Computational Paraphrase Recognition & Generation
Paraphrase applications

• Search
  • Summarization
  • Dialog
  • Question Answering Systems
• Machine Translation (MT) Evaluation
  • Simplification
• Natural Language Processing (NLP)
• Natural Language Generation (NLG)
  • Teaching/Education
  • Information Extraction
  • Author Identification
  • Plagiarism Detection
• Intelligent Tutoring Systems (ITS)
What is a paraphrase?

Alternative ways of expressing the same information in one language

Six *commonly mentioned* paraphrasing patterns:

1. **Synonym**: substitute a word with its synonym, e.g. help, assist, aid;
2. **Voice**: change the voice of sentence from active to passive or vice versa;
3. **Word-Form/Part-of-speech**: change a word into a different form, e.g. change a noun to a verb, adverb, or adjective;
4. **Break down Sentence**: break a long sentence down into small sentences;
5. **Definition/Meaning**: substitute a word with its definition or meaning;
6. **Sentence Structure**: use different sentence structures to express the same thing.
What changed?

The processors were announced in San Jose at the Intel Developer Forum.

They had published an advertisement on the internet on June 10, offering the cargo for sale, he added.

Dzeirkhanov said 36 people were injured and that four people including, a child, had been hospitalized.

The new processor was unveiled at the Intel Developer Forum 2004 in San Jose, California.

On June 10, the ship’s owners had published an advertisement on the internet, offering the explosives for sale.

Of the 36 wounded, four people including one child, were hospitalized, Dzheirkhanov said.
Finding a paraphrase corpus: paraphrase acquisition

- The immediate major obstacle for paraphrase work
  - Baseline statistical methods require large corpora
- Paraphrase Acquisition an independent field
  - RTE (Recognizing Textual Entailment)
    - Annual RTE Challenge
- Parallel Translations (Literature)
- Internet Corpus (Microsoft)
Paraphrase Acquisition

• Newspaper Article Comparison
  • Following Shinyama & Sekine, 2003
• 1. Find comparable sentences which report the same event from different newspapers.
  1) Article Level Matching
  2) Sentence Level Matching
• 2. Identify anchors in the comparable sentences.
  – Named Entities form the strongest anchors
    • Extended Named Entity Tagging
• 3. Extract corresponding portions from the sentences.
• 4. Generalize the obtained expressions to paraphrase templates.
Paraphrase Acquisition

• Newspaper Article Comparison
  • Article Level Matching
  • Topic Detection & Tracking
  • Named Entity Recognition
    • 1. “The government has announced that **two** people have died in Hong Kong after contracting the **SARS** virus and **61** new cases of the illness have been detected.” *(Reuters, Apr. 11)*
    • 2. “Hong Kong reported **two** more deaths and **fresh cases** of **SARS** Friday as governments across the world took tough steps to stop the killer virus at their borders.” *(Channel News Asia, Apr. 11)*

Figure 1: The overall procedure
Building a Corpus from Newswire

From Quirk, Brockett, and Dolan

- Collected news stories from several online news sites (yahoo, google, etc.)
- 117,095 articles in 8 months
- Grouped them into clusters (15–16 articles per cluster)
Paraphrase Acquisition

- Newspaper Article Comparison
  - Sentence Level Matching
  - Eliminate non-sentences
  - Part of Speech tagging to segment words
  - Vector space model for sentence matching
  - Extended Named Entity tagging
  - Dependency Analysis (fig. 2)

Figure 2: Extracting portions of sentences
Excerpts of three translations of *Madame Bovary* by Gustave Flaubert.

“Now Emma would often take it into her head to write him during the day. Through her window she would signal to Justin, and he would whip off his apron and fly to la huchette. And when Rodolphe arrived in response to her summons, it was to hear that she was miserable, that her husband was odious, that her life was a torment.”

“Often, even in the middle of the day, Emma suddenly wrote to him, then from the window made a sign to Justin, who, taking his apron off, quickly ran to la huchette. Rodolphe would come; she had sent for him to tell him that she was bored, that her husband was odious, her life frightful.”

“Often, in the middle of the day, Emma would take up a pen and write to him. Then she would beckon across to Justin, who would off with his apron in an instant and fly away with the letter to la huchette. And Rodolphe would come. She wanted to tell him that life was a burden to her, that she could not endure her husband and that things were unbearable.”
Paraphrase Recognition

- Distinct Applications (generation unnecessary):
  - Tutoring
  - Question Answering
  - Information Retrieval

- iSTART (Interactive Strategy Trainer for Active Reading and Thinking)
Paraphrase Recognition iSTART
Based on Boonthum 2004

- Analyzes a student's self-explanation of a sentence and then responds.
  - Old methods use a word-match algorithm ("Did they mention ___?")

- Two-part improvement:
  1) Construct an Internal Representation
  2) Recognize the Paraphrase
Paraphrase Recognition

iSTART

- **Constructing an Internal Representation**
  - Make a Link Grammar parse.
  ```
  (0=LEFT-WALL) (1=a) (2=walnut.n) (3=is.v) (4=eaten.v) (5=by) (6=a) (7=monkey.n) (8=.)
  [[0 8 (Xp)] [0 2 (Wd)] [1 2 (Ds)] [2 3 (Ss)] [3 4 (Pv)] [4 5 (Mvp)] [5 7 (Js)] [6 7 (Ds)]]
  ```
  - Convert the LG triplets to Conceptual Graph triplets.
  ```
  0 [0 8 (Xp)] -> #S4 -> - N/A -
  1 [0 2 (Wd)] -> #S4 -> - N/A -
  2 [1 2 (Ds)] -> #S4 ->
     [walnut.n] -> (Article) -> [a]
  3 [2 3 (Ss)] -> #M# S + Pv (4) # ->
     [eaten.v] -> (Patient) -> [walnut.n]
  4 [3 4 (Pv)] -> #M# Pv + MV(5)+O(6) # ->
     [eaten.v] -> (Agent) -> [monkey.n]
  5 [4 5 (Mvp)] -> #S4 eaten.v by
  6 [5 7 (Js)] -> #S4 monkey.n by
  7 [6 7 (Ds)] -> #S4 ->
     [monkey.n] -> (Article) -> [a]
  ```
Paraphrase Recognition
iSTART

• Preliminaries to Recognition
  • Each sentence has various content words, including stop words.
  • Each content word has a definition, synonyms, antonyms, and other Wordnet-provided relations.
Paraphrase Recognition
iSTART

- Recognizing Paraphrase Phenomena
  - Single-word synonyms
  - Compound-word synonyms
  - Idiomatic clause/phrase
  - Voice
  - Part of Speech
  - Breaking Long Sentence
  - Definition, Meaning
  - Sentence Structure
Paraphrase Recognition

iSTART

• Single-word synonyms

Both CG have the same pattern; check if words in the same position have the same meaning.

• “Jenny helps Kay”
  • [Help] → (Agent) → [Person: Jenny]
  • ←→ (Patient) → [Person: Kay]
  • vs.
  • “Jenny assists Kay”
  • [Assist] → (Agent) → [Person: Jenny]
  • ←→ (Patient) → [Person: Kay]
Paraphrase Recognition

iSTART

- **Compound-word synonyms**
  - A word matches its compound-word synonym, or vice-versa.
    - [Install] → (Object) → [Thing]
    - ≡ [Set-Up] → (Object) → [Thing]
    - ≡ [Put-In] → (Object) → [Thing]
Paraphrase Recognition

iSTART

- Idiomatic Clause/Phrase
  - For each idiom a CG is generated.
    - [Help] → (Patient) → [Person: x]
    - ≡ [Give] → (Patient) → [Person: x]
      - +→ (Object) → [Hand]
  - Which produces:
    - “Jenny gives Kay a hand”
    - [Give] → (Agent) → [Person: Jenny]
    - +→ (Patient) → [Person: Kay]
    - +→ (Object) → [Hand]
  - Contextual analysis may be needed to prevent literal interpretation.
Paraphrase Recognition

iSTART

- **Voice Change**
  - “Mario smashes the mushroom” and “The mushroom is smashed by Mario” generate the same CG.
    - [smash] → (Agent) → [person: Mario]
    - +→ (Patient) → [badguy: mushroom]
  - Modification needed to ensure a paraphrase and not simply a copy.
Paraphrase Recognition

• **Part-of-speech**
  • Keywords might have changed part of speech.

  • **Original sentence**: “All thunderstorms have a similar life history.”
  • **Student’s Explanation**: “All thunderstorms have similarity in their historical life story.”

• Search for same word or same base-form.
Paraphrase Recognition

iSTART

- **Breaking a long sentence**
- Integrate CGs of all input before comparing with original sentence (useful for limited applications).
  - Pronoun Resolution

- **Original sentence**: “All thunderstorms have a similar life history.”
  - [Thunderstorm: \( \forall \)] –
  - (Feature) \( \rightarrow \) [History] –
  - (Attribute) \( \rightarrow \) [Life]
  - (Attribute) \( \rightarrow \) [Similar]

- **Student’s Explanation**: “Thunderstorms have life history. It is similar among all thunderstorms”
  - [Thunderstorm] –
  - (Feature) \( \rightarrow \) [History] –
  - (Attribute) \( \rightarrow \) [Life]
  - [It] (pronoun) –
  - (Attribute) \( \rightarrow \) [Similar]
  - (Mod) \( \rightarrow \) [Thunderstorm: \( \forall \)] (among)
Paraphrase Recognition

iSTART

• Definition

• Definition CGs

**Original sentence:** “All thunderstorms have a similar life history.”

[Thunderstorm: ∀] \rightarrow [History] \rightarrow (Feature) \rightarrow [Life] \rightarrow (Attribute) \rightarrow [Similar]

**Student’s Explanation:** “Thunderstorms go through similar cycles. They will begin the same, go through the same things, and end the same way.”

[Go] \rightarrow (Agent) \rightarrow [Thunderstorm: #] \rightarrow (Path) \rightarrow [Cycle] \rightarrow (Attribute) \rightarrow [Similar]

[Begin] \rightarrow (Agent) \rightarrow [Thunderstorm: #] \rightarrow (Attribute) \rightarrow [Same] \rightarrow [Go-Through] \rightarrow (Agent) \rightarrow [Thunderstorm: #] \rightarrow (Path) \rightarrow [Thing: ∃] \rightarrow (Attribute) \rightarrow [Same] \rightarrow [End] \rightarrow (Agent) \rightarrow [Thunderstorm: #] \rightarrow (Path) \rightarrow [Way: ∃] \rightarrow (Attribute) \rightarrow [Same]
There is someone happy”, we can say “Someone is happy”, “A person is happy”, or “There is a person who is happy”, etc.

Despite similar triplets (“[Person: \(\exists\)] \rightarrow (Char) \rightarrow [Happy]”) in CG, such structures cannot be universally defined as synonymous.

Unresolved in source research; further study needed.
Approaches to Paraphrase Generation

- Statistical (monolingual) Machine Translation
  - Microsoft 2004
- Machine Learning
- Rule based methods
- Thesaurus based methods
- Natural Language Generation approach
- Statistical Paraphrase Generation
Paraphrase Generation

SMT

- Statistical Machine Translation
  - Microsoft Research (Quirk et al, 2004)
  - Monolingual Translation: English → English = Paraphrase
  - Noisy Channel method (Brown et al, 1993)
  - Main problem: finding a sufficient corpus
    - Addressed with methods already considered (comparable online news articles)
  - Human assessed
Paraphrase Generation
SMT

- **Word Alignment Ingredients**
  - Giza++ word alignment algorithms (state of the art)
    - IBM models 1-5 ("noisy channel" machine translation, Brown 1993)
    - Hidden Markov Models
  - By combining multiple unidirectional word alignments, a single bidirectional alignment is produced.
Recognition
From Malakasiotis

Strip the sentence down into word stems

They had published an advertisement on the internet on June 10, offering the cargo for sale, he added.

On June 10, the ship’s owners had published an advertisement on the internet, offering the explosives for sale.
Paraphrase Generation
SMT

- Example Giza++ Monolingual Word Alignment
- Lines indicate monodirectional links.
- Ticks indicate direction of the link.

```
Of
the
%%NUMBER%%% wounded
, four
people
, including
one
child
, were
hospitalized
Dzheirkhanov
said
%%NUMBER%%% people
were
injured
and
that
four
people
, including
a
child
Dzheirkhanov
said
had
been
hospitalized
```
Paraphrase Generation

SMT

• **Alignment Evaluation**
  
  - Word alignments were evaluated by two humans. Judged as:
    - SURE: necessary
    - POSSIBLE: allowed, but not required.
  
  - Gold standard produced only SURE when both annotators agreed; otherwise produced POSSIBLE.
  
  - Let the “SURE” alignments be considered
    - the gold standard. For $A$ total alignments,
    - $S$ “SURE” alignments, and $P$ union of “SURE”
    - And “POSSIBLE” alignments:

\[
\text{precision} = \frac{|A \cap P|}{|A|}, \quad \text{recall} = \frac{|A \cap S|}{|S|},
\]

\[
\text{AER} = \frac{|A \cap P + A \cap S|}{|A + S|}
\]
Paraphrase Generation
SMT

- **Phrasal Replacements**
  - Simple SMT phrase-based systems outperform more complex word-based systems.
  - For source and target sentences $S$ and $T$, consider them as word sequences $s_1...s_n$ and $t_1...t_n$.
  - “Cept”: the aligned subset of these tokens.
  - “Phrase Pair”: subset of the cept in which both the source and target tokens are contiguous.
  - Phrase Pairs are collected into a database that forms the heart of the channel model.
  - Limit: not expanded to inter-phrase reordering.
Paraphrase Generation

SMT

- Paraphrase Generation

1) Preprocessing (tokenization, lowercase, basic NE recognition, etc)

2) Decoding (Translation/Generation)
   1) Tap the phrasal translation database to generate a lattice of possible paraphrases
Paraphrase Generation
SMT

- Paraphrase Generation (cont.)
  2) Find optimal path through the lattice (Viterbi algorithm)
  3) Post-processing: clean up products, eliminate too-near (boring) paraphrases.
Paraphrase Generation
SMT

- Assessing Results

<table>
<thead>
<tr>
<th>Method</th>
<th>B&amp;L59</th>
<th>B&amp;L59 + 141</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR #1</td>
<td>54 / 59 = 91.5%</td>
<td>177 / 200 = 89.5%</td>
</tr>
<tr>
<td>PR #2</td>
<td>53 / 59 = 89.8%</td>
<td>168 / 200 = 84.0%</td>
</tr>
<tr>
<td>PR #3</td>
<td>46 / 59 = 78.0%</td>
<td>164 / 200 = 82.0%</td>
</tr>
<tr>
<td>PR #4</td>
<td>49 / 59 = 83.1%</td>
<td>163 / 200 = 81.5%</td>
</tr>
<tr>
<td>MSA</td>
<td>46 / 59 = 78.0%</td>
<td>46 / 59 = 78.0%</td>
</tr>
<tr>
<td>PR #5</td>
<td>44 / 59 = 74.6%</td>
<td>155 / 200 = 77.5%</td>
</tr>
<tr>
<td>WN</td>
<td>23 / 59 = 39.0%</td>
<td>25 / 59 = 37.9%</td>
</tr>
<tr>
<td>WN+LM</td>
<td>30 / 59 = 50.9%</td>
<td>53 / 200 = 27.5%</td>
</tr>
<tr>
<td>CL</td>
<td>14 / 59 = 23.7%</td>
<td>26 / 200 = 13.0%</td>
</tr>
</tbody>
</table>

Table 2. Human acceptability judgments

<table>
<thead>
<tr>
<th></th>
<th>MSA</th>
<th>PR#1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rearrangement</td>
<td>28 / 59 = 47%</td>
<td>0 / 100 = 0%</td>
</tr>
<tr>
<td>Phrasal alternation</td>
<td>11 / 59 = 19%</td>
<td>3 / 100 = 3%</td>
</tr>
<tr>
<td>Info added</td>
<td>19 / 59 = 32%</td>
<td>6 / 100 = 6%</td>
</tr>
<tr>
<td>Info lost</td>
<td>43 / 59 = 73%</td>
<td>31 / 100 = 31%</td>
</tr>
</tbody>
</table>

Table 3. Qualitative analysis of paraphrases

Although resultant paraphrases were not as dramatic as Barzilay's Multi-Sequence Alignment method, they were also not template-based and could handle a much broader range of materials.
Paraphrase Generation
Other Methods

- Application Driven Statistical Paraphrase Generation (Zhao et al, 2009)

- Uses special filters to adjust output to a certain application
  - Sentence Simplification
  - Sentence Compression
  - Sentence Similarity (for comparison/improvement)
Paraphrase Generation

Other Methods

- Paraphrase Generation with Monte Carlo Sampling (Chevelu et al, 2009)
  - Applies transformation rules
    - No need for Left-to-right decoding (as in SMT)
    - Any state can be a final state
  - Changing the transformation rules can allow diverse paraphrase results.
  - An NLP approach
  - So far has a tendency to lose meaning
Paraphrase Generation
Other Methods

- **Syntactic Approaches to Paraphrase Generation** (Madsen, 2006)
  - Applies Link Grammar and Part of Speech Tagging to statistical methods
  - Applied a Giza++ phrasal (as opposed to word) alignment model
  - Used B-I-O chunking methods to move noun phrases, prepositional phrases, relative clauses, and complement clauses.
Paraphrase Generation
Other Methods

- Syntactic Approaches to Paraphrase Generation (Madsen, 2006) cont.

Excellent fluency results (double the baseline)

<table>
<thead>
<tr>
<th></th>
<th>Baseline (SMT-only)</th>
<th>ParaMeTer (Syntax-based)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluent</td>
<td>53 / 201 = 26.37%</td>
<td>119 / 201 = 59.20%</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>27 / 201 = 14.14%</td>
<td>60 / 201 = 29.85%</td>
</tr>
<tr>
<td>Rearrangement</td>
<td>161 / 201 = 84.29%</td>
<td>167 / 201 = 83.08%</td>
</tr>
<tr>
<td>Phrasal alternation</td>
<td>140 / 201 = 73.30%</td>
<td>63 / 201 = 31.34%</td>
</tr>
<tr>
<td>Info lost</td>
<td>145 / 201 = 75.9%</td>
<td>125 / 201 = 62.19%</td>
</tr>
<tr>
<td>Info added</td>
<td>75 / 201 = 39.26%</td>
<td>41 / 201 = 20.40%</td>
</tr>
</tbody>
</table>

Table 2: Human judgment results.
Recognizing Textual Entailment (RTE)

- Not identical to paraphrase recognition
  - Findings not limited to strict semantic equivalence.
Recognizing Textual Entailment (RTE)

- RTE Challenge
  - [http://pascallin.ecs.soton.ac.uk/Challenges/RTE/](http://pascallin.ecs.soton.ac.uk/Challenges/RTE/)
  - Motivated by a need for improved semantic search capabilities.
  - Provided a text snippet $T$ and a hypothesis $H$ of some entailment of that snippet, programs must determine whether $T$ entails $H$.

<table>
<thead>
<tr>
<th>ID</th>
<th>TEXT</th>
<th>HYPOTHESIS</th>
<th>TASK</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>iTunes software has seen strong sales in Europe.</em></td>
<td>*Strong sales for <em>iTunes in Europe.</em></td>
<td>IR</td>
<td>True</td>
</tr>
<tr>
<td>2</td>
<td><em>Cavern Club sessions paid the Beatles £15 evenings and £5 lunchtime.</em></td>
<td><em>The Beatles perform at Cavern Club at lunchtime.</em></td>
<td>IR</td>
<td>True</td>
</tr>
<tr>
<td>3</td>
<td><em>American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union’s bid to block the job losses.</em></td>
<td><em>American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.</em></td>
<td>PP</td>
<td>False</td>
</tr>
<tr>
<td>4</td>
<td><em>The two suspects belong to the 30th Street gang, which became embroiled in one of the most notorious recent crimes in Mexico: a shootout at the Guadalajara airport in May, 1993, that killed Cardinal Juan Jesus Posadas Ocampo and six others.</em></td>
<td><em>Cardinal Juan Jesus Posadas Ocampo died in 1993.</em></td>
<td>QA</td>
<td>True</td>
</tr>
</tbody>
</table>

Table 1: Examples of Text-Hypothesis pairs
International Competition on Author Identification

Tasks

Authorship clustering/intrinsic plagiarism
Given a text, cluster the paragraphs into exactly two clusters: one that includes paragraphs written by the "main" author of the text and another that includes all paragraphs written by anybody else.

Sexual Predator Identification
Given chat logs involving two (or more) people, determine who is the one trying to convince the other participants(s) to provide some sexual favour. You will also need to identify the particular conversation where the person exploits his bad behavior.
# International Competitions on Author Identification

## Results

<table>
<thead>
<tr>
<th>Participant main run</th>
<th>Retrieved</th>
<th>Relevant</th>
<th>P</th>
<th>R</th>
<th>F(β=1)</th>
<th>F(β=0.5)</th>
<th>Rank F(β=0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>villatorotello-run-2012-06-15-2157g</td>
<td>204</td>
<td>200</td>
<td>0.9804</td>
<td>0.7874</td>
<td>0.8734</td>
<td>0.9346</td>
<td>1</td>
</tr>
<tr>
<td>snider12-run-2012-06-16-0032</td>
<td>186</td>
<td>183</td>
<td>0.9839</td>
<td>0.7205</td>
<td>0.8318</td>
<td>0.9168</td>
<td>2</td>
</tr>
<tr>
<td>parapar12-run-2012-06-15-0959j</td>
<td>181</td>
<td>170</td>
<td>0.9392</td>
<td>0.6693</td>
<td>0.7816</td>
<td>0.8691</td>
<td>3</td>
</tr>
<tr>
<td>morris12-run-2012-06-16-0752-main</td>
<td>159</td>
<td>154</td>
<td>0.9686</td>
<td>0.6063</td>
<td>0.7458</td>
<td>0.8652</td>
<td>4</td>
</tr>
<tr>
<td>eriksson12-run-2012-06-15-1949</td>
<td>265</td>
<td>227</td>
<td>0.8566</td>
<td>0.8937</td>
<td>0.8748</td>
<td>0.8638</td>
<td>5</td>
</tr>
<tr>
<td>peersman12-run-2012-06-15-1559</td>
<td>170</td>
<td>152</td>
<td>0.8941</td>
<td>0.5984</td>
<td>0.7170</td>
<td>0.8137</td>
<td>6</td>
</tr>
<tr>
<td>grozea12-run-2012-06-14-1706b</td>
<td>215</td>
<td>163</td>
<td>0.7581</td>
<td>0.6417</td>
<td>0.6951</td>
<td>0.7316</td>
<td>7</td>
</tr>
<tr>
<td>sitarz12-run-2012-0615-1515</td>
<td>218</td>
<td>159</td>
<td>0.7294</td>
<td>0.6260</td>
<td>0.6737</td>
<td>0.7060</td>
<td>8</td>
</tr>
<tr>
<td>vartapetian12-run-2012-06-15-1411</td>
<td>160</td>
<td>99</td>
<td>0.6188</td>
<td>0.3898</td>
<td>0.4783</td>
<td>0.5537</td>
<td>9</td>
</tr>
<tr>
<td>kontostathis-run-2012-06-16-0317e</td>
<td>475</td>
<td>170</td>
<td>0.3579</td>
<td>0.6693</td>
<td>0.4664</td>
<td>0.3946</td>
<td>10</td>
</tr>
<tr>
<td>kang12-run-2012-06-15-0904b</td>
<td>930</td>
<td>203</td>
<td>0.2183</td>
<td>0.7992</td>
<td>0.3429</td>
<td>0.2554</td>
<td>11</td>
</tr>
<tr>
<td>kern12-run-2012-06-18-1827b</td>
<td>1172</td>
<td>177</td>
<td>0.1510</td>
<td>0.6969</td>
<td>0.2482</td>
<td>0.1791</td>
<td>12</td>
</tr>
<tr>
<td>bogdanova12-run-2012-06-14-1117</td>
<td>2109</td>
<td>55</td>
<td>0.0261</td>
<td>0.2165</td>
<td>0.0466</td>
<td>0.0316</td>
<td>13</td>
</tr>
<tr>
<td>prasath12-run-2012-06-15-2122</td>
<td>10289</td>
<td>207</td>
<td>0.0201</td>
<td>0.8150</td>
<td>0.0393</td>
<td>0.0250</td>
<td>14</td>
</tr>
<tr>
<td>vilarino12-run-2012-06-14-2121b</td>
<td>5225</td>
<td>98</td>
<td>0.0188</td>
<td>0.3858</td>
<td>0.0358</td>
<td>0.0232</td>
<td>15</td>
</tr>
<tr>
<td>gomezhidalgo12-run-2012-06-15-1900</td>
<td>150</td>
<td>1</td>
<td>0.0067</td>
<td>0.0039</td>
<td>0.0050</td>
<td>0.0059</td>
<td>16</td>
</tr>
</tbody>
</table>
Table 2.3: Statistics for the PAN-PC-09 Corpus [Stein et al., 2009]

<table>
<thead>
<tr>
<th>Document Purpose</th>
<th>Document Statistics</th>
<th>Obfuscation Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Document Length</td>
</tr>
<tr>
<td>source documents</td>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>suspicious documents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>— with plagiarism</td>
<td>25%</td>
<td>15% paraphrasing</td>
</tr>
<tr>
<td>— without plagiarism</td>
<td>25%</td>
<td>15% automatic (low)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20% automatic (high)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10% translation (de, es to en)</td>
</tr>
</tbody>
</table>

Table 2.4: Statistics for the PAN-PC-10 Corpus [Potthast et al., 2010a]

<table>
<thead>
<tr>
<th>Document Purpose</th>
<th>Document Statistics</th>
<th>Document Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>source documents</td>
<td>50%</td>
<td>45% short (1-10 pp.)</td>
</tr>
<tr>
<td>suspicious documents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>— with plagiarism</td>
<td>25%</td>
<td>25% medium (10-100 pp.)</td>
</tr>
<tr>
<td>— without plagiarism</td>
<td>25%</td>
<td>15% long (100-1000 pp.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obfuscation</th>
<th>Case Length</th>
<th>Topic Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>40%</td>
<td>34% intra-topic cases</td>
</tr>
<tr>
<td>artificial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>— low obfuscation</td>
<td>20%</td>
<td>33% inter-topic cases</td>
</tr>
<tr>
<td>— high obfuscation</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>simulated</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>translated</td>
<td>14%</td>
<td></td>
</tr>
</tbody>
</table>
None Obfuscation (Artificial)

**Source:** The first agrarian movement after the enactment of lex Licinia took place in the year 338, after the battle of Veseris in which the Latini and their allies were completely conquered.

**Rewrite:** The first agrarian movement after the enactment of lex Licinia took place in the year 338, after the battle of Veseris in which the Latini and their allies were completely conquered.

<table>
<thead>
<tr>
<th>Low Obfuscation (Artificial)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source:</strong> Mr. Loring P. Rixford, Room 24, Menisini Building, 231 Post Street, San Francisco, is secretary. Brochure Series Competitions. From time to time, as opportunity offers, competitions in design will be conducted by THE BROCHURE SERIES. An upright or cabinet piano case, the subject of the first one, badly needs the attention of good designers.</td>
</tr>
<tr>
<td><strong>Rewrite:</strong> Loring P. Rixford, area 24, Menisini edifice, 231 position Street, San Francisco, is head. book serial game. From day to case, as opportunity offers, contestant in arrangement will be conducted by THE BROCHURE SERIES. An erect or cabinet softness humiliation, the subject of the first one, well necessitate the work of good designers.</td>
</tr>
</tbody>
</table>