MSE | 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)

GraphQA, a Deep Retrieval Chatbot

A Multi-hop Conversational Question-Answering Chatbot using Sub-Knowledge Graphs

Author:

Romain Claret

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

External expert:
Prof. Dr. Michael Ignaz Schumacher
HES-SO//Valais
Institute of Business Information Systems

Fribourg, HES-SO//Master, February 2, 2020

Project Specification

This project specification for the Master's Thesis has been accepted by Romain Claret (the student) and Jean Hennebert (the supervisor) on the 4th of October 2019 at HES-SO//Fribourg.

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 Natural Language Processing (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.

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).

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 Question Answering (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 HogarBenaich et al., 2019

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.

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?

Objectives

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.

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`

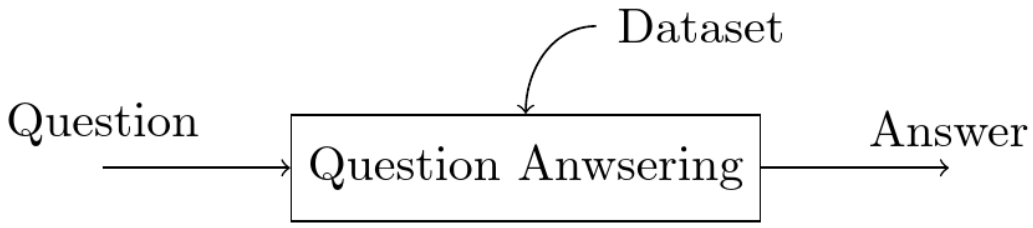

Figure 1: Suggested QA diagram

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.

Generative QA Chatbot

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

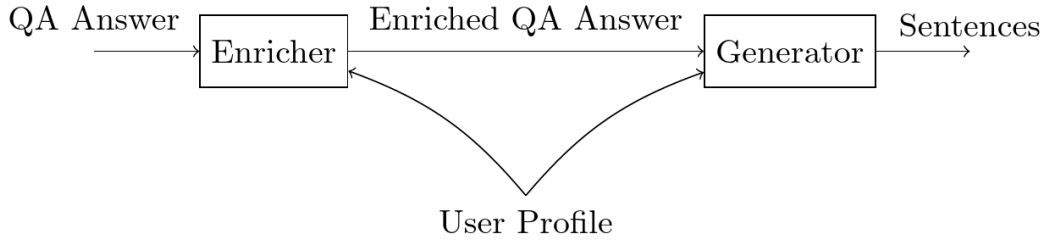

Figure 2: Suggested Generative QA diagram

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 AlgorithmRadford et al., 2018

Plan

Contraints

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

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 Natural Language Understanding (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

Gantt

The Figure 3 represents the chart for the initial plan.

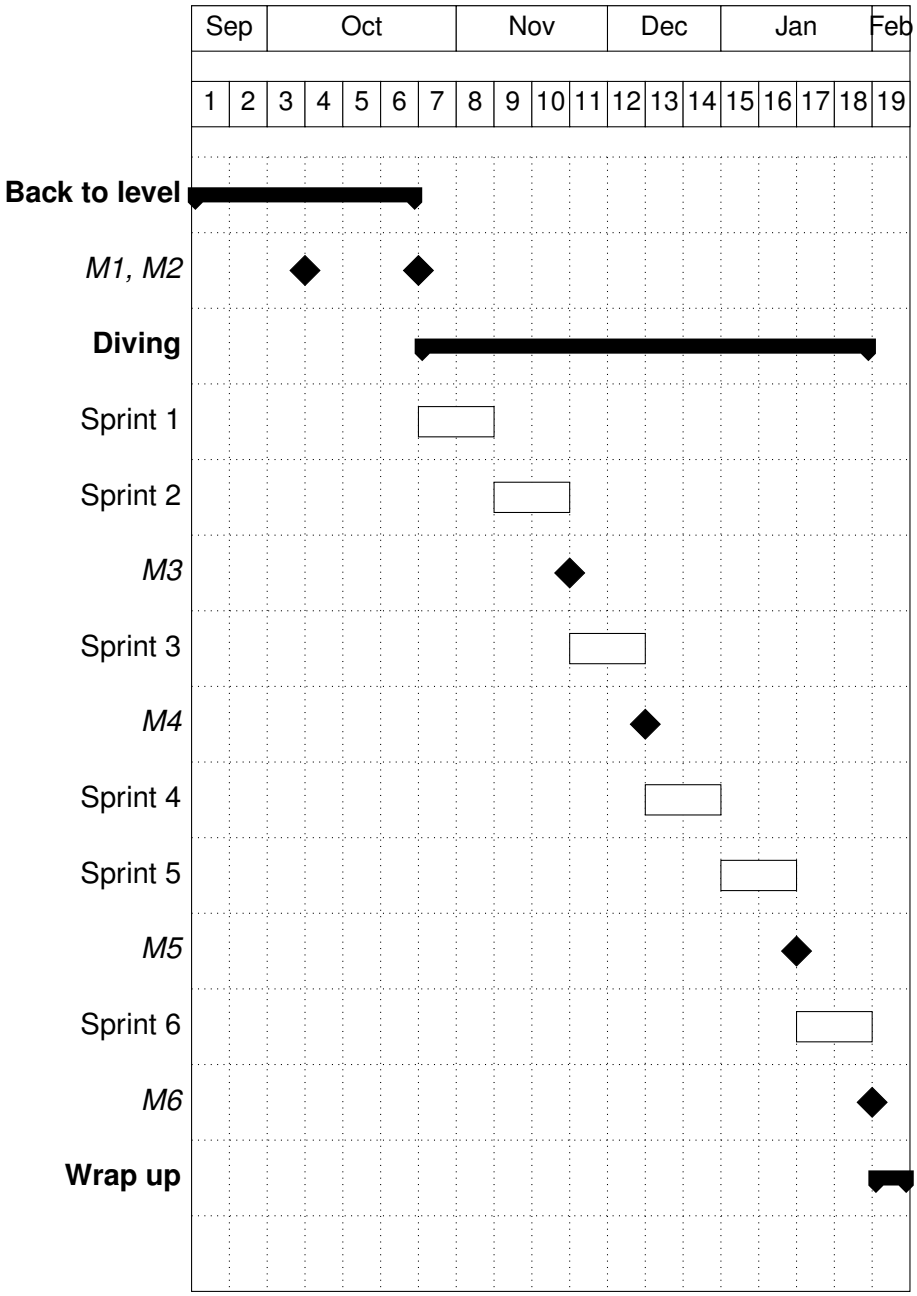

Figure 3: Project Specificiation Gantt Chart

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

Prof. Dr. Jean Hennebert, Master's Thesis Supervisor

Place, date: _____

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

Dedicate

To my family that believed in me, and still I don't know why.

Contents

Acknowledgments

I wish I could thank an AGI for doing my thesis.

Glossary

Adversarial Learning

In Machine Learning (ML), the concept of this technique relies on trying to fool models via malicious inputs. It can be interpreted as a game a model is playing with itself by modifying the input in such a way that the model will recognize it as another input then learn from its mistake.

Attention Mechanism

In Natural Language Processing (NLP), the Attention Mechanism is an algorithm used to calculate the relational weight between elements in a sequence of elements (most often words).

Bidirectional Language Model

In NLP, a Bidirectional Language Model represents a Language Model (LM) combining the forward pass and a backward pass of the same corpora.

Close-ended

A closed-ended question is designed to allow a limited amount of responses.

Encoder-Decoder

In Machine Learning (ML), Encoder-Decoder is two Neural Networks (NNs) that work in pair. The Encoder generates a fixed-size output vector from any sized vector input. And the Decoder generates from the Encoder output a vector that could be any size.

Few-Shot Learning

In Machine Learning (ML), Few-Shot Learning is technique used to solve tasks with a very small amount of training data.

Generative

In the context of the Thesis, we are using the generic word Generative as the ability concept of an algorithm able generating outputs in a meaningful but unpredictable manner from an input, which includes LMs and Generative Models.

Generative Model

In Machine Learning (ML), Generative Models are generating random outputs from a single input by using the probability of observing the output based on the input. In other words, it models the probability of observation for a given target.

Glossary

Ground Learning

In the context of the AI, Grounded Learning is based on the Grounded theory from the social sciences, which uses inductive reasoning. In the context of AI, it is the mechanism of combining structured and unstructured data as small conceptual parts to then apply machine reasoning.

Hop

In Question Answering (QA) Systems, a Hop is a quantitative measure of the number of combinations necessary between indirectly related pieces of information to provide an answer.

Knowledge Base

In Information Systems (IS), a Knowledge Base is a Knowledge Representation using a Linked Data database for storing and interlinking structured and unstructured data using a standard.

Knowledge Graph

In Information Systems (IS), a Knowledge Graph is a Knowledge Base (KB) organized as a graph using semantics.

Language Model

In Natural Language Processing (NLP), a Language Model is a Model trained to provide likelihood probabilities of the following sequence of words in addition at providing the probability for each sequences of words.

Linked Data

In Information Systems (IS), Linked Data is a structured interlinked database mainly used for semantic queries.

Machine Reasoning

In ML, Machine Reasoning represent the ability to apply reasoning for a given input by using knowledge representations and logic patterns such as inductions, analogies, or abductions.

Machine Understanding

In ML, Machine Understanding stays ambiguous in its definition. However, we use the term as the ability of representing knowledge at an atomic building blocks and fundamental relations.

Markov Decision Process

In the context of the Reinforcement Learning (RL), this process models the ability of a predicting the next state of a finite-state machine-like process, such as a game, with only the information contained in the present state.

Model

In Machine Learning (ML), a model is the representation of the assumptions made by the algorithm during the training phase. Models are used to output a result based on a provided input and the learned patterns.

Model Fine-Tuning

In Machine Learning (ML), Fine Tuning a Model is the technique of using a trained Neural Network (NN) model as a base and tune it for a specific task.

Multi-Hop

In Question Answering (QA) Systems, a Multi-Hop implies that the answer is within multiple Hop of the question. In other words, the answer requires a combination of different information to be answerable. Generally, extra qualifying Subject-Predicate-Object Tuple (SPO) are separating the question Subject and the answer Object.

Named-Entity Linking

In Natural Language Processing (NLP), Name-Entity Linking extends the Named-Entity Recognition by providing an unique identifier to each word allowing a mapping in various databases (useful in translations).

Named-Entity Recognition

In Information Extraction (IE), Named-Entity Recognition is a technique used to extract from unstructured text words predefined in a vocabulary.

Open Domain

In Information Retrieval (IR), the support of Open Domain questions provides a no restrictions for the theme of the question asked.

Part of Speech

In Natural Language Processing (NLP), Part of Speech is a technique used to categorize words that behave syntactically similarly.

Part of Speech Tagging

In Natural Language Processing (NLP), The Part of Speech Tagging is extending the Part of Speech by adding a label to the word depending on its context (the neighboring words).

Reinforcement Learning

In ML, this type of learning combines generally a Markov Decision Process (MDP) environment with an approach similar to Unsupervised Learning (UL) as it does not require labelled data. The particularity of this technique is that it uses a notion of rewards to predict the best next-step by running a large amount of simulation as training.

Sequence-to-Sequence

In Machine Learning (ML), a Sequence-to-Sequence or Seq2Seq is an Encoder-Decoder Neural Network (NN) that for a given sequence of elements as input, outputs another sequence of elements.

Shallow Neural Network

In ML, similar to Deep Learning (DL), Shallow Neural Networks have a Encoder-Decoder approach by having a single hidden layer, which often has a high amount of parameters.

Glossary

Single-Hop

In Question Answering (QA) Systems, a Single-Hop implies that the answer is within a single Hop of the question. Generally, a unique Predicate separates the question Subject and the answer Object.

Supervised Learning

In ML, this type of learning implies the uses of labelled datasets to perform the training.

Transformer

In Natural Language Processing (NLP), a Transformer is a Sequence-to-Sequence (Seq2Seq) architecture using the Attention Mechanism..

Unsupervised Learning

In ML, this type of learning implies the uses of unlabelled datasets to perform the training.

Word Embedding

In Natural Language Processing (NLP), the Word Embedding is a technique for word representation as vectors in an embedding matrix. Additionally, it has often the particularity of preserving the semantical analogies of word-vectors.

Zero-Shot Learning

In Machine Learning (ML), Zero-Shot Learning is technique used to solve tasks without training on examples.

Acronyms

AGI
Artificial General Intelligence.

AI
Artificial Intelligence.

AIML
Artificial Intelligence Markup Language.

AL
Adversarial Learning.

ANI
Artificial Narrow Intelligence.

ANN
Artificial Neural Networks.

BERT
Bidirectional Encoder Representations from Transformers.

biLM
Bidirectional Language Model.

CBOW
Continuous Bag of words.

CE
Character Embedding.

CWE
Context-based Word Embedding.

DL
Deep Learning.

DNN
Deep Neural Networks.

Acronyms

FAQ
Frequently Asked Questions.

GAN
Generative Adversarial Networks.

GPT-2
Generative Pre-Training 2.

ICT
Information and Communications Technologies.

IE
Information Extraction.

IR
Information Retrieval.

IS
Information Systems.

KB
Knowledge Base.

KG
Knowledge Graph.

LM
Language Model.

MDP
Markov Decision Process.

ML
Machine Learning.

MR
Machine Reasonsing.

MRR
Mean Reciprocal Rank.

MRU
Master Research Units.

MT
Master's Thesis.

MU
Machine Understanding.

NL
Natural Language.

NLG
Natural Language Generation.

NLP
Natural Language Processing.

NLU
Natural Language Understanding.

NN
Neural Network.

OOV
Out-of-Vocabulary.

POC
Proof of Concept.

QA
Question Answering.

RL
Reinforcement Learning.

Seq2Seq
Sequence-to-Sequence.

SL
Supervised Learning.

SNN
Shallow Neural Network.

SOTA
State of the Art.

SPO
Subject-Predicate-Object Tuple.

TF-IDF
Term Frequency-Inverse Document Frequency.

Acronyms

UL

Unsupervised Learning.

Weak AI

Weak Artificial Intelligence.

Abstract

We propose an innovative approach for question-answering chatbots to handle conversational contexts and generate natural language sentences as answers. In addition to the ability to answer open-domain questions, our zero-shot learning approach, which uses a pure algorithmic orchestration, provides a modular architecture to swap statically or dynamically task-oriented models while preserving its independence to training.

In the scope of this research, we realize the Proof-of-Concept of an Open-domain and Closed-ended Question-Answering chatbot able to output comprehensive Natural Language generated sentences using the Wikidata Knowledge Base.

To achieve the concept, we explore the extraction, and the use of sub-knowledge graphs from the Wikidata knowledge base to answer questions conversationally and to use the sub-graphs as context holder. Additionally, we are extracting Subject-Predicate-Object tuples from the graph and using Language Models to join the SPOs and extend the answers as natural language sentences.

The proof-of-concept architecture uses a combination of state-of-the-art and industry-used models with a fine-tuning strategy. As a motivational target, we use a Zero-Shot Learning approach, by combining various models with an algorithmic orchestrator and using pure algorithmic for the graph manipulation and answer extraction.

Finally, we evaluate the answers and compare the results with state-of-the-art Single-Hop and Multi-Hop question-answering systems on question-answering datasets. We find out that, aside from the computation time and the computational resources needed, our proof-of-concept performs similarly at question-answering compared to its competitors.

Keywords: Machine Learning (ML), Natural Language Processing (NLP), Single-Hop, Multi-Hop, Question Answering (QA), Wikidata, Wikipedia, Knowledge Graph (KG), Knowledge Base (KB), Word Embedding, Part of Speech Tagging, Named-Entity Recognition, Named-Entity Linking, Language Model (LM), Model Fine-Tuning, Graphs, Sub-Knowledge Graphs, Transformer, Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-Training 2 (GPT-2), Information Extraction (IE), Spacy, GloVe, DeepCorrect, Chatbot, Conversational, Information Retrieval (IR), Queries, Python

How to read this document

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

To be completed at the end of the work

Project preface

Introducing the project

State-of-the-art

In this part, we will be exploring the state of the art of NLP technologies as it is at the beginning of 2020.

Design and realization

Explaining how we got to build a proof of concept, what happened during the process of the initial plan, and how when came up with an innovative solution while solving and starting the design project from scratch.

Retrospective

The results are here; it's awesome what we accomplished!

Part I

Project preface

Chapter 1

Introduction

New technologies are revolutionizing the way humans access knowledge as a service from multiple platforms and providers. Thanks to the emergence of increasingly powerful AI algorithms, particularly in the field of NLP, conversational agents, commonly named 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 the chatbots are continually improving at providing the best outputs for specific tasks and also improving at providing meaningful human-like sentences, they still cannot generalize the tasks toward human-like conversations. The task of conversation, as humans are applying it, a complex integration of tasks including understanding, reasoning, context linking, context tracking, curiosity, initiatives, Few-Shot Learning or Zero-Shot Learning and learning on the fly, have yet to be accomplished. Nonetheless, as research progress, chatbots are improving with new technics and tools that are making them step by step closer to complete human-like discussions, slowly progressing towards AGI chatbots. As for the scope of the thesis, we are humbly focusing on the combination of few NLP tasks with a Zero-Shot Learning approach to help ML and NLP research getting closer to General QA Conversational Chatbots.

1.1 Aim of the Research

The initial goal of the thesis was to explore and combine State of the Art (SOTA) QA Systems and LMs to into an experimental POC of a Conversational QA Chatbots.

During our research journey, we discovered a new purpose to the project, and took a step into the unknown with a Zero-Shot Learning approach with sub-knowledge graphs.

1.1.1 Project's Overall Scope

We are focusing on the English language as an attempt to increase the number of compatible datasets and make community accessible solutions. We are exploring and combining two types of systems as an attempt to build QA chatbots. The first system will produce factual answers, and the second system will generate human-like sentences from the answers found by the primary system. For the factual answers, we will be evaluating the results of our combined system against SOTA QA

[1] The State of AI Report 2019 (Benaich et al., 2019)

systems on QA testing datasets. Humans will manually evaluate the answered sentences from our combined system. Finally, as the time allocated for the thesis is 19 weeks, the outcomes are narrowed at providing non-exhaustive research and a POC solution. On a side note, the review of the risks and ethical problems that could be raised by the development of such solutions are not part of this work.

1.1.2 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 scenario-based narrow chatbots.

1.1.3 Personal Interest

In harmony with the thesis subject, as the author is particularly interested in exploring the premises to AGI related technologies such as Zero-Shot Learning, Ground Learning, Machine Understanding, and Machine Reasoning for a Multi-Domain Task Generalization. The human-like QA frame of this project is particularly motivational.

1.2 Research Questions

We articulate here the initial 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.

- What are the components to make QA chatbots?
  - What is the SOTA of chatbots and QA systems?
  - 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 Generative QA chatbots?
  - Are Generative chatbots only as good as the data they consume?
  - Could Generative chatbots be a step toward AGI?

[2] `AINews.ch`

Part II

State-of-the-art

Chapter 2

Chatbots

Based on latest MMC's state of AI report[1], it appears that 26% of the AI-Startups studied by Gartner[2] are using or making chatbots, see Figure 2.1. The same study, made a year earlier, in 2018, shows that chatbots are not present as an application, which implies that either chatbots were not referenced as AI or that their popularity exploded within a year.

As it is at the beginning of 2020, based on The State of AI Report 2019 (Benaich et al., 2019) and the two previously mentioned studies, chatbots are commonly present but limited to narrow tasks. In most cases, they a scenario-based with sequences of if-else conditions that we classify as non-learning AI. Moreover, hard-coded scenarios are requiring an infinite amount of human power to create generic Chatbots able to maintain a conversation at a human level. However, progress in the field of ML and NLP is demonstrating that providing large corpora to an unsupervised algorithm is enough to maintain a passive conversation with users, which results into a shifting of the human power into data engineering. Increasingly complex algorithms and techniques are emerging at a monthly in the field, demonstrating a trend towards conversational performance improvements. Note that even if they are getting better at providing meaningful sentences, current Chatbots are still not able to orchestrate the generalization of all the tasks required to a human-like conversation. E.g., such as understanding and reasoning based on the context, initiatives to search and learn for missing information, initiate dialogue in a meaningful manner, intuition, and much more. As a side note, the generalization of those tasks would reduce the steps significantly towards general Chatbots.

From a user-centric point of view, chatbots are currently trending and rising global interest for various reasons. Big companies such as *Google* or *Apple* are believing in the technology and are making a lot of effort at pushing the chatbots into the mainstream. Even if the word "chatbot" is commonly used as a buzzword without a proper definition, people have at least a mental representation of its concept. Indeed, whether they call it "Digital Assistant", "Siri", "ok Google" or "Alexa", they all expect to a more or less human-like conversations after using those triggering keywords.

[1] The State of AI 2019: Divergence (Kelnar, 2019)

[2] 2'791 European AI Startups from the 2019 CIO Survey: CIOs Have Awoken to the Importance of AI (Rowsell-Jones et al., 2019)

It is interesting to note that the majority of the following sections could be included in the field of AI in general. The extrapolation of the chatbot subject to AI as a whole is worth further studying, but it not part of this work. Instead, the focus of this chapter is Chatbots; we provide a synthesis and classification of the different methods used to build chatbots. We will define the main categories identified and continue on the main sub-categories and conclude with a cartographical chart of our chatbot vision.

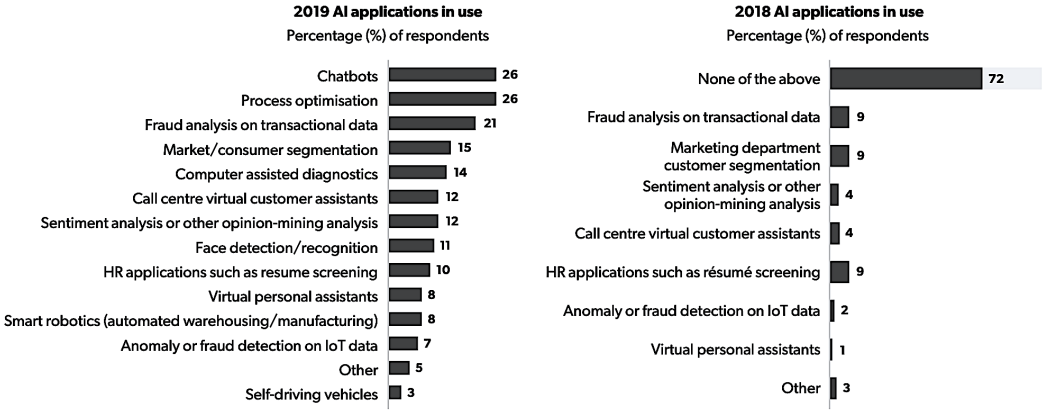

Figure 2.1: Figure 31 from *The State of AI 2019: Divergence (Kelnar, 2019)*. The top AI applications used in European AI Startup in 2019 are Chatbots and Process optimization.

2.1 Chatbot History

Not mentioning *Alan Turing* or *Joseph Weizenbaum*, both considered as the fathers of AI and chatbots, would not be fair to this research. Indeed, in 1950 they forecasted human-like communication with computers and proposed a test to differentiate humans from machines, the Turing Test (Turing, 1950). The test performs as follows: a supervisor asks a human to talk to a masked entity and determine rather he is talking to a human or a computer. If the human cannot recognize speaking to a computer, then the machine passes the Turing test.

In 1966,*Joseph Weizenbaum* wrote Eliza(Dunlop, 1999), a computer program simulating a psychotherapist, it is seen today as one of the first well-documented attempts to make a Chatbot designed at passing the Turing test. However, due to techniqueal restrictions, Eliza was not performing particularly well in all contexts. As for today, it is still possible to play with the chatbot on a dedicated website.

Since Eliza, a lot of progress has been made until 2020, From conditional IF-ELSE, Artificial Intelligence Markup Language (AIML), up to ML with Artificial Neural Networks (ANN) and Deep Neural Networks (DNN), the improvements in the field of chatbots increased drastically over the years. Each iterations delivering algorithms being continuously more sophisticated and better at using the Natural Language (NL), resulting in a new field of ML called NLP. As a reminder of the

chatbots history and progress from 1966 to 2016, the infographic(Futurism, 2016) from Futurism is particularly speaking.

2.2 Main Categories in the Chatbot Realm

While performing the state-of-the-art, we identified three main chatbots categories.

2.2.1 Conversational

We like to call them the Chatty bots, and they are great for interaction and structured replies, well designed for their ability to talk. E.g., *User*: “Hello, how are you?”, *Bot*: “Good, what about you?”.

2.2.2 Task-Oriented

The Task-Oriented bots are performing particularly well at specific tasks as smart-assistants. As their design is not toward generalization, their abilities are limited and will fail at off-tasks. A common workflow used by those bots is to detect the Intent and the Entities of the user request, often in NL, then apply a rule-based matching to perform the command intended by the user. E.g., *User*: “Book the next flight to Geneva from Zürich.”, *Bot*: “Alright! Your ticket number is 00XXYYZZ. Have a great flight!”

2.2.3 Dispatcher

The dispatcher acts as a middleware, who’s unique job is to categories the user input and forward the input to the task executor from any of the previous two categories that the user requested. E.g., If the user request the following "What is the weather in Geneva?", the dispatcher will categories the question as the task of providing the weather and sent it to the weather module. As a second example, if the user provides the following input "Hey! Let’s talk about random stuff!", the dispatcher will forward the request to the chatty module.

2.3 Retrieval Chatbots

As it is today, Retrieval-based Chatbots are popular in the industry. Indeed, a lot of tools are available, and they perform well for specific tasks. However, the response capabilities are limited to their databases and the retrieval algorithm used. Indeed, for a given input, the system is using heuristics to find the best output from the pre-defined responses. The choice of the algorithms is wide and depends on the task the chatbot is required to perform. Regardless of the heuristic used, from keywords matching up to DL, the output will always be retrieved from the database. Concerning the database itself, the data needs a pre-processing step to generate indexes linking the questions, answers, and apply pre-calculated scores. Pre-processing also implies that if the database is updated, a new pre-processing batch is required, which implies that the scalability or fine-tuning if compromised in the long run. We like to call this type of chatbots “Keywords-based”. See Figure 2.2.

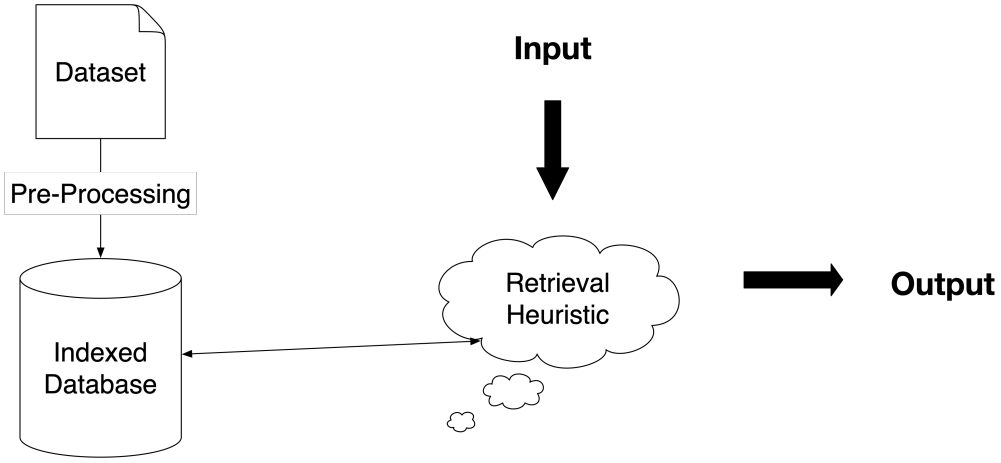

Figure 2.2: Illustrative representation of frequent retrieval chatbots architecture.

2.4 Rule-Based Chatbots

"Scenario-based", as we name it, is the oldest and relatively straightforward system for chatbots. The ElizaDunlop, 1999 Chatbot, as mentioned in the Chatbot History 2.1, is scanning the input text for keywords, calculates a ranking for each keyword, and finally goes through a series of conditions called rules, and some randomness to reach the best ending leaf. Usually, the bot also includes a default output if the matching process fails, which we can still nowadays see in chatbots: "Hmm, this is interesting, tell me more.". Such bots are often used for interactive chatbots, as it can, in a controlled environment, give a sense of deep meaning in the context of the conversation. Note that such systems require a lot of human power to build a frame for the bot to play in, and by this mean makes rule-based chatbots great for the specific scenario but is hard particularly hard to generalize. See Figure 2.3.

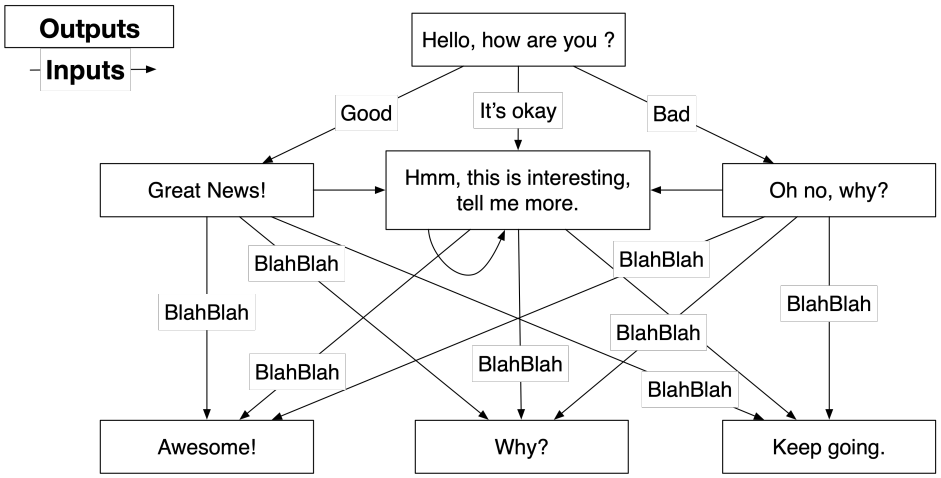

Figure 2.3: Illustrative representation of frequent rule-based chatbots process.

2.5 Generative Chatbots

As the current result of all the incredible innovations made in the past years in NLP, and is a premise to true conversational chatbots, generative methods are overcoming the limitations of the Retrieval 2.3 and Rule-Based 2.4 Chatbots, by its ability to generate new content. Either Supervised 2.5.1, Unsupervised UL or Adversarial 2.5.2, no pre-defined outputs are used, the models are trained on large corpora to learn the language patterns and outputs relatively meaningful responses to give inputs. Another particularity of generative chatbots, is that building a domain-oriented chatbot does not require the engineers to have the domain expertise, as the expertise is embedded into the data, which allows a relative scalability to new domains. However, even if the trained models can output responses at nearly no timespan, the data-engineering of the datasets and the training phase is most often long and complicated. As a final note, the responses generated by such chatbots are only as good as the data it was fed during the training.

2.5.1 Supervised Learning

Supervised Learning (SL) is probably the most common method used by Generative Chatbots, as it provides relative control over training. Sequence-to-Sequence (Seq2Seq) is commonly used as architecture for those chatbots, a NLP version of the Encoder-Decoder, which encodes the input words sequence and decode it into a words sequence as an answer into a framed conversation fashion. The training only requires a dataset containing a sentence and its desired response, the model will then map similar inputs with similar outputs. However, a clear limitation for this learning is that the model will for any input always have an answer, regardless of the overall meaning. Additionally, Seq2Seq will prioritize the highest word apparition probabilities, meaning that data duplicates and requiring sentences will create a trend during decoding. E.g., “I don’t know the answer.”. See Figure 2.4

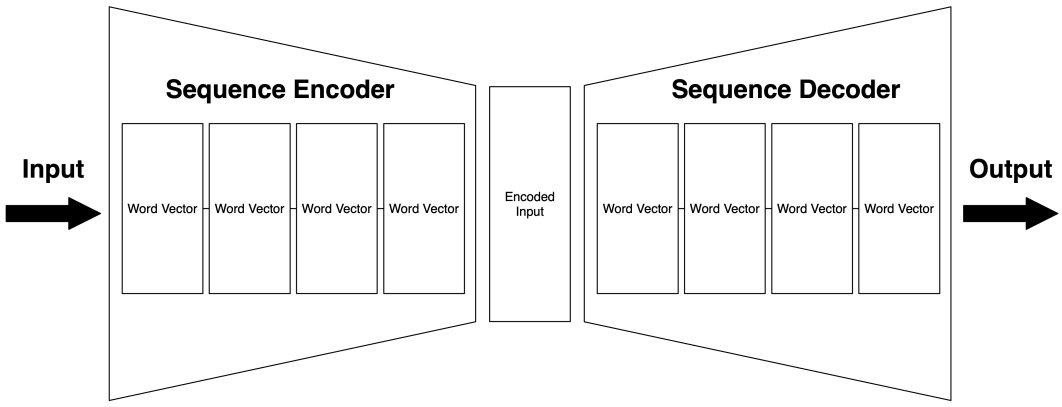

Figure 2.4: Illustrative representation of a Sequence to Sequence architecture.

2.5.2 Adversarial Learning

Adversarial Learning (AL) has driven attention thanks to Computer Vision Generative Adversarial Networks (GAN) (Karras et al., 2019) by proving that it is possible to generate realistic human face (Wang, 2019). In the chatbots context, it can be extrapolated into a futuristic version of the Turing Test 2.1, where machines are confronting themselves instead of humans. The concept implies the use of a training dataset containing human conversations, and compare them against the generated answer; the discriminator will then judge which is from a human and which is from an algorithm. Note that adversarial methods such as GAN are working well because of the nature of the data it plays with; indeed, pixels can be deeply noised, but words cannot be due to their discrete nature. See Figure 2.5

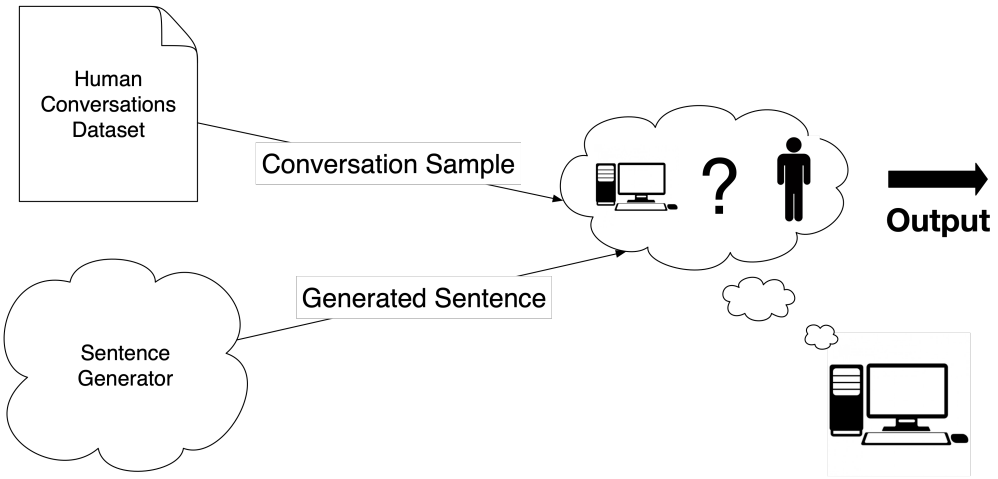

Figure 2.5: Illustrative representation of an adversarial architecture in a chatbot context.

2.5.3 Pre-trained Language Models

Language Models are currently the most recent and the most promising models due to their ability to model language itself instead of conversations and then tune the outputs as a chatbot would. It can been seen as semi-supervised learning, as it uses UL for training and supervised learning 2.5.1 for fine-tuning 2.5.4. We will dive into LM in the NLP chapter 3.3.

2.5.4 Model Fine-Tuning

With the specificity of Model Fine-Tuning, lm as it provides the tools to build chatbots based on the ground of the language itself and then customize the model into a specific manner by fine-tuning it on a dataset fitting the domain required by the chatbots. Indeed, it is relatively easy to fine-tune a QA dataset to a LM, making the model able to answer questions instead of descriptively filling sentences. The main downside to those models is the large memory size required to run them. However, due to their nature, they are trained once and then fine-tuned. Note that training requires an enormous amount of computational power. E.g, The largest form of BERT (Devlin et al., 2019) was trained on 16 TPUs for 4 days. Fine-tuning, on the

other hand, scales down to few hours on a single TPU, which makes it relatively scalable to new domains. See Figure 2.6

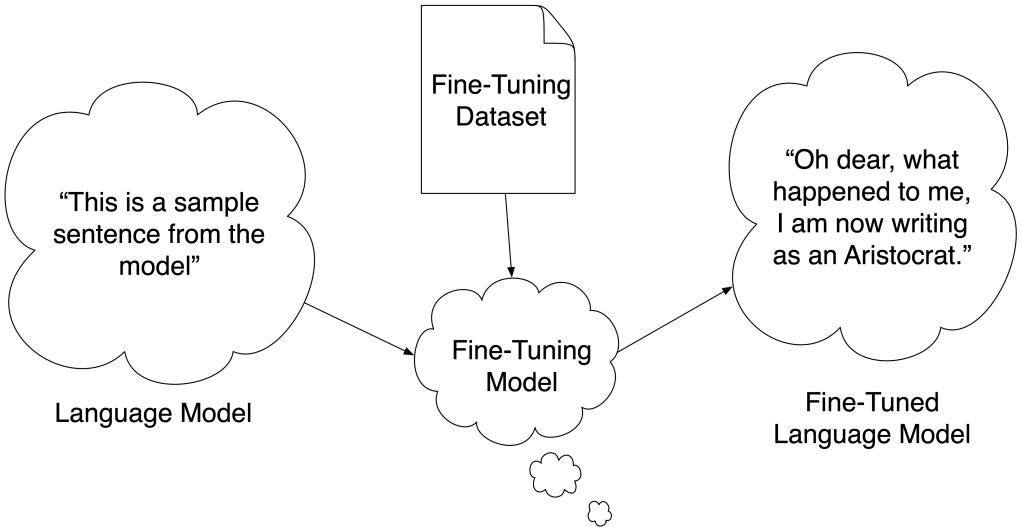

Figure 2.6: Illustrative representation of fine-tuning in a chatbot context.

2.5.5 Reinforcement Learning

RL is proven to be very powerful by the latest research made by *Open-AI* with its DOTA2 bot or *Google's Deepmind* with AlphaZero, so we believe that it is worth mentioning it. However, this type of learning requires a finite state similar to a MDP, which matches game cases but not conversations, and impacting by this means the motivation to export the technique to NLP. Indeed, this methodology requires that all information required for the next step are wrapped into a single state to predict it, which makes it hard to use the dialogue case. For now, NLP research does not provide a conclusion as if, with billions of simulations, RL Chatbots could reach comparative results to Generative Chatbots 2.5.

2.6 Grounded Chatbots

Falling in a particularly rare research field of ML and NLP, Ground Learning can be seen as the future of Machine Understanding (MU) and Machine Reasoning (MR). In a chatbot context, the goal is to simulate, based on the Grounded Theory from the social sciences, how humans are using inductive reasoning to create conversations with unstructured knowledge. The idea is to give the ability to the bot, for any given input, to gather information from any data sources and provide an inductive output. E.g., Combining Knowledge Bases with weather forecaster. Second e.g., For example, for the given input: "What is the color in autumn of a leaf in Switzerland?", 1) the bot would have first to identify the context keywords (color, leaf, autumn, switzerland), 2) the bot would select where to gather the information, 3) the would investigate the Wikidata Knowledge Base, Wikipedia, and The Weather Channel API, 4) the bot would formulate an answer based on the information it gathered. 2.7

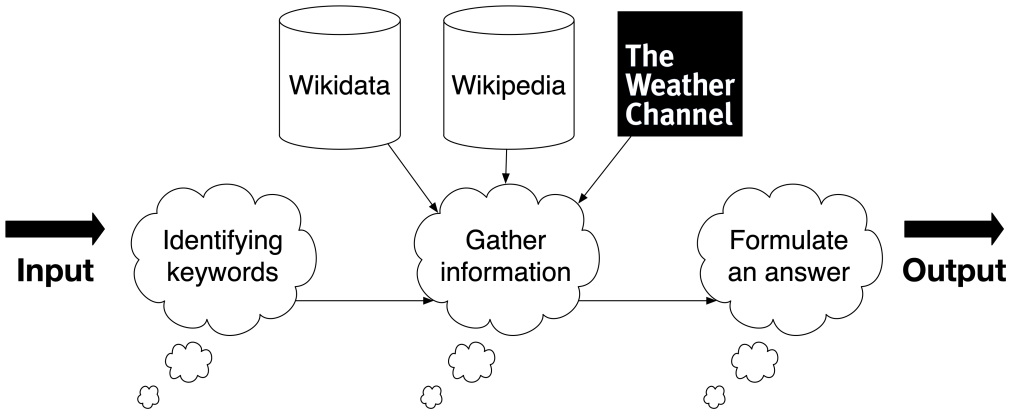

Figure 2.7: Illustrative representation of a grounded chatbot.

2.7 Question-Answering Chatbots

QA is a prevalent task for chatbots; indeed, they are widely used for questioning tasks in either Single or Open Domain, Open or Close-ended, Single or Multi-Hop with applications such as FAQs, Supports, help to find the meaning of life, and so on. Due to the broadness of the field, no defined methodology has been generalized; instead, it uses either one or multiple techniques described in the previous sections. It is interesting to note that the field of QA is raising a lot of interest in NLP research lately, and the benchmarking game of creating the new baselines, with increasingly complex datasets, is still in progress. In this section, we will overview some recent baselines.

Fine-Tunning Language Models Large LM such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2018) are often fine-tuned on QA datasets similar to SQUAD 2.0 (Rajpurkar et al., 2018) which are particularly tricky, even for humans.

Querying Models Based on QA datasets, a model is trained to fill structured templates. The generated output is a structured query for a particular querying language such as SPARQL for Wikidata.

Retrieval A popular approach in the industry is to use tools such as Elasticsearch for indexing and additional tools using ML heuristics to perform the queries.

2.8 Common Chatbot Features Overview

In this section, we are non-exhaustively naming a few recurring features appearing during our targeted research.

2.8.1 Context

Humans are intuitively and extensively relying on the context for conversational purposes, chatbots relying on dialogue as part of their task, requires the capacity to hold context. On a side note, one-way style dialogues such as commands or none-nested questions do not need to hold context to perform well.

Short term context Implying the ability for the bot to hold context for at least the current conversation, e.g., few keywords or on-the-fly Model Fine-Tuning.

Long term context Often, chatbots would use user-profiles as part of their architecture to remember information such as the favorite pizza flavor of a client.

2.8.2 Proactivity

Simulating personalized interest as a human would do is not new to chatbots, as it has been proven by becoming a standard in marketing and customer support chatbots. Messages such as “Hey, you are on our web store for a while, can I help you?”, are carrying a sense of proactivity; however, beyond asking general pre-made questions, limitations are clear, and not much progress has been made yet in the field. Indeed, human-like proactive chatbots imply algorithms capable of initiating conversations by initiating a dialogue or asking information in a meaningful manner based on the long and short term context.

2.8.3 Narrow vs General Chatbots Scope

Beyond the three main categories 2.2 identified during the study, in general, chatbots can additionally be classified within a scope starting at Narrow Chatbots up to General Chatbots. To position them, we defined a two axes classification using Tasks and Knowledge as represented on Table 2.1.

Tasks Axis To name a few examples of task-oriented Chatbots: Talk, Frequently Asked Questions (FAQ), Customer Support, or Ordering.

Knowledge Axis Non-exhaustively, as follows, a few knowledge-centric examples for chatbots: Health, Weather, or Customer Service.

Narrow Chatbots Narrow chatbots are limited by the range of tasks they can accomplish and the knowledge they can use. By design, they are very good at a particular task for a particular knowledge requirement.

General Chatbots They are neither limited by the range of tasks they can accomplish or the knowledge they can use. However, they often have an average performance for any task or knowledge. We go in more details at section 2.8.4.

Tasks \ Knowledge	
Expert in a specific Field Expert at all Tasks	**General Chatbots** Expert in all Fields Expert at all Tasks
Narrow Chatbots Expert in a specific Field Expert at specific Task	Expert in all Fields Expert at specific Task

Table 2.1: This table represents categories in Narrow and General Chatbots in a Tasks versus Knowledge format.

2.8.4 General Chatbots

As research progress in the NLP field, chatbots are improving as an effort to perform simultaneously well in various tasks and multi-knowledge bases. As a contemporary goal, in addition to any chatbot related tasks and broad knowledge expertise, General Chatbots must not be limited to their current capabilities, but on the contrary, be able to learn new tasks and subjects continuously. As far as we as this study went, we could not find SOTA general chatbots as defined. However, companies like *Amazon* are selling to a large public a feel to general chatbots with Alexa. Indeed, apart from ordering goodies from *Amazon* and roughly conversing with Alexa, users can command their smart homes, use it as a personal assistant, or even program *skills* to perform custom actions.

2.9 Chatbots Cartography

As a result of this chapter, we created a chart on Figure 2.8 representing the current state of chatbots from our point of view. Note that a particular use-case could be in multiple leafs.

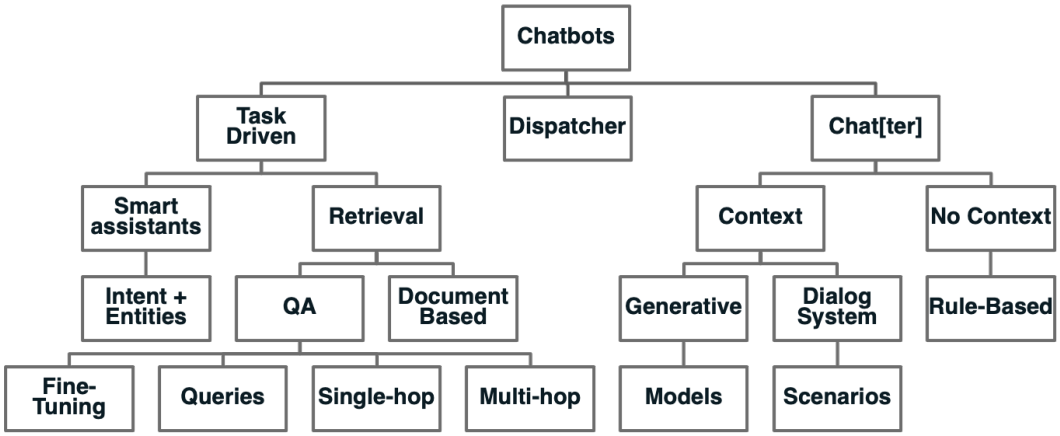

Figure 2.8: Represents the chatbots cartography as conclusion to the chatbot state-of-the-art chapter.

Chapter 3

Natural Language Processing

It is often challenging to realize the complexity behind Natural Language (NL), even to experts. First of all, Language is an academic field of study, implying multi-disciplinary skills. And secondly, staying up to date with evergrowing tools and new SOTA algorithms proves to be challenging. NL is the fundamental communication element for humans, NLP is the field of ML studying NL with the goal of providing the ability to machines to handle and mimic NL to create human-like verbal interactions. Beyond words and grammar rules, NL is a complex orchestration of subtleties, intuitively handled by humans, but not for easily handled by machines. Nonetheless, NLPs technologies are massively used in our daily lives, including information extraction, summarization, and conversation simulation. However, even if machines are given the same language rules as humans, they do not yet understand the manipulation they are processing, as humans would do. Indeed, NLP algorithms are applying pre-defined or multiple examples-based learned rules, which may result in ambiguities while applying NL. Using a rule-based approach 2.4 to build a NL model would result into near to infinite amount of conditions, this is the main reason for NLP to be particularly present ML, particularly in DL.

3.1 Word Embeddings

Commonly used as the first data pre-processing in DL NLP. Those Unsupervised Learning (UL) algorithms capture syntactical and semantical words representation from large unlabelled corpora datasets as vectors by building a multi-dimensional matrix. On average, dimensions are held in a scope of 100 to 400, and thanks to its the vectorized nature captured words, geometrical operations can be applied, such as the cosine functions to calculate word similarities. Another feature related to word embeddings, is the ability to apply analogical operations such as *'king' - 'man' + 'woman' = 'queen'*, which popularize Word2Vec 3.1.1 and gave credits to the method, even if the justification to this effect has been theorietized 4 years later [1] by stating that the compositionality is only seen when assumptions are held, in particular when words are uniformly distributed in the embedding space.

3.1.1 Word2Vec and GloVe

Published by *Google* in 2013, Word2Vec (Mikolov et al., 2013), and its competitor GloVe (Pennington et al., 2014) published by the *University of Standford* in 2014,

[1] Skip-Gram - Zipf + Uniform = Vector Additivity (Gittens et al., 2017)

both use a Shallow Neural Network (SNN), as illustrated on Figure 3.1, similarly to SL by feeding as input a text corpora, and outputting word vectors with a given vocabulary. Training and testing is straightforward but painful tweaking make it hard to build good generalized word embedding representations. Even if the SNN could remind a DL approach, it is only has one hidden layer; however, the output word vectors are particularly useful for DNN as input.

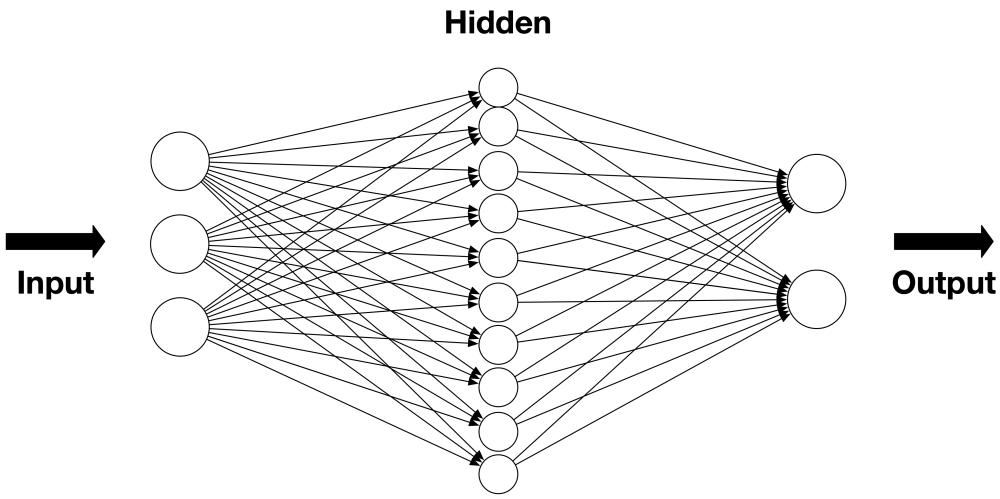

Figure 3.1: Illustrative representation of a Shallow Neural Network

3.1.2 Out of Vocabulary Problem

A common issue in Word Embedding is related to the vocabulary itself when words are unknown, called the Out-of-Vocabulary (OOV) issue. The issue occurs when post-training the model is requested to provide a vector representation that it never seen before. A solution could be to handle the exception by forwarding it to a default or pre-defined error vectors such as a series of zeros. We could approach the problem sophisticatedly, by defining on-the-fly OOV words with at a high learning rate as the sum of word-vectors contextualizing the OOV (Herbelot et al., 2017). Another solution would be to fallback to 3.2 by either training a model to compositional map characters to words (Pinter et al., 2017), or using Character Embedding (CE) as a whole instead of Word Embedding 3.2.

3.2 Character Embeddings

Additionally to Word Embedding similar abilities to capture semantics and syntactic relations, CE handles by design OOV issues 3.1.2, which is common for rich vocabularies languages. Instead of using words as vocabulary, CE uses individual characters and semantics embeds words using the characters compositionally, which avoids word segmentation and makes it useful for language such as Chinese (Chen et al., 2015a). Moreover, CE can also perform complementary NLP tasks such as Part of Speech Tagging (Santos et al., 2014), Named-Entity Recognition (Ma et al., 2016), Sentiment Analysis (Hao et al., 2017) and LM (Kim et al., 2015). As it is at the time of writing, *FastText* based on the a morphologically-rich skip-gram approach (Bojanowski et al., 2016) as been popularilized due to its ability to

be scalabely trained on large corpora fast, and effectively.

3.3 Pre-trained Language Models

Beyond complex semantics and syntaxes provided by Word Embedding 3.1 and CE 3.2, LMs handles Context-based Word Embedding (CWE) by additionally capturing the polysemy across multiple contexts. Indeed, it was discovered that a distributed semantic, such as Word Embedding and CE are not sufficient to infer context within the embeddings (Lucy et al., 2017). A solution is to combine overall word representations from Word Embedding with *ELMo* (Peters et al., 2018), as its authors suggests, a Bidirectional Language Model (biLM) able to build deep contextual word embeddings by handling multiple word representations. As mentioned in the study, handling polymesy is just one of the LMs features as they are theoritized to capture meaningful NL traits used in NLU and Natural Language Generation (NLG). To increase the LM quality, defined by language syntactic and semantical complexities captured, UL on large corpora is popularly used, as no labeled data is required.

3.4 Transformers

Currently defining the SOTA for multiple NLP tasks with its parallelized attention 3.4.1 architecture, multi-directional pre-trained LM such as Generative Pre-Training 2 (GPT-2) or the Bidirectional Encoder Representations from Transformers (BERT) family, additionally to their ability to capture features at sentence level, out-performs by a large margin previously mentioned NLP technics at tasks such as QA by performing fine-tuning.

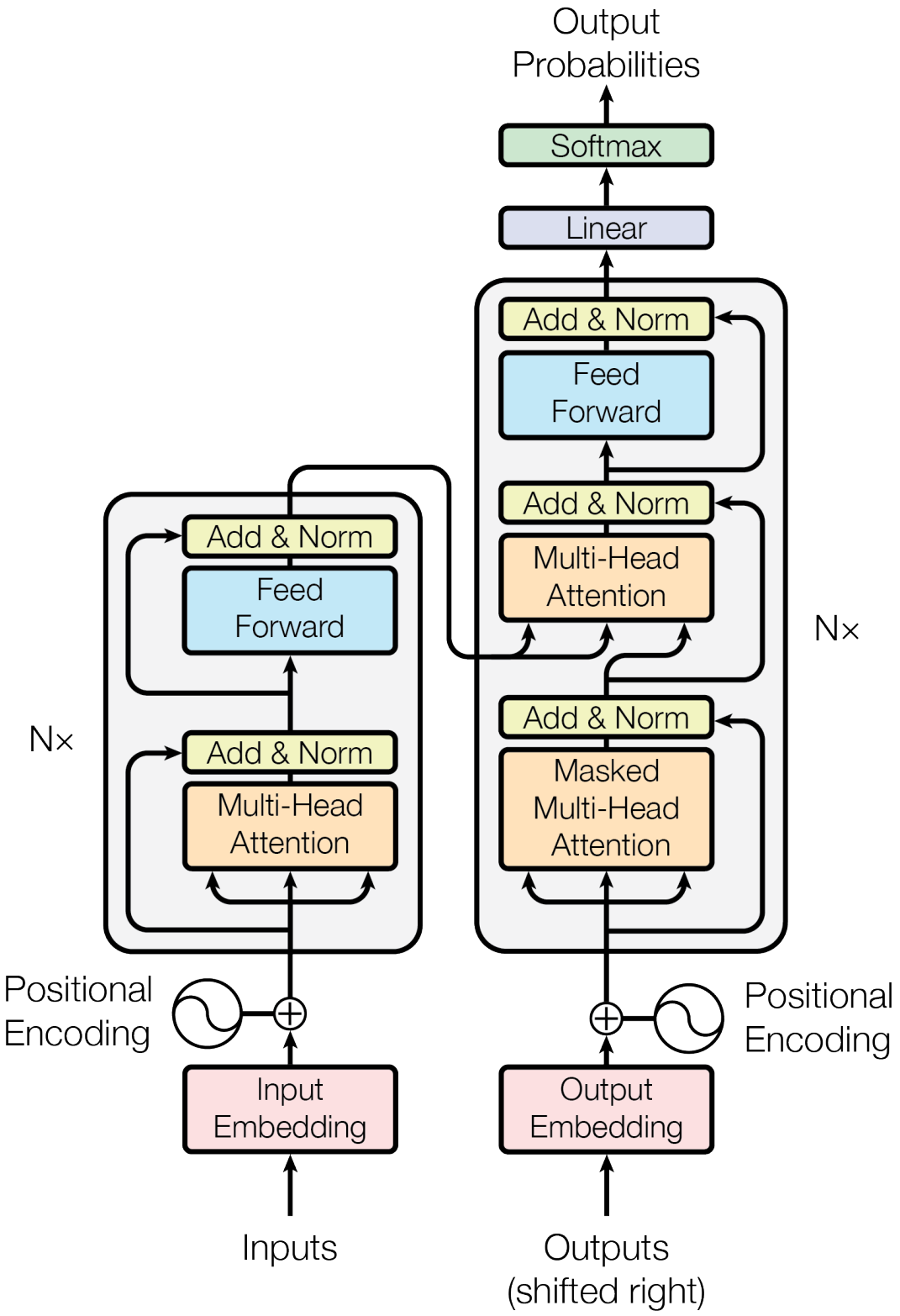

Figure 3.2: Represents a Transformer, consisting of an encoder and a decoder stacking layers with has two sub-layers comprising multi-head attention layer (Vaswani et al., 2017a)

3.4.1 Attention Mechanism

Introduced in 2017 (Vaswani et al., 2017a), The Attention Mechanism has been

The described approaches for contextual word embeddings promises better quality representations for words. The pre-trained deep language models also provide a headstart for downstream tasks in the form of transfer learning. This approach has been extremely popular in computer vision tasks. Whether there would be similar trends in the NLP community, where researchers and practitioners would prefer such models over traditional variants remains to be seen in the future.

3.4.2 Deep Learning

What are the problems such as bias, and how it can accumulate? Used to predict the next step. Various techniques, such as generative, recursive, etc.

3.4.3 Unsupervised Deep Learning

How and when unsupervised training is helpful for NLP. What are the techniques. Talk about fine tuning and re-training.

3.4.4 Generative Deep Learning

Talk about the Variational Autoencoders and how they work. What are the problems to generate realistic sentences? Talk about the latest work made with VAE (2017?) and what is the outcome. Obviously talking about the Generative Adversarial Networks, and how they work. Talk about how to evaluate the generated outputs (with an oracle?) or with BLEU scores.

3.5 Honorable Mentions

3.5.1 Recursive Neural Network

Say that in the context of NLP, it's a tree that represents every word as terminal nodes and pieces of phrases as non terminal node. Indeed, we could intuitively say that language shows some recursive structure and a hierarchy. It can be useful for sentiment analysis, particularly with the negative word like not, that inverse the meaning of a word. Not a lot of work has been done in the field, however one of the latest work in 2015 (Tai et al., 2015), showed that recusivity can be used for LSTM to avoid the gradient vanishing problem.

3.5.2 Convolutional Neural Networks

Explain what is a CNN, and how it is commonly used in sentence modeling. Say what the problems are. Compare with the Window Approach, which is working with the neighboring words. Talk about what is a Dynamic CNN and how DCNN can solve problems from CNN. Say that in the field of QA, the Multi-Column CNN (Dong et al., 2015) approach is working on multiple aspects of a question to create a representation, which is working with Freebase (ancestor of wikidata and the knowledge graphs). Furthermore, in 2016, Severyn and Moschitti (Severyn et al., 2016) proposed a QA model for Question and Answer sentences, which proposes to handle a form of relational information by matching words between question and answer pairs. Talk about the 2015 Dynamic Multi-Polling CNN (Chen et al., 2015b), which incorporates events that trigger information for the polling layer. Talk about the Conditional Random Field. Explain the Time-delayed Neural Network. Conclude that CNNs, are good at mining semantics, however, they are very heavy models, plus they are not good are modeling long-distance information from a context point of view, which is a problem also for keeping a sequential order.

3.5.3 Recurrent Neural Networks

Explain what is an RNN, basic, long short-term memory, gated recurrent units. RNN over CNN have a memory from previous computation and take advantage of it. Based on (Yin et al., 2017) say that there is no clear winner between RNN over CNN in NLP in performance, it depends on the global semantics and the task itself. Say that it's widely used in NLP tasks such as Language Modelling, Word and Sentence Classification, Machine Translation, Text generation, Image Captioning, Speech Recognition, and probably more. Say how it is useful in the previously named applications. Talk about the attention mechanism (Bahdanau et al., 2014) and its parallelized version: The Transformer (Vaswani et al., 2017b)

3.5.4 Memory Networks

Compare with the hidden vectors from encoders and decoders used in the attention mechanism, where here the hidden vectors are the input of the model. Talk about the Dynamic Memory Networks and how it is used in QA, POS, sentiment analysis, visual signals and probably more.

3.6 Common Natural Language Processing Features

Most of the following technics have been introduced to NLP by the Information Retrieval (IR) field. Indeed

- Term Frequency-Inverse Document Frequency (TF-IDF): Used to set the word importance in corpora.
- Continuous Bag of words (CBOW) [**??**]: Counts the words occurrences throughout in corpus.

- Skip-Grams [**??**]: Counts the occurrences of the character throughout in corpus.
- Topic modeling: Text clustering providing meaningful information to discover hidden structures via text chunking to identify the parts of the sentence in relation to each other.
- Segmentation: Split corpus into predefined parts, such as: *sentences, paragraphs, chapters, etc.*
- Tokenization: Split the sentences into words.
- Tagging: Based on a pre-made dictionary, it gives a new layer of meaning to the word, such as: *verb, adverb, noun, people name, locations, number, etc.*
- Dictionary: Use of tokenized words to build a dictionary, which could contain the word occurrences.
- Stop Words: Ignoring words only used as liaisons, and not containing information, such as: *and, or, etc.*
- Stemming: Uniformizing words to their root by removing the prefix and suffix, such as: *remake and loveable*.
- Lemmatization: Replace the words to their base form, such as *conjugated verb*.

Chapter 4

Generative Models

Chapter 5

Question Answering Systems

Chapter 6

Datasets

6.1 QA

Standford Question Answering Dataset (SQuAD), 100k questions posed by crowdworkers based on wikipedia atricles. bAbI. Farbes.

6.2 Dialogue Systems

Twitter conversation Triple Dataset (BLEU). Ubuntu Dialogue dataset.

6.3 Evaluation

As a beginner, it is difficult to get started at evaluating a model, and moreover, there are no simple answer about what metric to use. Here we need to find the tools to mainly evaluate QA systems and sequence to sequence generated texts.

6.4 QA

QA could provide quantifiable tasks to test a system reasoning.

2016 Ubuntu Dialogue Corpus. 1 Million multi turn dialogues, 7 million utterances and 100 million words.

SQuAD

2018 SQuAD 2.0 over 2016 1.0 includes 50'000 unanswerable questions similar to answerable ones (adversarially). It's a closed dataset that gives the answers to question to a given context. It focuses on extreme confusing questions. However it doesn't contain questions that require common sense or reasoning. Wikipedia based. Keeps track of the previous interactions. Yes/No answers accepted for abstractions.

TriviaQA

2017. 650k question-answer-evidence triples.

MS MARCO

2016. (Machine Reading Comprehension). Has been used to evaluate generative models from Seq2Seq, Memory Networks, and Disciminative models. Uses Real and anonymized user queries from Bing. Context is given by a real web documents. All answers are human generated. Subsets are multiples anwsers or no as. All queries are tagged with segment informations.

CoQa

2019. Wikipedia based and 6 other domains. Keeps track of the previous interactions. Yes/No answers accepted. Allows crowd to add an

QuAC

2018. Focus on missing information. Wikipedia based. Keeps track of the previous interactions. Yes/No answers accepted for abstract questions. They doesn't allow the crowd to see the context before formulating the question.

A Qualitative Comparision of CoQA, SquAD 2.0 and QuAC, 2019

Test of the Unanswerable questions, multi-turn interactions, and abstractive answers. Responses are produced by crowd about a paragraph of text, and required to provide a span of text validating their answers. None of the datasets are providing

Natural Questions Corpus

2019. Google dataset (Natural Questions: a Benchmark for Question Answering Research). It's goal is to provide an appropriate training and testing set for QA. It pairs real user queries to what they self call high quality annotations of answers in documents. It also provide metrics to evaluate the performances. Mainly based on a wikipedia, it provides the page, a long and a short answer, and additionally statistics.

GLUE

2019. Banchmarking tool based on existing datasets, which 4 of them are private.

Dataset	Segment	Query Source	Answer	# Queries	# Documents
MCTest	N	Crowdsourced	Multiple choice	2640	660
WikiQA	N	User logs	Sentence selection	3047	29.26K sentences
CNN/Daily Mail	N	Cloze	Fill in entity	1.4M	93K CNN, 220K DM
Children's Book	N	Cloze	Fill in the word	688K	688K contexts, 108 books
SQuAD	N	Crowdsourced	Span of words	100K	536
MS MARCO	Y	User logs	Human generated	100K	1M passages, 200K+ doc.

Table 1: Comparison of some properties of existing datasets vs MS MARCO. MS MARCO is the only large dataet with open ended answers from real user queries

Figure 6.1: TMP comparative table took from MS-MARCO

Chapter 7

Evaluation

7.1 Generative

Not a lot of work has been done in the domain of computer generated text, taking appart machine translation. So far I found a paper (GLTR: Statistical Detection and Visualization of Generated Text) that talks about a technic to detect generated content. And I found an article about the Readers' perception of computer-generated news(: Credibility, expertise, and readability) which says that people like to read computer generated content because it's what it is, and humans are amazed about it, the long term have to be explored.

BLEU

The Bilingual evaluation understudy was originally made to measure machine translation and now is used for Natural language generation. The goal is to compare the machine generated output text string to its expected output, the algorithm is fast and easily computable compared to human translators. Plus, it's benchmarkable. However, it's not made to take the meaning of the sentence into account (attention), it's not taking into account the sentence structure, it's doesn't work with rich languages and it doesn't take into account how humans are actually interpreting the sentences. Basically to used only in the machine translation context to evaluate an entire corpus.

ROUGE

Adaptation from BLUE to focus on the recall instead of Precision. It checks the the reference translation to the output.

performances

7.2 QA Systems

7.3 Generative Systems

7.4 Conversational Agents

7.5 Convex Dataset

7.5.1 Data augmentation

Example:

Who is the author of the Harry Potter series? OR Who wrote Harry Potter? What was the year of publication for the first book? OR When was it first published? Title of the first book? OR The first book was called what? What country was the book set in? OR It was set in what country? Which book has the highest page count? OR What's the longest book?

7.5.2 Human errors

Mechanical Truck gathering mistakes during the dataset building due to humans not respecting the standard format. Implying 32 wrong question-answers for a single mistake. When did the first The Fast and the Furious film come out? TO "answer": `"https://www.wikidata.org/wiki/Q155476"` instead of 22 June 2001 however: "answer_text": "The first film came out 22 June 2001."

7.5.3 Data inconstancy

"question": "When was he born?", "answer_text": "1 August 1819" same as answer

Sometimes it is binary, and some times it is in NL

"question": "When was it published?", "answer_text": "The book came out 30 June 1997 in the UK."

This is an important inconstancy biaises for training.

7.5.4 Wrong answers

When did the first The Fast and the Furious film come out?

GraphQA answers 1955, which is the date of publication of the original The Fast and the Furious movie. And non of the competing qa systems answer correctly to the question, neither to the provided false answer, neither the correct one.

We didn't take time to go thru the whole dataset because time was missing, but funnily, GraphQA most often triggered warnings when the dataset had such errors, that's why we saw them.

This all implies that GraphQA, could even perform better than the concurrents, but more exhaustive evaluations on additional datasets are required. But in the current version GraphQA is very time and computing resources consuming, which made it hard to evaluated it on multiple datasets in parallel to development

7.5.5 Don't trust Mechanical Trucks

7.6 What we learned from the project

7.6.1 Don't trust someone else word without proving it yourself

Preprints: some good and mostly bad Published articles: some good, but mostly interative research with name dropping Published in conferences: some good, but be careful at where it is published, china is worrying Sadly everything looks alike with time Never trust what's written, always cross check the results and the given datasets or code if any Be critique with state-of-the-art and baseline clams as long as it was not reproduced.

7.7 What happens

Conversational, It tests each conversation with convex and graphqa as extension. Note that if no initial answer if found, no graph is built for platypus or qanswer, which skips the graph extension, as it's part of the nature of GraphQA and Convex

Chapter 8

Problems

Generative: Exfiltrating copyright notices, news articles, and IRC conversations from the 774M parameter GPT-2 data set. Sofist Machines

Chapter 9

Conclusions

Bibliography

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua, 2014. Neural Machine Translation by Jointly Learning to Align and Translate, pp. 1–15. Available from arXiv: 1409.0473.

BENAICH, Nathan; HOGARTH, Ian, 2019. State of AI 2019, pp. 126. Available also from: \url{https://www.stateof.ai/}.

BOJANOWSKI, Piotr; GRAVE, Edouard; JOULIN, Armand; MIKOLOV, Tomas, 2016. Enriching Word Vectors with Subword Information. *CoRR*. Vol. abs/1607.04606. Available from arXiv: 1607.04606.

CHEN, Xinxiong; XU, Lei; LIU, Zhiyuan; SUN, Maosong; LUAN, Huan-Bo, 2015a. Joint Learning of Character and Word Embeddings. In: *Joint Learning of Character and Word Embeddings*. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 1236–1242. Available also from: http://ijcai.org/Abstract/15/178.

CHEN, Yubo; XU, Liheng; LIU, Kang; ZENG, Daojian; ZHAO, Jun, 2015b. Event extraction via dynamic multi-pooling convolutional neural networks. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. Vol. 1, pp. 167–176. ISBN 9781941643723.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina, 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available from DOI: 10.18653/v1/N19-1423.

DONG, Li; WEI, Furu; ZHOU, Ming; XU, Ke, 2015. Question answering over freebase with multi-column convolutional neural networks. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. Vol. 1, pp. 260–269. ISBN 9781941643723.

DUNLOP, Michal Wallace & George, 1999. *Eliza, the Rogerian Therapist* [http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm]. (Accessed on 10/09/2019).

Bibliography

FUTURISM, LLC, 2016. *The History of Chatbots Infographic* [https://futurism.com/images/the-history-of-chatbots-infographic]. (Accessed on 10/09/2019).

GITTENS, Alex; ACHLIOPTAS, Dimitris; MAHONEY, Michael W., 2017. Skip-Gram â^' Zipf + Uniform = Vector Additivity. In: *Skip-Gram â^' Zipf + Uniform = Vector Additivity. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 69–76. Available from DOI: 10.18653/v1/P17-1007.

HAO, Yazhou; ZHENG, Qinghua; LAN, Yangyang; LI, Yufei; WANG, Meng; WANG, Sen; LI, Chen, 2017. Improving Chinese Sentiment Analysis via Segmentation-Based Representation Using Parallel CNN. In: CONG, Gao; PENG, Wen-Chih; ZHANG, Wei Emma; LI, Chengliang; SUN, Aixin (eds.). *Advanced Data Mining and Applications*. Cham: Springer International Publishing, pp. 668–680. ISBN 978-3-319-69179-4.

HERBELOT, Aurélie; BARONI, Marco, 2017. High-risk learning: acquiring new word vectors from tiny data. *CoRR*. Vol. abs/1707.06556. Available from arXiv: 1707.06556.

KARRAS, Tero; LAINE, Samuli; AITTALA, Miika; HELLSTEN, Janne; LEHTINEN, Jaakko; AILA, Timo, 2019. Analyzing and Improving the Image Quality of StyleGAN. *CoRR*. Vol. abs/1912.04958.

KELNAR, David, 2019. The State of AI, pp. 151. Available also from: https://www.stateofai2019.com/summary/.

KIM, Yoon; JERNITE, Yacine; SONTAG, David A.; RUSH, Alexander M., 2015. Character-Aware Neural Language Models. *CoRR*. Vol. abs/1508.06615. Available from arXiv: 1508.06615.

LUCY, Li; GAUTHIER, Jon, 2017. Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. *CoRR*. Vol. abs/1705.11168. Available from arXiv: 1705.11168.

MA, Yukun; CAMBRIA, Erik; GAO, Sa, 2016. Label Embedding for Zero-shot Fine-grained Named Entity Typing. In: *Label Embedding for Zero-shot Fine-grained Named Entity Typing. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 171–180. Available also from: https://www.aclweb.org/anthology/C16-1017.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey, 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, pp. arXiv:1301.3781. Available from arXiv: 1301.3781 [cs.CL].

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D., 2014. GloVe: Global Vectors for Word Representation, pp. 1532–1543. Available also from: http://www.aclweb.org/anthology/D14-1162.

PETERS, Matthew E.; NEUMANN, Mark; IYYER, Mohit; GARDNER, Matt; CLARK, Christopher; LEE, Kenton; ZETTLEMOYER, Luke, 2018. Deep contextualized word representations. *CoRR*. Vol. abs/1802.05365. Available from arXiv: 1802.05365.

PINTER, Yuval; GUTHRIE, Robert; EISENSTEIN, Jacob, 2017. Mimicking Word Embeddings using Subword RNNs. *CoRR*. Vol. abs/1707.06961. Available from arXiv: 1707.06961.

RADFORD, Alec; WU, Jeffrey; CHILD, Rewon; LUAN, David; AMODEI, Dario; SUTSKEVER, Ilya, 2018. Language Models are Unsupervised Multitask Learners.

RAJPURKAR, Pranav; JIA, Robin; LIANG, Percy, 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In: *Know What You Don't Know: Unanswerable Questions for SQuAD*. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 784–789. Available from DOI: 10.18653/v1/P18-2124.

ROWSELL-JONES, Andy; HOWARD, Chris, 2019. *2019 CIO Survey: CIOs Have Awoken to the Importance of AI* [https://www.gartner.com/en/documents/3897266/2019-cio-survey-cios-have-awoken-to-the-importance-of-ai]. (Accessed on 10/09/2019).

SANTOS, Cícero Nogueira dos; ZADROZNY, Bianca, 2014. Learning Character-level Representations for Part-of-Speech Tagging. In: *Learning Character-level Representations for Part-of-Speech Tagging. ICML*. JMLR.org, vol. 32, pp. 1818–1826. JMLR Workshop and Conference Proceedings. Available also from: http://dblp.uni-trier.de/db/conf/icml/icml2014.html#SantosZ14.

SEVERYN, Aliaksei; MOSCHITTI, Alessandro, 2016. Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks. Available from arXiv: 1604.01178.

TAI, Kai Sheng; SOCHER, Richard; MANNING, Christopher D., 2015. Improved semantic representations from tree-structured long short-Term memory networks. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. Vol. 1, pp. 1556–1566. ISBN 9781941643723. Available from arXiv: arXiv:1503.00075v3.

TURING, A. M., 1950. Computing Machinery and Intelligence. *Mind*. Vol. 59, no. 236, pp. 433–460. ISSN 00264423. Available also from: http://www.jstor.org/stable/2251299.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia, 2017b. Attention is all you need. *Advances in Neural Information Processing Systems*. Vol. 2017-December, no. Nips, pp. 5999–6009. ISSN 10495258. Available from arXiv: arXiv:1706.03762v5.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Lukasz; POLOSUKHIN, Illia, 2017a. Attention Is All You Need. *CoRR*. Vol. abs/1706.03762. Available from arXiv: 1706.03762.

WANG, Phil, 2019. *This Person Does Not Exist* [https://www.thispersondoesnotexist.com]. (Accessed on 10/09/2019).

Bibliography

YIN, Wenpeng; KANN, Katharina; YU, Mo; SCHÜTZE, Hinrich, 2017. Comparative Study of CNN and RNN for Natural Language Processing. Available from arXiv: 1702.01923.

List of Figures

List of Tables

Appendix

.1 Worklog

.2 Jupyter Notebooks

.3 Spreadsheet

.4 Meeting Notes