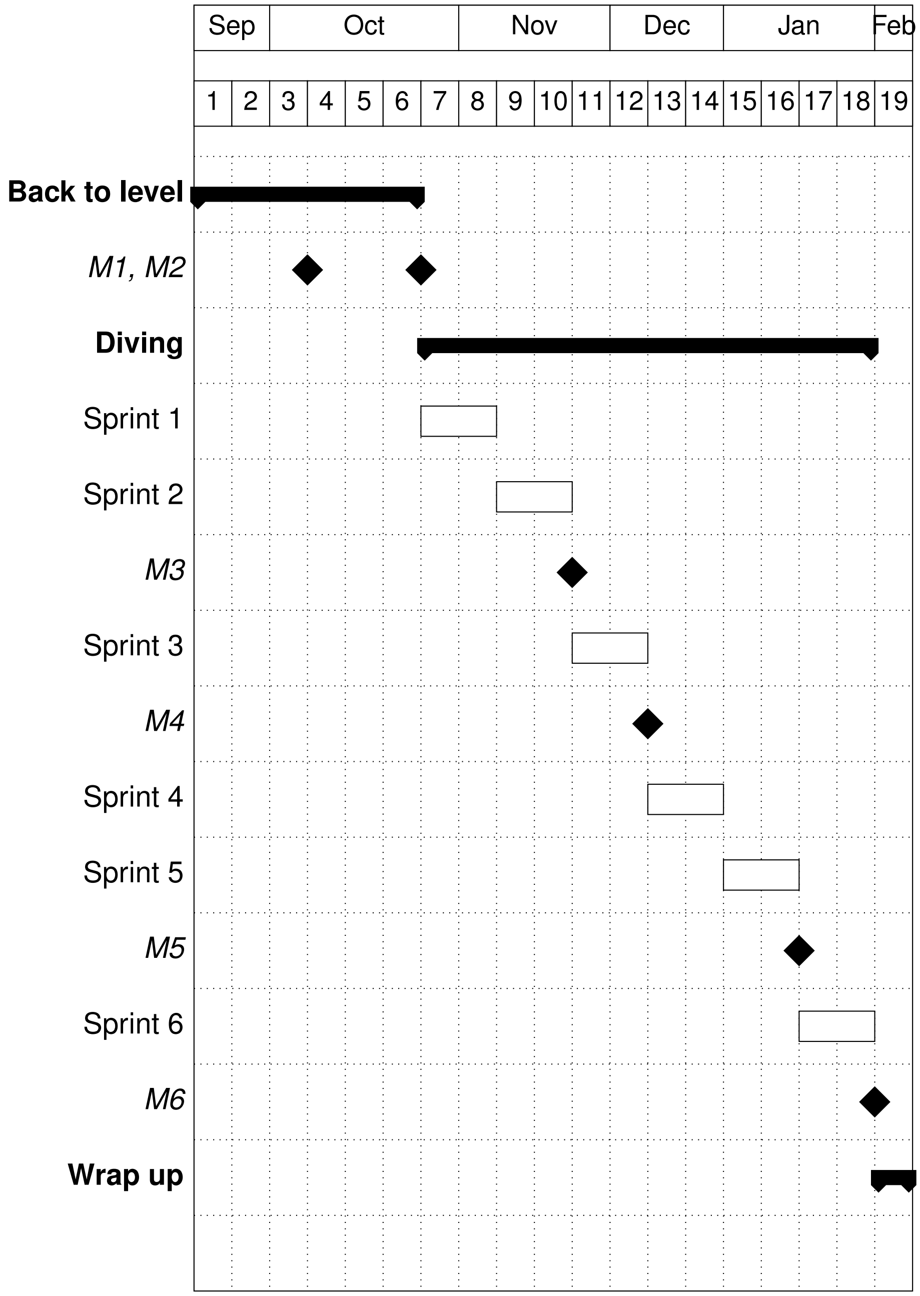

Figure 1.3: Initial Gantt Chart

Chapter 2

Tasks

2.1 Initial Tasks

2.1.1 Primaries

1. AI in journalism state of the art
2. NLP and NLU state of the art
3. Find relevant datasets
4. Find existing projects and papers responding to the questions
5. Explore documents' topics extraction
6. Explore the Wikidat and knowledge graphs
7. Explore question-answering technologies and technics
8. Evaluate by comparing to similar systems

Milestones

1. Initial MT plan and project specification
2. Overview topics extraction technics
3. Overview Wikidata and knowledge graphs technics
4. Overview text transformative and generative technics
5. Mindmap of the current NLP and NLU technologies
6. Pytorch hands-on

2.1.2 Secondaries

- Explore AI implications in journalism
- Explore AI personalization implications
- Explore text generative technologies
- Explore profile-based customization
- Explore text transformative technologies
- Explore the attention mechanism
- Explore text summarization

- Explore text flavoring to write as a specific author
- Explore news extraction from social media
- Explore news baseline extraction
- Explore news drafts and briefs generation
- Explore text adapted suggestions for journalists
- Explore knowledge graphs as content enrichment
- Explore multiple sources cross-checking to reduce fake news
- Explore tracker for the original source
- Explore autonomous knowledge gathering
- Explore machine-generated factual discussions
- Explore machine self-training
- Explore chain reasoning
- Explore artificial common sense
- Explore artificial intuition
- Explore on the fly translations
- Make overall improvements

Milestones

- Basic topic extraction from documents
- Basic conversational agent
- Basic journalistic agent

Chapter 3

State of the art

3.1 Chatbots

3.1.1 Retrieval Chatbots

3.1.2 QA Chatbots

3.1.3 Technologies

3.1.4 Evaluations

Bibliography

[1] Nathan Benaich and Ian Hogarth. “State of AI 2019”. In: (2019), p. 126. URL: https://www.stateof.ai/.

[2] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2018).