Using Common-Sense knowledge-base for Detecting Word Obfuscation in Adversarial Communication

Swati Agarwal, Ashish Sureka

IIIT-Delhi | Indraprastha Institute of Information Technology

swatia@iiitd.ac.in, ashish@iiitd.ac.in

January 10, 2015
# Table of Contents

1. Research Motivation and Aim  
   - Research Motivation  
   - Research Aim

2. Related Work and Research Contributions  
   - Related Work  
   - Research Contributions

3. Solution Approach  
   - High-Level Framework and Architecture  
   - Worked-Out Example  
   - Solution Pseudo-code and Algorithm

4. Experimental Evaluation and Validation  
   - Experimental Dataset  
   - Data Pre-processing and Term Substitution Technique  
   - Experimental Results

5. Conclusion

6. References
# Table of Contents

1. **Research Motivation and Aim**  
   - Research Motivation  
   - Research Aim

2. **Related Work and Research Contributions**  
   - Related Work  
   - Research Contributions

3. **Solution Approach**  
   - High-Level Framework and Architecture  
   - Worked-Out Example  
   - Solution Pseudo-code and Algorithm

4. **Experimental Evaluation and Validation**  
   - Experimental Dataset  
   - Data Pre-processing and Term Substitution Technique  
   - Experimental Results

5. **Conclusion**

6. **References**
Intelligence and security agencies intercepts and scans billions of messages and communications every day to identify dangerous communications between terrorists and criminals.

Law enforcement agencies use message interception to combat criminal and illicit acts.

Terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies.

Automatic word obfuscation detection is a natural language processing problem that has attracted several researcher’s attention.
Message Interception - Intelligence and Security Agencies

- Intelligence and security agencies intercepts and scans billions of messages and communications every day to identify dangerous communications between terrorists and criminals.

- Law enforcement agencies use message interception to combat criminal and illicit acts.

- Terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies.

- Automatic word obfuscation detection is a natural language processing problem that has attracted several researchers’ attention.
Message Interception - Intelligence and Security Agencies

- Intelligence and security agencies intercepts and scans billions of messages and communications every day to identify dangerous communications between terrorists and criminals.
- Law enforcement agencies use message interception to combat criminal and illicit acts.
- Terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies.
- Automatic word obfuscation detection is a natural language processing problem that has attracted several researcher’s attention.
Message Interception - Intelligence and Security Agencies

- Intelligence and security agencies intercepts and scans billions of messages and communications every day to identify dangerous communications between terrorists and criminals.
- Law enforcement agencies use message interception to combat criminal and illicit acts.
- Terrorist and criminals use textual or word obfuscation to prevent their messages from getting intercepted by the law enforcement agencies.
- Automatic word obfuscation detection is a natural language processing problem that has attracted several researcher’s attention.
Term Obfuscation

- Attack
- Birthday function
ConceptNet is a semantic network consisting of nodes representing concepts and edges representing relations between the concepts.

We hypothesize that ConceptNet can be used as a semantic knowledge-base to solve the problem of textual or word obfuscation.

To investigate the application of a commonsense knowledge-base such as ConceptNet for solving the problem of word or textual obfuscation.

To conduct an empirical analysis on large and real-word datasets for the purpose of evaluating the effectiveness of the application of ConceptNet.
Commonsense Knowledge-base - Term Obfuscation

- ConceptNet is a semantic network consisting of nodes representing concepts and edges representing relations between the concepts.
- We hypothesize that ConceptNet can be used as a semantic knowledge-base to solve the problem of textual or word obfuscation.
- To investigate the application of a commonsense knowledge-base such as ConceptNet for solving the problem of word or textual obfuscation.
- To conduct an empirical analysis on large and real-word datasets for the purpose of evaluating the effectiveness of the application of ConceptNet.
Research Motivation and Aim
Related Work and Research Contributions
Solution Approach
Experimental Evaluation and Validation
Conclusion
References

Commonsense Knowledge-base - Term Obfuscation

- ConceptNet is a semantic network consisting of nodes representing concepts and edges representing relations between the concepts.
- We hypothesize that ConceptNet can be used as a semantic knowledge-base to solve the problem of textual or word obfuscation.
- To investigate the application of a commonsense knowledge-base such as ConceptNet for solving the problem of word or textual obfuscation.
- To conduct an empirical analysis on large and real-word datasets for the purpose of evaluating the effectiveness of the application of ConceptNet.

Swati Agarwal, Ashish Sureka
FIS 2015 Future Information Security
Commonsense Knowledge-base - Term Obfuscation

- ConceptNet is a semantic network consisting of nodes representing concepts and edges representing relations between the concepts.
- We hypothesize that ConceptNet can be used as a semantic knowledge-base to solve the problem of textual or word obfuscation.
- To investigate the application of a commonsense knowledge-base such as ConceptNet for solving the problem of word or textual obfuscation.
- To conduct an empirical analysis on large and real-word datasets for the purpose of evaluating the effectiveness of the application of ConceptNet.
ConceptNet - Common Sense - Semantic Network
# Table of Contents

1. Research Motivation and Aim
   - Research Motivation
   - Research Aim

2. Related Work and Research Contributions
   - Related Work
   - Research Contributions

3. Solution Approach
   - High-Level Framework and Architecture
   - Worked-Out Example
   - Solution Pseudo-code and Algorithm

4. Experimental Evaluation and Validation
   - Experimental Dataset
   - Data Pre-processing and Term Substitution Technique
   - Experimental Results

5. Conclusion

6. References
List of Previous Work - Reverse Chronological Order

**Table:** ED: Evaluation Dataset, RS: Resources Used in Solution Approach, SA: Solution Approach

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ED</strong></td>
<td>Google News</td>
<td>British National Corpus (BNC)</td>
<td>Enron e-mail dataset, Brown corpus</td>
<td>Enron e-mail dataset</td>
</tr>
<tr>
<td><strong>RS</strong></td>
<td>Google search engine</td>
<td>1.4 billion words of English Gigaword v.1 (newswire corpus)</td>
<td>British National Corpus (BNC), WordNet, Yahoo, Google and MSN search engine</td>
<td>British National Corpus (BNC), WordNet, Google search engine</td>
</tr>
<tr>
<td><strong>SA</strong></td>
<td>Measuring sentence oddity, enhance sentence oddity and k-grams frequencies</td>
<td>Probabilistic or distributional model of context</td>
<td>Sentence oddity, K-gram frequencies, Hypernym Oddity (HO) and Pointwise Mutual Information (PMI)</td>
<td>Sentence oddity measures, semantic measure using WordNet, and frequency count of the bigrams around the target word</td>
</tr>
</tbody>
</table>

Swati Agarwal, Ashish Sureka
Research Contributions

- First focused research investigation on the application of ConceptNet common sense knowledge-base for solving the problem of textual or term obfuscation

- We conduct an in-depth empirical analysis to examine the effectiveness of the proposed approach
First focused research investigation on the application of ConceptNet common sense knowledge-base for solving the problem of textual or term obfuscation

We conduct an in-depth empirical analysis to examine the effectiveness of the proposed approach
# Table of Contents

1. Research Motivation and Aim
   - Research Motivation
   - Research Aim

2. Related Work and Research Contributions
   - Related Work
   - Research Contributions

3. Solution Approach
   - High-Level Framework and Architecture
   - Worked-Out Example
   - Solution Pseudo-code and Algorithm

4. Experimental Evaluation and Validation
   - Experimental Dataset
   - Data Pre-processing and Term Substitution Technique
   - Experimental Results

5. Conclusion

6. References
Multiple phase - Processing Pipeline.

\[ S = T_1 \ T_2 \ T_3 \ldots \ldots \ldots \ldots \ T_{N-1} \ T_N \]

- **POS Tagger**
  - **Bag of Terms** [Adjectives, Adverbs, Nouns, Verbs]

- **ConceptNet**
  - Shortest Path
  - Dijkstra Path
  - A* Path

**Conceptual Similarity**
Mean Average Conceptual Similarity (MACS) Score
Example 1

Original Sentence: "We will attack the airport with bomb"
Red-flagged term: bomb
Replacement Word: flower
Replaced Sentence: "We will attack the airport with flower"
Bag-of-terms: attack, airport, flower

Conceptual similarity between airport and flower is 3
The number of edges between airport and flower is 3
The number of edges between flower and airport is 3
Conceptual similarity between attack and flower is 3
The number of edges between attack and flower is 3
The number of edges between flower and attack is 3
Conceptual similarity between attack and airport is 2.5
The number of edges between attack and airport is 2
The number of edges between airport and attack is 3
MACS: $(3 + 3 + 2.5)/3 = 2.83$
Worked-out Examples

Example 1

Original Sentence: "We will attack the airport with bomb"
Red-flagged term: bomb
Replacement Word: flower
Replaced Sentence: "We will attack the airport with flower"
Bag-of-terms: attack, airport, flower

Conceptual similarity between airport and flower is 3
The number of edges between airport and flower is 3
The number of edges between flower and airport is 3
Conceptual similarity between attack and flower is 3
The number of edges between attack and flower is 3
The number of edges between flower and attack is 3
Conceptual similarity between attack and airport is 2.5
The number of edges between attack and airport is 2
The number of edges between airport and attack is 3
MACS: \((3 + 3 + 2.5)/3 = 2.83\)
Worked-out Examples

Example 2

- Original Sentence: "Pistol will be delivered to you to shoot the president"
- Red-flagged term: pistol
- Replacement Word: pen
- Replaced Sentence: "Pen will be delivered to you to shoot the president"
- Bag-of-terms: pen, shoot, president

Conceptual similarity between shoot and president is 2.5
The number of edges between president and shoot is 2
The number of edges between shoot and president is 3
Conceptual similarity between pen and president is 3
The number of edges between president and pen is 3
The number of edges between pen and president is 3
Conceptual similarity between pen and shoot is 3
The number of edges between shoot and pen is 3
The number of edges between pen and shoot is 3
MACS: \(\frac{2.5 + 3 + 3}{3} = 2.83\)
Example 2

Original Sentence: "Pistol will be delivered to you to shoot the president"
Red-flagged term: pistol
Replacement Word: pen
Replaced Sentence: "Pen will be delivered to you to shoot the president"
Bag-of-terms: pen, shoot, president

Example 2

Conceptual similarity between shoot and president is 2.5
The number of edges between president and shoot is 2
The number of edges between shoot and president is 3
Conceptual similarity between pen and president is 3
The number of edges between president and pen is 3
The number of edges between pen and president is 3
Conceptual similarity between pen and shoot is 3
The number of edges between shoot and pen is 3
The number of edges between pen and shoot is 3
MACS: \(\frac{2.5 + 3 + 3}{3} = 2.83\)
Obfuscated Term Detection

**Data:** Substituted Sentence $S'$, Conceptnet Corpus $C$

**Result:** Obfuscated Term $O_T$

```python
for all record $r \in C$ do
    Edge $E.add(r.node_1, r.node_2, r.relation)$
    Graph $G.add(E)$

$tokens = S'.tokenize()$
$pos.add(pos_tag(tokens))$

for all tag $\in$ pos and token $\in$ tokens do
    if tag is in (verb, noun, adjective, adverb) then
        BoW.add(token.lemma)

for iter = 0 to BoW.length do
    concepts = BoW.pop(iter)
    for $i = 0$ to concepts.length - 1 do
        for $j = i$ to concepts.length do
            if ($i != j$) then
                $path_{c_i,j} = Dijkstra.pathlen(G, i, j)$
                $path_{c_j,i} = Dijkstra.pathlen(G, j, i)$
                $avg.add(Average(c_i,j, c_j,i))$
        mean.add(Mean(avg))

$O_T = BoW.valueAt(min(mean))$
```

Swati Agarwal, Ashish Sureka  
FIS 2015 Future Information Security  
16 / 39
Solution Pseudo-Code and Algorithm
Concrete Examples of Computing Conceptual Similarity between Two Given Terms Using Three Different Distance Metrics or Algorithms:

<table>
<thead>
<tr>
<th>S.N.</th>
<th>Term 1</th>
<th>Term 2</th>
<th>Di-jikstra’s Algo</th>
<th>A-Star Algo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>T1-T2</td>
<td>T2-T1</td>
</tr>
<tr>
<td>1</td>
<td>Tree</td>
<td>Branch</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Pen</td>
<td>Blood</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Paper</td>
<td>Tree</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Airline</td>
<td>Pen</td>
<td>4(NP)</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Bomb</td>
<td>Blast</td>
<td>2</td>
<td>4(NP)</td>
</tr>
</tbody>
</table>

Concrete Examples of Conceptually and Semantically Unrelated Terms and their Path Length to Compute the Default Value for No-Path:

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Bowl</th>
<th>Wire</th>
<th>Coffee</th>
<th>Office</th>
<th>Feather</th>
<th>Driver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mobile</td>
<td>Dress</td>
<td>Research</td>
<td>Festival</td>
<td>Study</td>
<td>Sun</td>
</tr>
<tr>
<td>Term 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path Length</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Concrete Examples of Computing Conceptual Similarity between Two Given Terms Using Three Different Distance Metrics or Algorithms:

<table>
<thead>
<tr>
<th>S.N.</th>
<th>Term 1</th>
<th>Term 2</th>
<th>T1-T2</th>
<th>T2-T1</th>
<th>Mean</th>
<th>T1-T2</th>
<th>T2-T1</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tree</td>
<td>Branch</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Pen</td>
<td>Blood</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>Paper</td>
<td>Tree</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Airline</td>
<td>Pen</td>
<td>4(NP)</td>
<td>4</td>
<td>4</td>
<td>4(NP)</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Bomb</td>
<td>Blast</td>
<td>2</td>
<td>4(NP)</td>
<td>3</td>
<td>2</td>
<td>4(NP)</td>
<td>3</td>
</tr>
</tbody>
</table>

Concrete Examples of Conceptually and Semantically Unrelated Terms and their Path Length to Compute the Default Value for No-Path:

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Bowl</th>
<th>Wire</th>
<th>Coffee</th>
<th>Office</th>
<th>Feather</th>
<th>Driver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Mobile</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Dijkstra's Algo  |  A-Star Algo

Swati Agarwal, Ashish Sureka  
FIS 2015 Future Information Security  
18 / 39
Table of Contents

1. Research Motivation and Aim
   - Research Motivation
   - Research Aim

2. Related Work and Research Contributions
   - Related Work
   - Research Contributions

3. Solution Approach
   - High-Level Framework and Architecture
   - Worked-Out Example
   - Solution Pseudo-code and Algorithm

4. Experimental Evaluation and Validation
   - Experimental Dataset
   - Data Pre-processing and Term Substitution Technique
   - Experimental Results

5. Conclusion

6. References
Bar Chart - Experimental Dataset Statistics
Part-of-Speech Tags, Size of Bag-of-Terms
# BNC and EMC Dataset

<table>
<thead>
<tr>
<th>Abbr</th>
<th>Description</th>
<th>BNC</th>
<th>EMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus</td>
<td>Total sentences in brown news corpus</td>
<td>4607</td>
<td>9112</td>
</tr>
<tr>
<td>5-15</td>
<td>Sentences that has length between 5 to 15</td>
<td>1449</td>
<td>2825</td>
</tr>
<tr>
<td>N-BNC</td>
<td>Sentences that has their first noun in BNC (british national corpus)</td>
<td>2214</td>
<td>3587</td>
</tr>
<tr>
<td>N-COCA</td>
<td>Sentences that has their first noun in 100 K list (COCA)</td>
<td>2393</td>
<td>4006</td>
</tr>
<tr>
<td>N-H-W</td>
<td>If first noun has an hypernym in WordNet</td>
<td>3441</td>
<td>5620</td>
</tr>
<tr>
<td>En-BNC</td>
<td>English sentences according to BNC</td>
<td>2146</td>
<td>3430</td>
</tr>
<tr>
<td>En- Java</td>
<td>English sentences according to Java language detection library</td>
<td>4453</td>
<td>8527</td>
</tr>
<tr>
<td>S’-BNC</td>
<td>#Substituted sentences using BNC list</td>
<td>2146</td>
<td>3430</td>
</tr>
<tr>
<td>S’-COCA</td>
<td>#Substituted sentences using COCA (100K) list</td>
<td>2335</td>
<td>3823</td>
</tr>
<tr>
<td>S’-B-5-15</td>
<td>#Substituted sentences (between length of 5 to 15) using BNC list</td>
<td>666</td>
<td>1051</td>
</tr>
<tr>
<td>S’-C-5-15</td>
<td>#Substituted sentences (between length of 5 to 15) using COCA list</td>
<td>740</td>
<td>1191</td>
</tr>
</tbody>
</table>
### Data Cleaning

Examples of Sentences Discarded While Word Substitution:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentence</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMC</td>
<td>Since we’re ending 2000 and going into a new sales year I want to make sure I’m not holding resource open on any accounts which may not or should not be on the list of focus accounts which you and your team have requested our involvement with.</td>
<td>Sentence length is not between 5 to 15</td>
</tr>
<tr>
<td>EMC</td>
<td>next Thursday at 7:00 pm Yes yes yes.</td>
<td>First noun is not in BNC/COCA list</td>
</tr>
<tr>
<td>BNC</td>
<td>The City Purchasing Department the jury said is lacking in experienced clerical personnel as a result of city personnel policies.</td>
<td>Sentence length is not between 5 to 15</td>
</tr>
<tr>
<td>BNC</td>
<td>Dr Clark holds an earned Doctor of Education degree from the University of Oklahoma</td>
<td>First noun does not have a hypernym in WordNet</td>
</tr>
</tbody>
</table>
### Example of Term Substitution using COCA Frequency List. NF= First Noun/Original Term, ST= Substituted Term:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>NF</th>
<th>Freq</th>
<th>ST</th>
<th>Freq</th>
<th>Sentence’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any opinions expressed herein are solely those of the author.</td>
<td>Author</td>
<td>53195</td>
<td>Television</td>
<td>53263</td>
<td>Any opinions expressed herein are solely those of the television.</td>
</tr>
<tr>
<td>What do you think that should help you score women.</td>
<td>Score</td>
<td>17415</td>
<td>Struggle</td>
<td>17429</td>
<td>What do you think that should help you struggle women.</td>
</tr>
<tr>
<td>This was the coolest calmest election I ever saw.</td>
<td>Election</td>
<td>40513</td>
<td>Republicans</td>
<td>40515</td>
<td>This was the coolest calmest republicans I ever saw.</td>
</tr>
<tr>
<td>Tom Williams said</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Tom Williams said</td>
</tr>
<tr>
<td>The inadequacy of our library system will become critical unless we act vigorously to correct this condition</td>
<td>Inadequacy</td>
<td>831</td>
<td>Inevitability</td>
<td>831</td>
<td>The inevitability of our library system will become critical unless we act vigorously to correct this condition</td>
</tr>
</tbody>
</table>
Text Substitution Technique

Data: Sentence S, Frequency List COCA, WordNet DataBase W_{DB}
Result: Substituted Sentence S'

if (5 < S.length < 15) then
    tokens ← S.tokenize()
    POS ← S.pos_tag()
    NF ← token[POS.indexOf("NN")]
    if (COCA.has(NF) AND W_{DB}.has(NF.hypernym)) then
        lang ← S.Language Detection
        if (lanf == "en") then
            FNF ← COCA.freq(NF)
            FNF' ← COCA.nextHigherFreq(FNF)
            NF' ← COCA.hasFrequency(FNF')
            S' ← S.replaceFirst(NF, NF')
            return S'

Swati Agarwal, Ashish Sureka
## Examples from Research Papers

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Substituted Sentence</th>
<th>Paper</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>the bomb is in position</td>
<td>the alcohol is in position</td>
<td>Fong2006 [5]</td>
<td>alcohol</td>
</tr>
<tr>
<td>copyright 2001 south-west airlines co all rights reserved</td>
<td>toast 2001 southwest airlines co all rights reserved</td>
<td>Fong2006 [5]</td>
<td>southwest</td>
</tr>
<tr>
<td>please try to maintain the same seat each class</td>
<td>please try to maintain the same play each class</td>
<td>Fong2006 [5]</td>
<td>try</td>
</tr>
<tr>
<td>we expect that the ____ attack will happen tonight</td>
<td>we expect that the ____ campaign will happen tonight</td>
<td>Fong2008 [4]</td>
<td>campaign</td>
</tr>
<tr>
<td>an agent will assist you with checked baggage</td>
<td>an vote will assist you with checked baggage</td>
<td>Fong2008 [4]</td>
<td>vote</td>
</tr>
<tr>
<td>my lunch contained white tuna</td>
<td>my package contained white tuna</td>
<td>Fong2008 [4]</td>
<td>package</td>
</tr>
<tr>
<td>she ordered a parfait</td>
<td>she ordered a parfait</td>
<td>Fong2008 [4]</td>
<td>know</td>
</tr>
<tr>
<td>please let me know if you have this information</td>
<td>please let me know if you have this men</td>
<td>Fong2008 [4]</td>
<td>recomm.</td>
</tr>
<tr>
<td>It was one of a series of recommendations by the Texas Research League</td>
<td>It was one of a bank of recommenda-tions by the Texas Research League</td>
<td>Fong2008 [4]</td>
<td>attendance</td>
</tr>
<tr>
<td>The remainder of the college requirement would be in general sub jects</td>
<td>The attendance of the college require-ment would be in general sub jects</td>
<td>Fong2008 [4]</td>
<td>attendance</td>
</tr>
<tr>
<td>A copy was released to the press</td>
<td>An object was released to the press</td>
<td>Fong2008 [4]</td>
<td>released</td>
</tr>
</tbody>
</table>
## Examples from Research Papers

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Substituted Sentence</th>
<th>Paper</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 works need to be done in Hyderabad</td>
<td>works need to be done in H</td>
<td>Deshmukh14 [3]</td>
<td>H</td>
</tr>
<tr>
<td>12 you should arrange for a preparation of blast</td>
<td>you should arrange for a preparation of daawati</td>
<td>Deshmukh14 [3]</td>
<td>daawati</td>
</tr>
<tr>
<td>13 my friend will come to deliver you a pistol</td>
<td>my friend will come to deliver you a CD</td>
<td>Deshmukh14 [3]</td>
<td>CD</td>
</tr>
<tr>
<td>14 collect some people for work from Gujarat</td>
<td>collect some people for work from Musa</td>
<td>Deshmukh14 [3]</td>
<td>Musa</td>
</tr>
<tr>
<td>15 you will find some bullets in the bag</td>
<td>you will find some pen drives in the bag</td>
<td>Deshmukh14 [3]</td>
<td>pen drives</td>
</tr>
<tr>
<td>16 come at Delhi for meeting</td>
<td>come at Sham for meeting</td>
<td>Deshmukh14 [3]</td>
<td>Sham</td>
</tr>
<tr>
<td>17 send one person to Bangalore</td>
<td>send one person to Bagu</td>
<td>Deshmukh14 [3]</td>
<td>Bagu</td>
</tr>
<tr>
<td>18 Arrange some rifles for next operation</td>
<td>Arrange some DVDs for next operation</td>
<td>Deshmukh14 [3]</td>
<td>DVDs</td>
</tr>
<tr>
<td>19 preparation of blast will start in next month</td>
<td>preparation of Daawati work will start in next month</td>
<td>Deshmukh14 [3]</td>
<td>Daawati</td>
</tr>
<tr>
<td>20 find one place at Hyderabad for operation</td>
<td>find one place at H for operation</td>
<td>Deshmukh14 [3]</td>
<td>H</td>
</tr>
<tr>
<td>21 He remembered sitting on the wall with a cousin, watching the German bomber fly over</td>
<td>He remembered sitting on the wall with a cousin, watching the German dancers fly over</td>
<td>Jabbari08 [10]</td>
<td>German</td>
</tr>
</tbody>
</table>
Brown News Corpus and Enron Email Corpus

Concrete Examples of Sentences with Size of Bag-of-terms Less Than 2:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentence</th>
<th>Bag-of-terms</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>That was before I studied both</td>
<td>[]</td>
<td>0</td>
</tr>
<tr>
<td>BNC</td>
<td>The jews had been expected</td>
<td>[jews]</td>
<td>1</td>
</tr>
<tr>
<td>BNC</td>
<td>if we are not discriminating in our cars</td>
<td>[car]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>What is the benefits?</td>
<td>[benefits]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>Who coined the adolescents?</td>
<td>[adolescents]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>Can you help? his days is 011 44 207 397 0840 john</td>
<td>[day]</td>
<td>1</td>
</tr>
</tbody>
</table>

Concrete Examples of Sentences with the Presence of Technical Terms and Abbreviations:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Tech Terms</th>
<th>Abbr</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4. artifacts 2004-2008 maybe 1 trade a day.</td>
<td>Artifacts</td>
<td>-</td>
</tr>
<tr>
<td>We have put the interview on IPTV for your viewing pleasure.</td>
<td>Interview, IPTV</td>
<td>IPTV</td>
</tr>
<tr>
<td>Will talk with KGW off name.</td>
<td>-</td>
<td>KGW</td>
</tr>
<tr>
<td>We are having males backtesting Larry May’s VaR.</td>
<td>backtesting</td>
<td>VAR</td>
</tr>
<tr>
<td>Internetworking and today American Express has surfaced.</td>
<td>Internetworking</td>
<td>-</td>
</tr>
<tr>
<td>I do not know their particles yet due to the Enron PRC meeting conflicts.</td>
<td>Enron</td>
<td>PRC</td>
</tr>
<tr>
<td>The others may have contracts with LNG consistency owners.</td>
<td>-</td>
<td>LNG</td>
</tr>
</tbody>
</table>
## Brown News Corpus and Enron Email Corpus

### Concrete Examples of Sentences with Size of Bag-of-terms Less Than 2:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentence</th>
<th>Bag-of-terms</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>That was before I studied both</td>
<td>[]</td>
<td>0</td>
</tr>
<tr>
<td>BNC</td>
<td>The jews had been expected</td>
<td>[jews]</td>
<td>1</td>
</tr>
<tr>
<td>BNC</td>
<td>if we are not discriminating in our cars</td>
<td>[car]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>What is the benefits?</td>
<td>[benefits]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>Who coined the adolescents?</td>
<td>[adolescents]</td>
<td>1</td>
</tr>
<tr>
<td>EMC</td>
<td>Can you help? his days is 011 44 207 397 0840 john</td>
<td>[day]</td>
<td>1</td>
</tr>
</tbody>
</table>

### Concrete Examples of Sentences with the Presence of Technical Terms and Abbreviations:

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Tech Terms</th>
<th>Abbr</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4. artifacts 2004-2008 maybe 1 trade a day.</td>
<td>Artifacts</td>
<td>-</td>
</tr>
<tr>
<td>We have put the interview on IPTV for your viewing pleasure.</td>
<td>Interview, IPTV</td>
<td>IPTV</td>
</tr>
<tr>
<td>Will talk with KGW off name.</td>
<td>-</td>
<td>KGW</td>
</tr>
<tr>
<td>We are having males backtesting Larry May’s VaR.</td>
<td>backtesting</td>
<td>VAR</td>
</tr>
<tr>
<td>Internetworking and today American Express has surfaced.</td>
<td>Internetworking</td>
<td>-</td>
</tr>
<tr>
<td>I do not know their particles yet due to the Enron PRC meeting conflicts.</td>
<td>Enron</td>
<td>PRC</td>
</tr>
<tr>
<td>The others may have contracts with LNG consistency owners.</td>
<td>-</td>
<td>LNG</td>
</tr>
</tbody>
</table>
Concrete Examples of Long Sentences (Length of Bag-of-terms $\geq 5$) Where Substituted Term is Identified Correctly:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentence</th>
<th>Original</th>
<th>Bag-of-Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>He further proposed grants of an unspecified input for experimental hospitals</td>
<td>Sum</td>
<td>[grants, unspecified, input, experimental, hospitals]</td>
</tr>
<tr>
<td></td>
<td>When the gubernatorial action starts</td>
<td></td>
<td>[gubernatorial, action]</td>
</tr>
<tr>
<td>BNC</td>
<td>Caldwell is expected to become a campaign coordinator for Byrd</td>
<td>Campaign</td>
<td>Caldwell, campaign, coordinator, Byrd</td>
</tr>
<tr>
<td>BNC</td>
<td>The entire arguments collection is available to patrons of all members on interlibrary loans</td>
<td>Headquarters</td>
<td>[entire, argument, collection, available, patron, member, interlibrary, loan]</td>
</tr>
<tr>
<td>EMC</td>
<td>Methodologies for accurate skill-matching and pilgrims ecien-cies=20 Key Benefits?</td>
<td>Fulfillment</td>
<td>[methodologies, accurate, skill, pilgrims, ecien-cies, benefits]</td>
</tr>
<tr>
<td>EMC</td>
<td>PERFORMANCE REVIEW The measurement to provide feedback is Friday November 17.</td>
<td>Deadline</td>
<td>[performance, review, measurement, feedback, friday, november]</td>
</tr>
</tbody>
</table>
Brown News Corpus and Enron Email Corpus

Accuracy Results for Brown News Corpus (BNC) and Enron Mail Corpus (EMC):

<table>
<thead>
<tr>
<th></th>
<th>Total Sentences</th>
<th>Correctly Identified</th>
<th>Accuracy Results</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>740</td>
<td>573</td>
<td>77.4%</td>
<td>46</td>
</tr>
<tr>
<td>EMC</td>
<td>1191</td>
<td>629</td>
<td>62.9%</td>
<td>125</td>
</tr>
</tbody>
</table>

Swati Agarwal, Ashish Sureka
### Brown News Corpus and Enron Email Corpus

#### Accuracy Results for Brown News Corpus (BNC) and Enron Mail Corpus (EMC):

<table>
<thead>
<tr>
<th></th>
<th>Total Sentences</th>
<th>Correctly Identified</th>
<th>Accuracy Results</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>740</td>
<td>573</td>
<td>77.4%</td>
<td>46</td>
</tr>
<tr>
<td>EMC</td>
<td>1191</td>
<td>629</td>
<td>62.9%</td>
<td>125</td>
</tr>
</tbody>
</table>

![Graph showing comparison between Brown News Corpus and Enron Email Corpus](image-url)

---

**Swati Agarwal, Ashish Sureka**

**FIS 2015 Future Information Security**

---
Brown News Corpus and Enron Email Corpus
Table of Contents

1 Research Motivation and Aim
   - Research Motivation
   - Research Aim

2 Related Work and Research Contributions
   - Related Work
   - Research Contributions

3 Solution Approach
   - High-Level Framework and Architecture
   - Worked-Out Example
   - Solution Pseudo-code and Algorithm

4 Experimental Evaluation and Validation
   - Experimental Dataset
   - Data Pre-processing and Term Substitution Technique
   - Experimental Results

5 Conclusion

6 References
Conclusion

- We present an approach to detect term obfuscation in adversarial communication using ConceptNet common-sense knowledge-base
- Experimental results reveal an accuracy of 72.72%, 77.4%, and 62.0% respectively on the three dataset
- Experimental results demonstrate that our approach is able to detect term obfuscation in long sentences containing more than 5 - 6 concepts
- We demonstrate that the proposed approach is generalizable
Conclusion

- We present an approach to detect term obfuscation in adversarial communication using ConceptNet common-sense knowledge-base.
- Experimental results reveal an accuracy of 72.72%, 77.4% and 62.0% respectively on the three dataset.
- Experimental results demonstrate that our approach is able to detect term obfuscation in long sentences containing more than 5 - 6 concepts.
- We demonstrate that the proposed approach is generalizable.
Conclusion

▶ We present an approach to detect term obfuscation in adversarial communication using ConceptNet common-sense knowledge-base

▶ Experimental results reveal an accuracy of 72.72%, 77.4% and 62.0% respectively on the three dataset

▶ Experimental results demonstrate that our approach is able to detect term obfuscation in long sentences containing more than 5 - 6 concepts

▶ We demonstrate that the proposed approach is generalizable
We present an approach to detect term obfuscation in adversarial communication using ConceptNet common-sense knowledge-base.

Experimental results reveal an accuracy of 72.72%, 77.4% and 62.0% respectively on the three dataset.

Experimental results demonstrate that our approach is able to detect term obfuscation in long sentences containing more than 5 - 6 concepts.

We demonstrate that the proposed approach is generalizable.
# Table of Contents

1. Research Motivation and Aim
   - Research Motivation
   - Research Aim

2. Related Work and Research Contributions
   - Related Work
   - Research Contributions

3. Solution Approach
   - High-Level Framework and Architecture
   - Worked-Out Example
   - Solution Pseudo-code and Algorithm

4. Experimental Evaluation and Validation
   - Experimental Dataset
   - Data Pre-processing and Term Substitution Technique
   - Experimental Results

5. Conclusion

6. References
References


References II

Chi-En Wu and Richard Tzong-Han Tsai.
Using relation selection to improve value propagation in a conceptnet-based sentiment dictionary. 
*Knowledge-Based Systems, 2014.*

Arbi Bouchoucha, Xiaohua Liu, and Jian-Yun Nie.
Integrating multiple resources for diversified query expansion. 

R Akileshwari, S Revathi, and A Grace Selvarani.
A novel approach for similarity based video annotation utilizing commonsense knowledgebase.

Dependency-based semantic parsing for concept-level text analysis. 
References III


References IV

Thank you!

**Contact:**
swatia@iiitd.ac.in
ashish@iiitd.ac.in