Table of Contents

1 Motivation
2 The tool
3 Experimental evaluation
2005: SCIgen, a nonsense CS papers generator

SCIgen - An Automatic CS Paper Generator

About Generate Examples Talks Code Donations Related People Blog

About

SCIgen is a program that generates random Computer Science research papers, including graphs, figures, and citations. It uses a hand-written context-free grammar to form all elements of the papers. Our aim here is to maximize amusement, rather than coherence.

One useful purpose for such a program is to auto-generate submissions to conferences that you suspect might have very low submission standards. A prime example, which you may recognize from spam in your inbox, is SCI/IIFS and its dozens of co-located conferences (check out the very broad conference description on the WMSCI 2005 website). There's also a list of known bogus conferences. Using SCIgen to generate submissions for conferences like this gives us pleasure to no end. In fact, one of our papers was accepted to SCI 2005! See Examples for more details.

We went to WMSCI 2005. Check out the talks and video. You can find more details in our blog.

Also, check out our 10th anniversary celebration project: SClpher!

Generate a Random Paper

Want to generate a random CS paper of your own? Type in some optional author names below, and click "Generate".

Author 1:
Author 2:
Author 3:
Author 4:
Author 5:

Generate Ripristina

SCIgen currently supports Latin-1 characters, but not the full Unicode character set.
Thanks, SClgen. I wrote a paper!

RPCs Considered Harmful

Alice Medvet, Andrea Medvet and Eric Medvet

Abstract
System administrators agree that probabilistic models are an interesting new topic in the field of networking, and information theorists concur. After years of practical research into neural networks, we disconfirm the construction of RAID, which embodies the essential principles of software engineering. This follows from the emulation of IPv6. Charon, our new methodology for checksums, is the solution to all of these obstacles.

1 Introduction
The implications of secure symmetries have been far-reaching and pervasive. To put this in per-

2 Related Work
Primarily, we motivate the need for SCSI disks. Next, to surmount this obstacle, we explore a scalable tool for improving DHTs (Charon), demonstrating that thin clients and cache coherence can cooperate to overcome this quagmire. Finally, we conclude.
What after SClgen?

Born as a toy: “We were fed up with all of the bogus journals and conferences that spam researchers and charge crazy fees for articles they don’t even read before accepting. So we created SClgen.”
What after SCIgen?

Born as a toy: “We were fed up with all of the bogus journals and conferences that spam researchers and charge crazy fees for articles they don’t even read before accepting. So we created SCIgen.”
Why?

- Pressure: publish or perish!
- Incentives: research evaluation driven by quantity rather than quality
- Opportunities: predatory journals, conferences

Pressure + Incentives + Opportunities + Misconduct + SCIIgen =

Image: The Guardian article titled "How computer-generated fake papers are flooding academia." The article discusses how more and more academic papers that are essentially gobbledygook are being written by computer programs and accepted at conferences.
The role of review(ers)

Ideally:

- assess quality, novelty, impact of research
- facilitate/determine distribution of limited resources: time, funds, space in journals
- a keystone in scholarly publishing!
The role of review(er)s

Ideally:

- assess quality, novelty, impact of research
- facilitate/determine distribution of limited resources: time, funds, space in journals
- a keystone in scholarly publishing!

But, on reviewer side:

- uncredited reviews, “credited” PC/board participation
- overwork
The role of review(er)s

Ideally:

- assess quality, novelty, impact of research
- facilitate/determine distribution of limited resources: time, funds, space in journals
- a keystone in scholarly publishing!

But, on reviewer side:

- uncredited reviews, “credited” PC/board participation
- overwork

And, on publisher side:

- profit!
Review: pressure, incentives, opportunities

What if a tool for automatic generation of reviews existed? (Reviews at no cost!)

- “scholars” could take part in many PC/boards without bothering in review activities
- predatory journals could inflate their credibility by sending many reviews to authors
- ...
Motivation

Review: pressure, incentives, opportunities

What if a tool for automatic generation of reviews existed? (Reviews at no cost!)

- “scholars” could take part in many PC/boards without bothering in review activities
- predatory journals could inflate their credibility by sending many reviews to authors

Is that tool feasible?

We built that tool!
Table of Contents

1. Motivation

2. The tool

3. Experimental evaluation
Problem statement

- Paper \( a \)
- Recommendation \( o \)
- Review generator
- Review \( r \)

\( r \) should:
- appear as generated by humans
- be specific for paper \( a \)
- express an overall recommendation \( o \) (reject, neutral, accept)
Problem statement

- appear as generated by humans
- be specific for paper $a$
- express an overall recommendation $o$ (reject, neutral, accept)

A corpus $R$ of real reviews is available.
Method overview

Given the input $a, o$:

1. use sentences in $R$
2. identify and replace specific terms
3. perform sentiment analysis
4. reorder sentences
Method overview

Given the input $a, o$:

1. use sentences in $R$
2. identify and replace specific terms
3. perform sentiment analysis
4. reorder sentences

Goals:
- appear as generated by humans
Method overview

Given the input $a, o$:

1. use sentences in $R$
2. identify and replace specific terms
3. perform sentiment analysis
4. reorder sentences

Goals:

- appear as generated by humans
- be specific for paper $a$
Method overview

Given the input $a, o$:

1. use sentences in $R$
2. identify and replace specific terms
3. perform sentiment analysis
4. reorder sentences

Goals:

- appear as generated by humans
- be specific for paper $a$
- express an overall recommendation $o$
Preprocessing

Only once, for each review $r$ in $R$:

- tokenize
- Named-entity recognition (NER)
- Part-of-speech annotation (POS)

At the end, each token of each sentence associated with one NER and one POS tag.

For the input paper $a$:

- concatenate title, abstract, and main content
- tokenize, NER, and POS
Specific terms identification

**Goal:** given a document $d$, find terms which are specific of $d$. 
Specific terms identification

Goal: given a document $d$, find terms which are specific of $d$.
A term is specific if:

- POS tag is noun or adjective
- length is $\geq 2$
- contains at least a letter
Specific terms identification

**Goal:** given a document $d$, find terms which are specific of $d$.

A term is specific if:
- POS tag is noun or adjective
- length is $\geq 2$
- contains at least a letter

**Example:** The problem has a multiobjective nature, we want a regular expression able to [...]
Specific terms replacement

**Goal:** given a review $r$ and a paper $a$, replace specific terms in $r$ with specific terms of $a$. 

1. Find specific terms $W_r$ in $r$ and specific terms $W_a$ in $a$.
2. $W_r := W_r \setminus W_a$; $W_a := W_a \setminus W_r$.
3. Split $r$ in sentences and, for each sentence $s$:
   1. Map $W_s$ to $W_a$, preserving POS and NER tags.
   2. If mapping not possible, discard $s$.
   3. Otherwise, replace mapped terms in $s$ and add $s$ to a set $S$.

**Example:**

Review $r$ in $R$ ("Inference of Regular Expressions for Text Extraction from Examples")

The problem has a multiobjective nature, we want a regular expression able [...]

Paper $a$ ("Pre-hospital delay in Vietnamese patients hospitalized with a first acute myocardial infarction: A short report")

Bartoli, De Lorenzo, Medvet, Tarlao (UniTs) Generation of Scientific Paper Reviews August 31st, 2016 14 / 26
Specific terms replacement

**Goal:** given a review $r$ and a paper $a$, replace specific terms in $r$ with specific terms of $a$.

1. Find specific terms $W_r$ in $r$ and specific terms $W_a$ in $a$
2. $W_r := W_r \setminus W_a$; $W_a := W_a \setminus W_r$
3. Split $r$ in sentences and, for each sentence $s$:
   1. Map $W_r^s$ to $W_a$, preserving POS and NER tags
   2. If mapping not possible, discard $s$
   3. Otherwise, replace mapped terms in $s$ and add $s$ to a set $S$
Specific terms replacement

**Goal:** given a review $r$ and a paper $a$, replace specific terms in $r$ with specific terms of $a$.

1. Find specific terms $W_r$ in $r$ and specific terms $W_a$ in $a$
2. $W_r := W_r \setminus W_a; W_a := W_a \setminus W_r$
3. Split $r$ in sentences and, for each sentence $s$:
   1. Map $W_r^s$ to $W_a$, preserving POS and NER tags
   2. If mapping not possible, discard $s$
   3. Otherwise, replace mapped terms in $s$ and add $s$ to a set $S$

**Example:**
Review $r$ in $R$ (“Inference of Regular Expressions for Text Extraction from Examples”)
The problem has a multiobjective nature, we want a regular expression able to [...] 

Paper $a$ (“Pre-hospital delay in Vietnamese patients hospitalized with a first acute myocardial infarction: A short report”)
Specific terms replacement

**Goal:** given a review \( r \) and a paper \( a \), replace specific terms in \( r \) with specific terms of \( a \).

1. Find specific terms \( W_r \) in \( r \) and specific terms \( W_a \) in \( a \)
2. \( W_r := W_r \setminus W_a \); \( W_a := W_a \setminus W_r \)
3. Split \( r \) in sentences and, for each sentence \( s \):
   1. Map \( W_s^r \) to \( W_a \), preserving POS and NER tags
   2. If mapping not possible, discard \( s \)
   3. Otherwise, replace mapped terms in \( s \) and add \( s \) to a set \( S \)

**Example:**

Review \( r \) in \( R \) ("Inference of Regular Expressions for Text Extraction from Examples")

The problem has a multiobjective nature, we want a regular expression able to [...]

Paper \( a \) ("Pre-hospital delay in Vietnamese patients hospitalized with a first acute myocardial infarction: A short report")

\( W_a = \{ \text{myocardial, middle, cause, problem, nature, ...} \} \)
Specific terms replacement

**Goal:** given a review $r$ and a paper $a$, replace specific terms in $r$ with specific terms of $a$.

1. Find specific terms $W_r$ in $r$ and specific terms $W_a$ in $a$
2. $W_r := W_r \setminus W_a$; $W_a := W_a \setminus W_r$
3. Split $r$ in sentences and, for each sentence $s$:
   1. Map $W_r^s$ to $W_a$, preserving POS and NER tags
   2. If mapping not possible, discard $s$
   3. Otherwise, replace mapped terms in $s$ and add $s$ to a set $S$

**Example:**

Review $r$ in $R$ (“Inference of Regular Expressions for Text Extraction from Examples”)

The problem has a **multiobjective** nature, we want a **regular expression** able to [...]  

Paper $a$ (“Pre-hospital delay in Vietnamese patients hospitalized with a first acute myocardial infarction: A short report”)

$W_a = \{\text{myocardial, middle, cause, ...} \}$
Specific terms replacement

**Goal:** given a review $r$ and a paper $a$, replace specific terms in $r$ with specific terms of $a$.

1. Find specific terms $W_r$ in $r$ and specific terms $W_a$ in $a$
2. $W_r := W_r \setminus W_a$; $W_a := W_a \setminus W_r$
3. Split $r$ in sentences and, for each sentence $s$:
   1. Map $W_s^r$ to $W_a$, preserving POS and NER tags
   2. If mapping not possible, discard $s$
   3. Otherwise, replace mapped terms in $s$ and add $s$ to a set $S$

**Example:**

Review $r$ in $R$ (“Inference of Regular Expressions for Text Extraction from Examples”)
The problem has a myocardial nature, we want a middle cause able to [...] 

Paper $a$ (“Pre-hospital delay in Vietnamese patients hospitalized with a first acute myocardial infarction: A short report”) $W_a = \{\text{myocardial, middle, cause, ...}\}$
Goal: retain only sentences in $S$ which are consistent with recommendation $o$. 

Assumption: “accept” reviews contain only positive sentences, “reject” contain negative, ...
Sentiment analysis

**Goal:** retain only sentences in $S$ which are consistent with recommendation $o$.

**Assumption:** “accept” reviews contain only positive sentences, “reject” contain negative, . . .

1. classify each sentence in $S$ as positive/neutral/negative
2. discard sentences not consistent with $o$
Sentiment analysis

**Goal:** retain only sentences in $S$ which are consistent with recommendation $o$.

**Assumption:** “accept” reviews contain only positive sentences, “reject” contain negative, . . .

1. classify each sentence in $S$ as positive/neutral/negative
2. discard sentences not consistent with $o$

We used a pre-trained sentiment classifier.
Sentences reordering

**Idea:** real reviews have a structure (opening, central considerations, final remarks) and a typical length (we assume 5 sentences).

**Goal:** resemble real reviews structure and length.
Idea: real reviews have a structure (opening, central considerations, final remarks) and a typical length (we assume 5 sentences).

Goal: resemble real reviews structure and length.

1. classify all sentences in $S$ as opening/central/final
2. randomly pick 1 opening, 3 central, 1 final
3. concatenate to form $r$
Sentences reordering

**Idea:** real reviews have a structure (opening, central considerations, final remarks) and a typical length (we assume 5 sentences).

**Goal:** resemble real reviews structure and length.

1. classify all sentences in $S$ as opening/central/final
2. randomly pick 1 opening, 3 central, 1 final
3. concatenate to form $r$

We used an ad hoc classifier.
Order classifier

- Based on the general purpose text classifier (by Stanford NLP group)
- Trained with all sentences in $R$, automatically labeled based on their position.
Table of Contents

1 Motivation

2 The tool

3 Experimental evaluation
Aims

- Does the tool generate reviews which appear human-generated and specific for the input paper?
- Can a generated review affect the decision (reject/accept)?
Aims

- Does the tool generate reviews which appear human-generated and specific for the input paper?
  - Intrinsic evaluation
- Can a generated review affect the decision (reject/accept)?
  - Extrinsic evaluation
Data and subjects

Data:
- 48 papers, 168 related reviews
- from F1000Research, Elifescience, Openreview, PeerJ, our lab records

Subjects:
- ≈ 7 experienced (professors, post-docs, PhD students)
- ≈ 3 intermediate (students)
- ≈ 5 novice (out of academy)
Baseline

- a 2nd order Markov chain on tokens
- trained on $R$
- inputs $a, o$ are not considered
Intrinsic evaluation

Many forms per subject, each consists of:

- Title of the paper
- 10 reviews randomly picked from:
  - Real reviews for the paper
  - Real reviews not for the paper
  - Generated for the paper and a random one
- For each review, the question "Does it appear as a genuine review written by a human reviewer for the paper with the shown title?"

Review 1:
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...]

Appear as been written by human for this paper? Y □ N □

Review 2:
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and r [...]

Appear as been written by human for this paper? Y □ N □

... 

Review 10: ...
Intrinsic evaluation

Many forms per subject, each consists of:
- title of paper

---

**Review 1:** The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...] 

*Appear as been written by human for this paper?* Y □ N □

**Review 2:** The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and r [...] 

*Appear as been written by human for this paper?* Y □ N □

... 

**Review 10:** …
Experimental evaluation

Intrinsic evaluation

Many forms per subject, each consists of:

- title of paper $a$
- 10 reviews randomly picked from:
  - real reviews for $a$ (5 sentences)
  - real reviews not for $a$ (5 sentences)
  - generated for $a$ (and a random $o$)
  - generated with the baseline method

A Language and an Inference Engine for Twitter Filtering Rules

Review 1: The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...] Appear as been written by human for this paper? Y □ N □

Review 2: The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and r [...] Appear as been written by human for this paper? Y □ N □

... Review 10: ...
Intrinsic evaluation

Many forms per subject, each consists of:

- title of paper \( a \)
- 10 reviews randomly picked from:
  - real reviews for \( a \) (5 sentences)
  - real reviews not for \( a \) (5 sentences)
  - generated for \( a \) (and a random \( o \))
  - generated with the baseline method

for each review, the question “does it appear as a genuine review written by a human reviewer for the paper with the shown title?”

Review 1:

The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and […]

Appear as been written by human for this paper? Y □ N □

Review 2:

The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and r […]

Appear as been written by human for this paper? Y □ N □

…

Review 10: …
Intrinsic evaluation: percentage of yes answers

- 30% of generated reviews appear human-generated and specific
- 3 × better than the baseline (9%)
- Real-specific is high (85%): small impact of shortening
- Experienced subjects less fooled than intermediate, novice

Bartoli, De Lorenzo, Medvet, Tarlao (UniTs)
Generation of Scientific Paper Reviews
August 31st, 2016 23 / 26
Experimental evaluation

Intrinsic evaluation: percentage of yes answers

- 30% of generated reviews appear human-generated and specific
- 3× better than the baseline (9%)
30% of generated reviews appear human-generated and specific
- 3× better than the baseline (9%)
- real-specific is high (85%): small impact of shortening
Intrinsic evaluation: percentage of yes answers

- 30% of generated reviews appear human-generated and specific
  - 3× better than the baseline (9%)
- real-specific is high (85%): small impact of shortening
- experienced subjects less fooled than intermediate, novice
Extrinsic evaluation

Many forms per subject, each consists of:

<table>
<thead>
<tr>
<th>Reviewer 1: accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewer 2: accept</th>
</tr>
</thead>
<tbody>
<tr>
<td>The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and [...]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewer 3: reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of [...]</td>
</tr>
</tbody>
</table>

Decision based on the 3 reviews: Accept □ Reject □
Experimental evaluation

Extrinsic evaluation

Many forms per subject, each consists of:

- title of paper \( a \)

A Language and an Inference Engine for Twitter Filtering Rules

Reviewer 1: accept
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...]

Reviewer 3: reject
The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of [...]

Reviewer 2: accept
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and [...]

Decision based on the 3 reviews: Accept □ Reject □
Extrinsic evaluation

Many forms per subject, each consists of:

- title of paper \( a \)
- 3 reviews randomly picked from:
  - real reviews for \( a \) (5 sentences)
  - real reviews not for \( a \) (5 sentences)
  - generated for \( a \) (and a random \( o \))

**A Language and an Inference Engine for Twitter Filtering Rules**

**Reviewer 1: accept**
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...] 

**Reviewer 3: reject**
The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of [...] 

**Reviewer 2: accept**
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and [...] 

*Decision based on the 3 reviews: Accept □ Reject □*
Extrinsic evaluation

Many forms per subject, each consists of:

- title of paper \( a \)
- 3 reviews randomly picked from:
  - real reviews for \( a \) (5 sentences)
  - real reviews not for \( a \) (5 sentences)
  - generated for \( a \) (and a random \( o \)) along with their recommendation

A Language and an Inference Engine for Twitter Filtering Rules

Reviewer 1: accept
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and [...]

Reviewer 3: reject
The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of [...]

Reviewer 2: accept
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and [...]

Decision based on the 3 reviews: Accept □ Reject □
Extrinsic evaluation

Many forms per subject, each consists of:

- title of paper \(a\)
- 3 reviews randomly picked from:
  - real reviews for \(a\) (5 sentences)
  - real reviews not for \(a\) (5 sentences)
  - generated for \(a\) (and a random \(o\))
- along with their recommendation
- the question: “basing on these 3 reviews, would you recommend to accept or reject the paper?”

---

**A Language and an Inference Engine for Twitter Filtering Rules**

**Reviewer 1: accept**
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and […]

**Reviewer 3: reject**
The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of […]

**Reviewer 2: accept**
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and […]

*Decision based on the 3 reviews: Accept □ Reject □*
Extrinsic evaluation

Many forms per subject, each consists of:
- title of paper \(a\)
- 3 reviews randomly picked from:
  - real reviews for \(a\) (5 sentences)
  - real reviews not for \(a\) (5 sentences)
  - generated for \(a\) (and a random one) along with their recommendation
- the question: “basing on these 3 reviews, would you recommend to accept or reject the paper?”

The subject is the editor in a simulated peer review process.

A Language and an Inference Engine for Twitter Filtering Rules

Reviewer 1: accept
The reporting has been done clearly and meets the standards of the journal. Topic has been introduced sufficiently and […]

Reviewer 3: reject
The foremost weak point of this paper is the presentation. The English is not satisfactory and there is a number of […]

Reviewer 2: accept
The paper presents an automatic technique to generate fake paper reviews on the basis of a small corpus of true reviews and […]

Decision based on the 3 reviews: Accept \(\square\) Reject \(\square\)
**Extrinsic evaluation: decision**

<table>
<thead>
<tr>
<th>Subject class</th>
<th>Subverted</th>
<th>Discordant</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>4</td>
<td>16</td>
<td>25.0</td>
</tr>
<tr>
<td>Intermediate</td>
<td>4</td>
<td>15</td>
<td>26.7</td>
</tr>
<tr>
<td>Novice</td>
<td>5</td>
<td>21</td>
<td>23.8</td>
</tr>
<tr>
<td>Overall</td>
<td>13</td>
<td>52</td>
<td>25.0</td>
</tr>
</tbody>
</table>

**Discordant** Forms with $\geq 1$ generated reviews and $\geq 1$ real reviews with discordant recommendation

**Subverted** Discordant forms where the subject’s decision is consistent with a generated review (hence opposite to a real review)
Extrinsic evaluation: decision

<table>
<thead>
<tr>
<th>Subject class</th>
<th>Subverted</th>
<th>Discordant</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>4</td>
<td>16</td>
<td>25.0</td>
</tr>
<tr>
<td>Intermediate</td>
<td>4</td>
<td>15</td>
<td>26.7</td>
</tr>
<tr>
<td>Novice</td>
<td>5</td>
<td>21</td>
<td>23.8</td>
</tr>
<tr>
<td>Overall</td>
<td>13</td>
<td>52</td>
<td>25.0</td>
</tr>
</tbody>
</table>

Discordant Forms with $\geq 1$ generated reviews and $\geq 1$ real reviews with discordant recommendation

Subverted Discordant forms where the subject’s decision is consistent with a generated review (hence opposite to a real review)

- in 25% of discordant forms, subject’s decision agrees with a generated review and disagrees with a real review
- an injected fake review can manipulate the outcome of the (simulated) peer review process
Conclusions

- a tool like SCIgen for reviews could exist
- another element in the ongoing debate on scholarly publishing system and on peer review, in particular
Thanks!