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)

# GenBot

Author:
Romain Claret[1]

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

External expert:
Dr. Christophe Gisler

Lausanne, HES-SO//Master, May 23, 2019

[1] romain.claret@master.hes-so.ch

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

Prof. Dr. Jean Hennebert, deepening project supervisor
Dr. Christophe Gisler, iCoSys, main expert

Place, date: _____

| | |
|---|---|
| Prof. Dr. Jean Hennebert | Prof. Dr. Philippe Passeraub |
| Supervisor | Dean, HES-SO//Master |

# Contents

# Acknowledgments

# Glossary

**AdaGrad**

Adaptive Gradient Algorithm that maintains a per-parameter learning rate that improves performance on problems with sparse gradients (e.g. natural language and computer vision problems)..

# Acronyms

**13C**

carbon-13.

**ACN**

Acetonitrile.

**CHCA**

Cyano 4 hydroxy cinnamic acid.

# Abstract

In the scope of this deepening project, and as the technology of NLP is in constant evolution, we will be focusing on the exploration of the word embedding algorithm Word2Vec, which is, at the beginning of 2019, commonly used as a fondation for Deep Neural Network Chatbots. As a result to this project, the student is demonstrating what is the Word2Vec technology, its extensions, and its applications.

**Keywords:** Word Embedding, Word2Vec, Natural Language Processing, NLP, Natural Language Understanding, NLU, Machine Learning, Data Engineering, Conversational Agent, Chatbot, Generic

# Chapter 1

# Introduction

Beginning of 2019, chatbots are everywhere but very limited to narrow tasks, and are, in most cases, sequences of if-else conditions resulting in a very weak AI. Indeed, hard-coded connections are requiring an infinite amount of human power to create generic Chatbots able to maintain a conversation at a human level. However, the progress in the field of machine learning 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. Multiple algorithms and technics are emerging monthly, which are demonstrating promising conversational performance improvement; however, they are all still narrow AIs. Indeed, even if they are getting better at providing meaningful sentences, they are still not able to generalize all tasks linked to a conversation, such as, understanding the context, search and learn for missing information, initiate conversation in a meaningful manner, be intuitive, and more. The generalization of those features would allow a significant step forward into general Chatbots.

As the driver, iCoSys, the Institut of Complex Systems at University of Applied Sciences and Arts at Fribourg, Switzerland, is interested into the result of this project as a study for their AI-News project, whose goal is to provide a chatbot as a tool to reader, to help them narrow their interests and deliver the right information. AI-News is in collaboration with the Swiss Innovation Agency from the Swiss Confederation, and La Liberté, the daily newspaper from Fribourg.

## 1.1 Aim of Study

In harmony with the author interest, the goal of this deepening semester project is to suggest and demonstrate strategic approaches as a premise to the Artificial General Intelligence (AGI) and getting a step closer to general Chatbots, which can initiate and maintain human-like conversations in a pro-active manner.

## 1.2 Scope and Study Borders

As a red line for this deepening project, the focus will be on the Word2Vec technology, from a research perspective. Indeed, this technology is seen as a foundation for the modern Natural Language Processing (NLP) and Deep Neural Network (DNN) Chatbots, which makes it an exciting vector of study about its current usage, its extensions, and potential evolution.

# Chapter 2

# Questions

As a methodology to help the student to find a red line to focus its research on, he was required work on the subject "What should be the initial questions to asks to start making Artificial General Intelligence Chatbots" as preliminary study before the beginning of deepening project itself and to write down the outcome as a set of questions related to his interests and the field of AGI Chatbots.

## 2.1 Initial and Broad Questions

As a result to the preliminary study, the following question were produced. Please take into account that those questions were not meant to be answered as part of the project itself, but as part of the process of appropriation of the field of study.

- Is the artificial neural network approach appropriate to represent the world?
- Can agents be made exclusively from a language?
- Can agents able to experience an environment?
- Is a narrative environment be enough to understand an environment?
- Is the language able to provide to an agent an understanding of the world?
- Is the knowledge of the language syntax enough to gain an understanding?
- Is the result of unsupervised learning enough to discover all nuances?
- Is the unsupervised learning sufficient to make sense to an environment?
- Is a descriptive explanation of the world be expressed in a language?
- Is the description good enough to catch all the nuances?
- Is the language good enough to explain?
- Can we augment or make a semantic language?
- Can we create a common symbolic language?
- Is the language multi-dimensional?
- How many dimensions are needed for a complex language?
- Is it possible to give a word equivalence to machines for human-specific words?
- Are all emotions describable into words?
- Are emotions altering language descriptions?
- Is an approximation of the real world enough to understand the environment?
- Would a the simulated world be a good approximation of the real world?

## 2.2 Narrowed Questions

In a second time, the student was asked to narrow the initial questions above into potential fields of study.

- Common human-machine language
  - Is it possible to create a multi-dimensional human-machine language, which includes a common semantic, symbolic, and emotion definition.
  - Is it possible to create an abstract world for machines to understand human symbolic based on a real world, and define fundamentals for machine representation of the language.
- Machine intuition
  - Is it possible to provide to machines an human-like intuition (inside voice), which would help to keep a long term context and specialize in specific fields.
- Evaluate human-machine communication
  - Is it possible to provide a protocol to test the communication skills and machine understanding.

## 2.3 Potential Red lines

From the potential fields above, the following suggested red lines were proposed.

- How to verify and quantify a chatbot understanding?
- What is the premise to make chatbots general with today's technology?
- How chatbot can be proactive?
- How to simulate human-like intuition in chatbots?

## 2.4 The Deepening Project Question and Red line

Based on reflective work and discussions, the concluding red line and question for this deepening project is:

- What is Word Embedding and how is it useful for chatbots?

# Chapter 3

# Plan

## 3.1 Contraints

**Timeframe:** 15 weeks
**Starting date:** 18.02.2019
**Ending date:** 31.05.2019

## 3.2 Initial Plan

As an initial milestone for the deepening project, the student were required to create an initial plan, with the purpose to help the student and the teacher to visualise the project main red line.

### 3.2.1 Tasks

1. Initial research about general chatbots
2. Determine the project target
3. Play with the subject
4. Explore the Word2Vec methodology
5. Explore the Word2Vec extensions
6. Combine and test ANN algorithms with Word2Vec
7. Explore ANN algorithm topology for the chatbot
8. Analysis of the chatbot intuition with parallel algorithms
9. Analysis of a protocol to evaluate proactive chatbots
10. Profile-based initiatives
11. Analyze and experiment profile nurturing
12. Analyze and experiment with chatbot initiatives with no profiles
13. Overall improvements
14. Autonomous data gathering
15. Make suggestions
16. Determine possible continuation and future outcomes for the project

### 3.2.2 Milestones

1. Initial deepening project plan and specification document
2. Basic multi-dimensional word embedding space
3. Basic conversational agent
4. Basic proactive chatbot
5. Deepening project report

### 3.2.3 Sprints

**18.02.19 to 08.03.19** (3 weeks)

- Do the initial research about general chatbots
- Determine the project target
- Play with the subject
- **DELIVERABLE:** Plan and Initial Specification document

**11.02.19 to 29.03.19** (3 weeks)

- Explore the Word2Vec methodology and its extensions
- Combine and test ANN algorithms with Word2Vec
- **MVP:** Basic multi-dimensional word embedding space

**01.04.19 to 19.04.19** (3 weeks)

- Explore ANN algorithm topology for the chatbot
- Analysis of the chatbot intuition with parallel algorithms
- Analysis of a protocol to evaluate proactive chatbots
- **MVP:** Basic conversational agent

**22.04.19 to 10.05.19** (3 weeks)

- Profile-based initiatives
- Analysis and experiment of the profile nurturing
- Analyze and experiment with chatbot initiatives with no profiles.
- **MVP:** Basic proactive chatbot

**13.05.19 to 31.05.19** (3 weeks)

- Overall improvements
- Autonomous data gathering
- Make suggestions
- Determine possible continuation and future outcomes for the project
- **DELIVERABLE:** Report + Sources

### 3.2.4 Gantt chart

## 3.3 Effective Plan

As expected the initial plan served as an initial model, and evolved iteratively based on the student and teach feedback while exploring the subject.

### 3.3.1 Tasks

1. Initial research about general chatbots
2. Determine the project target
3. Set the initial plan
4. Explore the Word2Vec subject
5. Explore the Word2Vec algorithm
6. Build a Word2Vec model on the latest english wikipedia dump
7. Explore Word2Vec parameters
8. Explore Word2Vec analogies
9. Explore Word2Vec sentence generation
10. Explore visual representations of Word2Vec vectors
11. Explore Word2Vec applications with chatbots
12. Writing the report

### 3.3.2 Milestones

1. Initial deepening project plan and specification document
2. Basic Word2Vec Word Embedding Model
3. Conclusions Word2Vec based chatbots
4. Ideas to make chatbots proactive
5. Deliver the report

### 3.3.3 Gantt chart

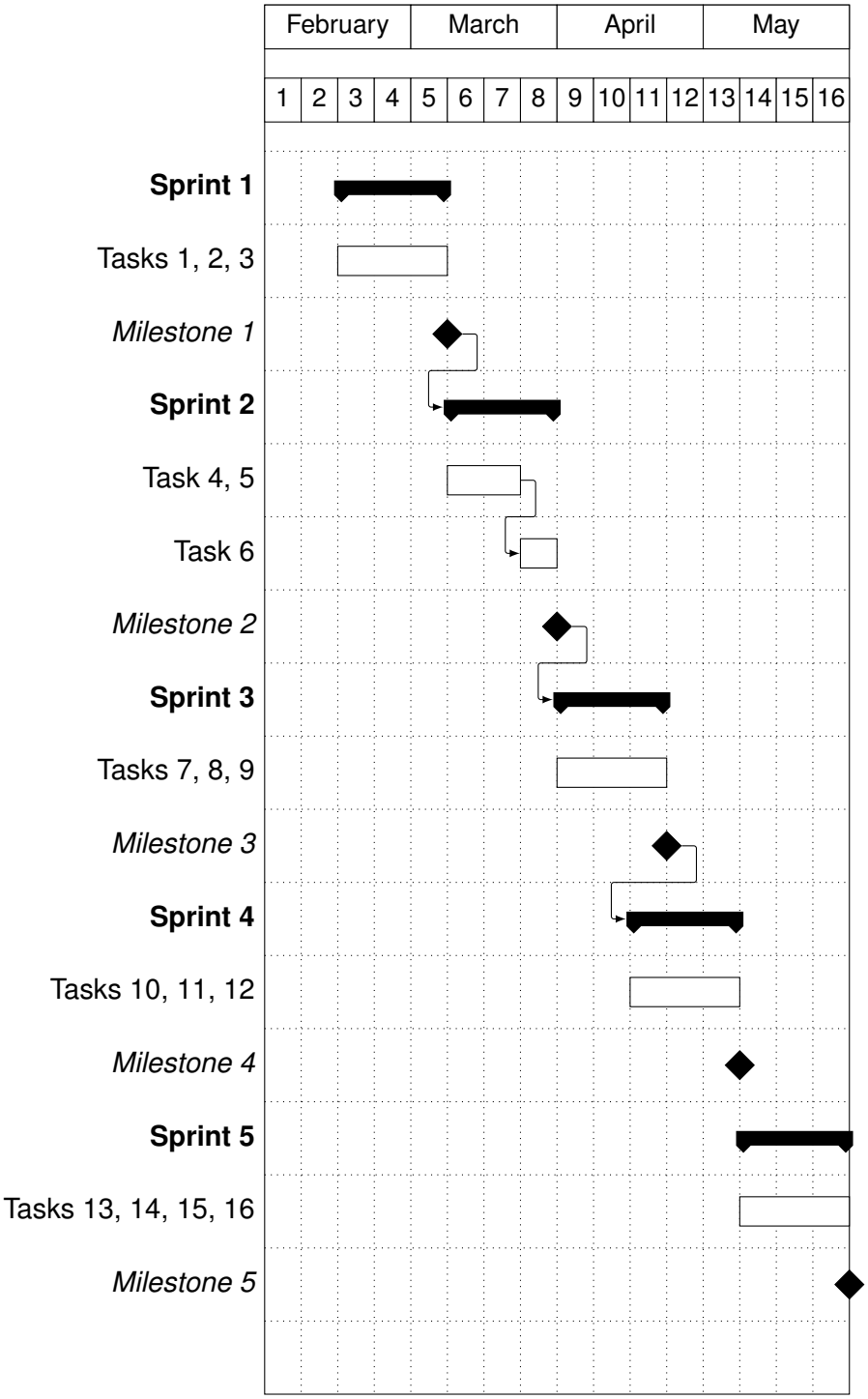


Figure 3.1: Initial Gantt Chart

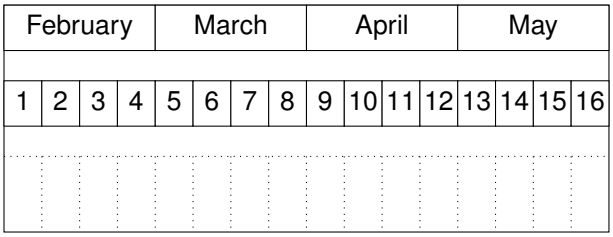


Figure 3.2: Effective Gantt Chart

# Chapter 4

# State of the art

## 4.1 Chatbots

### 4.1.1 What are Chatbots

### 4.1.2 Narrow

### 4.1.3 General

### 4.1.4 AIML

### 4.1.5 Deep Neural Networks

### 4.1.6 Retrieval

## 4.2 Word2Vec

### 4.2.1 What is Word2Vec

### 4.2.2 Gensim

### 4.2.3 Framworks

## 4.3 Word Embedding Alternatives

### 4.3.1 FastText

### 4.3.2 Glove

### 4.3.3 Word2Vec-f

### 4.3.4 Wang2vec

## 4.4 Sentence/Document Embedding Alternatives

### 4.4.1 Doc2vec

### 4.4.2 Skip-thought

### 4.4.3 Smooth Inverse Frequency

### 4.4.4 RNN

# Chapter 5

# Analysis

## 5.1 Natural Language Processing

## 5.2 Pipeline

## 5.3 Word2Vec

### 5.3.1 Bag of Words VS Skip-Gram

### 5.3.2 Dimensions

### 5.3.3 N-Grams

### 5.3.4 Epochs

### 5.3.5 Lemmatization

### 5.3.6 Normalization

### 5.3.7 Distance and Cosine Angle

### 5.3.8 Training

### 5.3.9 Retrain Model

### 5.3.10 Memory Issues

### 5.3.11 Analogies

### 5.3.12 Proverbs

### 5.3.13 Evaluation

### 5.3.14 Visual Representation

### 5.3.15 Benchmarks

### 5.3.16 CPU VS GPU

### 5.3.17 Datasets

## 5.4 Chatbot

## 5.5 Proactivity

# Chapter 6

# Experiments & Results

## 6.1 Environments

### 6.1.1 Jupyter Notebook

### 6.1.2 Local Machine

### 6.1.3 Amazon Web Services

### 6.1.4 iColab-gpu2

### 6.1.5 CPU Dedicated Machine

## 6.2 Gensim Framework

**Errors** Memory allocation with multi-core. The problem is occurring during the merge of the cores. Indeed, my current machine has 128GB ram, and the dataset weights about 16GB in the memory, and each core during merging is processing at least the same amount, plus the processed informations.

```
2019-03-25 08:31:18,867 : INFO : PROGRESS: pass 0, dispatched chunk #34 = documents up to #70000/4614519, outstanding queue
 size 31
Exception in thread Thread-1:
Traceback (most recent call last):
  File "/usr/lib/python3.5/threading.py", line 914, in _bootstrap_inner
    self.run()
  File "/usr/lib/python3.5/threading.py", line 862, in run
    self._target(*self._args, **self._kwargs)
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 366, in _handle_workers
    pool._maintain_pool()
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 240, in _maintain_pool
    self._repopulate_pool()
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 233, in _repopulate_pool
    w.start()
  File "/usr/lib/python3.5/multiprocessing/process.py", line 105, in start
    self._popen = self._Popen(self)
  File "/usr/lib/python3.5/multiprocessing/context.py", line 267, in _Popen
    return Popen(process_obj)
  File "/usr/lib/python3.5/multiprocessing/popen_fork.py", line 20, in __init__
    self._launch(process_obj)
  File "/usr/lib/python3.5/multiprocessing/popen_fork.py", line 67, in _launch
    self.pid = os.fork()
OSError: [Errno 12] Cannot allocate memory
```

Figure 6.1: Error 1

## 6.3 Materials

## 6.4 Products

```
2019-03-25 08:31:18,867 : INFO : PROGRESS: pass 0, dispatched chunk #34 = documents up to #70000/4614519, outstanding queue
 size 31
Exception in thread Thread-1:
Traceback (most recent call last):
  File "/usr/lib/python3.5/threading.py", line 914, in _bootstrap_inner
    self.run()
  File "/usr/lib/python3.5/threading.py", line 862, in run
    self._target(*self._args, **self._kwargs)
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 366, in _handle_workers
    pool._maintain_pool()
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 240, in _maintain_pool
    self._repopulate_pool()
  File "/usr/lib/python3.5/multiprocessing/pool.py", line 233, in _repopulate_pool
    w.start()
  File "/usr/lib/python3.5/multiprocessing/process.py", line 105, in start
    self._popen = self._Popen(self)
  File "/usr/lib/python3.5/multiprocessing/context.py", line 267, in _Popen
    return Popen(process_obj)
  File "/usr/lib/python3.5/multiprocessing/popen_fork.py", line 20, in __init__
    self._launch(process_obj)
  File "/usr/lib/python3.5/multiprocessing/popen_fork.py", line 67, in _launch
    self.pid = os.fork()
OSError: [Errno 12] Cannot allocate memory
```

Figure 6.2: Error 2

# Chapter 7

# Discussion

# Chapter 8

# Conclusion

[1][2][3][4] AdaGrad Acetonitrile (ACN) 13C Cyano 4 hydroxy cinnamic acid (CHCA)

Lausanne, May 23, 2019

Romain Claret

# References


[1] J. J. Li and E. J. Corey. “Name Reactions in Heterocyclic Chemistry”. In: *John Wiley & sons, Inc.* 2005.

[2] *2013 28th Annual ACM/IEEE Symposium on Logic in Computer Science* (New Orleans, Louisiana, June 25–28, 2013). IEEE Computer Society, 2013. ISBN: 978-1-4799-0413-6.

[3] Dirk Pattinson. “The Logic of Exact Covers. Completeness and Uniform Interpolation”. In: pp. 418–427.

[4] (New Orleans, Louisiana, June 25–28, 2013). IEEE Computer Society, 2013. ISBN: 978-1-4799-0413-6.

# Appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

### .0.1 Appendix

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.