POLITECNICO DI TORINO

Master Degree in Computer Engineering

Master Degree Thesis

Towards a Self-aware Intelligent Agent

Through Automated Output Interpretation and Improvisation

Supervisor
Maurizio Morisio

Candidate
Oliveri Isabeau

Co-supervisor
Giuseppe Rizzo

October 2019

To my home, to my loved M.

Summary

We propose and analyse a set of metrics that can be used as an assessment of artificial intelligence, known as Psychometric AI. In this particular experimental setup, our studies are applied to the linguistic field: therefore, these metrics can be used for the evaluation of internal content generated by intelligent conversational agents, better known as chatbots.

Intelligent conversational agents are becoming more and more pervasive in our lives. Despite being defined as intelligent, these agents are still unable to react correctly and adapt in contexts for which they have not been trained. This shows us that the artificial intelligence we are aiming for is still a very long way. The first research question of this thesis is: what is a stronger definition of true intelligence, and what are the mechanisms by which conversational agents can achieve it? The second question, more technical, is: how to objectively measure products of intelligence so defined?

Looking at the world of artificial intelligence, professor Searle proposes a strong concept of artificial intelligence, stating that an agent can be considered intelligent only if he is able to explain what he is processing, as an internal process. Experts in this field have brought these concepts back into explainability, self-awareness, and combinatorial creativity, as a tool to evaluate and create novel and right content.

Analysing state of the art, we find ourselves in a system without well defined shared techniques for the evaluation of conversational agents. This is even truer when we are looking for metrics that analyse the internal evaluation of the content respect previous knowledge. The arise of very

young field, Psychometric AI, pointing out that the scientific community of the AI needs to explore shared metrics on these unconventional aspects.

The aim of this thesis is propose five metrics that fill this gap: adherence, diversity, novelty, serendipity and magnitude. We collectively define them under the name of creativity index. To test these metrics, we had to build a knowledge base over which we can compute their values for new input content. In a first step, we extract data from Wikidata, an open RDF structured knowledge base, pre-processing selected pages as a document (D). Each one is composed by statements that we shape in simplified triples (T). We store and adapt them into a knowledge graph, ones of the best structure to store linguistic and abstract concepts, capable of showing easily links among them. In a second step, we evaluate documents using our metrics respect the knowledge graph previously created, tracing the results. Remembering that we are talking about conversational agent, the similarity function that is reported below in the formulas is a similarity function based on semantic similarity, that in our case exploits the power of word embeddings.

$$ci(D) = [ad, coe, div, ser, mag] \tag{1}$$

Adherence is the part of our indicator that represents our agent's ability to generate existing content shared and accepted by agent and referring knowledge base.

$$ade(D) = \frac{1}{n} \sum_{i=0}^{n} exist(T_i) \tag{2}$$

The diversity of a document is seen as the semantic difference between internal triples. $\#C(x, y)$ is the number of combinations without repetition.

$$div(D) = \frac{1}{\#C(n,2)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1 - similarity(T_i, T_j) \tag{3}$$

Novelty evaluates the novelty brought by the document with respect to what was previously expressed by those present in the knowledge graph, regardless of whether it is correct or incorrect.

$$nov(D) = \frac{1}{n} \sum_{i=1}^{n} 1 - similarity(D, D_i) \tag{4}$$

Serendipity is the part of our indicator that represents the quantity of novel, correct and not trivial data that agent generates. S represents the first s documents in a document ranked list ordered by novelty nov(D) function.

$$ser(D) = ade(D)\frac{\sum_{i=1}^{s} 1 - similarity(D, S_i)}{s} \tag{5}$$

With magnitude, we are going to evaluate the weight of triple components based on their frequency in the graph. In particular, we see how rare (function rank) are the items and properties.

$$mag(D) = \frac{1}{n}\sum_{i=1}^{n} \frac{rank(sbj) + rank(prop) + rank(obj)}{3} \tag{6}$$

Pareto optimum was used to analyse correlation between key components of our metrics, in order to identify parameters which imply particular agent behaviours. The final work was evaluated by comparing the results with user surveys on heterogeneous groups of people and our personal consideration. Different discrepancies arise, but they are identified and resolvable problem. Several aspects have to bearing in mind: incompleteness of a knowledge graph, context recognition, proper merge between graph and deep learning strategy, time dimension, bias, large amount of data, technical limit of our hardware, subjective opinions of the interviewees.

This study is an opportunity to investigate early introspective steps of future conversational agents, which should attempt to become self-aware, in order to further develop the capacity to understand information they store, with the goal to interact with users or other agents in a non-trivial way. To pursue this goal, future work may include incorporating our metrics directly within the generative phase of the conversational agent and creating a control ring capable of driving the agent in the autonomous creation of content.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 LINKS foundation

This work were made in collaboration with LINKS Foundation[1], founded in 2018 by Compagnia di San Paolo and Politecnico di Torino. Rose from the union of Istituto Superiore Mario Boella and SITI, it has of its main objective improving technological innovation and territorial development, conveying high competence and latest technology onto real case scenario. This turns it in a remarkable meeting point between research and socio-tecnologic renewal for companies. One of LINKS strengths is begin this process directly into university, by diving thesis student to face with high level question, going beyond state of the art: this work was developed within this visionary idea, and we take the opportunity to thank the whole team.

1.2 Conversational Agents

The aim of this introduction is providing an overall non-technical idea of the background scenario that concerns this work. We give a brief recap of crucial steps in artificial intelligence that brought us to the main topic of this writing, conversational agents CA. At the meanwhile, we write down the most relevant concepts to ensure that readers could easier understand the common thread of the following lecture. In the latter part, we clarify

[1] https://linksfoundation.com

what is intended with a "self-aware" agent, explaining how this may address some current issues of CA. Mentioned arguments will be further investigated in the next chapters in technically and deeply way.

1.2.1 A brief history

The scientific community over time has attempted to mimic the human being in several ways, from the mechanic point of view as far as what more elegant he owns, its intelligence. Nowadays, in computer science terms, such studies are known under the name of artificial intelligence (AI). Far from collective imagination, this macro discipline it is not a unique Pandora's box. It consists of several branches, that working in conjunction and, in their turn, working with other macro disciplines such as robotics, biology and psychology. Indeed, during the year, it was clear that this task is an overwhelming one considered in its entirety, even for most powerful mainframes and innovative algorithms.

Among branches of artificial intelligence, this thesis mainly relies and questioning on Natural Languages Understanding (NLU) and Natural Languages Generation (NLG) ones, both better known under Natural Language Processing (NLP). As you can guess, together enable an important point in computer science, the natural communication between machine and human. Frequently dreamt and narrated in literature, movie, comics, a lot of people have to opportunity to give different shapes to this concept. It is challenging to define what is not agreed in a human scenario even earlier. Give a precise technical definition could be a hard nut to crack: the nuances more akin to our purpose will be provided. So, what are more specifically NLU and NLG?

Natural Languages Understanding is the branch of NLP that address the problem of understanding the input of human users when they interact in a natural manner. Different format of input as to be considered: text, voice, gesture and so on. Someone might say that more data will be considered and matched, harder will be the task. This correct, but only partially: with understanding, we do not just indicate the task of simple isolated recognition of keywords but the whole meaning and purpose of the sentences. For this reason, strictly human information as tone of voice, for example, can be improving and resolve the understanding of some unclear

situations. This layer is also in charge to manage some errors, as misspelling and mispronunciations.

Natural Languages Generation on its side, handles mirror problem of NLG, which is producing understandably and effectively output sentences or internal content [1]. Like in the previous point, the output could be given in different formats. So NLG, after NLU part if present, effective chooses the content of the response and bounce it on the instrument which furnishes the output, as textual GUI or a voice synthesizer. In some case, the output not only has the task to answer questions merely but vehicle the conversation through a goal.

Both of these areas have to consider also psychological and environmental aspects, as locations, timing, intent and sentiments. Actually, all could match psychology aspects could be a strong ally for NLP. For these purposes, NLP often works with emerging fields that deal with these type of tasks as sentimental analysis [2], personal profiling or user satisfaction. These aspects are part of a notable problem that affects NLP and all mechanisms that deal with understanding: the context.

Can we take advantage of natural languages processing in our everyday lives? Absolutely. It is common to exploit the potential of NLP skills, to give this capacity to something called agent. The definition of a generic agent, born in separate term from NLP, is dear to computer science but still blurred. Say what precisely an agent is not a simple task, as we can see in proposed taxonomy by Franklin and Graesser [3]. From little to big thing, it depends on the specific purpose for which agent is designed. We prefer to do not report well-know definitions derived from old papers, because they lay on strict constraint on the agent's goal, only user-defined, notation from which we want to escape. Instead, we take into account The Maes Agent [4] and The Wooldridge-Jennings Agent [5], that propose concepts that concern best our studies, focusing on a particular slice of agents, Autonomous Intelligent Agent.

In the Maes Autonomous Agent, a basic on-the-spot definition is given: an agent is computational systems that percepts the complex environment that surrounds it, performs interactions autonomously and realise a set of

planned goals. Autonomy and interaction with complex environment are properties remarked and detailed in Wooldridge-Jennings agent as well. It is highlighted that an agent has to operate even without the direct intervention of humans. Indeed the autonomously actions have to be taking in "pro-activeness" fashion. Agents do not simply act in response to their environment or human request: they taking in account internal states and are able to exhibit self-goal-directed behaviour by taking the initiative. In perceiving the environment, different types of input are considered: expect the most know (physical world, GUI, other agents and human) a modern instrument is included, the internet. Internet, as will be noticed in the following chapters, open up new avenues to great things, as real-time information, big data input and connection with a lot of tools. Time dimension is another mentioned point, in term of reassembling human memory to real-time flow. The most interesting assert that appears in Wooldridge-Jennings definition is what they called "social ability": agents interact with other agents and humans with a chosen agent-communication language.

Steeps needed to realise an agent like this, are still many. Realise claimed intelligence discussed in the last paragraph, has multiple issues. Despite this, taking into account simple versions of these concepts, we reaching fascinating results. Referring to the last sentences, is quite clear that merge agent and NLP skills are meant to be a natural process. And it is what is called Conversational Agent.

Conversational Agent also know as dialogue system, is a type of agent that is focused on conversation with humans, but other agents are not excluded. They are often referred as chat-bot.

All related choices regarded agent and NLP are reflected in the dialogue system. Case scenario and typologies of interaction are always different, so it is not possible to define a fixed schema of components. But nevertheless, we can provide a general graphical example: we hypothesise a not embodied dialogue system that interacts through a textual GUI with human and environment. A representative schema so is given in Figure 1.1.

Now that we touch important definition, there are some key points in NLP and chat-bot history, that is worth mentioning. Know where we came from could help us to understand actual questions on the argument. Accordingly

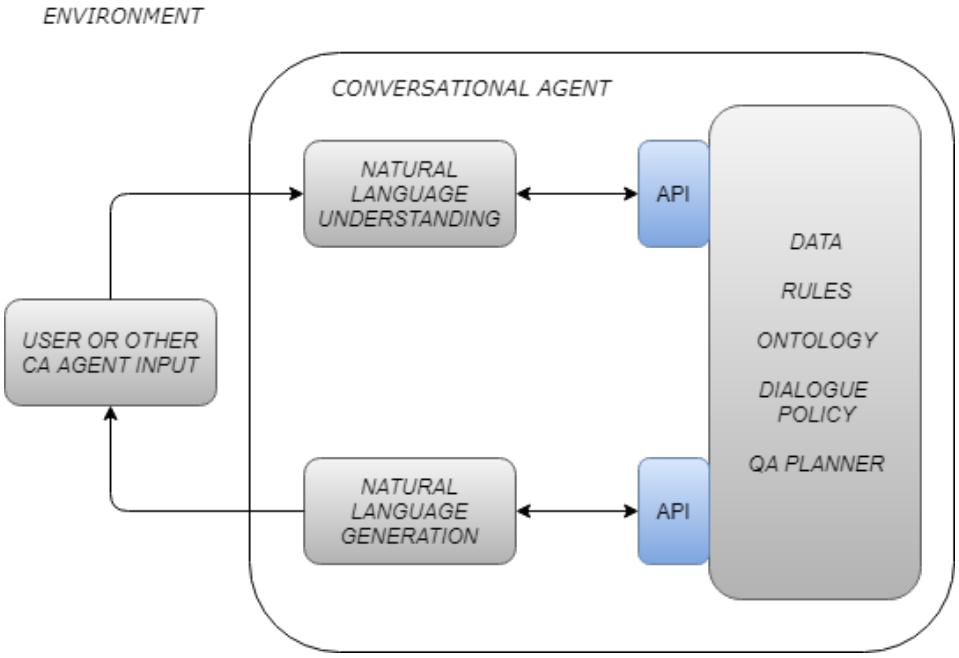

Figure 1.1. Dialogue system architecture

to many, we can identify in the Turing experiment not only the beginning of AI but NLP as well. Turing questioning himself if "Can machines think?" [6] introducing a deception experiment, where man has to guess if it is speaking with a computer or a woman (generally reported nowadays as another human, focused on the concept to recognise another similar). If the man fails the experiment, the computer was considered intelligent. Experiment which other chat-bots as Eliza [7], Parry, A.l.i.c.e. , subjecting themselves in next decades. Eliza and Parry are funny chat-bot that emulates a "doctor" and "schizophrenic patient". Initially developed separately, they have to deceive doctors - so specialised people of the argument - be in front of real doctors or real patients. Researcher enjoyed themselves brought them together in some experiments, talking to each other with catchy results. One of this can be found in this report [8], remarking the not only human-CA communication but also CA-CA communication. They fail the test in a more or less honourable manner, but despite this scientific community

of that time were enthusiastic about result achieved. We have to wait the beginning of this millennium, to pick up enough to notice a new concept of agents, Smartchild[2]. It belongs to instant messaging bots, and it can give quick data access. In their last year of life, it enhanced with the possibilities to customise conversation, also for that the company that, for example, want to sell and give information about specific products. Unfortunately, the technology of that time was too premature and the cost hamstringing the potential. We can consider it as the precursor of actual conversational agents, like Siri, first famous vocal assistant of Apple, running on iPhone devices.

Another mindful lecture is an experiment called Talking Heads, which steps were collected in the homonymous book[9], from Luc Steels. It makes us realise how incredible it can be this field. The object of these studies are agents that learn in a trial and error fashion to name shapes on a white screen with ad doc languages composed by random syllables. The teacher is human, but in the further steps could be other agents. But it is only the beginning. On this basis, with the most straightforward algorithm of optimisation, this population of the agent construct their languages, converge and agree on words, develop bilingualism, built skeletons of a grammar. These groups of agents are even capable of establishing the same linguistic dynamics that renewal of community produces. Like in the real world, taking off some old agent and put in new agents in the population, resulting in passing down and affirms new parts of the language. This experiment lasted two decades: this work has much more to say that we prefer and suggest to refer to the whole astonishing text.

1.2.2 Where we are?

So, where we are now? Nowadays, what is the task and level of chat-bot in our society? In the first fifteen year of 2000 we observe the emergence of

[2]https://en.wikipedia.org/wiki/SmarterChild

well know chat-bot such as Siri[3], Cortana[4], Google Now[5] and Alexa[6]. No more confined to just research field, conversational agent help us in daily operation as retrieval of information from the internet, turn on light, route planner, play music, general domestic manner. Companies application use chat-bot as predefined support of specific task, for FAQs or simple tasks for user services. Researchers use them to understand, format, extract and pull huge flow of information, fulfilling knowledge base. The most recent kick to this field was given by machine learning, that finally can exploit all of its concepts due to hardware computational power reached in this years. But not only. Machine learning most of the time is data-driven, so more information we have, we should perform better. Big data, with its enchanted information collection, feed machine learning's needs perfectly as never before.

However, nowadays agent belonging mostly to what we can define assistant agent, in our case conversational assistant agent. Why?

Assistant Agent, that we remind we see in conversational term, is a type of agent that provides support functions to the user. They are not necessarily supposed to take decision and action autonomously.

What it is and what are the limitations of these agents could be evident in two seconds thought. Every day chat-bot show us their weaknesses, shortcomings and is not authorised to do anything without our confirmations. They not provide real intelligence and not define support equals to humans one yet. They are not able to provide correct answers, because they miss and not understand the context commonly. Users do not fully use them, and retrieval of errors becomes harder for developers and analysts. Together with not perfect communication, they fall after a time under not friendly usages [10], left this instrument in contexts that are know well supported, confining lot of functionalities not exploited and tested. These could be improved by domain expertise in precise arguments, more training and data, but something still eludes us, and user experience is affected.

[3]https://www.apple.com/it/siri/

[4]https://www.microsoft.com/it-it/windows/cortana

[5]https://www.google.com/intl/it/landing/now/

[6]https://developer.amazon.com/it/alexa

In fact, sometimes this type of agents are capable of making advice, but how sincerely could be this advice is difficult to say. It is the result of powerful machine learning algorithms, net or some kind of intelligent observations? They could be considered as an upper step of assistant agents, advisor agent[7]. They keep information and context more effectively, collecting personal information and better context recognition, but that's not enough. We can not define them equals to us yet.

Here, at the end of this paragraph we found interesting for the reader showing ones on best Questioning-Answering framework by now, IBM DeepQA, that we can see at Figure 1.2. IBM DeepQA is the base of more popular IBM Watson: as we can see a lot of different component work in conjunction to achieve this process, algorithm, models, data sources, evaluating tools. IBM researcher assert to it as system that evolves from the continuous contribution of many different algorithms. In white paper[8] several challenge problems are discussed, to reinforce the multiple issues in that are to be addressed in this context.

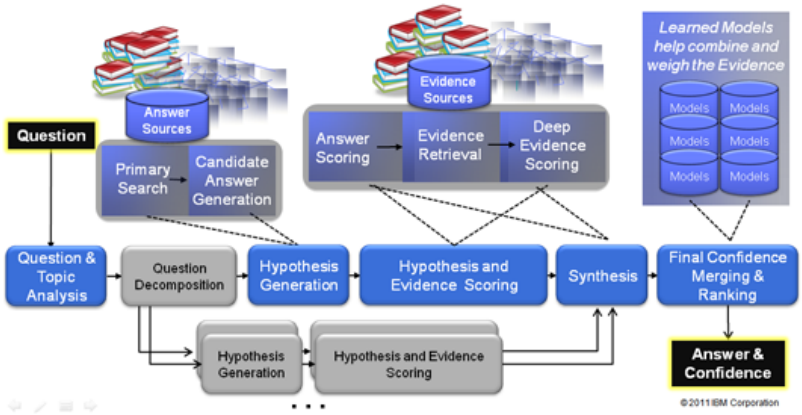

Figure 1.2. IBM DeepQA framework for IBM Watson

[7]https://www.slideshare.net/giusepperizzo

[8]https://researcher.watson.ibm.com/researcher/files/us-mike.barborak/rc24789.pdf

1.2.3 Self-aware agents

In last year we come a long way, and in collaboration with machine learning, we reach new heights, processing huge amount of data previously unimaginable. So, the question is: why conversational agents are still inaccurate? Why they are so identifiable and act with unnatural behaviour? How can we convince users that they talk to their equals, considering them truly intelligent?

To answer this question, let us begin with referring to an interesting counterexample of Turing idea of artificial intelligence, Chinese room argument [11]. This lecture was debated and critiqued, but above these opinions, it provides some interesting food for thought. In a nutshell, the author says that a machine can be considered intelligent only if *literally understand the content that manages*. In particular, in his argument, he supposes a box that takes in input Chinese characters and output Chinese characters. In order to pass the Turing test, this box has to convince people to known this language as a Chinese-speaking person. But now, think about to have a Chinese-English dictionary inside the box and processing input and output with a program that exploits it. Passing the Turing test, take into account Chinese symbol in such way, could be described as intelligence? In reality, the box doesn't know Chinese: it only holds a dictionary, but people outside the box are convinced to speak with someone that knows Chinese. So without understanding, intention and explainability, he states that machine doesn't think or own a mind in classical defined way, merely execute more or less complex algorithms.

The author and reviewer recently assert that intentionally as he defines can be aligned to the concept of self-consciousness [12][13]. Self-consciousness is defined as a bottom step of Self-awareness [14], but often are incorrectly used as synonymous. The difference is subtle in words but deep in concept. Self-consciousness is perceived as awareness of one's presence respect to surrounding environment; self-awareness is, like an inception of self-consciousness, understanding that one is aware of one's existence. Along with these concepts, creativity,improvisation, explainability are said as vehicles of correct self and external analysis, less poor judging of data. It's like to have an empty box, without characteristics that stand out from

the rest human being.

Hence the idea of intelligence is merged within the concept of consciousness, and at a higher level with awareness. It's clear that even before technical issues, the problem resides in our personal opinion that touches our strong beliefs. Every human can have a slightly different perception and convictions about intelligence. What one person can assume as intelligence, for another can not. For example, if we know nothing about an argument, people specialised in those fields can seem to us more clever respect ones that don not know anything about it. However, the last ones could be smart in an area never touch in previous discussions, or in way that we can not understand. Animals intelligence and cognition is another case of how opinion can be actually different. Now, let's imagine about how could be difficult award this capacity to an inanimate object. We are ready to accept the idea of an intelligent agent that can have it own true point of view, like us? This it strongly depends on the vision and consideration that we have of our humanity. Is it a question without an answer? Having said that, we can not just stand back and do nothing: we need choosing and finding next steps in artificial intelligence. What are the fundamental bricks, the elementary mechanism that triggers intelligence, how we can represent them, manage and agree on?

Therefore in the end, the problem practically did not solve. More question arises, but perhaps we detected new skills that may help us with agent capacity, like self-awareness and explainability. One of the biggest problem of conversational agents by now, and subsequently of agents in general, is that we probably have to escape the concept of convincing, deceive our self to something. We do not concentrate on programming them to elude, but we have to channel our effort to pick the essence of intelligence. We generalise more and better, but cases where we notice a genuine and spontaneous intelligence most of the time attributable to Eliza effect, that is "a subtle cognitive dissonance between the user's awareness of programming limitations and their behaviour towards the output of the program".

May we ask if it is only a research digression end in itself, without valid applications in real life. It's not true, and we could be sure noticing how

much companies invest in researching in these fields or how people are entertaining them self with dialogue systems. A well-designed, improved conversational agent could transform itself into considerable economic revenue. Halving the cost of the work - we are aware, debate if is a good point is always a dangerous line - and the time used to manage a significant number of users, also with complex problems, are only two of top practical aspects. A representative case is delivering fulfilling customer service, a daily practice that already looks at this untapped potential. Another example is the world of recommendation system: personalise communication selecting and proposing the right contents, is a method to avoid annoying people providing not trivial advice raising extra revenues. How would we appreciate suggestion from an agent that not seeming an invasive watcher, better yet trough reassuring interfaces? Many other examples could be given. Manage emergencies when a true intelligence is needed immediately to escape from schemes to save lives, recognise ill, symptoms. Elaborate massive quantity data and notice interesting thing where humans can not handle the flow, with the same critical eye or specialised knowledge (daily data is in the order of exabytes just considering internet). In education, to reduce costs in poor areas or improving instruments to satisfy curiosity.

Or simply need someone else to talk, facing with someone else idea, or loneliness.

And as in all fields, if these concepts are converted in products, who better performs, better reap the benefits in the financial world. Here, we wouldn not belittle philanthropic discussion, but only highlight the possible coexistence of commercial opportunity and enhancement of modern society.

Chapter 2

State of the art

Technical and psychology subjects touched in this thesis give boundless themes upon which converse. To be clear and get the point, we opted for selecting only salient aspects and expose state of the art directly. However, we will continuously provide the reader with a logical thread. In the first section, we talk about knowledge bases, data which an agent (and not only) relies on. In the second section, we report NLP techniques used in our implementation. In the last one, we leave the word to psychological topics, focusing on previously cited concepts, explain how as creativity, improvisation and explainability could be tools to reach self-awareness, and as a consequence true intelligence.

2.1 Knowledge Base

2.1.1 What is a Knowledge Base?

For the AI scientific community it was immediately obvious that store and access data in large quantities in a smarter way is needed to improve algorithm efficiency and propose new strategies: the infrastructure that accomplish this task was soon named knowledge base (KB). A knowledge base is a tool that takes care of storing information that will then potentially be used by a system, for various purposes. The information stored by a knowledge base, however, is not to be imagined as an infinite sequence of records like a normal database. The KB data must store information that is able to provide us with concepts, rules, facts about the world at multiple

levels of abstraction.

This clarification is important, as the use of the knowledge base has a very specific purpose: it presents itself as content and API, more or less structured, which allows the systems to make inferences and reasoning about the real world. In fact the knowledge base is born by the mind of artificial intelligence researcher in the eighties as a support to expert systems, that aim to reasoning in a certain specific field. Together, the knowledge base and inference algorithm form a knowledge-based system. A system of this kind is capable to use some type of logic to solve complex problems, derive new knowledge but also detect discrepancies.

In any case it is important to assert that the model we choose to build our KB influences our reasoning ability and vice versa. This statement derives directly from psychology: in fact, humans and their ways of solving problems derive in large part from the tools and formalism with which they store information. Reasoning in these terms can help us make complex systems easier to design and build.

One of the most used model, formalism, which could build upon a knowledge base is called "ontology": an ontology represent and group data with classes, sub classes and instances, highlight properties and relations between data. Clearly this hierarchic composition gives not only raw data but also helps the system to abstract concepts. Anyway there is not the only approach, indeed this model can be substitute or combined with others, such include conceptual graphs, logical assertions, semantic nets, frames, rules.

Knowledge representation so is mean to be at the service of reasoning, allowing new knowledge to be inferred. Most algorithms use a simple if-then paradigm but other complex approaches include the use of provers, logic programming, blackboard systems, rule-based language, inference engines.

Even using the best knowledge base we have to consider that it is affected by some problems, especially when talking about those that are set up to be the basis for more free and less structured engines. Generally, even the largest one is possibly inconsistent and incomplete. When trust and perceive KB as omniscient, we must remember that it derives from the knowledge that we insert as humans, which is by definition intrinsically inconsistent and incomplete. Subsequently, we need to manage these issue. In addition, we have to consider problems related to hardware and resource management and optimisation.

2.1.2 Conversational Agent and Knowledge Base

Given how we have defined the knowledge based systems, it is not difficult to imagine the conversational agents as belonging to these one, adapting reasoning part and knowledge base part to the language field. The reasoning part could be composed not only by simply inference algorithm, but also by natural languages processing algorithms, able to understand and create new data and conversations on the disposed knowledge base. The KB, on its side, should follow some shrewdness. In order to make a intelligence agent that can converse with humans using natural language and can elaborate answer and questions about the world, it is essential to represent knowledge in some way oriented and structurally predisposed to language. In fact the structure of KB itself will influence the agent's methods and ability to learn, as well as the quality of the intelligence that could be developed.

Considering the history of the conversational agents and the brilliant results obtained even before the advent of big data, machine learning and deep learning, is underlined how the knowledge base alone can be a fundamental instrument for reasoning. In fact, since the 1980s, the meeting of the cognitive revolution in psychology of and of AI has led to the development of expert systems and frame-based languages to solve complex tasks such as having a dialog in a natural language. Nowadays, the study of formalism for data supporting conversational agents is a hot topic. Not without reason, the companies mentioned in the introduction, that propose some of the best chat-bots on the market, are the same ones that are in possession of a huge source of data, which over the years were analysed and structured according to the most powerful and own developed formalities by highly specialised figures of the linguistic field.

The input data must therefore be formatted in the structure most suited to the task and inserted in the KB devoted according to the chosen formalism. Among all the possibilities, what can be a good formalism for the conversational agent knowledge base?

2.1.3 A formalism for Knowledge Base: Semantic network

One formalism to represent knowledge base is called semantic network. The reason why we decided to propose it among many become evident from the name. Semantics is the linguistic and philosophical study of meaning in

language, formal logic and semiotics: in particular it is focused relationship between symbols, like words, and concept we relate to them. For conversational agents, this is one of the best formalism to represent knowledge to achieve linguistic and reasoning ability.

Semantic network so represent semantic relationship between concepts in a network: this representation is made in practice with a directed or undirected graph, and for this reason are referred also with the name of knowledge graph. Typically this graph, if not showed graphically, is expressed in a standard manner by means of semantic triples. For clarity we can think of the association node-edge-node as more familiar one, subject-property-object. This type of formalism binds well to another, ontologies. Edges and nodes representing abstract concept could organised into a taxonomic hierarchy, represent inheritance and so on. Combining each other make a powerful data representation to conversation agent.

A representation of a semantic network can be found in Figure 2.1 (Wikidata implementation is showed). As we can see, it is represent by a graph composed of many triples connected to each other. One example of triple in this graph is pharmaceutical drug - instance of - chemical compound. In this particular case item are concrete object, but - instance of - expresses a condition of hierarchy that can be traced to the structures of an ontology, reinforcing concept previously stated.

2.1.4 Internet as Knowledge Base: Semantic web and RDF format

We choose from the state of the art semantic network formalism. By now, however, is how to have a beautiful car without oil. This structure have to be filled with data: data retrieval, especially for specific fields of knowledge, is the first and problematic issue of knowledge bases. Collecting a lot of data can require massive surveys or data collection from commercial products, often provided for a fee. Open source data sets exists, but may not fit our needs, much like ones those sold. Other potentially available data is protected by privacy. Even for research purposes it is difficult to access to this type of data.

Also potentially resolved the problem of the data retrieval there is a second phase not less important that consists in adapting the data to the structure of the chosen knowledge. Speaking of text expressed in natural

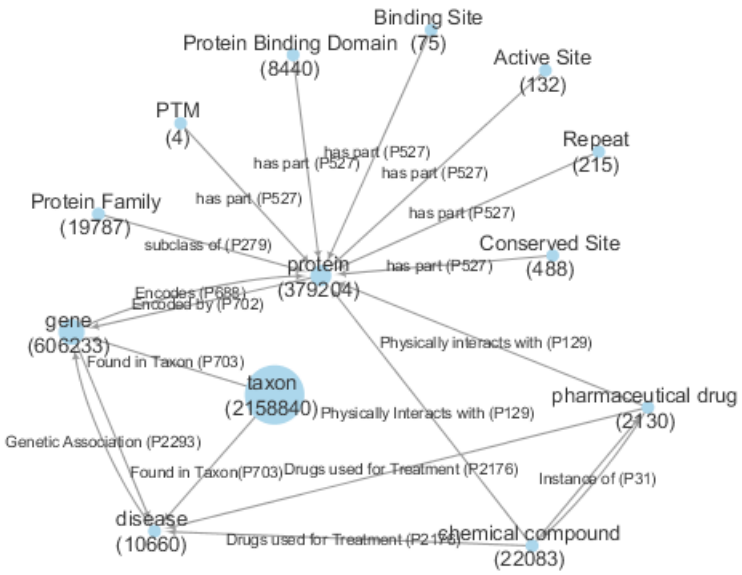

Figure 2.1. Example of semantic network

language and semantic net, it would be necessary for example to break the whole document into triples, taking care to maintain its original meaning.

Fortunately, in the last few decades a powerful instrument has proved to be useful ally, as well as necessitating the possible advantages, of the knowledge-base system, encouraging their development: the internet.

The internet could be consider one of existing more extensive collection of data: every day all type of information are poured in it, notice, images, research content. The problem is that most of time it have to deal with not structured data, composed by a heterogeneous and evolving contents that cannot never fit to a precise data model, but need some kind of extreme flexibility to organise and connect themselves. Knowledge-based systems like semantic net is ideal to store and do reason on data for this system, because allow to take in account different abstraction and connection of concepts of various nature. When this type of knowledge-based systems are applied to Internet we face what is named Semantic Web, sometimes also called Giant Global Graph.

The Semantic Web in few work aim to produce a layer of semantics

(meaning) on top of the current Internet, creates large ontology of concepts, provide a constantly dynamic network of knowledge. Internet is an open world and in order to achieve this result it is necessary to propose a set of standard framework, representations and conventions to which everyone can refer to work in a single direction. In figure 2.2 we can see proposed W3C stack that enable semantic web[1].

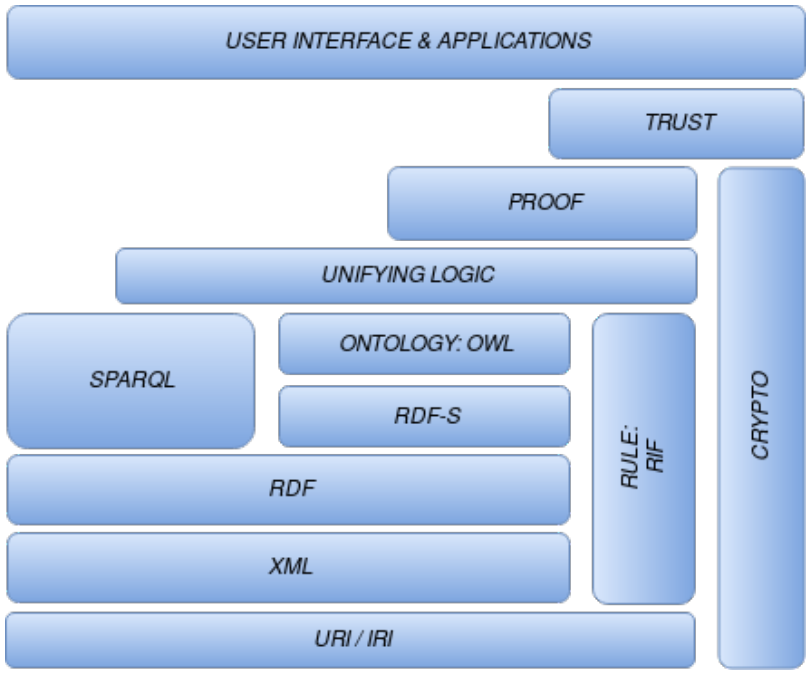

Figure 2.2. Semantic web stack

The various resources on the internet can be accessed by with an unique address called Uniform Resource Identifier (URI). Resource Description Framework (RDF) is a framework for creating statements in a form of triples, using as their component resources identified by URI. And example[2] is provide above (2.1.4). RDF work together with RDF Schema (RDFS), a

[1] https://www.w3.org/Consortium/techstack-desc.html

[2] https://www.w3schools.com/xml/xml_rdf.asp

basic vocabulary for RDF that is capable of creating hierarchies of classes and properties, giving the another level of abstraction. The syntax used to communicate among these part is defined by XML markup language, that enables creation of documents composed of structured data. Web Ontology Language (OWL) extends RDFS providing additional constructs to describe high lever of semantics of RDF statements, like constrains. Upper layer is compose by different engine and logic that is in charge to really work on these data, including validating phase when the result output from the stack are generated statements from scratch. Final layer is generally a user interface, that enable usage by humans of semantic web and application that exploit it.

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:si="https://www.w3schools.com/rdf/">

<rdf:Description rdf:about="https://www.w3schools.com">
  <si:title>W3Schools</si:title>
  <si:author>Jan Egil Refsnes</si:author>
</rdf:Description>

</rdf:RDF>
```

2.2 Word embeddings for Natural Language Processing

> *Such considerations led many people to believe that the ability to communicate freely using some form of natural language is an essential attribute of an intelligent entity.* - Fischler [15]

We have remembered what is a knowledge base and refer to it as first fundamental part of conversational agent. The second part is creating mechanisms to work on this data. We will not provide a deeply technical description of particular inference or reasoning mechanism, because this work is more focused on evaluating generated content. Sure, this metric could provide control strategies to an content generation algorithm. Despite this we consider it appropriate to introduce some basic state of the art concept

useful to our work related to the world of natural languages processing, as word embeddings.

In order to perform some metrics we need to mapping linguistic features to something that we can evaluate objectively an in an effective way: number, vector, matrix.

Word embeddings are one of the most used vector representation of words, that take in account semantics and syntactic information. Better to say, word embedding encloses a set of language model and feature learning methods in NLP where words or sentences from the vocabulary are transformed in vectors of numbers.

Word embeddings come after "hot encoding": the model of hot encoding consists of a dictionary of words. Every word is represented by a zero vector with only one at the index where it appear in the dictionary. This is not very efficient methods because incur in the curse of dimensionality and at the end of the day it says anything about relationship among words. Word embeddings goes partially beyond this problem using distributional semantics [16]. Distributional semantic highlights from semantic theory that words that appear in the same context could be considered similar and interchangeable, because tend to vehicle similar meanings. After this assumption, that obviously is not always true but not enough for us, each word is mapped to a dense vector with a fixed dimension, optimised respect to semantic distribution of the case: these optimisations reduce size of input arrays and avoid curse of dimensionality. These dense vectors allow us to visualise word in the space and do mathematical operation: because semantic and syntactic similarities of words are considered we can found analogies and similarity of words.

The next question is how these vectors are compute. These vectors are not always provided with the model by someone, but they have to be calculated. There are different approaches but surely the most used is the one based on the use of neural networks. These neural network take as input a text corpus, that is a wide structured set of texts. If we have enough data we can train our neural network that give us as output the embeddings. Usually train a neural network it take a long time and often we don not have a large corpus: if this is the case we could decide to use pre-trained embeddings, but we have to be sure that these modelled embeddings are fine for our particular application and data input.

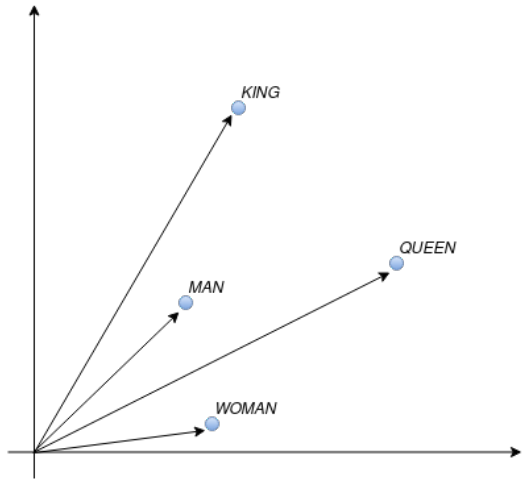

Figure 2.3. Example of word embeddings

2.2.1 Word2vec

Word2vec [17] is one of most known and faster possible methods to produce word embeddings. It consists of a simple two-layer artificial neural network that can support two model architectures: continuous bag-of-words CBOW and continuous skip-grams. The first model predicts the word considering the words that are close to it in the input vector, without considering the order, making the assumption that this does not influence the prediction. This is the faster method. The second model the current word is used to predict the set of possible other close words in the input vector. This model is slower but perform better when encounters infrequent words.

Extension of word2vec strategies from word to sentences, document[18] and graph are proposed and analysed under the name of sen2vec, doc2vec and graph2vec. In fact most of case we need to measure relationships between sentences and documents and not just between single word.

2.2.2 Cosine similarity for semantic similarity

Since we have transformed words into vectors, we can exploit all the related theory. To understand the semantic distance between two words, we could

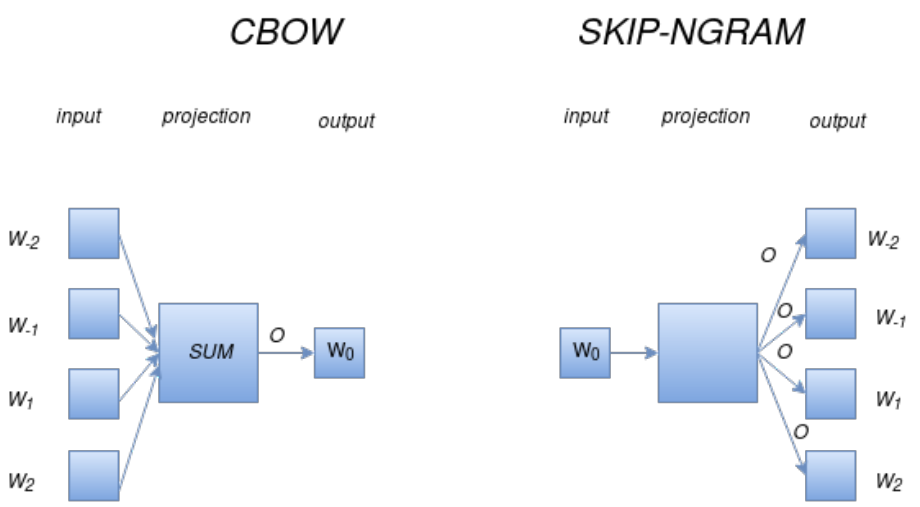

Figure 2.4. CBOW and Skip-gram

consider the distance between two vectors. There is not only one way to measure distance [19], but some are more used than others.

One of the simplest is cosine similarity. Cosine similarity measures the orientation of two n-dimensional sample vectors irrespective to their magnitude. It is calculated by the dot product of two numeric vectors, and it is normalised by the product of the vector lengths, so that output values close to one indicate high similarity. In our case semantic similarity between words. Cosine similarity is not a perfect metrics and its shortcomings are well treated [20], but despite this it is widely used over word embeddings.

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} \tag{2.1}$$

2.3 Explainability and creativity for self-awareness

"To succeed, planning alone is insufficient. One must improvise as well." - Isaac Asimov, Foundation

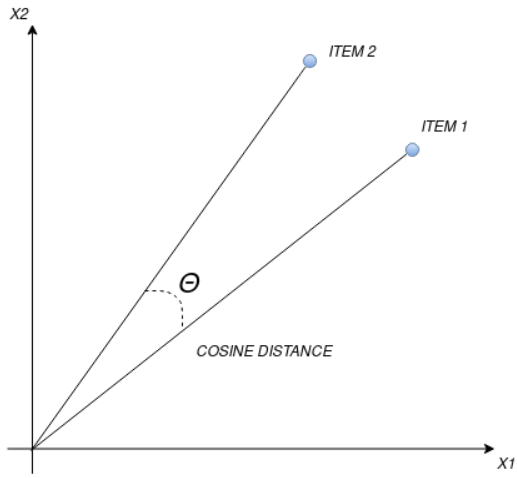

Figure 2.5. Cosine similarity

2.3.1 Explainability

When we have knowledge base and instrument to work on it, a third step is essential: measure and validate the result of this knowledge base system. Unfortunately the state of the art is poor in metrics that measure the aspects we investigate.

In our introduction we talked about self awareness and how this is a necessary step to enable an agent towards true autonomy and intelligence. But what are the tools that allow us to reach it? A solid point is certainly the one we introduce under the name of explainability, that is the ability to explain what is happening and why, a upper phase of understanding. A second point regard the capacity to use this understanding to generate content in non-trivial way. We can assert this to not only the capacity to replicate instilled algorithms, but create and improvise new concept and strategies, also in a restricted time. This process where something new and somehow valuable is formed are known under creativity and improvisation.

Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms: it more than interpretability, that is only the extent to which can be simply predicted, given different inputs. How we can see in next sections, we choose

to investigate as internal mechanics the concept of creativity. For us what we can explain is what we can create with little brick of knowledge base, whether it has already been seen previously or is totally new.

2.3.2 Creativity and improvisation

Concept of creativity and improvisation is widely studied in psychology, exploited in different situation as potential turning point in resolution of lots of situation on different scales. Most technical sources that talk about improvisation and creativity, refer to it in terms of theatrical or cinematographic improvisation, its effect in disaster or emergence situations [21] or in particular case of its application in business [22], politic, corporate scenario. There are very few papers that talk in general, with remarkable experiments that highlight in clear way final data consideration and outcomes, proposing innovative metrics. As also previously stated, is becoming more and more evident that we need it to reason on them to reach our raw idea of self awareness and intelligence in human way defined.

As a curse, once again almost nothing is ready to help us define these concepts on technical point of view - without forgetting that only in recent years this subject has been treated heavily in a technical manner. Stressing that it is always difficult to come into contact and learn about all the works present in the academic-scientific world, therefore it is possible that we have missed interesting readings.

From a general point of view, however, there seem to be some points in which it is agreed, potential bridges from psychological and technical resolution.

First of all, it clear that by now we are not interested in fall in not smart, destructive agent behaviour, because we have to thinking over research goals where bad behaviour has to taking in account and observed, about a product that have to perform correctly on the market. Therefore it is necessary to define and work for a "positive creativity". Furthermore, although there are cases of "destructive" improvisation and creativity in the human world, they are recognised as "good" if "appropriate, effective, novel, unexpected, useful" [23].

In accordance with what, all goals defined as creativity in this work have to be considered oriented in an enrichment of data owned, with good quality,

creating something new that can help the agent to survive and understand, improve and explain the data world around avoid unstable situation for a long time. Isolated episode of negative shall be accepted, because sometimes the data on which we base ourselves are wrong, and we need to avoid our beliefs or our modus operandi to get out of situations of stalemate or false correctness.

By now we talk about creativity and improvisation making no distinction. This practice could be considered acceptable? Really, no: so, what is the difference between them?

Improvisation, with respect to creativity, has a strong and specific temporal component, collapses the moment of thought and that of action in a single one [24]. A long paint session could recognise as creative process but it obviously stray from the 'improvise' act of a painter to cut the canvas or paint in a brief time something new. Now here we would have to deal with what "action" means for an conversational agent (like ours) that has no physical characteristics but only software, not embodied. By now, we can figure out as comedian improvisation, that in the same time think and tell a joke without script, for example.

Always because of this intrinsic limited time window, improvisation recognises an already present problem of the creative process. The objective to be reached as we have talked about it until now, is presented as well defined, but so obviously it is not. In fact, at the very moment of the improvisation, the objective may not yet be totally clear or understood in such a short time, as well as the actions that are to be implemented. This leads us to remember that even more than in the creative process, less rigid execution, a flexecution [25], may be implemented during an improvisation process, choice more similar to the human behaviour. Flexible execution is not intended as adapting strategies to reach goals but changing the goals themselves based on discoveries made during process. This type of reasoning are very close to concept to divergent thinking or counterfactual, "what if", where more path for a solution are taking in account at the same time or after one is chosen, improving evaluation of future situations or respond promptly when new information are available. But above all, is improvisation inexorably random?

2.3.3 Combinatorial creativity: a proposal for computational creativity

Different types of creativity have come, depending on the modalities and in the fields in which they occur, such as painting, theatre, and this is the analysis of creativity that most of us are used to dealing with. But this is not the kind of analysis we are interested in. What we are interested in is capturing its essence, elementary mechanisms in common at the base of all creative processes, seeing them as tools that enable it in more general way. When creativity is linked to a machine computation, often it is referred as Computational Creativity [26].

As reported in the article *What's old about new ideas* [27] of Thomas Ward, some evidence found shows us that in the act of creating new content, people subjected to various experiments moved in a structured and predictable manner, re-using known concepts, attributes and characteristics in a more or less artistic way. As far as the influence of hierarchical and group dynamics is concerned, for now, let's forget also because we are focusing on one agent at a time.

By following this path we can therefore say that improvisation and creativity depend on the tools available and on how the information is transmitted, memorised, perceived, culture, hierarchies. This means that there are limits also in the creation of something totally new: in our case intrinsic limitations due both to the representation of the data, to the tools (even starting only from how the code itself is written) that we give to our agent to cross them and combine them, but even if we could be perfect in that, the fact would remain that the influences mentioned above could be intrinsically present in the data itself. There is something to express this attitude in a scientific manner?

Umberto Eco, there is more known as novelist and philosopher, studies in deep semiotic and left us an interesting article about an idea of creativity that target this attitude, generally noted as Combinatorial Creativity [28]. In his article Eco, he defines creativity as the ability to combine knowledge in order to achieve a goal. The word "combine","combinatorial" in this article needs special attention, because is almost the pivot of this discourse. The creation of the new is not seen as an actual generation of what

could not have existed before, but only as a discovery and exploration of something that has always been there in nature, but never before discovered. Creativity is seen as a tool that proceeds in trial and error, for the exploration of what was not yet evident to us, of what has always been potentially existing. Following Pascal's suggestions, he defines creativity as an art, a combinatorial arrangement.

> [...] So we should say that what we consider absolutely creative is what is quite new, but we could have had something else, and we simply don't know [...]

However, he asserts that not all possible combinations are to be considered of good value but among these it is necessary to know how to select only the best possible ones. This ability is not entirely recognised as casual or innate, but rather it is a skill that must be trained over time, more precisely, by quoting (translated from Italian),

> [...] Finding a creative solution is a reward for the great "combinatorial" work done previously.[...]

We can see how the presence of previous internal and external work on data is admitted, of attempts, errors and choices within a wider range of possibilities, and that therefore creativity is not just pure randomness, but a little of both. A training in mixing, combining two or more existing things for a purpose, improving through the time. Everything is already present, elementary block, we have only to see them .

From the point of view of the concept of positivity of a creative process, if it wants to be considered scientifically relevant and consistent, in his speech Eco continues saying that creativity must be submitted to pass tests of experimental falsification. This point allows us to say therefore that positivity of something novel create with this process is not a concept that cannot be in any way objectively evaluated, but it is also possible to measure it and compare it with scientific instruments, at least in part.

But even more interest has developed in this reading when it comes to discussing therefore whether the creative process is something that only awaits a human mind. Defining creativity as a combinatorial process, it comes at least in small lines away from an idea of humanity, and therefore

possibly viable by any device capable of performing combinatorial processes, in any possible data world that we want consider: so not only our real world with our rules, but in any other possible world that we want consider. We have to raise ourselves from the conviction that the combinatorial creativity process can be implemented only by human minds: the brain is understood as a possible device among many to accomplish this task.

2.3.4 AI Psychometric Related work and look to the future

In the previous paragraph we were able to assert how the possibility of evaluating the goodness of a creative process is understood in the scientific world. Despite this, it is difficult to find clear and shared metrics in the literature, able to give us numerical values that allow us to analyse these behaviours quickly, objectively and on a large scale. In particular, in the chat-bot world the existing metrics rely on a human validation or usability, therefore subjective, of the external output of the conversational agent. Speaking of self-awareness we can instead say that we are looking for a metric for the evaluation of the internal elaborating process, that is subjective to the agent itself.

Examples of metrics for existing chat-bots, such as those selling a commercial product or providing a customer service, are In Messages, Miss Messages, Retention Rate, Goal Completion Rate, Fall Back Rate (FBR), User Satisfaction. These metrics, described implict by their name, point to an coldly evaluation of the interaction and only indirectly to what has been processed internally.

Analysing more scientific articles we came into contact with these review [29] [30] which in points selected recalls some of the aspects we want to analyse, enclosed under quality and quantity attributes.

Under the category of performance evaluation of a chat-bot, aspects as graceful degradation, robustness to unexpected input, avoiding inappropriate utterances and be able to perform damage control are taken into consideration. From the functional point of view, capacity of engage in on-the-fly problem solving responding to specific questions maintaining themed discussion is one of main goal. Humanity and affect attitudes are observed

thorough chat-bot identity, or with the capacity of increase realism, convincing, satisfying, convey natural interaction, convey personality, provide emotional information, inflection, expressivity, authenticity. Enable participant to enjoy the interaction, read and respond to moods of human participant, detect meaning or intent.

Furthermore, interesting past experiments are reported. Two of these could be of our interest.

Goh [31] measures effectiveness of question answering through precision, recall, and F1. For non-problematic situations, as well placed questioning-answering problems, these metrics turned out to be promising, but as soon as they left the comfort zone they quickly failed. The author point out that we have to work on metrics remembering that utility of responses is subjective, and topical domains have different and dynamic knowledge repositories.

Meira e Canuto [32] focusing on embodied emotional agents, in particular they investigated quality metrics per affective characteristics. Authors propose a measurement framework for goals of three different levels, namely conceptual level goals, operational level goals, and quantitative level goals. Together they examine quality of architecture and affective quality. Aspects measured among others are cohesion, coupling, size, cooperation, likeability, enjoyment, trust,naturalness, reduction of frustration, believability,and interestingness.

A more recent review [33], proposes an even more comprehensive and structured table of all attributes of chat-bot that have been evaluated in the state of the art, highlighting also the need to value the cost and the growing economic impact. It is also reported the example of PARAdigm for DIalogue System Evaluation (PARADISE), a general framework for evaluating spoken dialogue agents [34]. PARADISE estimates subjective factors by collecting user ratings through the questionnaires. Notably, it is and interesting framework that try to unify some concept in chat-bot evaluation, but as we said, also in review is remarked the necessity to unify metrics that can objectively, not subjective by human evaluation, investigate and compare chat-bot performance.

Another famous metrics in this field is BLEU [35], proposed by IBM researcher, that evaluates the quality of the text produced by a translation from one language to another by a machine. A text has a good quality if the text produced is similar to that which would produce a translation

performed in a professional manner, so comparing using a frequentist approach with a good quality reference translations. BLEU is an inexpensive metric and was one of the first to exploit the concept of human judgements of quality. METEOR is an improvement of this metrics.

Despite this, it is clear that these metrics take into account only the effect that data processing and language have on an external and output stage on a user, and do not go in any way to directly evaluate the internal processing of only conversational agent with himself. Although the external output is obviously conveyed primarily by internal behaviour, these methodologies do not allow us to fundamentally analyse the internal mechanisms and therefore be able to guide them towards an improvement also external.

Inevitably, we have once again approached the psychological world, remaining particularly affected and satisfied. In fact, in the 1800s, studies concerned with the theory and technique of psychological measurement has been born inside the world of psychology: psychometrics. Originally started with Charles Spearman for measuring intelligence, by now major focus of psychometrics is on personality testing. Among, this creativity.

Psychometrics addresses human abilities, attitudes, traits also considering internal states and creativity. Similar type of assessments over time have also been applied to non-human world, as animals world, under comparative psychology, or with a continuum between human and animals by evolutionary psychology. Despite this, it is considered appropriate to move on to metrics that can measure the same concepts regardless of the realm being analysed. This more integrated approach, under the name of universal psychometrics, has also been proposed [36]. The evaluation of abilities, traits and learning evolution of machines has been mostly unrelated to the case of humans and non-human animals with specific approaches in the area of artificial intelligence. However with this proposed approach, universal psychometrics brings us back to the concept set out by Eco, which states that a mind to carry out a creative process does not need to be human, but any device.

J. P. Guilford's group, which pioneered the modern psychometric study of creativity, constructed several tests to measure creativity [37]. On the top of these ideas, The Torrance Tests of Creative Thinking (TTCT) [38] is built. TTCT is a test of creativity that involved simple tests of divergent thinking and other problem-solving skills which were scored on four scales:

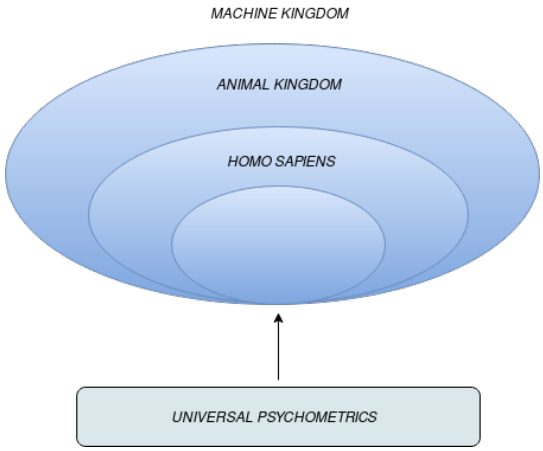

Figure 2.6. Universal Psychometric

fluency, flexibility, originality and elaboration. These are defined as:

- Fluency: the total number of interpretable, meaningful, and relevant ideas generated in response to the stimulus.
- Flexibility: The number of different categories of relevant responses.
- Originality: The statistical rarity of the responses.
- Elaboration: The amount of detail in the responses.

Psychometrics is devoted to systematically measuring psychological properties, usually via tests. Universal psychometrics, propose to extend this metrics to machine kingdom. Bringsjord [39] sums up concept in the name Psychometric AI (PAI): the field devoted to building a computational system able to score well on such tests.

> Psychometric AI is the field devoted to building information-processing entities capable of at least solid performance on all established, validated tests of intelligence and mental ability, a class of tests that includes not just the rather restrictive IQ tests, but also tests of artistic and literary creativity, mechanical ability, and so on.

Then, we are. It is clear that we need some Psychometric AI test and metrics to reach our concept of intelligence through self-awareness as creativity and explainability. In the next chapter, with the aid of other novel papers, we will define our PAI.

Chapter 3

Proposed Metrics

In this chapter, we will present the objectives pursued by this work, clearly and technically. We continue providing details of proposed metrics, that collectively define them under the name of creativity index: each part of this one will be explained, as well as the role it takes. We proposed Multi-objective optimisation, also known as Pareto optimisation, as the instrument to perform analysis of dependencies and compromises among the various index parts. We continue defining what we call improvisation, correlating it to our metrics. We end providing some examples of bias that can affect our metrics and our work.

3.1 Purpose of our studies

Analysing state of the art, we find ourselves in a system without well defined shared techniques for the evaluation of intelligent agents, including conversational ones. The goal of our studies is to investigate mechanisms that enable intelligence and provide a set of metrics to evaluate them. This in-depth study led us to the concept of self-awareness for intelligence, translated by us into explainability through combinatorial creativity. Focusing on conversational agents, we will apply our idea using the linguistic field as a training ground for our experiments.

Our metrics evaluates the goodness content generated by an agent over its linguistic knowledge graph (generating part is not addressed in this work). In the near future, the values returned can be used by an agent

to guide itself in this exploration and generation of knowledge, like in control loop. Not only perform a posteriori evaluation but conveying step by step towards defined positive behaviours.

3.1.1 Basic units of our work: triple and triple document

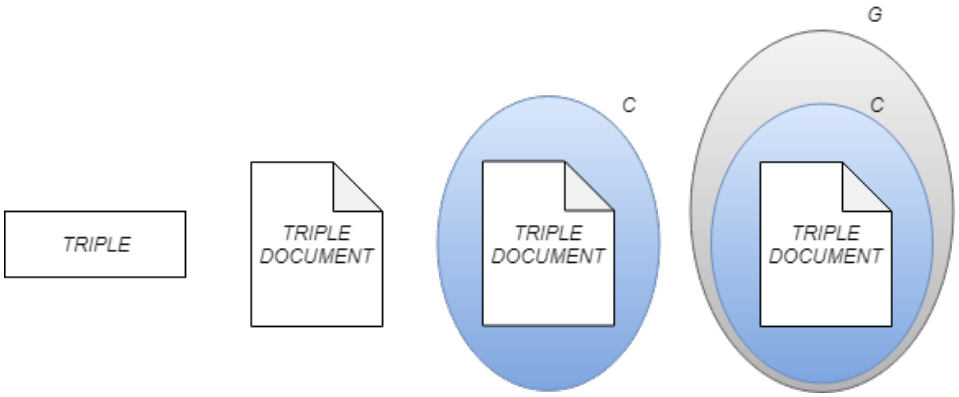

Figure 3.1. Basic units of our work

To avoid ambiguity we report the basic components of our work, which will be used in our algorithms: Agent Related:

- Triple is the atomic unit of this work, and consisting of a item subject, a applied property and item object (sbj,prop,obj).
- A triple document is a ordered group of triples T_i.
- A triple document can be evaluated within the graph triples that are more closer to the document, that we call C, to remind the concept of a context. We will think in the future to augment this C to all graph triple that can be related to the document, beside its proximity.
- A triple document can be evaluated within all graph triples, that we call G, that include C, in order to consider the document respect to everything agent knows.

External Actors Related:

- W represent the reference knowledge graph
- H represent human knowledge graph (human validation)

3.2 Creativity Index

We consider a conversational agent that triggers a creative process, thus generating new triples collected in new triple documents with respect to their mechanisms and the knowledge base available to them. Our metrics, adherence, diversity, novelty, serendipity, magnitude, aim to evaluate the agent output. These metrics are computed independently, but their joint analysis gives us a significant value of creativity index. An important thing to say for is that a single triple could be accepted as a document. We would have to examine this because it is evident that the length of a document influences the evaluation of the same.

$$ci(D) = [ad, coe, div, ser, mag] \tag{3.1}$$

3.2.1 Adherence

With adherence we define the fraction of triples within the content generated that exist in the first instance in the knowledge base of the agent or, secondly, in the reference knowledge base.

Adherence is the part of our index that represents our agent's ability to generate existing content shared and accepted by the actors involved (agent and knowledge base).

$$ade(D) = \frac{1}{n} \sum_{i=0}^{n} ex(T_i) \tag{3.2}$$

$$ex(T_i) = \begin{cases} 1 & \text{if } T_i \in W \\ 0 & \text{if } T_i \notin W \end{cases} \tag{3.3}$$

As we can see, the $ex(T_i)$ function returns a binary value. Future developments could define a weighted existence function, considered that not all triples have the same interest to be shared by all actors.

3.2.2 Diversity

Let us imagine the diversity of a document as the semantic difference between internal triples. This concept is adapted to linguistic field from definition related to recommended system filed[40] [41]. From a certain point of view, this can be seen as the first outline of cohesion. If computed with respect to G, documents that present very high diversity, have a strong possibility of belonging to disconnected graph points (not in the sense of graph theory).

$$div(D) = \frac{1}{\#C(n,2)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} 1 - similarity(T_i, T_j) \tag{3.4}$$

where similarity is a possible measure of distance. In our experiment, we use cosine similarity, treated in state of the art. Computationally, it risks being onerous. To fix this, in the experimental phase, we would reduce the number of triples per document.

3.2.3 Novelty

Concept of novelty, also reported in the previous references, deals with evaluating the novelty brought by the document with respect to what was previously expressed by those present in the knowledge graph, correct or incorrect. The similarity here is not calculated within the document, but between the entire document respect to all the others.

$$nov(D) = \frac{1}{n} \sum_{i=1}^{n} 1 - similarity(D, D_i) \tag{3.5}$$

In future steps we can think to compute this novelty on document bones, the sequence of only property, in order to identify patterns of dialogues.

3.2.4 Serendipity

The definition of serendipity is "always making discoveries, by accidents and sagacity, of things which they were not in quest of" [1]. With serendipity, we evaluate a likely novelty but eliminating most know document and

[1] https://www.oxfordlearnersdictionaries.com/definition/english/serendipity

information, highlighting non-trivial contents and taking in account only good data, using $ade(D)$ as factor of precision.

$S = first\ s\ documents\ in\ a\ ranked\ list\ ordered\ by\ novelty\ nov(D)\ function$

$$ser(D) = ade(D)\frac{\sum_{i=1}^{s} 1 - similarity(D, S_i)}{s} \tag{3.6}$$

3.2.5 Magnitude

With magnitude, we are going to evaluate the weight of triple components based on their frequency in the graph. In particular, we see how rare are the items and properties.

$$mag(D) = \frac{1}{n}\sum_{i=1}^{n}\frac{rank(sbj) + rank(prop) + rank(obj)}{3} \tag{3.7}$$

This part of the index could subsequently be extended with the use of Sentimental Analysis or taking into account if triples of the document decrease the shortest path between two items or are bridges. These concepts will be introduced directly into the implementation.

3.3 Correlation between parts of the index

We have noticed that until now all the listed metrics have been calculated independently. Nevertheless, it is obvious that the five metrics are correlated and that their perception is influenced with one another, at least in part. Creating from scratch a function f that takes in account five dependent variables, however, risks being destructive in thesis work with tight deadlines. Moreover we must think that this computation sometimes must to be done in conversational times: instants or few seconds. It is important to prefer simple and effective functions as far as possible. Many functions can be taken into consideration and the flexibility granted by providing arrays of five values for each computation leaves space for future experiments.

In the following section we will show the actual choice of f in our implementation, which will take into consideration significant mathematical concepts.

3.3.1 Pareto optimality

Binary tables with combinations linked to a label can be useful indications to classifiers, but analysis of this kind can be carried out where the values of the individual metrics are very close to one or zero. This situation is in no way guaranteed and it is expected that more than once intermediate values will appear during the execution. If we add more middle intervals, binary tables are no more useful and even doing this, the blurred concept of positive and negative is intrinsic to the problem.

Which function can therefore be chosen? Sums, multiplications, maybe weighted? How to assign weights? How to allow the user, the programmer or the agent to adjust these weights? What is the best compromise between all the metrics of the index?

Analysing the metrics of our index, the most desirable situation would be to always bring all the metrics close to one. This certainly could be seen as the maximisation of our index. The optimal compromise therefore consists in trying to maximise all the metrics simultaneously.

It is possible? This unfortunately in a real case is difficult to achieve, as it can be quite easy to have to sacrifice one component to increase another. Suffice it to say, that to develop novelty it is necessary most of the time to break away from conventional canons, which can be translated into: it is easy find yourself having to decrease coherence if you want to increase novelty.

This concept does not belong only to this case, but is widely known as Pareto efficiency or Pareto optimality. Pareto optimisation is also known as multi-objective optimisation and from now we will refer to it interchangeably. Multi-objective optimisation is an area of multiple criteria decision making that is concerned with mathematical optimisation problems involving more than one objective function to be optimised simultaneously. Where optimal decisions need to be taken in the presence of trade-offs between two or more conflicting objectives.

The Pareto optimum is a concept born in the socioeconomic sector, born from the mind of Vilfredo Pareto, a Italian engineer, sociologist and economist, but it is also heavily used in the world of finance, process optimisation and other areas. Pareto optimality states that:

> Pareto optimality is a state of allocation of resources from which it is impossible to reallocate so as to make any one individual

> or preference criterion better off without making at least one individual or preference criterion worse off.

In its first definition, as we can see, Pareto does not define in any way the preference for one or the other objective to be maximised. In fact in this light, each goal is treated fairly. In a socioeconomic case, for example, an Pareto optimum, where all the wealth is distributed on a single person, is recognised equivalent of Pareto optimum where wealth is redistributed over the entire population. As defined at the beginning, it only represents the impossibility in a problem of multi-objective optimisation to improve one goal without affecting another.

Over time this defect has been solved by giving the possibility of adding subjective preference information, that will be taken into account during the computation of the Pareto optimum. Therefore, to find a new Pareto optimal, called a new allocation, it is not enough to find another one that is simply equal. It has to be found one that improves one or more objectives without affecting the subjective preference information of the problem or other objectives, always considering the initial dispositions of the problem.

This upgrade is very useful, as it allows us to provide to the algorithm preferences respect the metrics of the index. So, the study of the Pareto optimum and of preference information help us to evaluate the progress of the various metrics of the index, selecting a set of these ones linked to positive attitudes. Recording them and reusing when it necessary could be essential in emergence situation, as an sudden decay of creativity index or quality of the agent knowledge base. Well aware that the same parameters cannot be equally positive for various situations, it is certainly a first method to avoid decaying into bankruptcy behaviour. We may want advantage novelty or combination of serendipity and adherence, or again prevent coherence from falling below a certain threshold. In some way we want advantage a particular behaviour of the agent. Thus defined this seems to orient the agent towards a goal user defined. It might be true, but if we encode this practice directly in the agent we can furnish to it instrument to guide his status imposing preferential information depending on the situation (for example, a banking conversational agent will have to have less freedom than an entertainment conversational agent)

It could be seen that it is possible to have more optimal combinations. This is possible because in problems which various objective functions are in conflict with each other there cannot be a single optimal solution that

maximise everything, as we have said. There may be an infinite number Pareto optimal solutions, that in turn benefit one or the other goal. Among all combination, these set of optimum combinations is called Pareto frontier. A summary of this concept can see at Figure 3.2, with in the case of multi-objective optimisation problem with two objective functions.

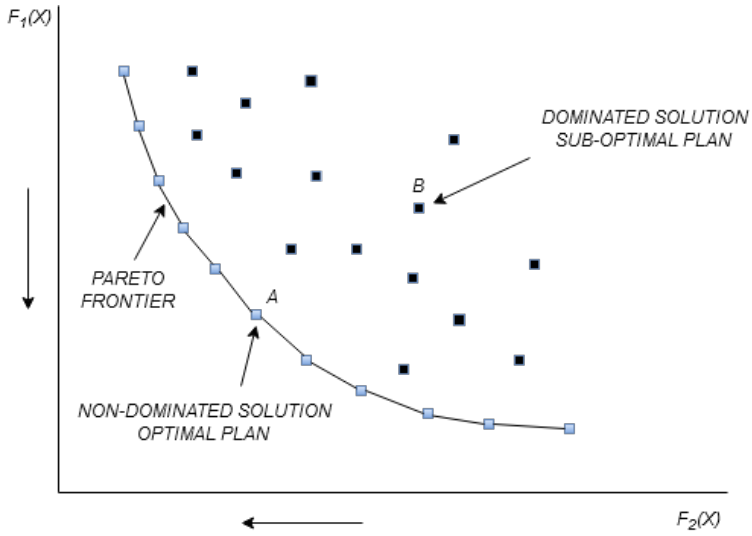

Figure 3.2. Example of Pareto frontier

There is not only one method to calculate Pareto optimum: the various methods differ for the moment and the frequency wherewith they are calculated and types of preference information provided to them and if they are used. No-preference methods are methods that don not take in account any preference method, instead a priori methods ask for a good amount of preference information before the computation of Pareto optimum. A posteriori methods producing all or a set of the Pareto optimal solutions, instead interactive methods produce one solution at time, iterate and improve the search with preference expressed at each iteration.

Finding therefore an Pareto optimality among all the possible combinations of creativity index that our agent computes, consists in selecting configurations that penalise the cohabitation of the various components as

little as possible. Each metrics of the index is seen as an objective function to be maximised.

Pareto solution and frontier are usually shown graphically for better understanding and in some methods to help the programmer or agent to provide new preference information. Up to three dimension we can easy provide a plot or other visualisation, but with high-order multi-objective optimisation problems, visualisation becomes non-trivial, due to the lack of spatial dimensions[42]. These studies require a strong in-depth study and we will don not use it but we believe it is right to notify them. In the dedicated chapter will provide some visualisations taking into consideration two objectives at a time to make discussions on the fly with this partial result.

3.4 Time considerations on creativity index: improvisation

As we have analysed in the state of the art the improvisation can be conceived collapse point between thought and action of new behaviours. In our case we can apply this concept monitoring creativity index over the time. In the agent, improvisation could correspond to a peak of creativity index on a very small time delta. In the mathematical field these points can be traced back to points of discontinuity. A figurative example could be see in Figure 3.3.

3.4.1 Time effects on behaviour perception

The delta t and the creativity index gap for which we can define the actual presence of improvisation must be defined in some way, by the programmer or by the agent, dynamically. In fact, time affects the perception of improvisation and creativity over the time in a continuous cycle of resizing and emphasise over the time of past behaviour. The time dimension is something that we cannot afford to don not take in account and is an evidence that can affects deeply evaluation of agent behaviour and series of other graph maintenance procedures.

The content generative phase of our agent depends on speed of the hardware device on which it is executed and on the complexity of the algorithm. These speeds obviously potentially differ from the speed of thought or the

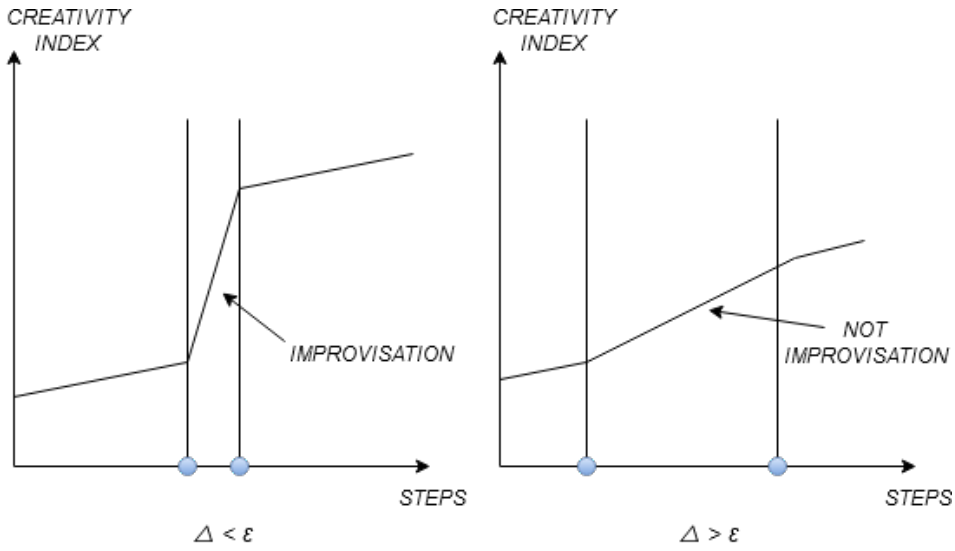

Figure 3.3. Improvisation: peak of creativity index within incremental time steps

speed with which we humans normally absorb new notions. Considering therefore the flow of the proportional time to the number of triples generated by the time zero of the agent, can be a good simplification but it would be extemporaneously away from human timing. So in evaluation and improvisation, what is the best measure for passing of time in our agent?

We do not claim that the agent for now follows exactly the timing rules set by humans, as we do not claim that the language is identical to ours. But if we want to achieve a good user experience and apply some human consideration, we need an internal clock that, together with other external trigger events (for example incoming of new information), can manage in some way this discrepancy and run maintenance procedures accordingly. As graph pruning, data compression but most important the refreshing procedures of creativity index linked to contents. This concept is related to Time Decay and will be better explained in implementation chapter.

Time dimension have to be considered in its concept also for the order of arrival of data, inside the triple documents but also considering agent's knowledge at is whole.

3.5 Bias in our work

This brief digression wants to be a remainder for the reader about aspects can affect the study of a new metric, the collection and reliability of the data, the analysis of the results.

In a simple way, the bias can be expressed as deviation of a subjective evaluation of something with respect to an objective one. There are various types of bias and reasons why this disproportion occurs. The bias can be incorporated into the evaluating structure itself, intrinsic, or absorbed by the circumstances and provided data. The types of bias are divided mainly into two branch, the cognitive and the statistical ones.

Cognitive bias is a systematic deviation from reality due to perceptual inaccurate judgement, representations or distortions. This translates into place the subject affected by bias in a different reality than the existing one. It is like to say that everyone live in a "reality" different to real one at least in part, and as consequence the same happen for our agent. There is a long list of cognitive bias: quoting them all through scientific surveys and papers would be verbose. Nevertheless, at this link[2] we can find an interesting map of cognitive bias and their groupings.

Statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameter being estimated. Among statistical biases we can find statistical bias selection, exclusion, bias of an estimator, educational measurement, reporting bias, recall bias, observer bias, uncertainly analysis, etc. It also know as systematic error.

Look at map given before, immediately clear cognitive biases that could afflict our work and agent itself stand out. To give an example of bias that could affect we as researchers, let's take "Experimenter's or expectation bias": it represent the attitude for a researcher to trust more on positive outcome of an experiment that agree with it conviction and consider less the conflict one of the same experiment. This type of bias it could be linked to a statistical bias called "Observer bias" that arises when the researcher influences the experiment due to cognitive bias without true intention. For our agent, instead, we can observe for example "Illusion of validity" that consists into believing in some digression when available data is consisted

[2] https://medium.com/better-humans/cognitive-bias-cheat-sheet-55a472476b18

and correlated with our knowledge.

Example of statistical bias cited before is in educational measurement, that can affect the metric itself. This bias represent the error in managing and evaluate outcome of a test that result in give lower or higher scores than their true ability of test subject would deserve.

Chapter 4

Dataset

4.1 Knowledge Graph

By choosing combinatorial creativity, our agent must dispose of a set of prior knowledge. In our experiment, we cannot launch our agent on an empty graph knowledge, but we have to feed it with a good number of triples, an initial dataset. Only at this point agent can start with its mechanisms and expand its graph with creative processes and explicit requests. Therefore, it is necessary to choose a data source and implement algorithms designed to request large chunks of information. To achieve this goal, we rely on Wikidata.

4.1.1 Wikidata

Wikidata[1] is a free and open knowledge base, part Wikimedia projects.

Its strength lies not in the simple concept of a knowledge base, but in the reason why it was created and structured. It not only contains a certain amount of information, but it centralises and connects information among Wikimedia projects, third-party sites and additional information. The result is a unique and not repeated bunch of data that overcoming lots of issues, as multilingualism. Furthermore, following the "Wiki" philosophy, it can be used and modified by humans. Even bots and other software

[1] https://www.wikidata.org/wiki/Wikidata:Introduction

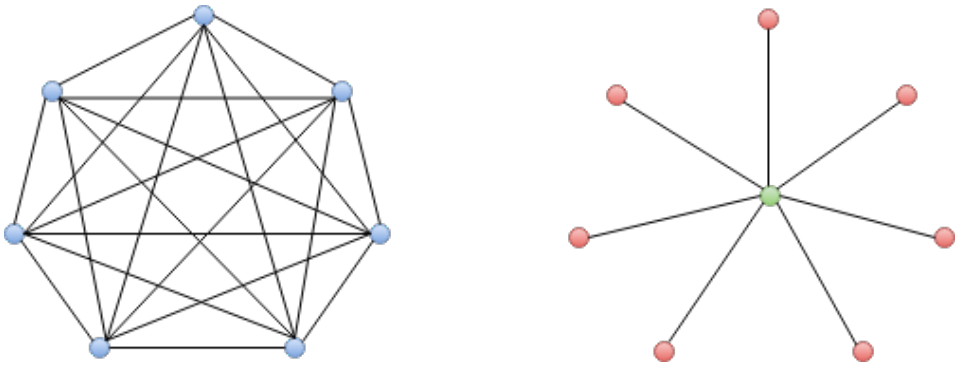

Figure 4.1. Wikidata centralising concepts through different languages

On the left we can imagine same concepts in different languages referring each other, on the right same concepts in different languages refer to one central concept point using Wikidata.

can easily modify content thanks to its structured format. Because of this, Wikidata is not a static source, but in continuous evolution and enrichment, making it harmonious and more similar to the human being.

We now enter into specifics to better understand how Wikidata works. The best way is to start by talking about its structured format. Each Wikidata page represents an item, uniquely identified by an identifier. Whatever in the world is represented in Wikipedia as an item. For each of these items, statements are listed, which describe this item in detail. Each statement consists of an item property with its item value.

A clarifying example of a Wikidata page can be found in Figure 4.2. In this example, the item considered is 'Colossus of Rhodes'[2], uniquely identified among the various languages by the identifier Q41553 within page's URI. At the bottom of the figure, we see an example of a statement. In this case, we have as subject 'Colossus of Rhodes' where the property 'instance of' is valued with 'lost sculpture'. As we can see, both property and

[2]https://www.wikidata.org/wiki/Q41553

value are hyper referring: in this way, Wikidata achieves connection through items, generating a graph of connected items that refer to each other. This last sentence is not always correct: by now, some objects are not yet an item in a narrow sense. Images, time, geographic coordinate, something that is not already edited as an item, are examples. In figure 4.3, we can see an example of a statement where property 'coordinate location' is not completed by an item but by a map. It gives some flexibility to Wikidata, that can express more than only textual based KB. In our work, we didn not take into account this type of information to approach the problem in more uncomplicated steps. Implementing management of this not-text related item in our algorithm could be a great future improvement. We said at the beginning of this section that Wikidata centralised information from various sites: in figure 4.4 and 4.5 we can see how this implemented. Related external source, as external sites links and a list of cross-languages version of the item, are reported. This improves cross-site uniqueness but also provide a direct bridge to more natural languages sources, that could be an interesting aspect for enhancing the NLP part.

The reader may also notice that property and value are treated equivalently at the Wikidata management level. It is true, but slightly difference obliviously exists. About properties, Wikidata holds a dense net of property constrains, that reflect their implication statements and their value. A common type of constraint could be "single value constraint", that means that for a particular property only one value is admitted. We would point out that also these constraints as seen as items in Wikidata, with their properties as well. One might say that it is a valuable push in our conversation agent: we have extra information about what the agent can accept or generate. But not all that glitters is gold: here the intrinsic incompleteness of knowledge act as a powerful brake. As we could not rely even on completeness of the KB, also we could not rely on this net of constrains: what if a constraint is valid of a set of items, but for similar items this constraint is invalid? It for an agent is not obvious when use or ignore these constraints. Concepts of T-Box and A-Box could better explain. In this type of KB, as Wikidata, we can recognise at high level two type of statements: one more generic, called T-box, useful to represent concepts, and another one, called A-box, with more specific statements within an object. Together, T-box and A-box provide a knowledge base. Wikidata, as it is

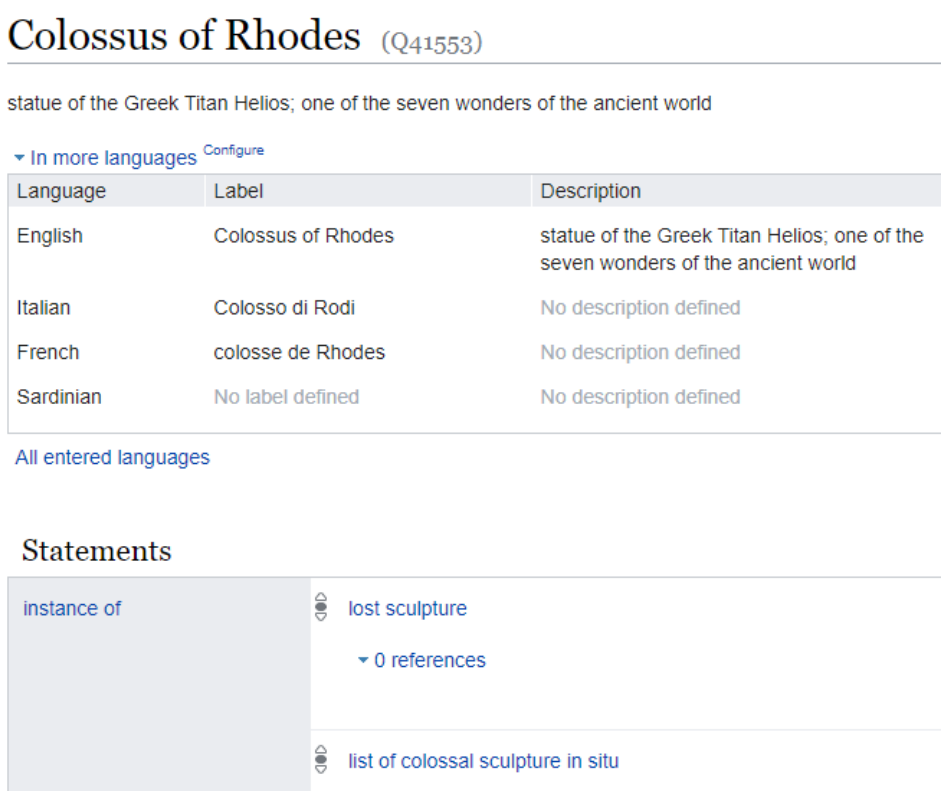

Figure 4.2. Example of Wikidata page

ontological nature, is more oriented in presenting this data through A-box, item and constrain. So we lack so much of T-box concept, so the capacity of generalising on Wikidata is left to grade of ontology and little else, and all depends on agent interpretation capacity. Because of this, we did not take into account constraint on Wikidata; we would try to give agent to understand autonomously "constrain" and "possible thing". Take into account this additional information could be worse than better on creativity behaviour of the agent . As the previous points, nothing can exclude in the future can take into consideration also provided constraints. In the first steps of our implementation, we were thinking about constraint matrices, but we have discarded them because of the ineffective result that for now

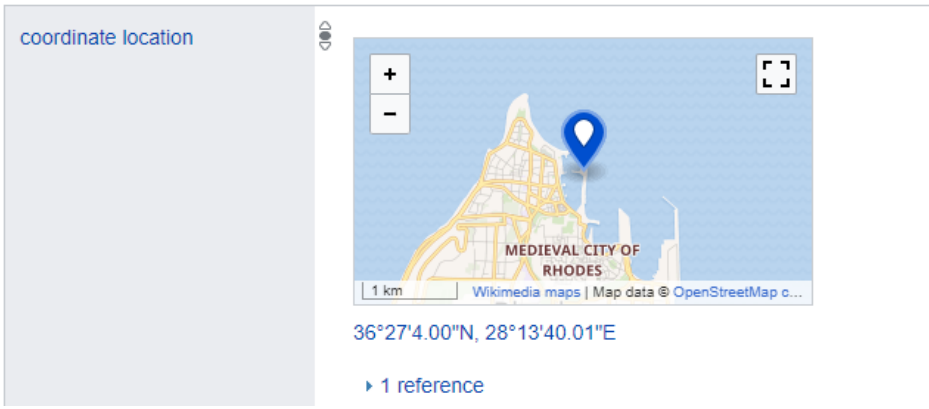

Figure 4.3. Example of Wikidata statements

Figure 4.4. Example of Wikidata identifiers

would have come out.

As the last point of this subsection, we want to point out that Wikidata format is practically identical to RDF format analysed in state of the art: it can be treated with the same techniques. In particular, its serialisation may be considered close to N-Triples style (subject, predicate, object) making Wikipedia perfectly adapted in a graph knowledge base for our purpose.

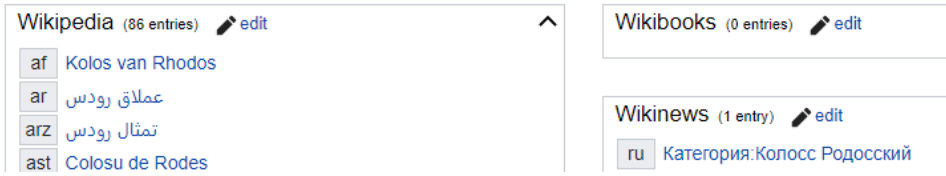

Figure 4.5. Example of Wikidata sitelinks

4.1.2 Wikidata statistic

Interesting statistics of Wikidata content, as statements per item, labels per item, references by type, are showed by a Wikidata tool at this link[3]. This tool provides with tables and plots a clear overview about distribution of data over Wikidata.

To provide an even more immediate idea of this distribution to the reader, we can see the refinements between the Wikipedia and Wikidata pages in Figure 4.6. This graph makes us immediately perceive the encyclopedic nature of Wikidata, where most of the items are catalogued as human, taxon or for example chemical compound. This could be a problem since these terminologies are not the basis of everyday conversations, but can be useful where a conversational agent should be particularly trained on a specific topic.

4.1.3 Wikidata dump

Wikidata provides various dumps of its data[4], which from a technical point of view, comes to us perfectly. We could import dump on a graph on which the agent could work. This method is preferable particularly if we want to start from an agent that can apply its algorithm on all available information: it can be capable of knowing lots of connections between the various items and, with a reasonable probability, perform at its best.

[3]https://tools.wmflabs.org/wikidata-todo/stats.php?

[4]https://www.wikidata.org/wiki/Wikidata:Database_download/en

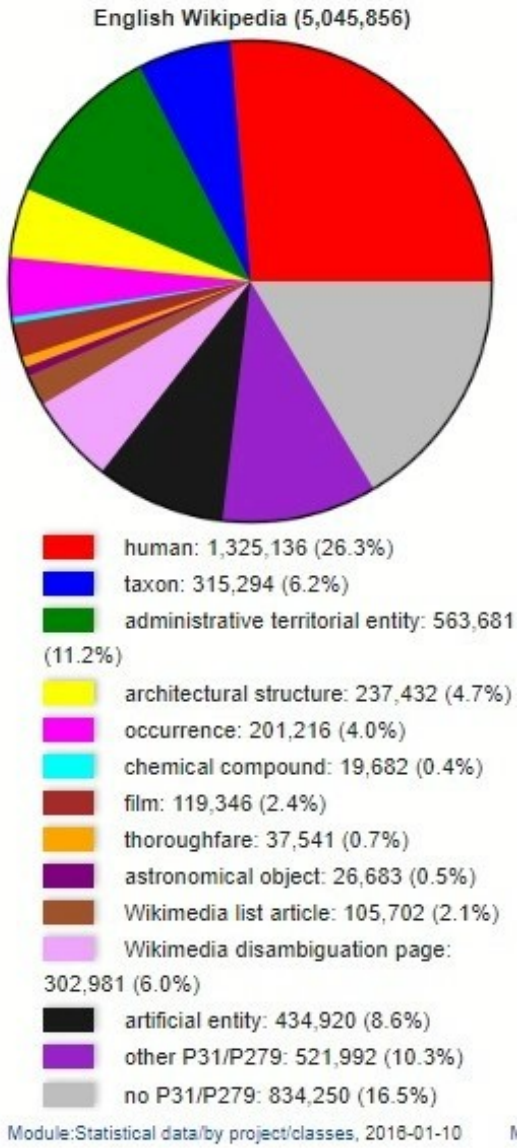

Figure 4.6. Wikidata referenced item by english Wikipedia

Nevertheless, it may not be the best idea to start from this dump. Looking at the final dump graph, we find ourselves managing knowledge of more than 50 gigabytes (compressed). Furthermore, considering the additional overhead created by our algorithms, the KB would reach a size that is hard manageable for the commonly used hardware. It is possible to hypothesise letting the agent reside and perform the various computations on more powerful hardware, displaying only the final results through a GUI. If we think to have more agents running on this hardware, we have to take into account that more KBs have to be managed. In fact, due to the inherent randomness that should occur in the exploration and generation phase, part of KBs could differ among agents. How powerful should we be? Import entire dump on a knowledge graph is not something that cannot be realised, but it indeed leads us to ask ourselves if there is another way. After all, it takes just a moment dive into Wikidata to realise that there are so many things that we ignore, but we consider our self intelligent even without knowing them. It means that we do not need to use all available data in the first step to reach good results.

4.1.4 Bottom-up query for topical domain dataset

The second strategy designed is selecting from Wikidata only specific topic domains. Considering that most of the elements are connected, cutting even a small part of this graph risks being fatal not only from a creativity point of view. In other words, the detected part risks being incomplete for agent reasoning, making the selection of a smaller part of Wikidata only move the problem to another. It is up to us to choose which one to suffer. We prefer the second strategy because it allows us making assessments on an easily controllable set of objects, avoiding too particular topics. Both in terms of time and hardware required, it makes us more skilled and faster in our test. Furthermore, in this first phase of combinatorial creativity, it would be advisable not to overload the agent with a thousand different information, but to limit the range of action to a minimum. Exploring the Internet, we can find up some Wikidata littler dumps regarding particular topics. However, if we decide to use it, we entrust everything owner capacity in his construction. If we want more control over this operation, as in our case, we can use the strategies that will be better explained in the implementation chapter: we can submit queries to Wikidata. The goodness of the dataset, therefore, moves into our ability to write the query. We are not for the

moment interested in very complicated queries, considering that we should provide the tools for possible automatic writing by the agent. How to do?

Analysing the structure of Wikidata, we can notice that starting from an item, passing through its properties and considering their value, we almost may reach only the parallel / bottom-up item graph. In other words, we can only directly go from a more specific thing to a more general item. The opposite is generally achievable just through external tool respect the item page selected. Consequently, we have chosen a straightforward strategy for the retrieval of dataset topic domain. Selected the desired level of specificity in a topic (an item that is placed at a certain depth of graph tree that suits us), we write launches a recursive query that iters on all the items valued for the selected item. We obtain an upward reading, which allows us to embrace the whole above graph concerning the selected item. We would like to underline that this choice is to be considered linked to the KB structure: other choices could be made on other KB.

This strategy is completely imperfect, not only for possible missing interesting items. Due to the high connectivity of Wikidata, data retrieved with this technique easily explodes for some items, so we have to pay attention to the selection phase respect given time requirements.

Nevertheless, during the experimental phase, it proved to be quite satisfactory, especially by introducing the possibility to choose the number of recursions. We need to consider that the agent, just like a human, needs to develop the ability to understand what to ask, what it needs. Broadly dataset could be required where there is still not much data about the topic. A more specific in-depth dataset may be requested where further details are necessary. A mixed approach is not excluded. New data may be required to escape in the case of poor agent behaviour.

4.1.5 Knowledge kernel

Perhaps it is too obvious to assume that consider generic concepts, there are kernels of information that are at the base of many others. There are some generic items, like colours or physical principles, that even if we change the topical domain will be the same. A subsequent analysis could show us, in a fairly predictable manner, that there is a base of shared minimum knowledge, which would be provided to agents for start from common bases and reduce time and effort for initials phases.

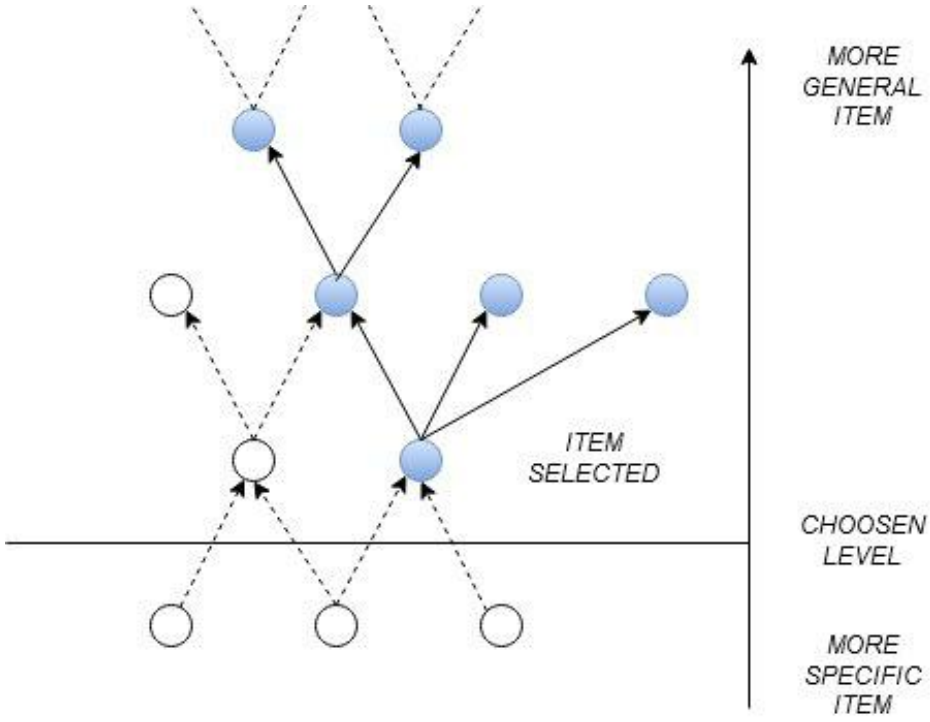

Figure 4.7. Query bottom-up approach

Blue point are items of the sub-graph selected from the chosen item.

That is not all. Other tools can be used to retrieve data from Wikidata. Some exploit ontologies: it is necessary to deepen the ontologies in Wikidata, which we know to be present for few fields. Unfortunately, the community says that they are still too incomplete.

Chapter 5

Approach

In this chapter, we describe in details the experimental setup. In section one we begin with clarifying overview: we show various components of our prototype and their interactions using a diagram, enriching it with meaningful captions. In section two, we proceed to introduce languages, model and tool used. Wikidata, source of our data, it was already presented in the previous chapter.

GitHub Repository The code written could be found in this repository `https://github.com/D2KLab/saiagent`. Through the text, we give only simple example of code about external tool : please refer to GitHub repository for details.

5.1 General Schema

Schema in Figure 5.1, represents the main components of our work and their interactions. As external actors, we have Wikidata and human users. Nothing excluded other interacting agents: on the contrary, is will be considered an added value. The presence of a textual GUI creates an interface that can be accessible indifferently by humans and agents, since agents are equipped with an NLG and NLU parts. But, not only: consider that agents talk the same meta-languages: NLP steps could be skipped, and data directly exchanged. Nevertheless this work is focused on internal evaluation of an agent: external humans or agents are here most seen as validation entities or sources of additional data. The textual GUI can be a simple terminal

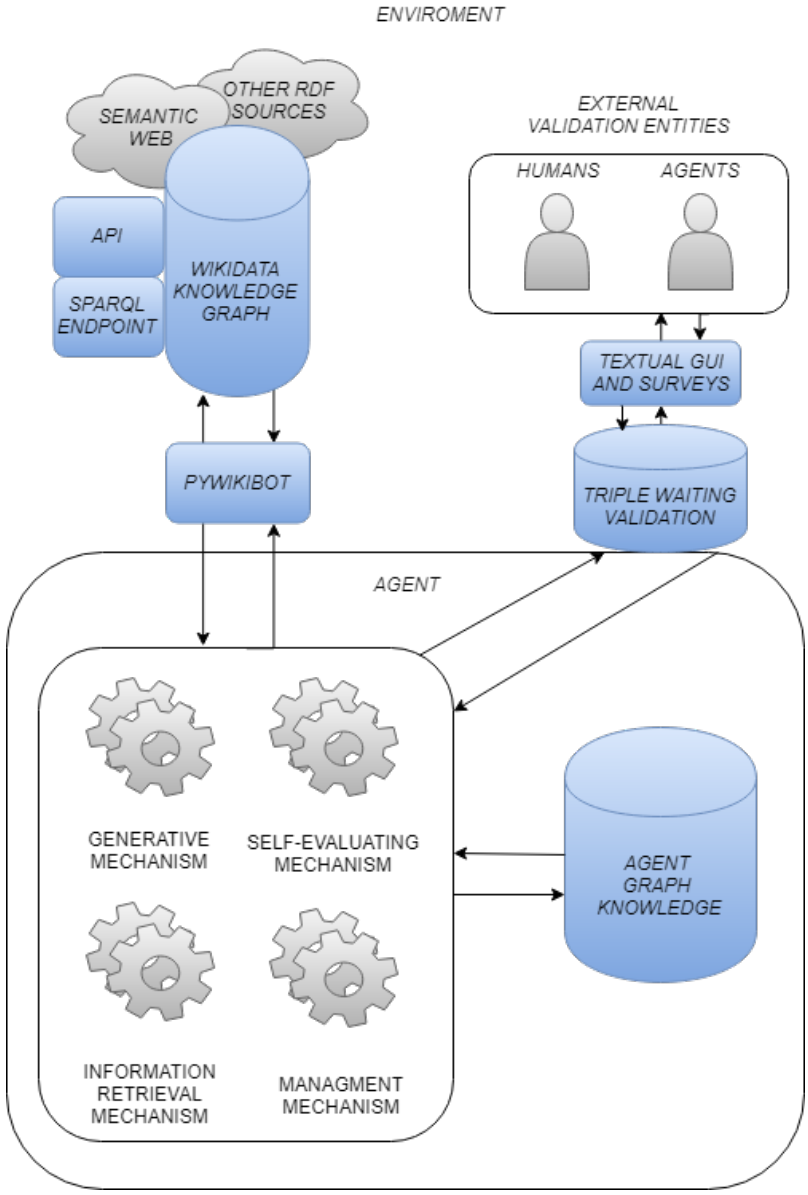

Figure 5.1. General schema of an agent

application: more appealing interface could be a task quickly addressed in a second time.

An agent is composed internally of different parts that work in conjunction. We can divide in generative part, evaluating part, request part, management part and personal graph knowledge. The generative part it is not analysed in this work, and we randomly produce some data to test the point of this thesis, evaluating part. Evaluating part compute our proposed metrics on input data, Wikidata data or self-generated content by agent. Management part is in charge to insert and update data correctly on the knowledge graph. It also has to fulfil some other tasks, as pruning not relevant triple and propagate effects of evaluating part and time decay. Request part asks for dataset or punctual request to Wikidata, through dereferenceable URIs or the MediaWiki API. We choose the last one given that is handily exploited by the usage of bots. In particular, we opted for Pywikibot, a Python library that enables us to access Wikidata automatically from our script. SPARQL queries can be exploited with Pywikibot too.

5.2 Used Tool and Languages

5.2.1 Python

The choice between the various programming languages is not always straightforward, and many aspects have to be taken into consideration when a new application is developed. Among the various possibilities, Python[1] comes as one of the best-known multi-paradigm programming languages, that make it one of the most flexible and dynamic instruments in our hands. This language has other particular advantages that make it an obvious choice, helping us to stay focused on our primal problems, cutting down significant efforts on coding. The wide availability of libraries, code readability, easy integration and full support in most devices, API and software packages, are just a few of the benefits, without forgetting an active support community.

[1] https://www.python.org/

Python is widely used in testing, scripting and maths problems. Practical examples could be machine learning and natural languages libraries used in this thesis.

High-level functions, structures and management may conduct the programmer not reasoning on low-level operations: it is clear that this type of programming style does not lead to precise and optimised management of memory and resources. If the amount of data to handle exponential grows, an issue that may certainly occur, this programming attitude may result in unacceptable computational time. Review and optimising the code could be considerable future work.

5.2.2 Pywibot and SPARQL

Pywikibot A knowledge base, in particular expressed in such a way as simile RDF format, can fall in a huge number of not human-readable information. Furthermore, in an experimental setup, validation operations could take a very long time. Each new request by our agent towards Wikidata is not predictable, and we cannot write each time exactly related queries. If all of these operations have to be managed manually by programmer, there is no exit. To overcome part of these problem we rely to Pywikibot[2], a Python library and collection of tools that automate work on MediaWiki sites[3][4].

Here we points that an approach like this is essential in our work:

- We are not in charge to perform this request manually, but our script using Pywikibot we can do this operation connecting automatically through internet furnishing a compliant request directly and storing answers, instead. We have only to feed it with item identifiers.

- Format problem is eliminated: we have only to choose content of our triple, Pywikibot can translate it directly in Wikidata simile RDF format. It is functional to coherency with Wikidata and easy integration of content but also to can write one script of all operation, with do not

[2]https://phabricator.wikimedia.org/diffusion/PWBC/browse/master/pywikibot/

[3]https://www.mediawiki.org/wiki/Manual:Pywikibot/Overview

[4]https://doc.wikimedia.org/pywikibot/master/

distinguish if this data is generated from Wikidata or the creativity phase of an agent.

- If we do not only need to make a punctual request, or we need to query most effectively, Pywikibot can submit SPARQL query to Wikidata. In this case, Pywikibot does not write the query for us, but coding skeletons of queries and feed it with the requested content, we can skip connection and GUI problem, obtain data directly. Pywikibot also has other query capacities, as download item for Wikidata Category.

- Python child: this bot could be used as Python library, so we do not have to integrate or exploit other languages, but we have all that we need in the same environment at the same time. Data is returned in different ways, among these useful dictionaries and structured Python data.

```
import pywikibot
from pywikibot import pagegenerators
site = pywikibot.Site()
cat = pywikibot.Category(site,'Category:Living people')
gen = pagegenerators.CategorizedPageGenerator(cat)
for page in gen:
    #Do something with the page object, for example:
    text = page.text
```

SPARQL We mentioned SPARQL[5] before and it is easy to understand why it is interesting for our work, paying attention to its recursive acronym: SPARQL Protocol and RDF Query Language. SPARQL is an RDF query language that performs semantic queries on knowledge bases in RDF format. It's well recognised as an excellent instrument in the semantic web, enough to make it a standard by W3C. In our work, enable us to download big chunks of data more smartly instead to perform single requests. Wikidata makes available an endpoint to launch this type of query on its data through SPARQL Wikidata Query Service[6]. This service can be accessed graphical via link reported or with direct requests, that we submit another time thanks to Pywikibot.

[5]https://www.w3.org/TR/rdf-sparql-query/

[6]https://query.wikidata.org

In real, SPARQL is not pretty different from other types of query languages once it gets inside the structure of data that you have. The real difficulty stays in a common problem, such write in an autonomous way as an agent the query. How does the agent know what it needs or wants? How can it compose the textual query? There are tools that can extract query from natural languages to SPARQL, but is not a level of difficulty that we can afford in a thesis and they do not work fine yet.

For the last point, we provide agents with simple prototypes of queries to agents to ask for data, without join, negation or group by and other complex operation. We choose to stay elementary and see agent play with the simplest instrument. The agent can modify items and property of the query according to its. For the first point is our tasks to see if our agent can learn properly how questioning itself. The instrument that we can give to agents can be added further in a modular way: we can construct a generic set of queries that agent could try to learn how to apply to escape minimum of creativity index and to extract relevant information.

Here we report one example of SPARQL query that we use to retrieve all items that have property valorized with a definite value. Here agent request for items that are 'instance of' (P31) 'chemical compound' (Q11173), in English language repository. This query will report almost 100 first result, showing and giving back the item identifiers and item labels of subjects retrieved.

```
SELECT ?item ?itemLabel WHERE {
  SERVICE wikibase:label
  { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  ?item wdt:P31 wd:Q11173.
}
LIMIT 100
```

5.2.3 NetworkX and GraphML

NetworkX To create and manage our knowledge graph we rely on one of the best and complete library for graphs write in Python, NetworkX[7].

[7] https://networkx.github.io/documentation/stable/

This library make available different type of graph structure and a wide set of functions. We choose to use a MultiDiGraph structure, a directed graph with self loops and parallel edges. This structure fit perfectly to Wikidata structure, that can have multiple value for the same property and different weight edges through different document. NetworkX edge can carry attribute of generic type: we exploit this possibility using them massively to store information about document and creativity index related to edge connections.

Visualize Knowledge Graph: GraphML NetworkX can save graph in GraphXML[8] format, an XML-based file format for graphs. The GraphML file format is an attempt to define a common format for graph structure data. It uses an XML-based syntax and supports the entire possible graph structure, as directed, undirected, mixed graphs, hypergraphs and relative attributes. This format we be easily exchanged and used to store out knowledge graph in a standard format, beside out implementation of edge attributes and our textual dumps. It is possible to visualise GraphML graph within a lot of tools. Best know are yEd[9], a graph editor that uses GraphML as its native file format, and Gephi[10], a graph visualisation software that supports a limited set of GraphML.

5.2.4 Used word embedding model: spaCy model

To take advantage of the word embeddings concept, we looked for a pre-trained model, as we don not have hardware and time to compute good word embeddings over 50 gigabyte of data. SpaCy is a free open-source library for Natural Language Processing in Python, that among it feature has linguistic Features, rule-based matching, processing pipelines, trained word vectors and similarity measures. We choose to use spaCy pre-trained model[11] that easily integrates with our python code. It is provided in two version, one smaller only include context-sensitive tensors, and one based

[8] https://projects.cwi.nl/InfoVisu/GraphXML/

[9] https://www.yworks.com/products/yed

[10] https://gephi.org/features/

[11] https://spacy.io/usage/models

on a larger vocabulary with more than 1 million unique word vectors[12]. We also use spaCy disposed similarity measures in computing our metrics: cosine similarity is used by default in spaCy similarity. This similarity can be computed over words and documents, but we think better investigate if document embedding are well supported.

5.2.5 Consideration about time dimension

Time dimension can deeply influence agent mechanisms, exactly how it can affect more or less the impact human perception of events and experiences. Information in the knowledge graph can change, become less important, relevant. This best know as problem of time decay.

To correctly evaluate links between nodes and edges, and therefore of items and properties, it is also necessary to analyse "how long" this information resides in our graph. But also how many times they have been used, when is the last time we took them in consideration and in what precise triple document.

The temporal dimension can help us to mitigate the creativity index over time. The agent can perceive the changing evaluation within a creative process, initially evaluating remarkably, that through the time has been reused over and over again (so may not be impressive anymore). It can also support graph managing part operations on possible pruning on information that has is minimally used over time.

For the timestamp of the triple discovery, NetworkX library helps us. Every edge requests to own unique identifier. By incrementally managing this identifier, we will be able to say accurately when a new triple has become part of the graph, compared to the last identifier available. But it is not all a bed of roses. This identifier does not solve everything: the last identifier available cannot represent the agent's lifetime. Why?

Someone could argue that using a unique identifier as a sign of time makes no sense: in fact, the agent may not generate new triples (and increment the identifier) for a long time. Nevertheless, "time is passing".

At this point, we might think of using our internal hardware clock as a discovery timestamp, managing the NetworkX needed incremental identifier separately. Unfortunately, this is not possible on most devices. The high

[12] https://spacy.io/models/en#en_vectors_web_lg

speed with which the agent can acquire or generate new triples, may not be followed by a guaranteed equivocally of the time returned by the function that accesses the clock of the device.

Another important consideration is the speed of computation and absorption of data by the agent. The speed with which these operations are performed are discrepant with respect to human ones.

In the end, to manage time dimension also for future improvement, we choose to keep track of incremental edge identifier plus discovery internal clock timestamp and last usage internal clock timestamp. To retrieve timestamp information, we use time python library.

5.2.6 Graph pruning, refreshing, quality

However, the refreshing of the graph is not to be attributed only to a temporal question. As previously mentioned, there must be a series of mechanisms designed to maintain the graph, or at least, to report any problems within the graph.

From this point of view, we consider techniques of pruning and evaluation of the quality of the graph. This information can also be used to provide agent status checkpoints, both to the programmer and to the agent itself to implement control strategies.

The pruning phase can be triggered at constant times or by particular events. In the decision to cut an edge and eliminate a node, aspects such as the number of times it has been used and its temporal attributes must be taken into consideration. These considerations relies also on the distribution and attributes of the data in the graph and context sub graph. The pruning phase can not always be accurate, and there is a risk to remove critical information. Analysing nodes and edges pruned, we can also identify unnecessary information previously to do not take them into consideration and don't put in the graph.

The graph pruning and graph refreshing phases are also linked to the quality of the graph. Since our goal is to create an agent that is able to connect and create new content properly, dispose of a large number of functional connections within the knowledge graph can be considered a possible advantage. In graph theory, this concept is called connectivity. Within

the information linked to our creativity index, connectivity can express the quality of our knowledge graph. Respect to the actual degree of connectivity, it would be better not to calculate one single value of all knowledge graph. Different arguments could have different connectivity because of their nature, without this means worse or better quality. It is better to calculate sub graph connectivity and compare with similar sub graph.

Evaluating the quality of the knowledge sub graph can also be a useful tool for identifying situations in which it is no longer possible to generate positive content, cause moving the agent in another part of the graph or trigger new data request. Or, at the contrary, notice changes, such as the arrival of new triples, which could trigger new creative processes in graph areas previously stagnant.

Other concepts that we use to investigate our knowledge graph, to highlight the importance of connectivity, are 'edge bridge' and 'shortest path'. An edge is a bridge if removing it disconnects the graph. In our case, an edge bridge is a triple that strongly relates two separate parts of a graph. It is like to highlight why and from what concept two parts of knowledge are uniquely connected. In the further part of this experiment, we can use this principle to try to generate triples that could be bridges. An edge can not be a bridge but decrease shortest paths, that is the minimum path between two nodes. In this case, it is like to notice that a more direct relationship exists between two items.

NetworkX provides various functions on evaluating these concepts. Among the various ones we use:

```
k_edge_augmentation(G, k[, avail, weight, ...])
Finds set of edges to k-edge-connect G.

is_k_edge_connected(G, k)
Tests to see if a graph is k-edge-connected.

is_locally_k_edge_connected(G, s, t, k)
Tests to see if an edge in a graph is locally k-edge-connected.

average_node_connectivity(G[, flow_func])
Returns the average connectivity of a graph G.

all_pairs_node_connectivity(G[, nbunch,...])
Compute node connectivity between all pairs of nodes of G.

edge_connectivity(G[, s, t, flow_func, cutoff])
Returns the edge connectivity of the graph or digraph G.
```

```
local_edge_connectivity(G, s, t[, ...])
Returns local edge connectivity for nodes s and t in G.

local_node_connectivity(G, s, t[, ...])
Computes local node connectivity for nodes s and t.

node_connectivity(G[, s, t, flow_func])
Returns node connectivity for a graph or digraph G.

bridge_components(G)
Finds all bridge-connected components G.

has_bridges(G[, root])
Decide whether a graph has any bridges.

local_bridges(G[, with_span, weight])
Iterate over local bridges of G optionally computing the span
```

Chapter 6

Results

6.1 Prototype of a user survey

To evaluate the quality of our metrics, we propose a qualitative analysis. We ask to a group of participants to assign values to the various metrics of the creativity index about triple documents, according to their opinion.

A sample of participant will take part in this survey without obtaining particular explanations on this experiment. Only brief descriptions of the various parts of the index will be provided, in order to not leave the interpretation of survey questions to users entirely.

Another sample of participant, will take part in this survey being informed about the dynamics of this work, and during their evaluations will be able to visualise actual knowledge graph.

Considering the subjective perception of creativity, for more comprehensive objectivity - strange to say - we will collect footprints of participant, through traditional background information, such as age, interests, employment and a few keywords about their personality. This report may help us in further consideration.

The collected values are compared with those generated by our metrics for the same documents, giving to us feedback about goodness of our metrics.

6.2 Results of a user survey

For this test we put in our knowledge graph few selected documents, that speak about colour, emotion, animals. Report here entire knowledge graph could be verbose: text version could be found at this link[1], while GraphML file of the same can be found here[2].

After this, we create eight documents. Four documents are random created over the knowledge graph. The other four are created by our self in order to test single parts of the metrics or users comprehension of the survey. This last four documents admit items not in knowledge graph yet. They could be found here[3].

We select sixteen participant, divided in two groups, aware e not aware of experiment details. Each group is provided with the eight documents that we have create. Participant are almost male engineer students, aged between twenty and twenty-five.

At this point we ask the participant to evaluates document respect the concept expressed by our metrics. The same metrics evaluate the documents in turn.

In the table below (Table 6.1), we report creativity index computed by the agent for the eight documents:

Document	ade	div	nov	ser	mag
1	0.1	0.33	0.61	0.025	0.077
2	0.2	0.47	0.61	0.05	0.14
3	0.0	0.29	0.65	0.0	0.23
4	0.2	0.27	0.65	0.74	0.24
5	0.6	0.41	0.58	0.19	0.8
7	1.0	0.43	0.44	0.24	0.9
8	1.0	0.68	0.58	0.31	0.88

Table 6.1. Computed values for the eight documents by our metrics

[1] https://github.com/D2KLab/saiagent/blob/master/dumpkb.txt

[2] https://github.com/D2KLab/saiagent/blob/master/graph.graphml

[3] https://github.com/D2KLab/saiagent/blob/master/testdocument.txt

Here are two of the eight most representative documents, on which we report the results.

First Document (fourth): random generated over previous knowledge graph.

```
kindness subclass of Template
kindness topic's main category light
kindness subclass of emotion
kindness subclass of emotion
kindness different from emotion
kindness has effect light
kindness different from qualia
kindness part of affect
kindness different from Ottuv slovnik naucny
kindness part of color
```

Second Document (eighth): generated by us to evaluate participant metrics understanding and test diversity metric.

```
gold instance of chemical element
Ludovico einaudi place of birth Turin
Mead material used honey
```

	fourth document	eighth document
ade	0.48	0.78
div	0.46	0.75
nov	0.46	0.75
ser	0.43	0.75
mag	0.38	0.62

Table 6.2. Not aware participants result

	fourth document	eighth document
ade	0.51	0.81
div	0.39	0.73
nov	0.57	0.65
ser	0.53	0.63
mag	0.46	0.85

Table 6.3. Aware participants result

As we can see from the results of the first group (Table 6.2), values are affected by the subjectivity of users. Users have more comprehensive knowledge graph not comparable to the agent one. Due to technical limitation, we provide a minimal knowledge graph to the agent, and so the agent can evaluate document on a smaller set of triples and documents. There is a high probability of meeting erroneous triple, or users may already encounter in their life the triples equivalent and do not perceive creativity behaviour.

This suggests that by providing the agent with a more extensive knowledge graph document evaluations may be more similar to those provided by humans. Technical limitations of devices used in our experiment, could be solved by running these computation on dedicated clusters, which own extra computational power.

Instead, as we could imagine, the second group provides values closer to those of the agent o even higher (Table 6.3). With the possibility of looking inside the agent knowledge graph, users can put themselves in the same optics as the agent, resulting in be more sensitive to simple creative and new information, as if they could put themselves in the agent's shoes.

Looking at the values computed for adherence, concerning the random example documents, we can note a defect in the adherence formula. Adherence takes as reference the agent knowledge graph and Wikidata. Wikidata, due to its encyclopedic nature, is very restrictive regarding terminology: equivalent humanly perceived phrases, for Wikidata, and consequently for the agent, instead may would be non-adherent. Evaluation of serendipity it is immediately influenced.

An idea to solve this problem is validate triples, not found neither in Wikidata or knowledge graph, by humans, before apply our metrics. It could be onerous, so another solution is change our exist function with a similarity function over Wikidata statement. Both solutions had already

been hypothesised in our general scheme (Figure 5.1) and can be implemented through a poll of temporarily nonexistent triples.

Analysing Pareto frontiers between combinations of creativity index parts, certainly the one we are most interested in is the relationship between adherence and novelty. The graphic obtained (Figure 6.1) is slightly different from what we expected. Despite the fact that no point has reached the maximum limit of both sides, we would have expected a more curved front instead of a nearly straight one. We believe that the number of documents and the issue already discussed for adherence metrics strongly affect this Pareto front. The Pareto front should be recalculated and analysed after the appropriate changes previously discussed.

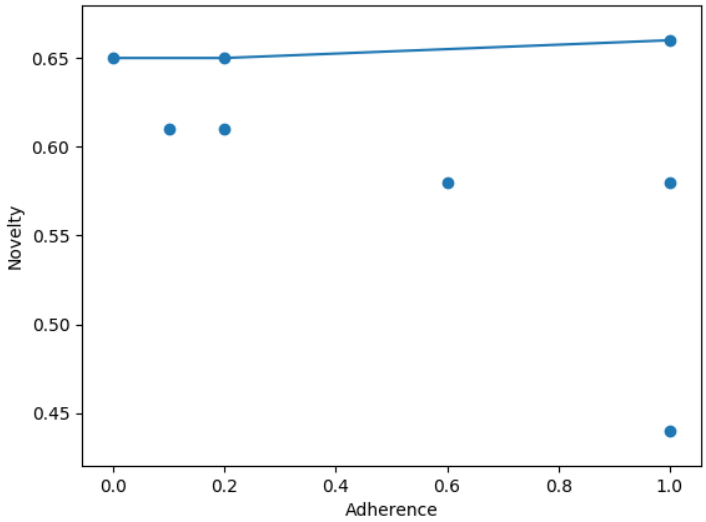

Figure 6.1. Pareto frontier: correlation between adherence and novelty

Unfortunately, the survey is not trivial, and it was challenging to get users into the right perspective: they had to be personally assisted during the survey. This has not allowed us to increase the number of people involved. Considering that the low cardinality of users, accuracy, recall, precision and F1 is affected: for such measures is better to have a higher number of users, to have a proper distribution of possible and mean results.

We prefer to collect more data before computing them.

What we can conclude from these results, is that the metrics reflect a certain degree of subjectivity with respect to prior knowledge. As soon as the participants take consciousness of the knowledge graph available to metrics, the parameters become immediately closer to those compute by metrics. This shows that the interpretation of content is profoundly influenced by personal internal content perception, even for an agent. To come to this observation, was enough to enter few documents into the knowledge graph. Therefore, we can infer that metrics could be effective also over little set of data, although it is evident that in this experiment we provided too few documents. What we believe is that one most relevant points is the distribution of the data of the knowledge graph. Observing the novelty metric, this becomes obvious: novelty values are almost constant ones, and this is probably due to equal distribution of topics in the knowledge graph in this experiment, that not represent a human one. Next experiments have to be unquestionably conducted on a well-distributed knowledge base.

Chapter 7

Conclusion

Analysing state of the art, we find ourselves in a system without well defined shared techniques for the evaluation of intelligent conversational agent. This is a completely understandable problem, since the aspects to be taken into consideration are endless, and same definition of human intelligence is debated. The arise of very young field, Psychometric AI, pointing out that the scientific community of the AI need to explore shared metrics on these unconventional aspects. The metric we advanced is undoubtedly part of this field. Comparing metrics result with user surveys, different discrepancies arise, but they are identified and resolvable problems. In fact, several aspects have to bearing in mind: incompleteness of a knowledge graph, context recognition, proper merge between graph and deep learning strategy, time dimension, bias, large amount of data, technical limit of our hardware, subjective opinions of the interviewees. This shows us that the artificial intelligence we are aiming for is still a very long way, but studies like this are essential to take the world of artificial intelligence to the next level. Only by thinking about this elaborate, there are many future works which we can already imagine.

7.1 Future works

The secondary goal of this metric, but not less important, is to provide in the near future not only a tool that the programmer can use to evaluate the agent, but to give the same tool to the agent itself. The next substantial step of this work will be to make this metric available to the agent as a self-control

system for the generative phase (Figure 7.1). This would make the agent capable of generating not existent triples according to goodness computed by our metric, allowing it to construct its own personal knowledge.

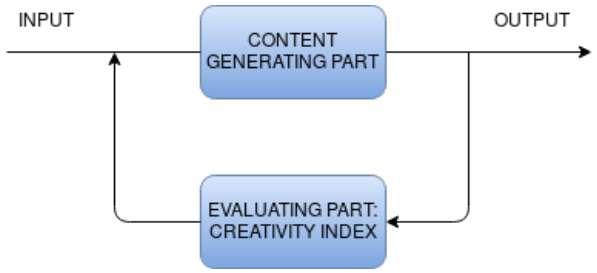

Figure 7.1. Self control loop of a future agent.

The similarity values returned by spaCy are computed on their model, which have surely been trained on texts of a discrepant structure from those contained in Wikidata. Wikidata due to its ontological nature it is not completely comparable to a free text or has same words frequencies. The idea is transforming Wikidata pages into triple documents and use these to train a model that can provide us with more relevant similarity measures.

About the reference knowledge base, Wikidata has a large number of not often used lemma and item: this is due to its purely encyclopedic nature. These triples affect evaluation and may be a problem for the generative phase. A solution is filter useless triples or change reference knowledge base with another RDF based that fits better our purposes.

It is obvious that our metrics, methodologies, algorithms, validations and analysis performed are affected by bias. The same argument we deal with is by definition affected by bias. We will may define a group of bias that are strongly influencing and construct a bias function that makes its contribution to creativity index function.

In our algorithms, there are numerous parameters fixed by us: for example, the depth of the tree requested in the data retrieval phase. Fine tuning or dynamic choice of a set of these parameters, according to needs, can be a substantial improvement.

As we work with graphs, we can think of deepening the graph theory to take advantage and inspiration for possible optimisations and algorithms.

Much more functions than those shown in the state the art are made available already by the NetworkX library.

As noted, one of the main problems is feasibly handling large graphs of both terms of time and memory. But to improve agent behaviour we have to feed it with a larger set of data. Improving the efficiency of the code and minimising online requests, run code on a cluster as well as the impeccable use of the libraries provided, is a beginning. Further elaboration of the proposed format could lead to a better compromise between online calculation and memory, also using hash tables with frequent used values. In the case of massive graphs, we could think of an algorithm that manages its decomposition between ram and secondary memory, bringing into ram only the parts that are used at that moment.

Is our interest also investigate graph embeddings, in the wake of word and document embedding, that could be an interesting compact mathematical representation of our agent knowledge graph, an "agent print". We will be curious to try to exchange entire dialogues by sending the embedding graph of the desired sub-graph directly. Unfortunately, the embedding operation almost always results in a loss of information, so there are some difficulties to overcome.

A last essential and continuous improvement we can make is constantly investigating state of the art: analyse new metrics, read about AI psychology, learn new tools and paradigms for natural language processing.

Bibliography

[1] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1), January 2018.

[2] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[3] Stan Franklin and Art Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *Proceedings of the Workshop on Intelligent Agents III, Agent Theories, Architectures, and Languages*, ECAI '96, pages 21–35, London, UK, UK, 1997. Springer-Verlag.

[4] Pattie Maes. Designing autonomous agents. *Robot. Auton. Syst.*, 6(1-2):1–2, June 1990.

[5] Jörg P. Müller, Markus Pischel, and Michael Thiel. Modeling reactive behaviour in vertically layered agent architectures. In *Proceedings of the Workshop on Agent Theories, Architectures, and Languages on Intelligent Agents*, ECAI-94, 1995.

[6] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(October):433–60, 1950.

[7] Joseph Weizenbaum et al. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[8] PARRY encounters the DOCTOR. RFC 439, January 1973.

[9] Luc L. Steels. *The Talking Heads experiment: Origins of words and meanings*. http://langsci-press.org, 2015-05-11.

[10] Petter Brandtzaeg and Asbjørn Følstad. Why people use chatbots. 11 2017.

[11] John R. Searle. Minds, brains, and programs. *Behavioral and Brain*

Sciences, 3(3):417–424, 1980.

[12] David Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, Inc., New York, NY, USA, 1996.

[13] John R. Searle. *Consciousness and Language*. Cambridge University Press, 2002.

[14] Philippe Rochat. Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition*, 12(4):717 – 731, 2003. Self and Action.

[15] Martin A. Fischler and Oscar Firschein. *Intelligence: The Eye, the Brain, and the Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1987.

[16] Alessandro Lenci. Distributional semantics in linguistic and cognitive research.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[18] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[19] Anna Huang. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[20] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*, 2016.

[21] Margaret J. Trotter, Paul M. Salmon, and Michael G. Lenné. Improvisation: theory, measures and known influencing factors. *Theoretical Issues in Ergonomics Science*, 14(5):475–498, 2013.

[22] Christine Moorman and Anne S. Miner. The convergence of planning and execution: Improvisation in new product development. *Journal of Marketing*, 62(3):1–20, 1998.

[23] Robert B. McLaren. The dark side of creativity. *Creativity Research Journal*, 6(1-2):137–144, 1993.

[24] Alfonso Montuori. The complexity of improvisation and the improvisation of complexity: Social science, art and creativity. *Human Relations*

- *HUM RELAT*, 56:237–255, 02 2003.

[25] Gary Klein. Flexecution as a paradigm for replanning, part 1. *IEEE Intelligent Systems*, 22, 2007.

[26] Amílcar Cardoso, Tony Veale, and Geraint A Wiggins. Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine*, 30(3):15–15, 2009.

[27] Thomas B Ward. What's old about new ideas. *The creative cognition approach*, pages 157–178, 1995.

[28] Umberto Eco. Combinatoria della creatività. *Lecture, Florence, September*, 15, 2004.

[29] Nicole M Radziwill and Morgan C Benton. Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*, 2017.

[30] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. On evaluating and comparing conversational agents. *ArXiv*, abs/1801.03625, 2018.

[31] Ong Sing Goh, Cemal Ardil, Wilson Wong, and Chun Che Fung. A black-box approach for response quality evaluation of conversational agent systems. 2007.

[32] Marcilio de Oliveira Meira and Anne M. P. Canuto. Evaluation of emotional agents ' architectures : an approach based on quality metrics and the influence of emotions on users.

[33] Dijana Peras. Chatbot evaluation metrics. *Economic and Social Development: Book of Proceedings*, pages 89–97, 2018.

[34] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 1997.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[36] José Hernández-Orallo, David L. Dowe, and M.Victoria Hernández-Lloreda. Universal psychometrics. *Cogn. Syst. Res.*

[37] J.P. Guilford. *The nature of human intelligence*. McGraw-Hill series

in psychology. McGraw-Hill, 1967.

[38] Kyung Hee Kim. Can we trust creativity tests? a review of the torrance tests of creative thinking (ttct). *Creativity research journal*, 18(1):3–14, 2006.

[39] Selmer Bringsjord and Bettina Schimanski. What is artificial intelligence? psychometric ai as an answer. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, IJCAI'03, pages 887–893, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[40] Diego Monti, Enrico Palumbo, and Giuseppe Rizzo. Sequeval: An offline evaluation framework for sequence-based recommender systems. *Information*, 10:174, 05 2019.

[41] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA, 2005. ACM.

[42] Gautam Agrawal, Christina Bloebaum, Kemper Lewis, Kevin Chugh, Chen-Hung Huang, and Sumeet Parashar. Intuitive visualization of pareto frontier for multiobjective optimization in n-dimensional performance space. In *10th AIAA/ISSMO multidisciplinary analysis and optimization conference*, page 4434.