Communication, Society and Media
ISSN 2576-5388 (Print) ISSN 2576-5396 (Online)
Vol. 3, No. 4, 2020
www.scholink.org/ojs/index.php/csm

Original Paper

An Investigation about Entailment and Narrative by AI Techniques (Generative Models)

Paolo Marocco[1] & Roberto Gigliucci[2*]

[1] Department of Economics, University of Genova, Italy

[2] Sapienza University of Rome, Italy

Received: October 28, 2020 Accepted: November 7, 2020 Online Published: November 16, 2020

doi:10.22158/csm.v3n4p61 URL: http://dx.doi.org/10.22158/csm.v3n4p61

Abstract

Many storytelling generation problems concern the difficulty to model the sequence of sentences. Language models are generally able to assign high scores to well-formed text, especially in the cases of short texts, failing when they try to simulate human textual inference. Although in some cases output text automatically generated sounds as bland, incoherent, repetitive and unrelated to the context, in other cases the process reveals capability to surprise the reader, avoiding to be boring/predictable, even if the generated text satisfies entailment task requirements. The lyric tradition often does not proceed towards a real logical inference, but takes into account alternatives like the unexpectedness, useful for predicting when a narrative story will be perceived as interesting. To achieve a best comprehension of narrative variety, we propose a novel measure based on two components: inference and unexpectedness, whose different weights can modify the opportunity for readers to have different experiences about the functionality of a generated story. We propose a supervised validation treatment, in order to compare the authorial original text, learned by the model, with the generated one.

Keywords

entailment, narrative, linguistic generation, transformers, narratology, AI

1. Introduction

To start with, let us look to discuss with what we consider to be an existing disconnection between the study of Artificial Intelligence and the analysis of storytelling. During the decade of 2000-2010, there was a period of confluence, say a mutual interest, more or less, between the two domains of research mentioned above. As a testament to what has previously been said, we may mention some bibliographical datum; for example the work by Cavazza and Pizzi (2006) about *Narratology for Interactive Storytelling* (narrative representations, Trees, etc.), or the proposal of *Advancing*

Computational Models of Narrative by Richards and Finlayson (2009). Other voices in Italy: Basili, Gigliucci, Marocco (2010) (where we proposed a mix between LSA and classification using syntagmatic features); Gigliucci and Marocco (2006); Gangemi (2006). We feel that it is fair to say that these premises have not been maintained.

Interestingly, from the nineties onward we have not experienced significant technological improvements in the field of studies about Artificial Intelligence and storytelling: the models have not changed to a great extent. These models have continued to adopt a mixture of classification, decision making and logic rules. Not so different, one could say, from the classical AI patterns of the eighties (Expert Systems) that was to be echoed and reverberated in Literature over the next two decades.

A bibliographical testimony of that is the book by Bringsjord and Ferrucci (1999). The two authors showed a good mathematical background, and their challenge was to demonstrate «that logic is forever closed off from the emotional world of creativity». Their approach was typical of the first chatbots that have been simulating the Turing Test, namely, to build a knowledge repository where an "agent" answered a human person: the agent ran through Decision Trees to query the repository. Alternatively, more agents interacted with each other—without human support—and they gave life to narrative paths.

When implementing these deductive-inductive schemes, which are enclosed and self-sufficient, we specifically limit the magnitude of events and stories to be narrated. However, these systems can easily be plunged into crisis in response to unexpected questions, whereby they need continuous extensions and generalizations, as well as ways of appending new rules and new events and stories: it is—you pardon the pun—a never-ending story.

Despite considerable advances in neural language modeling during the last twenty years (Note 1), the more hopeful branches toward the text generation and automated storytelling have simply addressed the statistical learning domain. The first goals of general computational linguistics of the nineties have been long forgotten. Indeed, promises of a machine, that tries to dive deeply into a text comprehension (using logics and others linguistic models), have not been kept. By adopting large training on millions of webpages and heterogeneous documents, the current language models pursue this law: more information ingested, more knowledge (reasoning abilities) acquired. An example is offered by OpenAI's GPT2 model, followed by its successor GPT3, which has recently overpassed the Turing Test, in some lite versions. These models are based on the assumption that the probability's distribution of a words sequence can be decomposed into the product of conditional next word distributions. Simply put, the text generation guidelines try to discover the more probable distributions of words chains, with respect to the global knowledge learned (paradigmatic level), and locally inferred by the previous sentence (syntagmatic level). In these cases, the inference is drawn by assigning a most probable maximum score to words' context and relationships among pieces of text.

The current metrics, used for measuring text-generated quality, follow the same concept. Essentially,

they use a similarity score, calculated in some vector space. They adopt lexical-based syntactic or semantic matches; recently, semantic similarities among sentences had been increasing, without being able yet to deeply investigate the meaning. The results are not very convincing, because, in linguistics and literature tradition, the meaning depends on other factors too, for example "Who addresses whom in what surrounding". Moreover, an entailment-based test considers only the semantic consistency of antecedent (premise) with consequent (entailment hypothesis), exploiting in any case the same geometrical-probabilistic level.

Some authors observe: "Quality human language does not follow a distribution of high probability next words. In other words, as humans, we want generated text to surprise us and not to be boring/predictable. The authors show this nicely by plotting the probability, a model would give to human text vs. what beam search does" (Note 2).

In addition to this kind of perplexity, we find case reports in which the fault is more pronounced, and the inference tests show lack of comprehension and common sense, in spite of the fact that the model used is GPT-3, the top current model. That has been announced as a result that would allow a machine to reason broadly in a manner similar to humans without having to train for specific tasks. Let us show an example of inference driven by GPT-3:

You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to…? Remove the door. You have a table saw, so you cut the door in half and remove the top half *(Note 3).*

Although the output is grammatical, and even impressively idiomatic, its comprehension of the world is strange, and the appeal is nonsensical. The machine offers a seriously a costly solution to destroying the home, for finding a solution to a trivial problem. This is absolutely at opposition to the effort minimizing suggestion of such a problem-solving machine (also automatic). It is not a singular case: many texts generated, when understandable, are curious and seem rich in random irony. Probably useful for creative storytelling, but definitely inadequate for logical entailment, as the NLP metrics attempt to measure.

In light of this, the purpose of our research is to extend the entailment, providing for other linguistic connectives, because the narrative and lyric tradition often does not proceed towards a real logical inference. The storytelling coherence and readability are conditioned by a complex way of reasoning, pointing out a spectrum of solutions, starting from an antecedent statement, completely untied from the next sentence, then trying to infer the truth of the antecedent (hypothesis). It is evident that this spectrum collects a rich variety of sentence relations, some inferred by cause-effect, some not. It is different from the simple opposition between entailment and contradiction, like the current evaluation systems insist on measuring. When we work from this point of view, it is extremely difficult to reduce all the relations which do not belong to the cause-effect inference to an easy "not-entailed-label", as

required by traditional NLP metrics. Primarily because we lose all the properties of narrative variety, and also the ability to capture the reader's attention. To achieve this scope, we propose a novel measure of narrative inference, and, more generally, the concatenation of a pair of sentences based on two components: *inference* and *unexpectedness*, whose different weights can modify the opportunity for readers to have different experiences about the functionality of a story.

2. Method

2.1 Related Literature

It is not simple to collect the related literature that has aided our research, for several reasons. From one perspective, the new horizons of NLG (Natural Language Generation) research, was born in 2017, with the Transformers models, probably too short on time to allow the use of new measures of narrative functionality. BLEU score (Note 4), and derivative metrics like BLEURT (Note 5) are used to evaluate the consistency of a text generated from a trigger (the previous words), but they fall into the trap of adopting the same technological level applied to language (the story to put in exam) and meta-language (the metric to measure the properties of the story). Since the automatic approach is limited by the lack of metalevel, the human evaluation is often preferred, but limited only to classical inference proof.

From another perspective, as we anticipated above, during the last decade we have assisted with a detachment from NLP (Natural Language Processing) research, performed basically by computer scientists who are unrelated to digital humanities fields and are not interested in literature studies and narrative experience. Despite the names, NLG and Computational Creativity (CC) convey different meanings: the first relies on some advanced and specific algorithms of Neural Network and Statistical Learning, whereas CC (Note 6) is oriented on the creativity from a less formalized point of view, drawing from psychology, philosophy and sometimes from neurosciences.

In this panorama, Mark Riedl (Note 7) is an author who merges the two areas, engaged in Intelligent Narrative Technology (Note 8) and in the new improvements of NLP. In "Computational Narrative Intelligence: Past, Present, and Future" (Note 9) he traces the insights of this approach: "*Narrative intelligence* is the ability to craft, tell, understand, and respond effectively to stories". The purpose is to highlight how the knowledge, required to understand a story, can be used to create new stories. Once the domain model has been ingested, a story generation system can produce an infinite number of stories involving characters, places, and actions that are known to the system. In "Scheherazade: Crowd-Powered Interactive Narrative Generation" (Note 10) the author uses crowdsourcing information to automatically learn the domain knowledge that is needed to construct and to understand stories about daily activities, such as going to a restaurant or to a movie theater. The approach is similar to the basic principles, implemented in the pre-trained model of NLG, but the difference is essentially given by the different level of human contribution. A system like Scheherazade uses a plot graph for

predicting possible stories' development, driven by human supervisors (Note 11), whereas the systems like GPT-2 learn by texts and generate new texts through a completely automatic way, driven only by statistical distribution laws.

In a recent paper, Riedl (Note 12) has been improving his research, with the goal of integrating NLG perspective, of decomposing neural plot generation in two issues: a) the generation of a sequence of events (event-to-event) and b) the transformation of these events into natural language sentences (event-to-sentence), so as to offer a solution to the current limits of NLG (the capability to produce grammatically correct but semantically inconsistent sentences). The paper, remarkable for measuring the topic alignments of sentences, fails to adopt a suitable evaluation of the inference problem, limiting the measure to BLUE score. In another paper (Note 13), Riedl focuses the attention about the skillfulness to learn normative natural language descriptions, exploiting predetermined narrative events, but without measuring in detail the entailment.

At the end of the story, except for particular cases, like the quoted studies of Riedl and his collaborators, we observe a fairly small intersection between NLG veins and Narrative Intelligence (or other sub-stream of Digital Humanities). The first branch continues to remain out of the field of narrative theories, and it avoids directing itself into the deep role of inference (and its variations) of the literary tradition shows. The second branch is often not updated to the Deep Learning models and is too much tied to the rules that are drawn by humans, thus introducing a bias that confounds a model evaluation.

In addition, recent *Digital Humanities* studies, specifically concerning literary texts, are definitely stopped at the stage of perhaps more sophisticated forms of querying textual *corpora* (that is a development of the "linguistica computazionale" in Italy, arising from the pattern DBT by Picchi-Stoppelli) and little else, as *e. g.*, examples of digital philology, critical editions of texts available online that are pretty much interactive. We have an evident testimony of this backwardness, when we read an insight into this topic which is now offered by an outstanding review such as *Italian Studies*: what should be—and it is, actually—an overview concerning "Italian Studies and the Digital" (Armstrong & Patti 2020). This gives us a very thorough frame of the existing international debate. Nonetheless, in this paper we find only a single reference to Artificial Intelligence as a future possibility, related to DH, presented as almost a science fiction projection (p. 206) (Note 14). *The rest is silence*, a silence that is overflowing with bibliography which has truly little to do with New Digital perspectives.

2.2 Research Question

Our research is focused on evaluating the narrative functionality of a story (generated through some AI tools). To achieve this scope, we propose a novel measure of narrative inference, extended to all the concatenation modes of a pair of sentences. For *functionality* we mean the encoding of a sequence of dynamic components (also in the diachronic meaning) that aim to achieve a *successful value judgement*, approved and confirmed by a community of readers. In layman's terms, with *narrative functionality* we

identify basically the propriety of a story (written as well as in audiovisual etc.) to be *successful* for much of its public. There is no need here to make matters more complicated, quoting the concept of *Erwartungshorizon* and the Reception Theory (*Rezeptionsästhetik*), or the thought of Gadamer and so on. We prefer to remain on an elementary level, at most evoking and invoking the common-sense theory.

Our attention is driven by the so-called NLG task: "prompt generated text". In simple terms: the human premise is followed by a consequent sentence generated by the machine, through a statistical learning mechanism. The purpose of this mechanism is to predict the next probable word based on the earlier sequence of words. The quality of this prediction depends on several factors, mainly based on the model (the neural network that implement the text generation) and on the features that are exploited by the model, to learn from a training set of data texts. In any case, the tool must meet stringent requirements, about building understandable stories, so that the reader satisfaction can be reached. There is therefore the requirement for an evaluation measure.

The method that we propose is based on two components: *inference* and *unexpectedness* (Note 15), whose different weights can modify the opportunity for readers to have different experiences about the functionality of a story. The inference part is related to the prediction capability of the reader, while the unexpectedness conveys a feeling of surprise, of astonishment and curiosity about the storytelling evolution. Our idea is supported by the belief that this functionality could be described by feasible features, quantitatively measurable by weights: $< \tau_i$: inference, τ_u: unexpectedness>, applied to the concatenation of two contiguous sentences (x,y). Better said, a weight τ ($0 < \tau < 1$) sets the leverage on the story, provided respectively by the inference and the unexpectedness of y, given x. The sum of the two-weights $s= \tau_i + \tau_u$ represents a sort of aggregative index, able to capture different opportunities for generating a story. For example, $\tau_i + \tau_u \approx 0$ sounds like a free concatenation of the sentences, with a lack of useful information for a story, whereas $\tau_i + \tau_u \approx 2$ represents a sort of contradiction, because it is a concatenation marked by high inference (involving high predictability) and high unpredictability at the same time. The best equilibrium for the readability of the story is to assess around $\tau_i + \tau_u \approx 1$ (see just below).

The challenge is the evaluation of *inference-unexpectedness* attributes, reaching writer's style, with respect to the case of the machine's prediction style. The evaluation is based on the comparison between one pair of sentences (x,y), with another pair (x,z), where x,z are written by the author, and y is generated by the machine. In this way, we can explain a quality of text generation, and its stickiness to a replicability style, from the point of view of narrative functionality. Otherwise, from a theoretical point of view, it is simple to generalize this mechanism, expanding more chains (x_1, …, x_n) where (x,y) = (x_j, x_{j+1}) for some j $0 < j <= n-1$.

Let us give some examples of the first case (relating to texts written by an author).

To simplify, let us consider only the average of our two indexes: <mean τ_i, mean τ_u> related to some pairs (x,y) captured from reference texts. In the novel *The Catcher in the Rye* by Salinger (Note 16) we expect average values of inference lower than those of unexpectedness (f. i. 0.3 < 0.7), because this novel belongs to a narrative strand which we can call post-modernist, where causality is often undetermined—in fact, in some cases, it is replaced by manifest contradictoriness. On the other hand, if we take such an example as Dickens' *David Copperfield* (Note 17), we expect some inference levels greater than unexpectedness (f. i. 0.6 > 0.4). That happens because a lot of events in the latter novel are logically caused by other events. Furthermore, the semantic relationship antecedent → consequent will coincide with mere contiguous couples (previous → following). But, still in relation to this author, if we consider another novel by him, a masterwork of humor, enriched with nonsenses, *Pickwick Club* (Note 18), we are now expecting: less inference than unexpectedness, as in *The Catcher*'s case, even though we shall find probably less obvious differences (0.4 < 0.6), since *Copperfield* is a 19th Century novel, that is more "classic", from the narratological point of view. Take another example, *Alice in Wonderland* by Carrol. The analysis may be quite intriguing, because the logical writing process of that novel will almost surely generate causal consistencies but, at the same time, the unrealistic world Carrol describes shall determine something unexpected and amazing. Perhaps, the amounts of <mean τ_i, mean τ_u> in *Alice* will be similar to *Pickwick*'s ones, but the "variance" (for instance variance in each chapter) will be diversified, because *unexpectedness* in Carrol's novel is not the same as that in *Pickwick*: in fact, it has a different distribution.

2.3 Methodological Proposal

In order to explain an idea of method, we now propose a classification with five possible cases. This allows us to make a partition of the range [0,1] in three levels for each index: small ($\approx$ 0), medium ($\approx$ 0.5), large ($\approx$ 1). This trisection allows a supervised evaluation in the light of a prior approximation.

1) <inference $\approx$ 0, unexpectedness $\approx$ 1>
2) <inference $\approx$ 1, unexpectedness $\approx$ 0>
3) <inference $\approx$ 0.5, unexpectedness $\approx$ 0.5>
4) inference $>$ unexpectedness (0.7 > 0.3)
5) inference $<$ unexpectedness (0.3 < 0.7)

Some examples:

1) I go to kiss my mother -> and my mother stabs me
2) I go to kiss my mother -> and she hugs me
3) I go to kiss my mother -> but she walks away, irritated
4) I go to kiss my mother -> and she invites me to lunch
5) I go to kiss my mother -> and she breaks down in tears

Certainly, the narrative context makes the sentences clearer; to be less abstract, we propose another sequence of examples, in the same order, with materials extracted by the summary of a renowned tale by Raymond Carver (Note 18):

1) Scotty goes to school with his mate and he wonders what gift he will receive -> then a car hits him
2) If the baker takes the order by keeping silent -> then the baker is abrupt
3) Scotty's mother goes to the bakery to buy a birthday cake for the son -> and she chooses it among those in the shop window
4) She sees a radio in the backroom of the bakery -> and she hears country-western music.
5) Scotty falls with his head in the gutter and his legs out in the road -> and his eyes are closed, but his legs moved back and forth

Logic-narrative sequences: our experiments seek to explain the cause-effect typologies; they range from ordinary to exceptional, by means of a neural network, with causal chains in the form of antecedent-consequent. The fuzzy cases shown before, like inference $\gtrsim$ unexpectedness (0.7 > 0.3) and the inverse values (0.3 > 0.7), set faint levels of the basic ones mentioned below.

Examples:

a) determined cause-effect $\langle \tau\text{-i}, \tau\text{-u} \rangle \approx \langle 1,0 \rangle$

c) verisimilar cause-effect $\langle \tau\text{-i}, \tau\text{-u} \rangle \approx \langle 0.5,0.5 \rangle$

d) unlikely cause-effect $\langle \tau\text{-i}, \tau\text{-u} \rangle \approx \langle 0,1 \rangle$

2.4 The Model

Nowadays there are two possible approaches for generating stories from NLG perspective. The first, and easier, is by adopting pre-training models, such that they are available through an open-access provider (like Hugging Face, Google and so forth). The second one concerns the creation of a dedicated model from scratch, otherwise customizing a pre-training model via a transfer learning process (transferring knowledge across tasks or domains), in order to make the machine learn the contents of a writer and his style.

In terms of the scope of our research, we essentially require a pre-trained model, that has learned from a default set of data texts (e.g., blogs or web pages) and a mechanism of fine-tuning of this model using new information. In such a way, we can compare some results of text generation, using the default model, with other results, captured by the transfer learning method. Our attempt is to show an improvement of the customized model, with respect to the trustworthiness of generating text in accordance with the original writer's style.

Basically, within the NLP neural networks we have two options to transfer new knowledge: a) using task-driven information or b) using domain-driven information. The original pre-trained network is

trained on a general domain of data for different source tasks: we must specialize the original knowledge to assist the target task (in our case: learning supervised information about the classification of sentence pairs, with respect to the two indexes: <inference, unexpectedness>) and the domain task (leveraging the original model and adding the contents of the reference writer submitted to our attention).

Let us discuss some examples:

Task driven: this learns from labeled pairs like: <premise → hypothesis> for example: A *soccer game with multiple males playing → Some men are playing a sport*; this takes into account the supervised label from an open source dataset (*e. g.*, **https://nlp.stanford.edu/projects/snli/**):

- **E**ntailment (inference),
- **C**ontraddiction (opposite inference),
- **N**eutral (the output does not rely on premises and consequences).

In a similar way, we proceed to learn the unexpectedness of the task. Since it is not a standard evaluation in the traditional NLP process, and it is hard to find it navigating the public repositories, we will compile a dedicated one.

- Domain driven: It also embodies the reference novels (*The Catcher in the Rye* and possibly the other works of Salinger) to map on the neural network.

We observe that the simple task-driven process is not enough to allocate the inference/unexpectedness structure of a story, because it has been set on the linguistic experience derived from a writer. It captures the insight of the inference property, compliant to natural language understanding systems, but it also needs to move this abstract knowledge toward concrete examples of the specific author utterances, depicted through the domain-driven task.

The problem of the making of the library (dataset) to achieve our *fine-tuning* is a particularly delicate task. The literary texts we need, to put together a homogeneous and consistent corpus for successful queries, must be at least—if not totally—oriented towards a narrative which is open to unexpectedness or, in any case, somehow distant from a classical (Note 19), traditional way of storytelling.

Furthermore, we have decided to include also poetic texts and dramas. The former subgroup of texts is the result of a choice of lyrical and epic poems whose main characteristic is to surprise the readers through unexpected relationships and metaphors, or unforeseen events. The latter sub-set has been chosen among quite “original” dramatists’ pieces that are renewals, more or less, of the classical theatrical traditions.

Almost one hundred thousand pages of this databank can be enough to result in interesting findings, in our opinion. Obviously, we shall implement, if necessary, the corpus in question.

3. Result

We intend to measure the weights of the two indexes <inference, unexpectedness> using samples extracted from different datasets: the pre-trained model, the fine-tuned model and the original samples of author's books. The original samples have to set the baseline of the writer's style, while the two levels of text generation (pre-trained and fine-tuned) need a comparison with respect to the baseline. The comparison will be estimated through principal statistics, like mean, variance, median and the Probability Density Distribution function (PDF) of the evaluation of N pairs of sentences in each section of partition (Note 20) (where: A. text generated by pre-trained model, B. text generated by fine-tuning; C. is the original text).

In more detail, the sequence of experiments as follows:

A. Using a pre-trained model of an open source library (e.g., GPT-2 or XLNet of HuggingFace framework). Extracting N sample (e.g., 1000) of pairs of sentences <**premise, hypothesis**> prompt-based generated (e.g., *they are always asking you to do them a big favor -> because you are a very handsome guy*).
    a. Submitting the sample to a supervised evaluation of weight level for each index < τ-inference, τ-unexpectedness> taking in account three levels small (≈ 0), medium (≈ 0.5), big (≈ 1).
    b. Making some descriptive statistics about the results.
    c. Calculating mean, variance, median and estimate the PDF function.
B. Getting Task-driven and Domain-driven information, then fitting the fine-tuned model. Extracting N samples of pairs among the new results and evaluating these, like in the basic step of pre-training.
C. Repeating the evaluation mechanism about pairs extracted from the original test of the author.

The C. test applied to the original text can suggest interesting scores about the writer's preference to privilege strong or fuzzy entailment, and about the editing style used to create sentences' concatenation. The statistical values, delivered by A. and B. test the different degree of <τ-inference, τ-unexpectedness> in a simple sentence concatenation, and ensure the ways in which a machine can learn about an original text and generate some variations, are more or less readable and understandable by a generic user, compliant to the original style of the author (Note 21).

4. Discussion

So far, we have been trying to measure the coefficients of inference and unexpectedness: a maximum of the first coefficient—say 0.9—implies a superminimal presence of the second, and vice versa. So, we have two extreme ends of a spectrum where the two coefficients find a reciprocal distribution.

This is not the place to draw a historical diagram of western narrative which can be grounded upon the

average measure of the equilibrium between the two coefficients, age upon age. However, we can say that, for example, the so-called Baroque period has had a particular orientation towards the *meraviglia* (anything that is met with curiosity or wonder, or is unusual, odd, rare, etc.), and this statement, with its clarifications and details, is essentially basic knowledge. In a post-modernist narrative we may find something of similar, but obviously not identical. We are seeing an increase in the desire for surprise and for a kind of "functional inconsequentiality"—we use the term *functional* in the practical meaning: in short, when a tale or a novel *functions*, *i. e*., works. We may exemplify with Pynchon's or De Lillo's novels, with Carver's and Salinger's stories, with Foster Wallace's inventions or, within the cinematographic field, with Cohen Brothers' movies, or with Lynch's masterworks. In fact, such an artist as David Lynch has accustomed his viewers—in the TV series of *Twin Peaks*—to narrative possibilities which are actually unexpectedness-oriented, and he has managed to familiarize the average audience with what, on the face of it, appears to be oddity or absurdity.

Obviously, the balancing act between the two coefficients proposed in this work has been a constant over the centuries. Nevertheless, possible objections could be proposed: is it true that all readers/viewers always aspire to something unexpected? The answer, we think, is not at all certain. Often, in fact, the pleasure for the audience grows from the solace given by a verisimilar order of cause-effect (remember Aristotle). Then, regardless of historical epochs and tastes of the times, the equilibrium/disequilibrium between inference and unexpectedness cannot be subjected to a deterministic law. We can say, to provisionally conclude, that the periods of "high classicistic orientation" privilege the entailment coefficient, whereas "baroque-oriented" ages (let's say) prefer to enhance the unexpectedness coefficient.

Our research stands on the line of recent studies about neural network applications for classifying humanities and art works. A computer scientist, who collaborates with museums and galleries to classify artwork from the point of view of creativity, is Ahmed Elgammal (Note 22). Conversely, we do not notice on the side of History of literature and of literary movements studies, an analogue improvement. Our preliminary work aims to bridge for this gap.

References

Ammanabrolu, P., Tien, E., Cheung, W., Zhaochen L., Ma, W., Martin, L. J. & Riedl, M. O. (2019). Guided Neural Language Generation for Automated Storytelling. In *Proceedings of the Second Storytelling Workshop* (pp. 46-55), Florence. https://doi.org/10.18653/v1/W19-3405

Armstrong, G., & Patti, E. (2020). Italian Studies and the Digital. *Italian Studies*, *75*(2), 194-208. https://doi.org/10.1080/00751634.2020.1744867

Basili, R., Gigliucci, R., & Marocco, P. (2010). Un approccio geometrico all'analisi dei testi letterari. In S. Bassi (Ed.), *Seminari Signum 2006* (pp. 61-80). Pisa: Scuola Normale Superiore.

Carver, R. (1983), *Cathedral*. N.Y, F. Knopf.

Cavazza, & Pizzi. (2006). Narratology for Interactive Storytelling: A critical introduction. *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*. Berlin-Heidelberg: Springer. https://doi.org/10.1007/11944577_7

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181-253. https://doi.org/10.1017/S0140525X12000477

Dessalles, J.-L. (2010). Have you anything unexpected to say? The human propensity to communicate surprise and its role in the emergence of language. In A. D. M. Smith, M. Schowstra, B. De Boer, & K. Smith (Eds.), *The Evolution of Language*. Singapore: World Scientific. https://doi.org/10.1142/9789814295222_0013

Dessalles, J.-L. (2015). L'imaginaire de la narration: Une approche cognitive. In P. Musso, S. Coiffier, & J.-F. Lucas (Eds.), *Pour innover, modéliser l'imaginaire—Regards croisés d'industriels et de chercheurs* (pp. 154-167). Paris: Editions Manucius.

Dessalles, J.-L., & Dimulesco, A. (2010). Understanding Narrative Interest: Some Evidence on the Role of Unespectedness. Retrieved from https://www.researchgate.net/publication/228667128_Understanding_Narrative_Interest_Some_Evidence_on_the_Role_of_Unexpectedness

Dickens, C. (1837). *Pickwick Club*. London: Chapman & Hall.

Dickens, C. (1850). *David Copperfield*. London: Bradbury & Evans. https://doi.org/10.1093/oseo/instance.00121331

Elgammal, A., & Babak, S. (2015). *Quantifying creativity in art networks*. From arXiv:1506.00711

Gangemi, A. (2006). A Semiotic Metamodel for Bridging Lexical and Formal Semantics. In S. Bassi (Ed.), *Seminari Signum 2006* (pp. 21-48). Pisa: Scuola Normale Superiore.

Gigliucci, R. (2020). If…then: Narratologia per terne inferenziali. *Studi (e testi) Italiani*, in press.

Gigliucci, R., & Marocco, P. (2006). Traduzione e creazione letteraria. In A. di Stefano (Ed.), *Cyberletteratura: tra mondi testuali e mondi virtuali* (pp. 85-89). Roma: Edizioni Nuova Cultura.

Holtzman, Ari, & all. (2019). *The curious case of neural text degeneration*. From: arXiv preprint arXiv:1904.09751

Li, B. Y., & Riedl, M. (2015) Scheherazade: Crowd-powered interactive narrative generation. *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Retrieved from https://nil.cs.uno.edu/publications/papers/li2015scheherazade.pdf

Marneffe, C.-M. de, & all. (2006). Learning to distinguish valid textual entailments. *Second Pascal RTE Challenge Workshop* (Vol. 62). Retrieved from https://nlp.stanford.edu/pubs/rte2-report.pdf

Moretti, F. (2013). *Distant Reading*. London-NY: Verso.

Nahian, M. S. A., Frazier, S., Riedl, M., & Harrison, B. (2020, February). Learning Norms from Stories: A Prior for Value Aligned Agents. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 124-130). https://doi.org/10.1145/3375627.3375825

Richards, W., Winston, P. H., & Finlayson, M. A. (2009). Advancing Computational Models of Narrative. *Computer Science and Artificial Intelligence Laboratory—Technical Report*. Cambridge (Mass.): MIT.

Salinger. (1951). *The Catcher in the Rye*. Boston: Little, Brown & Co.

Spratt, E. L., & Elgammal, A. (2014). *Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science*. Retrieved from https://arxiv.org/pdf/1410.2488.pdf

Sitographic References

https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html.

https://nlp.stanford.edu/projects/snli/

https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html

http://computationalcreativity.net/iccc20/

http://eilab.gatech.edu/mark-riedl

https://www.aaai.org/Library/Workshops/ws17-20.php

http://www.di.unito.it/~rossana/INT10/index.html,

https://www.dropbox.com/s/evivh66wq5zzu16/int10.pdf?dl=0

https://www.aclweb.org/anthology/W19-3405.pdf

Notes

Note 1. Natural Language Processing (aka NLP) is a field of Artificial Intelligence focused on the ability of the machines to comprehend language and interpret messages. a set of algorithms designed to explore, recognize and utilize text-based information and identify insights for the benefit of the business operation. As such, natural language processing and generation algorithms form a backbone for the majority of computerized processes. To put it simply, NLP gives the computer the skills to: understand informally written queries; extract the meaning out of it; generate the responses of its own; perform requested tasks.

Note 2. Holtzman, Ari, et all. (2019).

Note 3. The stressed bold words are the response of the machine. It is interesting the comments of the authors of the test (Gary Markus and Ernest Davis): [This is one confusion after another. The natural solutions here would be either to tip the table on its side (often sufficient, depending on the specifics of

the geometry) or to take the legs off the table, if they are detachable. Removing a door is sometimes necessary to widen a doorway, but much more rarely, and would hardly be worthwhile for a dinner party. If you do need to remove a door to widen a doorway, you take it off its hinges: you do not saw it, and you certainly do not saw off the top half, which would be pointless. Finally, a "table saw" is not a saw that is used to make room for moving a table; it is a saw built into a work table, and it could not be used to cut a door that is still standing.] See https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ and https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html.

Note 4. Automatic metrics often take two sentences as input, a candidate, and a reference, and they return a score that indicates to what extent the former resembles the latter, typically using lexical overlap. A popular metric is BLEU, which counts the sequences of words in the candidate that also appear in the reference. BLEU: Bilingual Evaluation Understudy Score. The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0. About entailment: Marneffe et all. (2006).

Note 5. https://ai.googleblog.com/2020/05/evaluating-natural-language-generation.html

Note 6. http://computationalcreativity.net/iccc20/

Note 7. http://eilab.gatech.edu/mark-riedl

Note 8. https://www.aaai.org/Library/Workshops/ws17-20.php

Note 9. http://www.di.unito.it/~rossana/INT10/index.html, https://www.dropbox.com/s/evivh66wq5zzu16/int10.pdf?dl=0

Note 10. https://nil.cs.uno.edu/publications/papers/li2015scheherazade.pdf

Note 11. In https://www.aclweb.org/anthology/W19-3405.pdf see Star Wars 3 example…. un sistema a metà tra regole e varianti statistiche: queste sono troppo rigide o troppo lasche, quindi generano varianti minimali, inutili per una storia proponibile nel mercato dell'entertainment o in quello delle arti letterarie.

Note 12. Ammanabrolu, Tien, Cheung, Zhaochen, Ma, Martin & Riedl (2019).

Note 13. Nahian, M. S. A., Frazier, S., Riedl et all. (2020).

Note 14. Furthermore, in the paper we do not find any quotation of the works by Franco Moretti: see now the collected essays in Moretti (2013).

Note 15. Published in volume in 1951.

Note 16. Publ. in vol. in 1850.

Note 17. Publ. in vol. in 1837.

Note 18. Namely: A small, good thing according to the version of Cathedral (1983), available online too: http://www.classicshorts.com/stories/sgthing.html.

Note 19. Note that also in neuroscientific fields some researchers use the term classical vs non-classical to identify a normal event in the environment vs an unexpected one that arouses the brain's reaction to implement its framework of prediction-reaction: see e. g. Clark (2013): 184.

Note 20. PDF function is referred to: in x-axis are set the three levels values of each index, and on y-axis the frequency.

Note 21. For a more detailed discussion on this point see Gigliucci (2020).

Note 22. Elgammal & Babak (2015); Spratt & Elgammal (2014).