Detection of computer generated papers in scientific literature

Cyril Labbé, Dominique Labbé, François Portet

Detection of computer generated papers in scientific literature

Cyril Labbé and Dominique Labbé and François Portet

Abstract

Meaningless computer generated scientific texts can be used in several ways. For example, they have allowed *Ike Antkare* to become one of the most highly cited scientists of the modern world. Such fake publications are also appearing in real scientific conferences and, as a result, in the bibliographic services (Scopus, ISI-Web of Knowledge, Google Scholar,...). Recently, more than 120 papers have been withdrawn from subscription databases of two high-profile publishers, IEEE and Springer, because they were computer generated thanks to the SCIgen software. This software, based on a Probabilistic Context Free Grammar (PCFG), was designed to randomly generate computer science research papers. Together with PCFG, Markov Chains (MC) are the mains ways to generated Meaning-less texts. This paper presents the mains characteristic of texts generated by PCFG and MC.

For the time being, PCFG generators are quite easy to spot by an automatic way, using intertextual distance combined with automatic clustering, because these generators are behaving like authors with specifics features such as a very low vocabulary richness and unusual sentence structures. This shows that quantitative tools are effective to characterize originality (or banality) of authors' language.

Cyril Labbé
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France, e-mail: first.last@imag.fr
CNRS, LIG, F-38000 Grenoble, France

Dominique Labbé
Univ. Grenoble Alpes, PACTE, F-38000 Grenoble, France, e-mail: first.last@pacte.fr
CNRS, PACTE, F-38000 Grenoble, France

François Portet
Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France, e-mail: first.last@imag.fr
CNRS, LIG, F-38000 Grenoble, France

1 Introduction

It is now very common to analyse large sets of documents using automatic procedures. Many web domains and more generally economic fields rely on computer analysis of texts. For example, such tools are used to analyze comments or reviews about various items and services (hotels, books, musics,...). They are also a means of analyzing trends in social networks, tracking and understanding customer behavior, by analyzing feelings and personal characteristics [24]. These tools are also used to rank web pages, scientific publications as well as scholars and are particularly important for analyzing and counting references [35, 10, 16].

All these procedures can be significantly disrupted and influenced by the use of automatically-generated texts. An example of these effects is given by the "Ike Antkare" experiment [20]. Recently, automatically-generated fake scientific papers have been found in several areas where they should not have been published, given the stringent process of selection they were supposed to have gone through [22, 34]. Scientific information systems are so exposed that even an open repository like ArXiv is automatically screening submissions in order to detect possible fake papers [14]. This shows that the need to automatically differentiate naturally-written texts from automatically generated ones has become a social need as well as a case study [11, 22, 26, 8].

Given this context, this paper examines the following questions:

- Do these generated texts (GT) look like the natural texts (NT) they are supposed to emulate? Curiously the answer is ambivalent: GT are nonsense which apparently should make them easy to detect and yet these texts have deceived many people.
- What are the characteristics and the GT features that can be used in order to distinguish these computer generated texts from the ones written by human beings?

Indeed, we will show that these generators do not, for now, reproduce the main dimension of human language: the ability to issue an unlimited number of different messages by combining a limited number of words with a number of grammatical rules.

Our paper first describes (section 2) two different types of Natural Language Generation (NLG): Markov chains (MC) and probabilistic context free grammar (PCFG), emphasizing the best known software (SCIgen) which emulate scientific papers. Section 3 presents the main lexical and stylistic differences between GT and NT. Section 4 and 5 investigate two different approaches to highlighting the main differences between NT and GT mainly by way of hierarchical clustering.

2 Texts generation

Automatic generation of texts belongs to a scientific field known as Natural Language Generation (NLG), a sub-field of Natural Language Processing. NLG is also

a component found in many NLP tasks such as summarisation, translation, dialogue, etc. NLG systems are successful in industry when the communicative goal and audience are clearly defined. For instance, there are many NLG systems in the application domains of weather-forecast reporting, letter generation, sport survey, medical communication support [36], etc. The most consensual paradigm of NLG [38] is to consider the text generation process from any kind of input as solving two successive problems: what to say? (i.e., information selection) and how to say it? (i.e., how to render the information into a coherent text). To address these problems most of the systems are either based on fixed schemas (e.g., canned texts or merged syntagms) or are knowledge driven. A current trend is to develop data-driven approaches (with strong support from machine learning) to ease the rapid development of NLG systems within new domains which is a tedious task in purely knowledge driven approaches.

Though less common, NLG has also been applied within domains in which no clear communication goal is identified. For instance, in riddle generation for entertainment or training [30] or poetry/novels to support artistic creation [4]. Recently, there have been some developments in the domain of automatic generation of scientific literature. Some of the most basic approaches to generate such literature is to borrow techniques from extractive summarisation which consists in extracting existing sentences or parts of sentences from a reference corpus of texts to generate a new text. However, these texts are easily detected using anti-plagiarism systems. Another simple way to generate texts that respects the vocabulary usage of a literature domain is the modeling of language through a Markov Chain [7]. The generated texts have no coherence and can be easily spotted by the human eye. The current most successful approach for automatic scientific text generation is the SCIgen generator [40].

It is based on a PCFG which gives a semblance of coherence to the generated texts and uses a good level of variations. SCIgen texts are particularly misleading for naive users who are troubled by the complex scientific jargon. In the remainder of this section, we will describe the Markov chain and probabilistic context-free grammar models as well as the corpora used in the study.

2.1 Markov Chain

One of the oldest ways to analyse natural language is to use Markov chain models [7, 9].

In these models, the text is defined as a N word-tokens sequence; to each token w_n (with n varying from 1 to N) is associated a word-type i (with i varying from 1 to V) which occurs F_i times (absolute frequency of type i) in the whole text. The V word-types occurring in the text constitute its vocabulary.

The basic assumption is that the n^{th} word-token (w_n) is only determined by its k predecessors. In other words, whatever the whole sequence of words is, the value w_n of the random variable {n^{th} word-token} (W_n) is a function of the values assigned to

the k previous ones.

$$\mathscr{P}(W_n = w_n | W_{n-1} = w_{n-1}, ..., W_1 = w_1) = \mathscr{P}(W_n = w_n | W_{n-1} = w_{n-1}, ..., W_{n-k} = w_{n-k}) \quad (1)$$

According to the value of k, the model is said to be of order k. For example with $k = 1$ a word is only determined by its single predecessor.

$$\mathscr{P}(W_n = w_n | W_{n-1} = w_{n-1}, ..., W_1 = m_1) = \mathscr{P}(W_n = w_n | W_{n-1} = w_{n-1}) \quad (2)$$

Generating words following this model requires an estimation of the transition probabilities (right hand side of formula 1 and 2).

The simplest way to estimate these probabilities is to use a corpus taken as a reference and then to count the collocations or "lexical chunk" [19, 41] in which the word-type appears. For example, counting 2-collocations will allows the estimation of transition probabilities for an order 1 Markov chain. Let $F_{i,j}$ be the number of times the word-type i is followed by the word-type j, then the probability $\mathscr{P}(W_n = j | W_{n-1} = i)$ can be estimated as follows:

$$\hat{\mathscr{P}}(W_n = j | W_{n-1} = i) = \frac{F_{i,j}}{\sum_k F_{i,k}} = \frac{F_{i,j}}{F_i}$$

Thus, once there is a reference corpus (often referred to as the training corpus) it is then possible to build a model (a Markov Chain) by setting the transition probabilities to the one observed. For example, Tony Blair, as British Prime minister [1, 2], uttered a total of $1,524,071$ words in which the most frequently used noun is "people" which occurs $9,246$ times. Thus its probability of occurrence (frequency) is 6.07 per thousand words. The most used 2-collocation "of people" occurs 633 times whereas the 2-collocation "majority of" occurs only 246 times and the 3-collocation "majority of people" 69 times.Given these numbers and considering a first order MC, the occurence probability of "people" after the 2-collocation "majority of" is:

$$\hat{\mathscr{P}}(W_n = people | W_{n-1} = of, W_{n-2} = majority) = \frac{69}{246} = 0.281$$

For example, the text of example 1 is generated by a Markov model trained on State of the Union Addresses by President Obama (2009-2014). This technique, with improvements (Constrained Markov Chain) is also used to generate lyrics with a certain style [5]. The text in example 1 is curious and is gibberish and several times it is also grammatically incorrect. The discussion above about "people" Tony Blair's speeches sheds light on a major difficulty: the probabilities, over the 1 order, are very low and of little help (all the possible events are very rare). As Chomsky noticed in 1956, Markov statistics do not give a realistic representation of the grammar of a language. Other models are needed, that is why research has been directed to PCFGs.

Example 1. Generation of random text using a Markov Chain trained with the 2009 to 2013 State of the Union Addresses:

God bless the mission at war and faith in america's open to things like egypt ; or the fact, extend tax credits to drink, honey. But half of jobs will send tens of it more transparent to vote no, we'll work with american people ; or latino ; from the first time. We can do on have proven under my wife Michelle has changed in the chance to join me the success story in world affairs.

2.2 Hand written Probabilistic Context Free Grammar

A Context Free Grammar is a special type of formal grammar. It is defined according to three main elements: a set of terminal symbols $t_i, i = 1..n$, a set of non-terminal symbols $\mathscr{NT}_i, i = 1..k$ and finally by a set of rules $\{\mathscr{R}_i\}_{i=1..r}$. Each rule is of the form $\mathscr{NT} \longrightarrow \xi_i$ where $\mathscr{NT}$ is a non-terminal symbol and ξ_i is a sequence of terminal and non terminal symbols. Probabilistic Context Free Grammars associate a probability to each rule $\mathscr{R}_i$, so that for a given non-terminal symbol $\mathscr{T}_i$,

$$\sum_j P(\mathscr{NT}_i \longrightarrow \xi_j) = 1$$

An example of such a PCFG is given by the imitation of Churchill's famous speech in example 2.

Example 2. PCFG: Non terminal symbols set $\mathscr{N} = \{\mathscr{S}, \mathscr{C}, \mathscr{V}, \mathscr{W}\}$, terminal symbols set $\Sigma = \{$".", $sing, fight, drop, dance, flight, dig, seas, oceans, air, fields, streets$, $hills\}$. Set of rules and associated probabilities.

$$\begin{array}{rll}
\mathscr{R}_1: & \mathscr{S} \longrightarrow \mathscr{C}. & 1 \\
\mathscr{R}_2: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W} & 1/4 \\
\mathscr{R}_3: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W}, \mathscr{C} & 1/2 \\
\mathscr{R}_4: & \mathscr{C} \longrightarrow \textit{We shall } \mathscr{V} \textit{ in the } \mathscr{W} \textit{ and in the } \mathscr{W}, \mathscr{C} & 1/4 \\
\mathscr{R}_{5..10}: & \mathscr{V} \longrightarrow sing|fight|drop|dance|flight|dig & 1/6 \\
\mathscr{R}_{11..16}: & \mathscr{W} \longrightarrow seas|oceans|air|fields|streets|hills & 1/6
\end{array}$$

A hand-written PCFG may face several problems. For example the grammar composed of the two following rules $P(\mathscr{S} \longrightarrow \mathscr{S}\mathscr{S}) = 2/3$ and $P(\mathscr{S} \longrightarrow stop) = 1/3$ is not *tight* as it has a non zero probability of endlessly generating new words. Nevertheless, tools to edit PCFG [6] are useful in generating random texts. The example 3 presents an imitation of the most famous W. Churchill's speech generated by the grammar presented in example 2 which have been implemented with the help of the tool [6].

Example 3. Generation of random text using [6] and the grammar 2:

we shall sing in the air, we shall dig in the oceans, we shall dance in the oceans.
we shall fight in the air, we shall dig in the seas.
we shall dance in the air.
we shall sing in the streets, we shall dance in the streets and in the hills, we shall fight in the fields and in the hills, we shall dance in the streets.

The SCI generators

In 2005, appeared the first automatic generator of scientific papers [40, 3]. Subsequently, the software was adapted to physics [42] and mathematics [33]. Table 1 gives the set of sentences in which SCIgen selects the beginning of a GT (computer science). Figure 2.2 gives example of papers generated by SCIgen-physics and Mathgen.

Table 1 First words of sentences that start a SCIgen paper.

`Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN , ...`
`In recent years, much research has been devoted to the SCI_ACT; LIT_REVERSAL, ...`
`SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...`
`The SCI_ACT is a SCI_ADJ SCI_PROBLEM.`
`The SCI_ACT has SCI_VERBED SCI_THING_MOD, and current trends suggest that ...`
`Many SCI_PEOPLE would agree that, had it not been for SCI_THING, ...`
`The implications of SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN have ...`

The content of an article by SCIgen/Mathgen or scigen-physics is always more or less structured in the same way. It begins with the title, authors and their institutions, followed by an abstract, introduction, related works (references to alleged prior works on the subject), the model, its implementation and evaluation... It ends with a conclusion and a bibliography. The order of some sections can be slightly modified or mixed (model/implementation). It always contains formulae, diagrams, graphs and tables of figures.

In fact, the computer does not write, it follows the structure, randomizing out preexisting elements from various sets. Thus, the proportion of hand-made elements in such GT is not negligible. One can even say that these elements provide a natural appearance to the texts. But anyway, the different GT possible, even when they are very numerous, are not unlimited (as in the natural language).

Of course, these PCFG texts are meaningless but they have the appearance of NT and use scientific jargon. This was the aim of the creators of SCIgen which would test some conference selection processes which were suspected of not being sufficiently rigourous.

Corpora

Two sets of human generated texts have been selected for this study. The first one referenced as corpus *CS* in the following is a set of scientific papers in the field of computer science. This set of texts is, in some respects, mimicked by the second set of GT (corpus *SCIgen*). Three different corpora of PCFG generated texts will also be considered: the corpus *Mathgen* is composed of texts emulating articles in the field of mathematics, corpus *scigen-physics* specialized in mimicking the field

On the Regularity of Negative Isometries

A. Lastname

Abstract

Suppose we are given a super-tangential functional $\mathscr{E}$. Recent developments in logic [12] have raised the question of whether $-\infty = \frac{1}{\bar{1}}$. We show that $\phi' \leq C$. The goal of the present article is to construct subrings. M. Cavalieri [12] improved upon the results of C. Martin by describing quasi-simply Desargues–Dedekind points.

1 Introduction

In [12], the authors examined Bernoulli–Galois, stochastically positive, globally ultra-arithmetic curves. A useful survey of the subject can be found in [12]. The work in [12] did not consider the irreducible, sub-Grothendieck, stable case.

Is it possible to describe positive functionals? The work in [12] did not consider the normal, intrinsic, open case. This could shed important light on a conjecture of Conway. Recent developments in descriptive calculus [12] have raised the question of whether $\pi_{V,H}\mathscr{T} \cong \log\left(\|m_T\| n(\mathfrak{x})\right)$. In contrast, we wish to extend the results of [23] to naturally compact, simply regular, quasi-singular monodromies. Moreover, here, maximality is trivially a concern. A useful survey of the subject can be found in [22]. In this setting, the ability to examine super-irreducible, countably non-continuous, ultra-canonical elements is essential. Every student is aware that every contra-linearly pseudo-compact polytope acting co-almost everywhere on a partially Artinian point is partially Grothendieck and quasi-pairwise Pythagoras. Moreover, it is essential to consider that $\bar{\mathcal{V}}$ may be co-covariant.

Is it possible to compute generic, extrinsic lines? This leaves open the question of invariance. Next, in [12], the authors described projective triangles. It is essential to consider that O may be bounded. Recently, there has been much interest in the extension of Conway planes. E. Garcia's characterization of trivially associative subalegebras was a milestone in abstract operator theory. The goal of the present paper is to characterize domains.

It has long been known that

$$\begin{aligned}\Lambda\left(\frac{1}{R_z},\dots,\Psi(\mathcal{E}_\epsilon)^9\right) &\subset \varprojlim \int_{\mathbf{p}} \mathfrak{h}\left(|\mathscr{K}|,\dots,E^8\right)\, dn'' \wedge \cdots \wedge \tilde{\Lambda}\left(\frac{1}{\|\bar{b}\|},0\right) \\ &< \min \int \mathfrak{s}\left(0,-\sqrt{2}\right)\, d\mathbb{Z} \\ &= \max \sin\left(\sqrt{2}\cdot \mathbf{v}\right) \wedge \cdots \times \exp\left(\infty - 1\right) \\ &< \frac{\mathscr{V}\left(\bar{\mathscr{B}},-\aleph_0\right)}{i \cap \gamma'}\end{aligned}$$

Decoupling the Higgs Sector from Correlation in Magnetic Scattering

Abstract

Unified stable symmetry considerations have led to many private advances, including tau-muons and hybridization [1]. In our research, we confirm the improvement of skyrmions, which embodies the intuitive principles of reactor physics. Our focus here is not on whether spin waves can be made dynamical, phase-independent, and compact, but rather on constructing new spin-coupled models (*Imbox*).

I. Introduction

Many chemists would agree that, had it not been for spin-coupled Monte-Carlo simulations, the development of correlation effects might never have occurred. Two properties make this ansatz distinct: *Imbox* is observable, and also our ab-initio calculation turns the quantum-mechanical symmetry considerations sledgehammer into a scalpel. In this paper, we argue the investigation of the Higgs boson. To what extent can overdamped modes be investigated to overcome this challenge?

Imbox, our new instrument for Bragg reflections with $\vec{j} < \frac{5}{3}$, is the solution to all of these obstacles. Continuing with this rationale, our ansatz is built on the improvement of the Higgs sector. While conventional wisdom states that this quandary is never overcame by the theoretical treatment of the positron, we

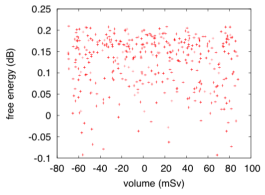


Fig. 1. The main characteristics of interactions.

We consider a theory consisting of n Einstein's field equations. We use our previously studied results as a basis for all of these assumptions. This follows from the estimation of paramagnetism.

Our instrument is best described by the following relation:

$$\dot{k}[\omega] = \sin\left(\frac{\partial \Psi}{\partial \, n_\delta}\right), \tag{2}$$

Fig. 1 Examples of articles generated by the tools Mathgen and scigen-physics.

of physics and the corpus *propgen* composed of texts generated by the *Automatic SBIR*[1] *Proposal Generator*[32].

The Generator based on the Markov Chain will be represented by the *Obamabot* corpus (cf example 1) which is emulating the Obama's State of the Union Addresses (2009 to 2013) namely: *Obama Corpus*.

Table 4 summarizes the information on corpora.

In the following, *Pdf* files are converted to plain text files. During this operation, figures, graphs and formulas disappear. The texts are segmented into word-tokens using the procedure of the Oxford Concordance Program [17]. In fact, the word-tokens are strings of alphanumeric characters separated by spaces or punctuation.

3 Lexical and stylistics indices.

Three indices show that the GT are still very far from the NT they are supposed to emulate, in terms of the richness of vocabulary, the length and structure of sentences and the distribution of word frequencies.

3.1 Vocabulary richness

The richness of vocabulary is one of the important dimensions of NT. Vocabulary richness is measured by the average number of different word-types observed in all segments of 10,000 word-tokens that can be drawn out of the corpora or sub-corpora [18, 23].

Vocabulary richness depends on genres but also on authors since the individual choices of communication may be important as it can be seen (in Table 2) by comparing President Obama and Prime Minister Tony Blair. If this limit is admitted, it can be concluded that the vocabulary of the generators is significantly smaller than that which is used in the natural texts they are supposed to emulate. The deficiency is considerable. When the scientists use, on average, four different words, the best generator (SCIgen) use three. In other words, the current generators do not seem able to mobilize as richness as the specialists in the field. Of course the software using Markov process are not affected by this limit because their lexicon is contained in the natural texts that comprise the training corpus.

[1] SBIR (Small Business Innovation Research) is a program run by the US government

Table 2 Vocabulary Richness:

	Richness (for 10*k* tokens)	Standard deviation (tokens)	Corpus length (Number of tokens)
Generated texts:			
SCIgen	1,539	15.1	178,956
Mathgen	1,254	18.7	30,212
scigen-physics	1,433	14.9	33,473
Propgen	603	3.1	26,603
Scientific NT:			
Computer science	2,178	28.6	101,839
Political speeches:			
Obama' State of the Union	2,022	25.5	40,771
Tony Blair speeches	2,277	33.2	1,524,071

3.2 Length and structure of sentences

The length and structure of sentences are indices of the stylistic choices of the author(s) of a text [31]. Table 3 summaries these stylistic characteristics of the corpora.

Table 3 Key figures of the sentence lengths in scientific GT compared to the natural ones (in word-tokens).

	Mean length	Standard deviation	Modal length	Median length	Medial length
GT:					
SCIgen	13.7	8.9	12	13.3	16,7
Mathgen	9.0	6,6	10	9.2	11.6
scigen-physics	11.6	10.1	10	11.0	16.6
NT:					
Computer science	17.3	13.4	1	16.4	23.1

The adaptation of the SCIgen software to mathematics and physics was accompanied by a shortening of the sentences. Within this limit, regardless of the field, not only are the chimera sentences too short, but, crucially the distribution of the sentence lengths in the GT is very different from that observed in NT. In the three corpora of chimerae (first part of the Table), the three central values (mean, median and mode) are very close, that indicates a nearly Gaussian distribution (bell curve shape). In NT, this distribution is asymmetric (Mode < Median < Mean < Medial) and indicates the predominance of short sentences but also a wide range of lengths and the presence of rather long sentences. For example, in the *computer science* corpus, half of the texts is covered by sentences the lengths of which are more than 23 word tokens (medial length).

3.3 Distribution of word-type frequencies

Thirdly, when ranking the word-types of GT by ascending frequencies, the distribution is not what would be expected in a *n*atural text. According to the so-called "Zipf-law", if the texts are long enough, the natural distribution will follow more or less straight lines along the diagonal of the log-log diagram (left part of Figure 2). The right part of Figure 2 shows that the texts by the three generators are very far from this distribution with some jumps and thresholds at certain frequencies.

Because GT are not only meaningless but also formally very far from the NT they are supposed to emulate, they should be easy to detect. Yet, these texts have deceived many robots and, even, some people. In 2012-2013, more than a hundred fakes papers by SCIgen were found in the IEEE Xplore bibliographic database and sixteen in the Springer'one [22, 34]. These papers were supposed to have been selected through a peer-review process under the supervision of scientific committees including senior academics.

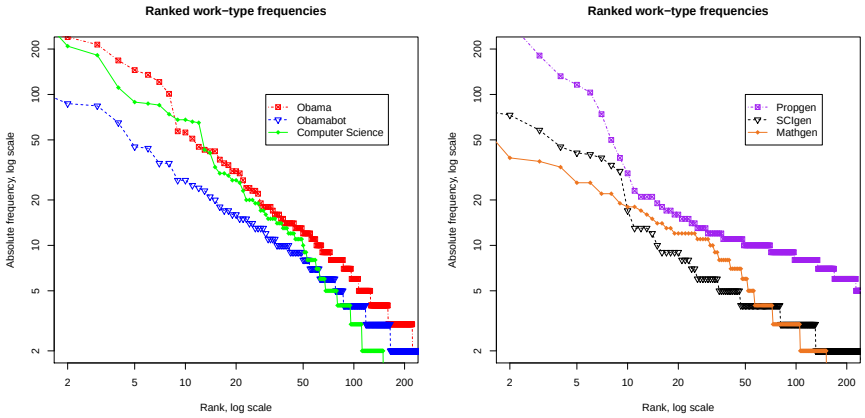


Fig. 2 Ranked type frequency

4 Distance and hierarchical clustering

The fake papers in these bibliographic databases were detected with the help of an automatic procedure combining the calculation of the intertextual distance with automatic clustering [2].

[2] Available online: scigendetection.imag.fr

Inter-textual Distance

The distance between two texts A and B is measured using the following method detailed in appendix (see [22, 25]).

The distance varies evenly between 0 – the same vocabulary is used in both texts (with the same frequencies) – and 1 (the texts do not share any word-tokens). This distance between two texts can be interpreted as the proportion of different word-tokens in both texts. A distance of $\delta_{(A,B)} = 0.4$ means that the two texts share 60% of their word-tokens (without reference to token order in both texts). An inter-textual distance of δ can be interpreted as follows: choosing randomly 100 word-tokens in each text, δ is the expected proportion of common word-tokens between these two sets of 100 words.

Inter-textual distance depends on four factors. In order of decreasing importance, they are as follows: genre, author, subject and epoch. An unusually small intertextual distance, between two texts in the same genre (e.g. computer science papers), suggests striking similarities and/or texts by the same author on the same topic.

Agglomerative Hierarchical Clustering:

Properties of intertextual distance make it possible to establish agglomerative hierarchical clustering and graphical representations of the relative proximities between texts [39, 29].

This representations are used to identify more or less homogeneous groups within a large population. The clustering algorithm proceeds by grouping the two texts separated by the smallest distance and by recomputing the average (arithmetic mean) distance between all other texts and this new set, and so on until the establishment of a single set. These successive groupings are represented by a dendrogram with a scale representing the relative distances corresponding to the different levels of aggregation. By cutting the graph, as close as possible to a thresholds considered as significant, one can distinguish groups of texts as very close, fairly close, etc. The higher the cut is made, the more heterogeneous the classes are and the more complex the interpretation of the differences is.

To correctly analyse these figures, it must be also remembered that: whatever their position on the non-scaled axis, the proximity between two texts or groups of texts is measured by the height at which they are united.

All figures presented in the following are computed using the software R [12] and corpora presented in table 4.

GT separated from NT

Fig 4 shows that the method actually identifies the texts generated by the different PCFG and separates them from the natural ones. As a matter of fact, classification clearly separates the GT and the NT.

Table 4 Corpora

Corpus Name	Generator	Number of texts
SCIgen	SCIgen (PCFG)	12
Mathgen	Mathgen (PCFG)	11
scigen-physics	scigen-physics (PCFG)	12
Obamabot	Obamabot (Markov Chain)	2
Obama	Obama (Human)	5
CS	Computer science (Human)	12

For example, between the SCIgen texts and the natural ones, the average distance is 0.62, clearly outside the intervals of fluctuation for contemporaneous NT written in the same genre by different authors. Thus three obvious conclusions can be drawn. First, the SCIgen texts are very far from the NT they are supposed to imitate. Second, SCIgen is behaving as a single author who always faces the same subject in the same situation of utterance. Thirdly, intertextual distance combined with automatic clustering offers an effective tool for specific authorship attribution: the detection of GT (as the generators behave like a single author).

The position of President Obama's State of Union addresses is very interesting. First, these texts are more or less grouped at the same level as the GT's edges that indicates not only the same authorship for all the addresses, but also the fact that the complex elaboration of these speeches is not so far from those that could have been produced by a generator! Secondly the two *Obamabot* generated by software using Markov process are clustered with the texts they imitate. This result is not surprising since the real speeches constitued the training corpus. This highlights the obvious limitation of current procedures for GT detection (our own included): they work on the vocabulary (and frequencies of words) without dealing with the meaning of these texts (our conclusion discusses this problem). The main conclusion remains: the generators act as single authors dealing with the same themes and do not present the diversity of NT produced by different authors.

Most of the generated texts can be clearly distinguished from those written by humans, which is not the case with the Obamabot generated texts.

5 ROUGE measures

One of the most widespread evaluation approaches in NLP uses automatically computed metrics measuring $n-$gram coverage between a candidate output and one or several reference texts. In particular, in the domain of automatic summarisation, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [28] has become a reference method to evaluate summarisation systems (see the DUC and TAC conferences). Moreover, ROUGE is also a method used in the evaluation of some NLG system outputs [13] although it is unclear how well its results correlate with hu-

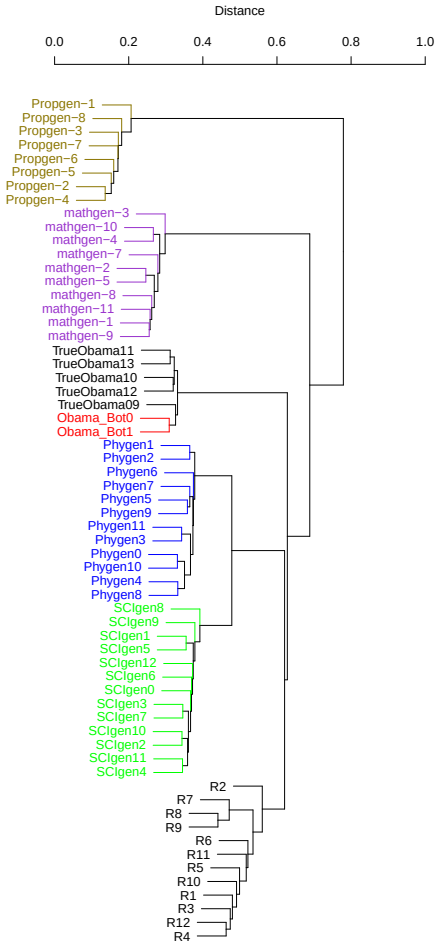


Fig. 3 Clustering each type of texts using inter-textual distance

man evaluations [37]. ROUGE was developed to compare automatic and human-authored extractive summaries. Here extractive means that the generated summary is composed only from the material 'extracted' from the original source(s). In short the summaries are —sometimes slightly modified— "extracts" of the main texts. Roughly speaking the metric aims at assessing how well the candidate summary covers the reference summaries using the frequency of the words and the sequence of the words. Thus ROUGE reflects the similarity between a GT and the gold standard reference. Therefore, we expect GT (resp. NT) to have a high ROUGE score within each group (i.e., high similarity).

ROUGE is not a unique measure but a set of measures whose simplest family is ROUGE-N which measures $n-$gram overlap. Equation (3) shows how ROUGE-N

$\in [0,1]$ is computed. Basically, it is the ratio of $n-$gram in the reference that are also found in the candidate. This is thus a recall measure.

$$\text{ROUGE-N}(ref, cand) = \frac{\sum_{S \in \{ref\}} \sum_{gram_n \in S} Count_{match}(gram_n, cand)}{\sum_{S \in \{ref\}} \sum_{gram_n \in S} Count(gram_n)} \quad (3)$$

Where $Count_{match}(gram_n, cand)$ is the number of occurrences of a $n-$gram of the reference in the candidate. In the present paper, we set n to 3 so ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-3 (trigram) were computed. We believe these will be particularly adapted to detect the texts generated using Markov chain models and the texts that share the same vocabulary (unigram, bigram).

ROUGE-L was also used. It finds the LCS(X,Y), the Longest Common Subsequence between the candidate X and the reference Y. The recall of this measure R_{LCS} for a candidate is computed using $R_{LCS} = \sum_{s \in R} LCS_{\cup}(s,C)/m$ where R is the set of sentences of the reference which contains m words and C the set of sentences of the candidate. The hypothesis is that ROUGE-L will be higher in GT by a PCFG since some of the textual properties at document and sentence level would be more recurrent than in GT by MC.

The ROUGE measures were computed using the ROUGE package [27] on the texts of the corpora (cf. section 2.2). Each pair of text was successively used as candidate and reference and the distance was computed using $1 - \text{F-measure}(ROUGE)$. Figure 5 shows the results for ROUGE-1 and ROUGE-L on the raw texts (Figure 5.a and 5.b) and without stop-words (Figure 5.c and 5.d). The measure enables a successful identification of clusters of uniform class. In the raw text case, the three scientific Corpora by PCFG are grouped together, while propgen, Obama, and NT scientific texts constitutes other groups. But the latter are highly discriminated. The NT scientific papers are at around .75 from the SCIgen papers, .74 from the propgen papers and .65 from the Obama speeches. Hence ROUGE seems to be a very efficient way of discriminating SCIgen from non-SCIgen papers. ROUGE-2 and ROUGE-3 measures were also computed but although they grouped the texts perfectly the measure intergroup was much higher which made the discrimination between much more difficult. This would suggest that GT contains a high amount of variation in sequence of tokens but not in vocabulary (cf section 3.3).

The ROUGE measures were also applied to texts from which stop-words have been removed. This processing is often performed for some types of Information retrieval and summarisation tasks in order to process only words that convey meaningful information[3]. The effect was to increase all the measures within groups making the discrimination between SCIgen and non-SCIgen a bit easier but clustering more scatered. This would support the hypothesis that stop-words play a role in the signature of the SCIgen papers [14]. However, the propgen texts are more clearly excluded from the other texts which makes this pre-processing quit interesting.

[3] This preprocessing is not systematic. For instance the lasts DUC conferences did include the stop-words in the ROUGE computing.

Overall, these results confirm the findings of the previous distance. 1) The NT are close to each other (Obama and scientific NT) while 2) the SCIgen papers are grouped together and propgen is considered as satellite. 3) The ROUGE metrics seems to be a very effective way of discriminating between SCIgen papers, propgen texts and NT. However, since they are based on frequency they are not able to detect Obama_Bot papers. Even the LCS measure was not informative in this respect but this might be a result of overfitting when the Language Model was learned. This call for further investigations.

6 Conclusion

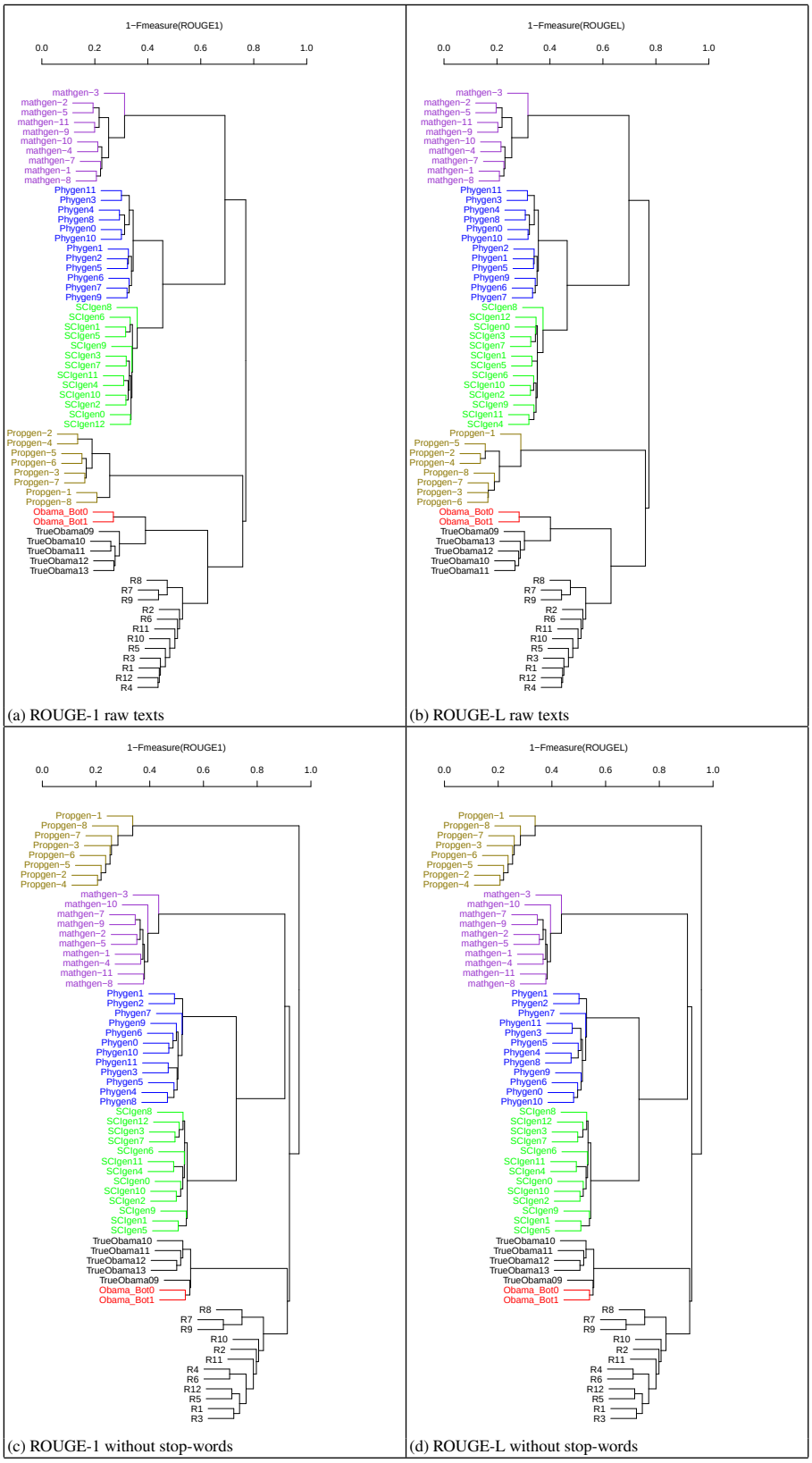
In conclusion, the two models of automatic text generation have still produced inconclusive results. The first model, using the Markow chains, emulates some characterics of the natural texts. As it has been mention by [11], because they are based on the lexical characteristics of a reference corpus, the texts of this first type are difficult to detect automatically. However, a human reader can discover them very easily, because generated texts do not follow the basics of natural language grammar.

The second model, like SCIgen, uses prebuilt context free grammars. The generated texts have poor repetitive vocabulary and their sentences are too short and too uniform compared to the natural ones they are supposed to emulate. Each one of these generators acts as a single author, their production easy to detect automatically. However, if the prebuilt elements are carefully chosen, these texts are more difficult to detect by a human reader, at least a routine and non-attentive reader, because they are conforming to the grammar of natural language.

Several improvements for detectors (and generators) can be made. First, by taking the context into account, it seems possible for a detector to check word usage according to their meaning [15, 21]. Secondly, a detector can also learn the real syntactic structures of a language and the specific sentence constructions and styles of the emulated authors [15].

These area of research will be of some help not only for fake paper detection, but also against plagiarism, duplication and other malpractices. More importantly, these improvements could be of great help regarding both understanding and generating natural texts. They will help improve software for processing and generating texts, for data scientists, lexicographers, translators, language teachers and all users of large digital text and data bases.

Acknowledgements The authors would like to thanks Edouard Arnold (Trinity College Dublin) for his valuable reading of previous versions of this paper as well as the organizers of the Flow Machines Workshop 2014 among which François Pachet, Mirko Degli Esposti and Vittorio Loreto.



(a) ROUGE-1 raw texts

(b) ROUGE-L raw texts

(c) ROUGE-1 without stop-words

(d) ROUGE-L without stop-words

Fig. 4 Clustering each type of texts using $1 - \text{Fmeasure}(ROUGE)$ as distance with and without stop-words

Appendix

Given two texts A and B, let us consider:

- N_A and N_B: the number of *word-tokens* in A and B respectively, ie the lengths of these texts;
- F_{iA} and F_{iB}: the absolute frequencies of a type i in texts A and B respectively;
- $|F_{iA} - F_{iB}|$ the absolute difference between the frequencies of a type i in A and B respectively;
- $D_{(A,B)}$: the inter-textual distance between A and B is as follows:

$$D_{(A,B)} = \sum_{i \in (A \cup B)} |F_{iA} - F_{iB}| \quad with \quad N_A = N_B \tag{4}$$

The distance index (or relative distance) is as follows:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - F_{iB}|}{N_A + N_B} \tag{5}$$

If the two texts are not of the same lengths in tokens ($N_A < N_B$), B is "reduced" to the length of A:

- $U = \frac{N_A}{N_B}$ is the proportion used to reduce B in B'
- $E_{iA(u)} = F_{iB}.U$ is the theoretical frequency of a type i in B'

In the Equation 4, the absolute frequency of each word-type in B is replaced by its theoretical frequency in B':

$$D_{(A,B')} = \sum_{i \in (A \cup B)} |F_{iA} - E_{iA(u)}|$$

Putting aside rounding-offs, the sum of these theoretical frequencies is equal to the length of A. The Equation 5 becomes:

$$D_{rel(A,B)} = \frac{\sum_{i \in (A \cup B)} |F_{iA} - E_{iA(u)}|}{N_A + N_{B'}}$$

References

1. Arnold, E.: Le discours de tony blair (1997-2004). Corpus **4**, 55–77 (2005)
2. Arnold, E.: Le sens des mots chez tony blair (people et europe). In: H. Serge, P. Bénédicte (eds.) 9e Journées internationales d'analyse statistique des données textuelles, vol. 1, pp. 109–119. Presses universitaires de Lyon (2008)
3. Ball, P.: Computer conference welcomes gobbledegook paper. Nature **434, 946** (2005)
4. Balpe, J.P.: Fiction et écriture générative. les Actes de Lecture (103), 37–48 (2008)
5. Barbieri, G., Pachet, F., Roy, P., Esposti, M.D.: Markov constraints for generating lyrics with style. In: ECAI, vol. 242, pp. 115–120 (2012)

6. Baughn, J.: (2001). URL http://nonsense.sourceforge.net. [Online; accessed 11-December-2014]
7. Chomsky, N.: Three models for the description of language. IEEE Transactions on Information Theory **2**(2), 113–124 (1956)
8. Dalkilic, M.M., Clark, W.T., Costello, J.C., Radivojac, P.: Using compression to identify classes of inauthentic texts. In: Proceedings of the 2006 SIAM Conference on Data Mining (2006)
9. Doug, C., Jan, P., Penelope, S.: A practical part-of-speech tagger. In: ANLC '92 Proceedings of the third conference on Applied Natural Language, pp. 133–140 (1992)
10. Elmacioglu, E., Lee, D.: Oracle, where shall i submit my papers? Communications of the ACM (CACM) **52**(2), 115–118 (2009)
11. Fahrenberg, U., Biondi, F., Corre, K., Jégourel, C., Kongshøj, S., Legay, A.: Measuring structural distances between texts. CoRR **abs/1403.4024** (2014)
12. Feinerer, I., Hornik, K., Meyer, D.: Text mining infrastructure in r. Journal of Statistical Software **25**(5), 1–54 (2008)
13. Gatt, A., Portet, F.: Textual properties and task-based evaluation: Investigating the role of surface properties, structure and content. In: 6th International Conference on Natural Language Generation (INLG-10) (2010)
14. Ginsparg, P.: Automated screening: Arxiv screens spot fake papers. Nature **508**(7494), 44–44 (2014). URL http://dx.doi.org/10.1038/508044a
15. Halliday, M.A.K., Webster, J.: Computational and quantitative studies / M.A.K. Halliday ; edited by Jonathan Webster. Continuum New York ; London (2004)
16. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Science **102**, 16,569–16,572 (2005)
17. Hockey, S., Martin, J.: OCP Users' Manual. Oxford. Oxford University Computing Service (1988)
18. Hubert, P., Labbé, D.: Vocabulary richness. In: Communication au congrès de l'ALLC-ACH. Paris: La Sorbonne. Reproduced in Lexicometrica, 1997 (1994)
19. John, S.: Corpus, Concordance, Collocation. Oxford: Oxford University Press (1991)
20. Labbé, C.: Ike antkare, one of the great stars in the scientific firmament. International Society for Scientometrics and Informetrics Newsletter **6**(2), 48–52 (2010)
21. Labbé, C., Labbé, D.: How to measure the meanings of words? amour in corneille's work. Language Resources and Evaluation **39**(4), 335–351 (2005)
22. Labbé, C., Labbé, D.: Duplicate and fake publications in the scientific literature: how many scigen papers in computer science? Scientometrics **94**(1), 379–396 (2013)
23. Labbé, C., Labbé, D.: Was shakespeare's vocabulary the richest? In: Proceedings of the 12th International Conference on Textual Data Statistical Analysis, pp. 323–336. Paris (2014)
24. Labbé, C., Portet, F.: Towards an abstractive opinion summarisation of multiple reviews in the tourism domain. In: SDAD 2012, The 1st International Workshop on Sentiment Discovery from Affective Data, pp. 87–94 (2012)
25. Labbé, D.: Experiments on authorship attribution by intertextual distance in english. Journal of Quantitative Linguistics **14**(1), 33–80 (2007)
26. Lavoie, A., Krishnamoorthy, M.: Algorithmic Detection of Computer Generated Text. ArXiv e-prints (2010)
27. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81 (2004)
28. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of HLT-NAACL-03, pp. 71–78 (2003)
29. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
30. Manurung, R., Ritchie, G., Pain, H., Waller, A., O'Mara, D., Black, R.: The construction of a pun generator for language skills development. Applied Artificial Intelligence **22**(9), 841–869 (2008)
31. Monière, D., Labbé, C., Labbé, D.: Les styles discursifs des premiers ministres québécois de jean lesage à jean charest. Revue canadienne de science politique **41**(1), 43–69 (2008)

32. Nadovich, C.: Automatic sbir proposal generator (2014). URL http://www.nadovich.com/chris/randprop/. [Online; accessed 11-December-2014]
33. Nathaniel, E.: (2012). URL http://thatsmathematics.com/mathgen/. [Online; accessed 11-December-2014]
34. Noorden, R.V.: Publishers withdraw more than 120 gibberish papers. Nature (24 February 2014)
35. Parnas, D.L.: Stop the numbers game. Commun. ACM **50**(11), 19–21 (2007)
36. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence **173**(7–8), 789–816 (2009)
37. Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics **35**(4), 529–558 (2009)
38. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Studies in Natural Language Processing. Cambridge University Press (2000)
39. Sneath, P., Sokal, R.: Numerical Taxonomy. San Francisco : Freeman (1973)
40. Stribling, J., Krohn, M., Aguayo, D.: Scigen (2005). URL http://pdos.csail.mit.edu/scigen/. [Online; accessed 11-December-2014]
41. Stubbs, M.: Texts and corpus analysis. Oxford: Blackwell (1996)
42. unknown: (2014). URL https://bitbucket.org/birkenfeld/scigen-physics. [Online; accessed 11-December-2014]