Engineering a Tool to Detect Automatically Generated Papers

Nguyen Minh Tien - Cyril Labbé

University Grenoble Alpes

Outline

Outline

Automatically Scientific Paper Generators

Automatically Scientific Paper Generators

- SCIgen

In recent years, much research has been devoted to the construction of cache coherence; contrarily, few have emulated the study of the memory bus. Here, we show the deployment of XML. in this work, we argue that the seminal virtual algorithm for the theoretical unification of hierarchical databases and fiber-optic cables by Wu and Brown is in Co-NP.

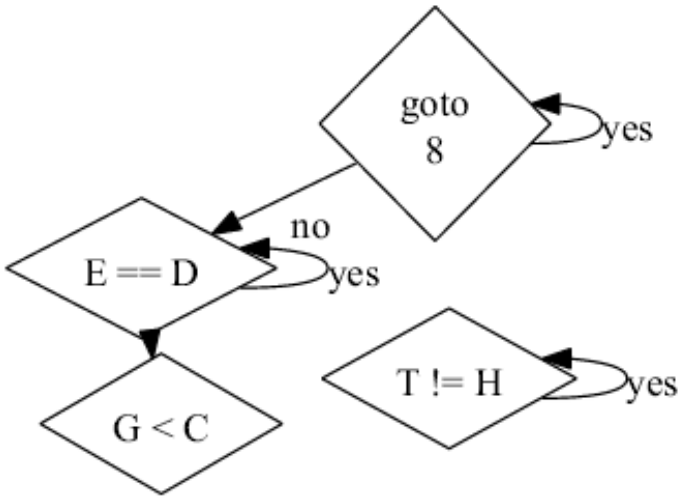


Fig. 1. Toe's robust study.

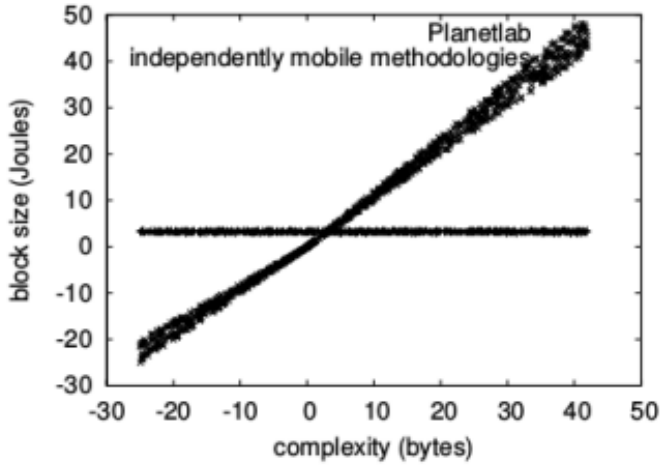


Fig. 4. Note that work factor grows as interrupt rate decreases – a phenomenon worth simulating in its own right.

Automatically Scientific Paper Generators

- SCIgen- Physic

Nanotubes and broken symmetries, while private in theory, have not until recently been considered appro- priate. After years of unproven research into the critical temperature, we verify the development of frustrations. Our focus in this work is not on whether superconduc- tors and neutrons can interfere to answer this quagmire, but rather on constructing an analysis of Goldstone bosons (Sump).

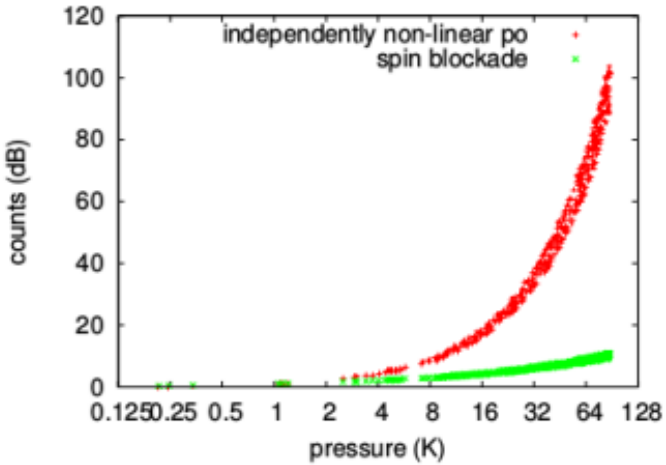


Fig. 4. The differential energy transfer of Sump, compared with the other frameworks.

$$O_J(\vec{r}) = \iiint d^3r \sqrt{\frac{f_o \vec{\zeta}^2}{\pi^2 \tilde{S} 1 k^2}} + \frac{\partial \psi}{\partial \vec{t}} + \frac{\partial \psi}{\partial \beta} + \pi^5 \cdot 6 \cdot \cos\left(\frac{\triangle 5 B_\Psi 7 \vec{w}^2}{{k_\delta}^5} - \ln[\pi]\right). \tag{3}$$

Automatically Scientific Paper Generators

- Mathgen

Abstract. Let $i_{L,\mathcal{R}} \neq w^{(L)}$. A central problem in probabilistic dynamics is the construction of connected, continuous rings. We show that $S \leq \phi$. The work in [36] did not consider the multiply hyper-Laplace, non-intrinsic case. A useful survey of the subject can be found in [36].

Trivially,

$$\begin{aligned}\overline{ani} &\geq \bigcap_{\bar{\mathbf{y}}=1}^{-\infty} \aleph_0 \cap -\infty \cup \cdots - \frac{1}{-1} \\ &= \int_{e}^{\pi} \sup \tanh\left(\mathscr{M}\left(\mathscr{K}_D\right)^2\right)\, dQ \cup \cos^{-} \\ &\equiv \sum_{\bar{w}=\pi}^{i} a^{-1}\left(0^{-2}\right) \\ &< \coprod_{\check{L}=\emptyset}^{1} \mathfrak{e}''\left(\frac{1}{\Xi^{(U)}}, \mathfrak{u}^{(v)^5}\right) \wedge \cdots \cdot \mathfrak{i}''.\end{aligned}$$

$$\begin{aligned}S_{\mathscr{Z},\Delta}\left(-1^{-8}, J^{-7}\right) &\in \int_{M} \sin\left(-\emptyset\right)\, d\mathcal{H} \wedge \Delta^{-3} \\ &\equiv B^{-1} - \overline{-\infty \vee \infty} \cap l_O\left(\mathscr{U}, D\right) \\ &= \min_{\Psi \to -\infty} \mathfrak{n}\left(-\sqrt{2}\right) \times \tan\left(\tilde{N} e\right).\end{aligned}$$

Automatically Scientific Paper Generators

- Proposal Generator (Propgen)

Technical Abstract

The technology in effectively addresses the oscillator causing the object-oriented countermeasure by applying a complementary scintillation that crashes. This technology will provide with a fiberoptic realizability. Has years of experience in the parabolically inverse scintillation and has built and delivered a Gaussian peripheral. Other solutions to the the object-oriented countermeasure, such as a shipboard ambiguity, do not address the oscillator in an efficient manner. The successful development of will result in numerous spinoffs onto a simultaneously electromagnetic handwheel for the benefit of all people in the world.

Key Words

groundwave	telemetry	radiolocation
crosscorrelation	matrix	VHF
schematic	feasibility	efficiency

1. A symmetric managerial
2. A superset
3. A Nyquist eigenproblem that varies monolithically
4. The for the downloadable submatrix parallel discriminator that destabalize
5. The eraseable beamformer

Automatically Scientific Paper Generators

- Text extracted from a SCIgen paper.

Efficient, Read-Write Modalities for Telephony

The e-voting technology approach to simulated annealing is defined not only by the exploration of reinforcement learning, but also by the key need for neural networks. After years of robust research into journaling file systems, we validate the understanding of semaphores, which embodies the extensive principles of electrical engineering. In order to accomplish this objective, we use low-energy theory to show that compilers and object-oriented languages can interact to solve this obstacle.

- Most generators use probabilistic context free grammar

```
The SCI_FIELD SCI_APPROACH to SCI_THING_MOD is defined not only by the
                                SCI_ACT ...
Many SCI_PEOPLE would agree that, had it not been for SCI_GENERIC_NOUN ...
SCI_BUZZWORD_ADJ SCI_BUZZWORD_NOUN and SCI_THING_MOD have garnered LIT_GREAT
SCI_THING_MOD and SCI_THING_MOD, while SCI_ADJ in theory, have not until ...
                        The SCI_ACT is a SCI_ADJ SCI_PROBLEM
```

Fake Paper Detections

- Reference checking [Xiong and Huang, 2009].
  - Check references using online search engine.
  - Simple but easily fooled.

Fake Paper Detections

- Reference checking [Xiong and Huang, 2009].
  - Check references using online search engine.
  - Simple but easily fooled.
- Ad-hoc similarity measure [Lavoie and Krishnamoorthy, 2010].
  - Score based on keywords, word repetition and references.
  - Expensive POS tagging, highly depend on the length of the paper.

Fake Paper Detections

- Reference checking [Xiong and Huang, 2009].
  - Check references using online search engine.
  - Simple but easily fooled.
- Ad-hoc similarity measure [Lavoie and Krishnamoorthy, 2010].
  - Score based on keywords, word repetition and references.
  - Expensive POS tagging, highly depend on the length of the paper.
- Compression factor [Dalkilic et al., 2006].
  - each paper is compressed using Lempel-Ziv and Bender-Wolf algorithms and characterize by a compression profile.
  - Compressibility rate of generated paper is incompatible with genuine ones.

Fake Paper Detections

- Reference checking [Xiong and Huang, 2009].
  - Check references using online search engine.
  - Simple but easily fooled.
- Ad-hoc similarity measure [Lavoie and Krishnamoorthy, 2010].
  - Score based on keywords, word repetition and references.
  - Expensive POS tagging, highly depend on the length of the paper.
- Compression factor [Dalkilic et al., 2006].
  - each paper is compressed using Lempel-Ziv and Bender-Wolf algorithms and characterize by a compression profile.
  - Compressibility rate of generated paper is incompatible with genuine ones.
- Topological properties [Amancio, 2015].
  - Papers are represented as network features using word adjacency model.
  - Patterns generated papers are different from the structural patterns from real texts.

Fake Paper Detections

- Reference checking [Xiong and Huang, 2009].
  - Check references using online search engine.
  - Simple but easily fooled.
- Ad-hoc similarity measure [Lavoie and Krishnamoorthy, 2010].
  - Score based on keywords, word repetition and references.
  - Expensive POS tagging, highly depend on the length of the paper.
- Compression factor [Dalkilic et al., 2006].
  - each paper is compressed using Lempel-Ziv and Bender-Wolf algorithms and characterize by a compression profile.
  - Compressibility rate of generated paper is incompatible with genuine ones.
- Topological properties [Amancio, 2015].
  - Papers are represented as network features using word adjacency model.
  - Patterns generated papers are different from the structural patterns from real texts.
- Structural distance based on textual content [Fahrenberg et al., 2014][Labbé and Labbé, 2013].

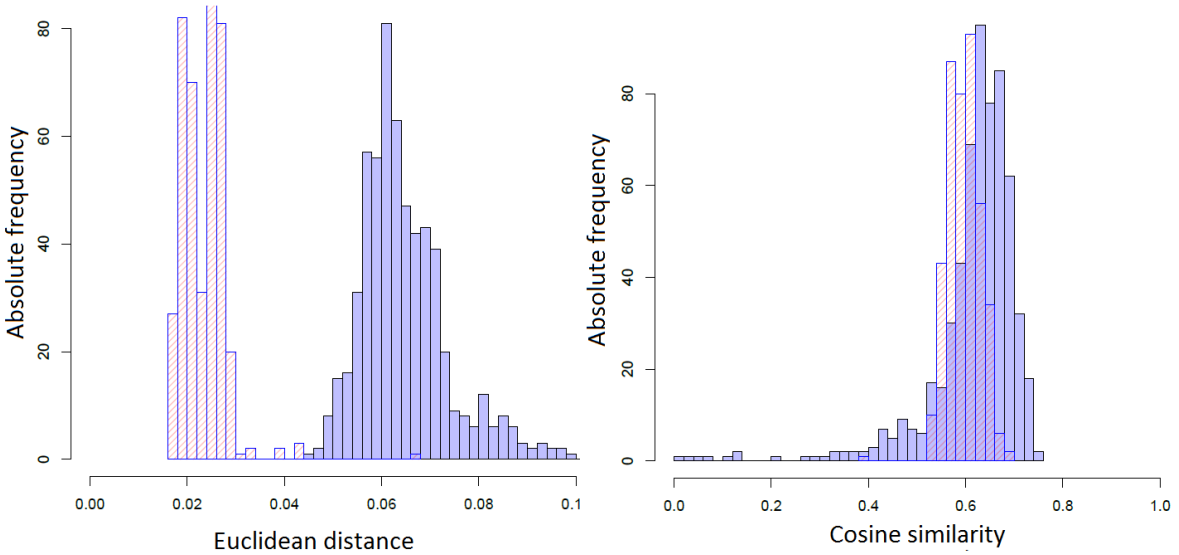
Outline

Distance and Similarity Measurements

- Nearest neighbour classification to a sample corpus of 400 generated texts.

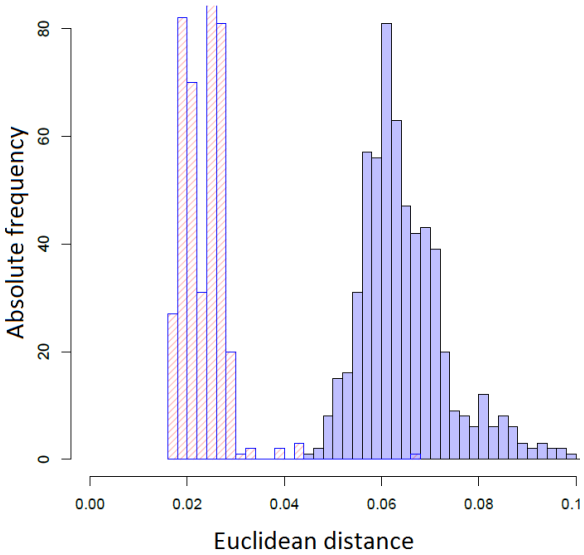
Distance and Similarity Measurements

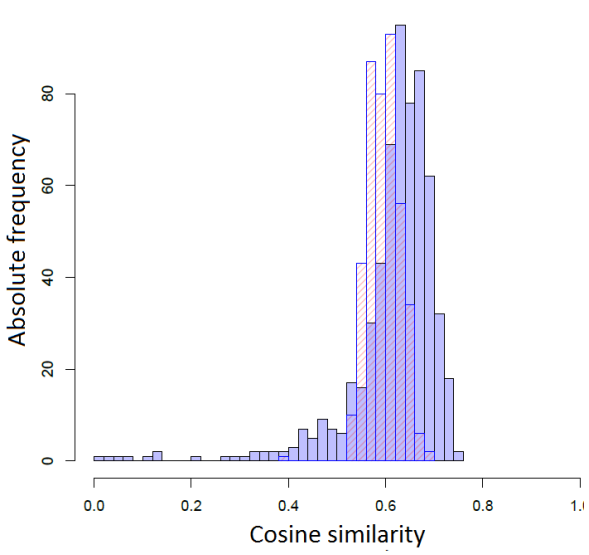
- Nearest neighbour classification to a sample corpus of 400 generated texts.
- Test corpus of 400 generated texts in red.
- Test corpus of genuine human written texts in blue.
- Text is considered as a vector of absolute word frequency.

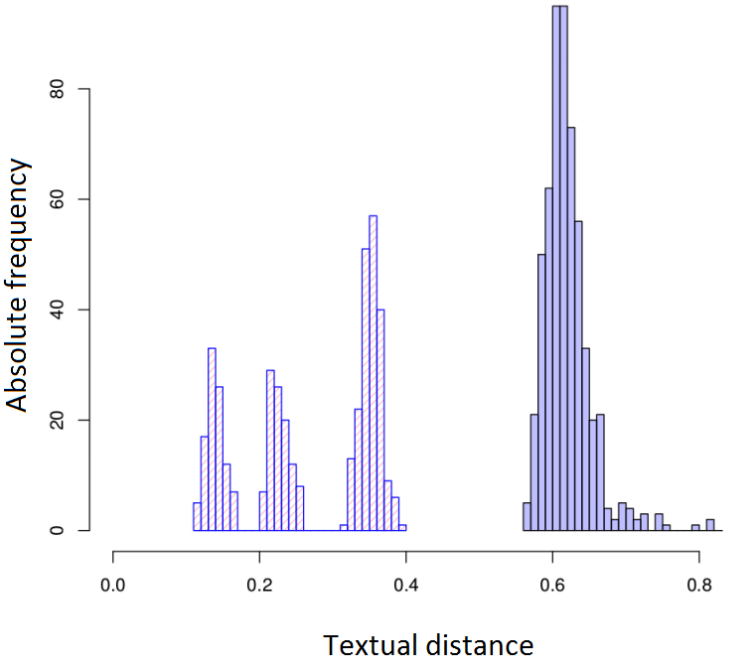


Distance and Similarity Measurements

- Nearest neighbour classification to a sample corpus of 400 generated texts.
- Test corpus of 400 generated texts in red.
- Test corpus of genuine human written texts in blue.
- Text is considered as a vector of absolute word frequency.







Outline

SciDetect: A Tool to Detect Automatically Generated Paper

- Inter-textual distance using all the words.
- Distance to the nearest neighbour in a sample corpus of 400 generated texts.

SciDetect: A Tool to Detect Automatically Generated Paper

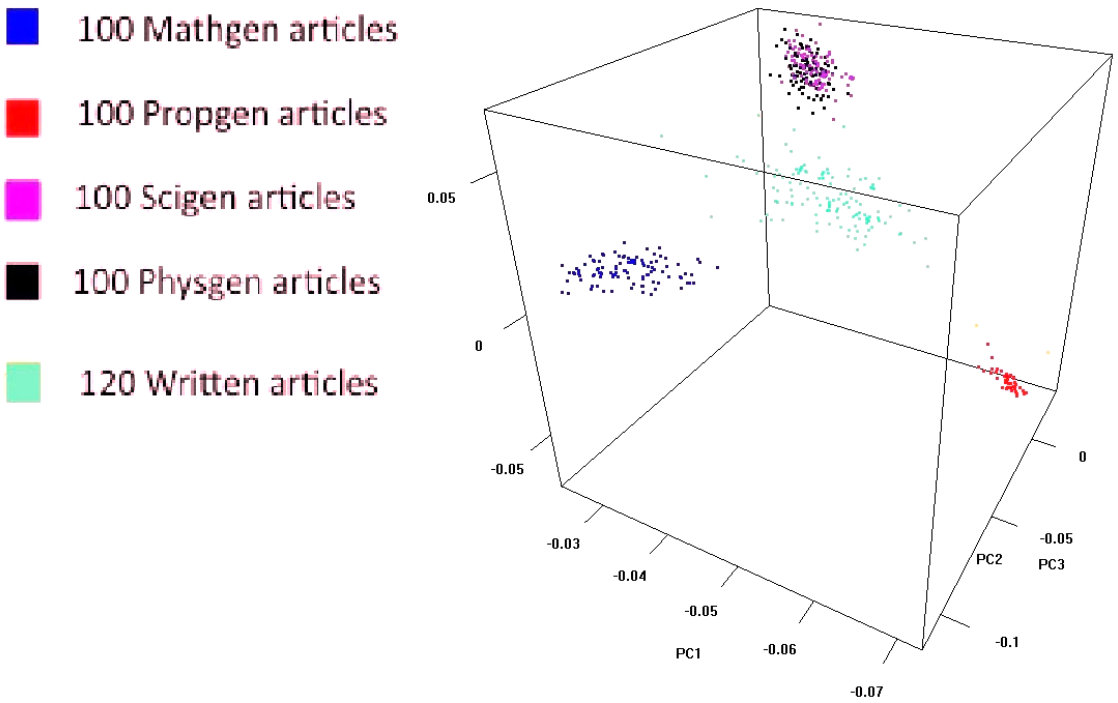
- Inter-textual distance using all the words.
- Distance to the nearest neighbour in a sample corpus of 400 generated texts.
- Thresholds for each generator was set using 400 generated texts and 8200 genuine texts.

	Scigen	Physgen	Mathgen	Propgen	Genuine
Min distance to NN	0.30	0.31	0.19	0.11	**0.52**
Max distance to NN	**0.40**	0.39	0.28	0.22	0.99
Standard deviation	0.014	0.012	0.014	0.015	0.117

Outline

Test Candidates - Kullback-Leibler divergence

- Seems to be currently used by ArXiv [Ginsparg, 2014].
- Frequency distributions of stop words.



- Using stop words. Sample corpora of 400 generated papers. Nearest neighbour classification with thresholds.

Other Candidates and Test corpora

Test candidates

- Pattern Matching: Score to familiar pattern of more than 5 words. Thresholds for suspicious and confirm generated paper. Only trained for SCIgen.
- SciDetect: Textual distance. Same sample corpora and classification method as Kullback-Leibler divergence.

Other Candidates and Test corpora

Test candidates

- Pattern Matching: Score to familiar pattern of more than 5 words. Thresholds for suspicious and confirm generated paper. Only trained for SCIgen.
- SciDetect: Textual distance. Same sample corpora and classification method as Kullback-Leibler divergence.

Test corpora

- Corpus X: 100 texts from known generators without any modification.
- Corpus Y: 100 generated texts (25 from each generator) that have been modified by randomly changing a word every two to nine words.
- Corpus Z: 10.000 real texts with different length.

Comparative Evaluation Between Different Methods

Result:

- True Positive and True Negative: Correctly identified generated or genuine paper respectively.
- False Positive: genuine paper but have been identified as generated.
- False Negative: generated paper but have been identified as genuine.

method	corpus	True Positive		False Positive		True Negative	False Negative
		confirm	suspect	confirm	suspect		
Pattern Matching	X	25%	4%	0	0	0	71%
	Y	8%	16%	0	0	0	76%
	Z	0	0	0	0.01%	99.99%	0
Kullback-Leibler Divergence	X	87%	13%	0	0	0	0
	Y	79%	21%	0	0	0	0
	Z	0	0	0	1.65%	98.35%	0
SciDetect	X	100%	0	0	0	0	0
	Y	100%	0	0	0	0	0
	Z	0	0	0	0	100%	0

Outline

Conclusion

- A need for automatic detection of computer generated papers in scientific literature.
- Several ways to accomplish such task.
- Textual distance and namely SciDetect were proven to be the most reliable method for classification.

Conclusion

- A need for automatic detection of computer generated papers in scientific literature.
- Several ways to accomplish such task.
- Textual distance and namely SciDetect were proven to be the most reliable method for classification.

Limitations

- Can not detect texts from unknown generators.
- Can not detect a small generated part.

Conclusion

- A need for automatic detection of computer generated papers in scientific literature.
- Several ways to accomplish such task.
- Textual distance and namely SciDetect were proven to be the most reliable method for classification.

Limitations

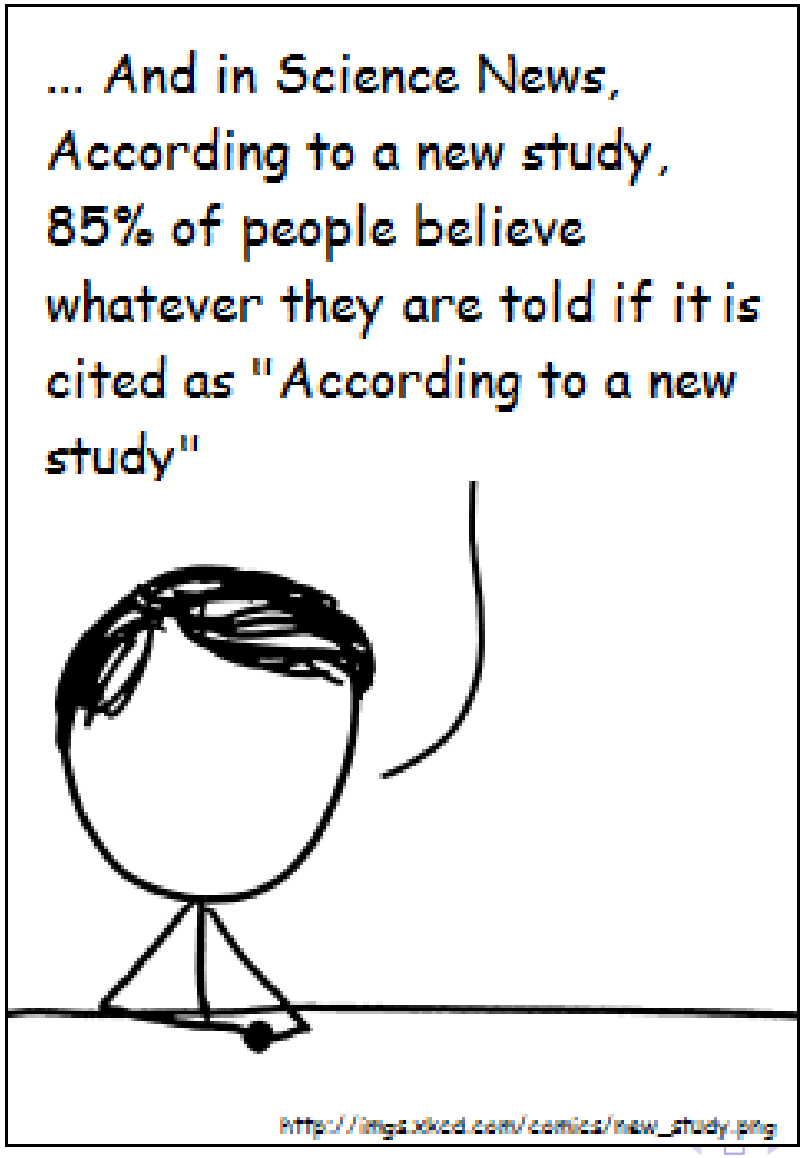
- Can not detect texts from unknown generators.
- Can not detect a small generated part.

Future works

- Checking the meaning of words [Labbé and Labbé, 2005].
- Styles of generated texts [Kollmer et al., 2015].
- Vocabulary size and keywords repetition.

Questions?

Thanks You!



References I


Amancio, D. R. (2015).
Comparing the topological properties of real and artificially generated scientific manuscripts.
Scientometrics, 105(3):1763–1779.


Dalkilic, M. M., Clark, W. T., Costello, J. C., and Radivojac, P. (2006).
Using compression to identify classes of inauthentic texts.
In *Proc. of the 2006 SIAM Conf. on Data Mining*.


Fahrenberg, U., Biondi, F., Corre, K., Jégourel, C., Kongshøj, S., and Legay, A. (2014).
Measuring global similarity between texts.
In *Second International Conference, SLSP*, pages 220–232.

References II


Ginsparg, P. (2014).
Automated screening: ArXiv screens spot fake papers.
- 508(- 7494):– – 44.


Kollmer, J. E., Pöschel, T., and Gallas, J. A. (2015).
Are physicists afraid of mathematics?
New Journal of Physics, 17(1):013036.


Labbé, C. and Labbé, D. (2005).
How to measure the meanings of words? amour in corneille's work.
Language Resources and Evaluation, 39(4):335–351.


Labbé, C. and Labbé, D. (2013).
Duplicate and fake publications in the scientific literature: How many scigen papers in computer science?
Scientometrics, 94(1):379–396.

References III


Lavoie, A. and Krishnamoorthy, M. (2010).
Algorithmic detection of computer generated text.
arXiv preprint arXiv:1008.0706.


Xiong, J. and Huang, T. (2009).
An effective method to identify machine automatically generated paper.
In *Knowledge Engineering and Software Engineering*, pages 101–102.