Learning Analytics

Dr. Bowen Hui
Computer Science
University of British Columbia Okanagan

Exercise: 5 minutes

- Write a coherent English paragraph (between 75–100 words) on the following topic:
  - "Why study Computer Science?"
  - Target at 6 year olds

- Use *any* source you want (book, friend, Internet, etc.)

Exercise: 5 minutes

- Write a coherent English paragraph (between 75–100 words) on the following topic:
  - "Why study Computer Science?"
  - Target at 6 year olds
- Use *any* source you want (book, friend, Internet, etc.)
- Submit it on Canvas today by 2:30pm
  - +3 pts bonus towards A6

Plagiarism: Natural Language Example

Student Writer A:

Long ago, when there was no written history, these islands were the home of millions of happy birds; the resort of a hundred times more millions of fishes, sea lions, and other creatures. Here lived innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government.

Image taken from www.bowdoin.edu

Source:

"In ages which have no record these islands were the home of millions of happy birds, the resort of a hundred times more millions of fishes, of sea lions, and other creatures whose names are not so common; the marine residence, in fact, of innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government."[1]

[1]A.J. Duffield, The Prospects of Peru (London: Newman, 1881) 78.

- How to detect cheating automatically?

Plagiarism: Natural Language Example

Student Writer A:

Long ago, when there was no written history, these islands were the home of millions of happy birds; the resort of a hundred times more millions of fishes, sea lions, and other creatures. Here lived innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government.

Source:

"In ages which have no record these islands were the home of millions of happy birds, the resort of a hundred times more millions of fishes, of sea lions, and other creatures whose names are not so common; the marine residence, in fact, of innumerable creatures predestined from the creation of the world to lay up a store of wealth for the British farmer, and a store of quite another sort for an immaculate Republican government."[1]

[1]A.J. Duffield, The Prospects of Peru (London: Newman, 1881) 78.

Image taken from www.bowdoin.edu

- How to detect cheating automatically?
  - Analyze text reuse between text excerpts

According to Harvard Guide…

- Verbatim plagiarism
- Mosaic plagiarism
- Inadequate paraphrase
- Uncited paraphrase
- Uncited quotations
- Using material from another student's work

A closer look at these

Ex: Mosaic Plagiarism

Image taken from http://isites.harvard.edu/icb/icb.do?keyword=k70847&pageid=icb.page342054

Plagiarized version

In order to advocate the use of the sitcom *Scrubs* as part of the medical education system, it is also important to look at the current bioethical curriculum. Medical school curriculum does not focus adequately on the moral issues that doctors face in the clinic. In fact, in more than 3500 hours of training that students undergo in medical school, only about 60 hours are focused on bioethics, health law, and health economics. It is also problematic that students receive this training before they actually go on to their hands-on, clinical training (Persad et al, 2008). Most of these hours are taught by instructors without current publications in the field.

By watching episodes of *Scrubs*, however, medical students would have the chance to watch people and hear their voices, providing a much better test of clinical skills of observation than you can get from reading words on a page. One must see a patient's body language and hear her tone of voice if one is to become a better observer, and watching the patients on television would provide a good opportunity for medical students to do so. Perhaps even more significantly, medical students would be introduced to certain issues, and while the experiences may not be their own, they would be effective in helping them to understand those experiences as they empathize with the characters.

The information in this sentence is drawn directly from Persad, but because the student ends the citation of Persad above, this sentence appears to be the student's own idea.

Everything up to this point in the paragraph is either paraphrased or taken verbatim from Spike, but the student does not cite Spike. As a result, readers will assume that the student has come up with these ideas himself.

The student has come up with the idea about the role of empathy on his own, but because nothing in the paragraph is cited, it seems to be part of a whole paragraph of his ideas, rather than the point that he is building from Spike's ideas.

Ex: Inadequate Paraphrase

Image taken from http://isites.harvard.edu/icb/icb.do?keyword=k70847&pageid=icb.page342054

Source material

So in *Romeo and Juliet*, understandably in view of its early date, we cannot find that tragedy has fully emerged from the moral drama and the romantic comedy that dominated in the public theaters of Shake>speare's earliest time. Here he attempted an amalgam of romantic comedy and the tragic idea, along with the assertion of a moral lesson which is given the final emphasis—although the force of that lesson is switched from the lovers to their parents. But tragedy is necessarily at odds with the moral: it is concerned with a permanent anguishing situation, not with one that can either be put right or be instrumental in teaching the survivors to do better.

--Leech, Clifford. "The Moral Tragedy of *Romeo and Juliet*." *Critical Essays on* Romeo and Juliet. Ed. Joseph A. Porter. New York: G.K. Hall, 1997. 20. Print.

Plagiarized version

In his essay, "The Moral Tragedy of *Romeo and Juliet*," Clifford Leech suggests that rather than being a straight tragedy, *Romeo and Juliet* is a mixture of romantic comedy and the tragic idea, and that it asserts a moral lesson which is given the final emphasis. The impact of the moral lesson is switched from the lovers to the parents (20).

This is an inadequate paraphrase because the student has only replaced a few words ("mixture" for "amalgam"; "asserts a moral lesson" for "assertion of a moral lesson"; "impact" for "force") while leaving the rest of Leech's words intact.

Plagiarism: A Complex Problem

Image taken from www.facultyfocus.com

- With N submissions, how many comparisons do you have to make to check for plagiarism?

Everyone is Weird in Their Own Way

- **Assumption**: Individuals have idiosyncrasies
  - Humans have peculiar behaviours
  - Different people express a concept in different ways

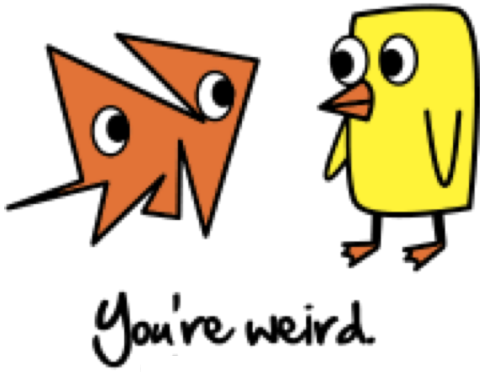


Image taken from spreadshirt.ca

Exploiting Idiosyncrasies

- **Claim: When the exact same expression is observed, there is a high chance of plagiarism**

- Consider a question from your parents:
  - What do you learn in COSC 111?

Exploiting Idiosyncrasies

- **Claim: When the exact same expression is observed, there is a high chance of plagiarism**

- Consider a question from your parents:
  - What do you learn in COSC 111?
    “Programming” vs. “Coding”

Exploiting Idiosyncrasies

- **Claim: When the exact same expression is observed, there is a high chance of plagiarism**

- Consider a question from your parents:
  - What do you learn in COSC 111?
    "Programming" vs. "Coding"

- Think about the overlap of subsequences
  - Subsequences of length >= n
  - As n gets bigger, overlap becomes less likely
- Simple method: n-gram similarity measures (Clough, Gaizauskas, Piao, & Wilks 2002)

Word N-Grams

- An n-gram is a sequence of n contiguous items
- A word n-gram is a sequence of n contiguous words
- In computational linguistics, n-grams are used to model language
  - E.g., Predict next word based on previous words typed
- Commonly used: bigrams (n=2), trigrams (n=3)

Word Bigram Example

- Example sentence: “Sam Sheep can’t sleep”
- Bigrams from sentence:
  - Sam Sheep
  - Sheep can’t
  - Can’t sleep

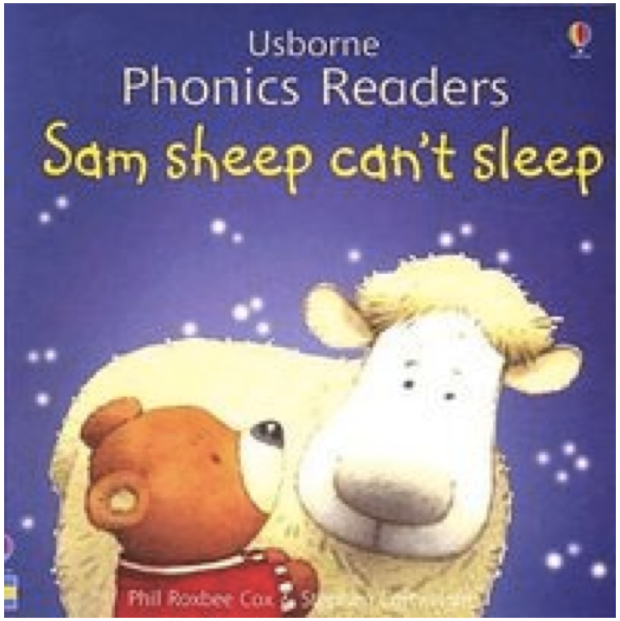


Image taken from www.goodreads.com

Exercise

- Sentence: “fat cat can sleep for weeks and weeks” (N = 8)
- List the unigrams, bigrams, and trigrams in this sentence

Unigrams: **Bigrams:** **Trigrams:**

Exercise

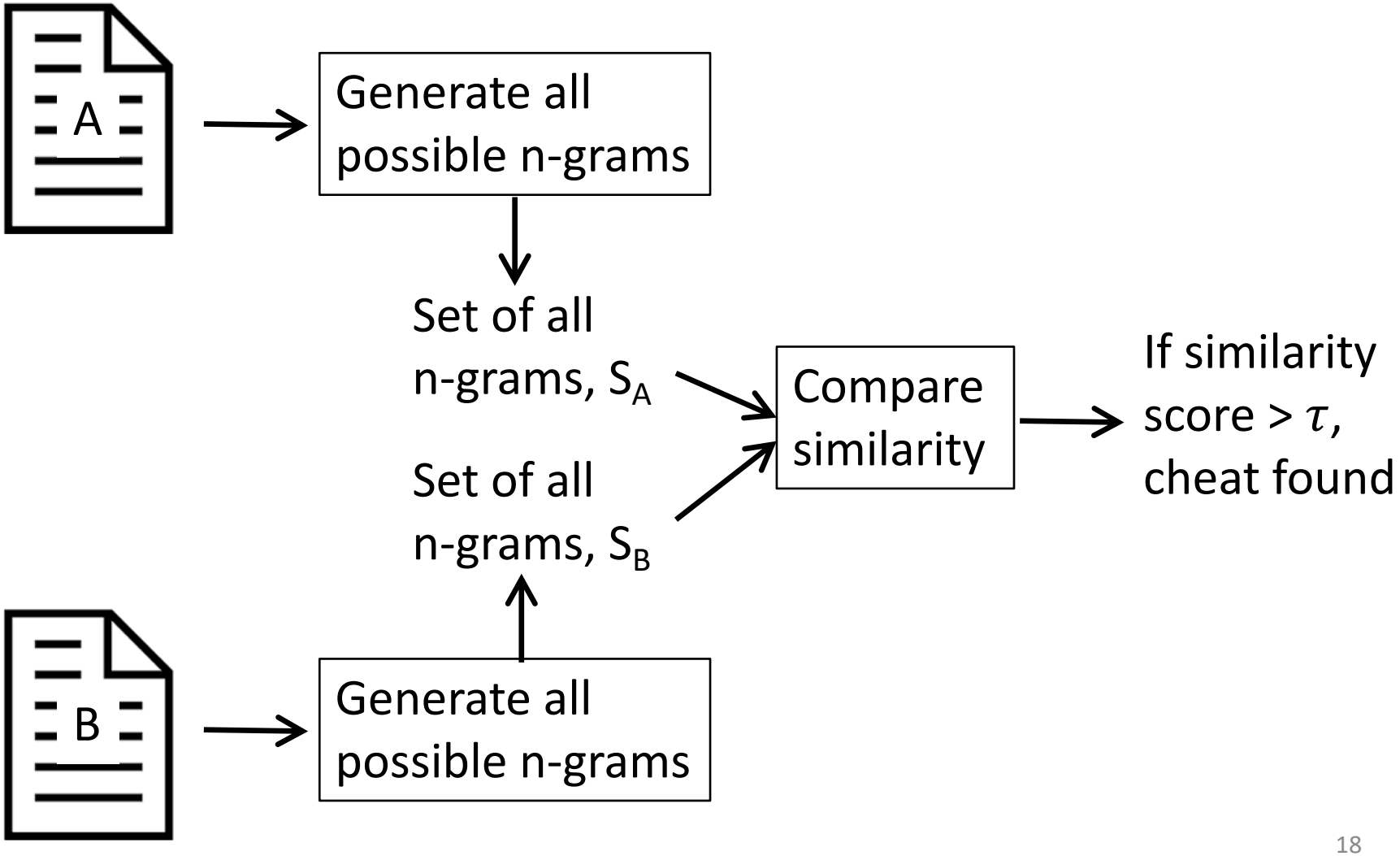
- Sentence: “fat cat can sleep for weeks and weeks” (N = 8)
- List the unigrams, bigrams, and trigrams in this sentence

Unigrams:
fat
cat
can
sleep
for
weeks
and
weeks

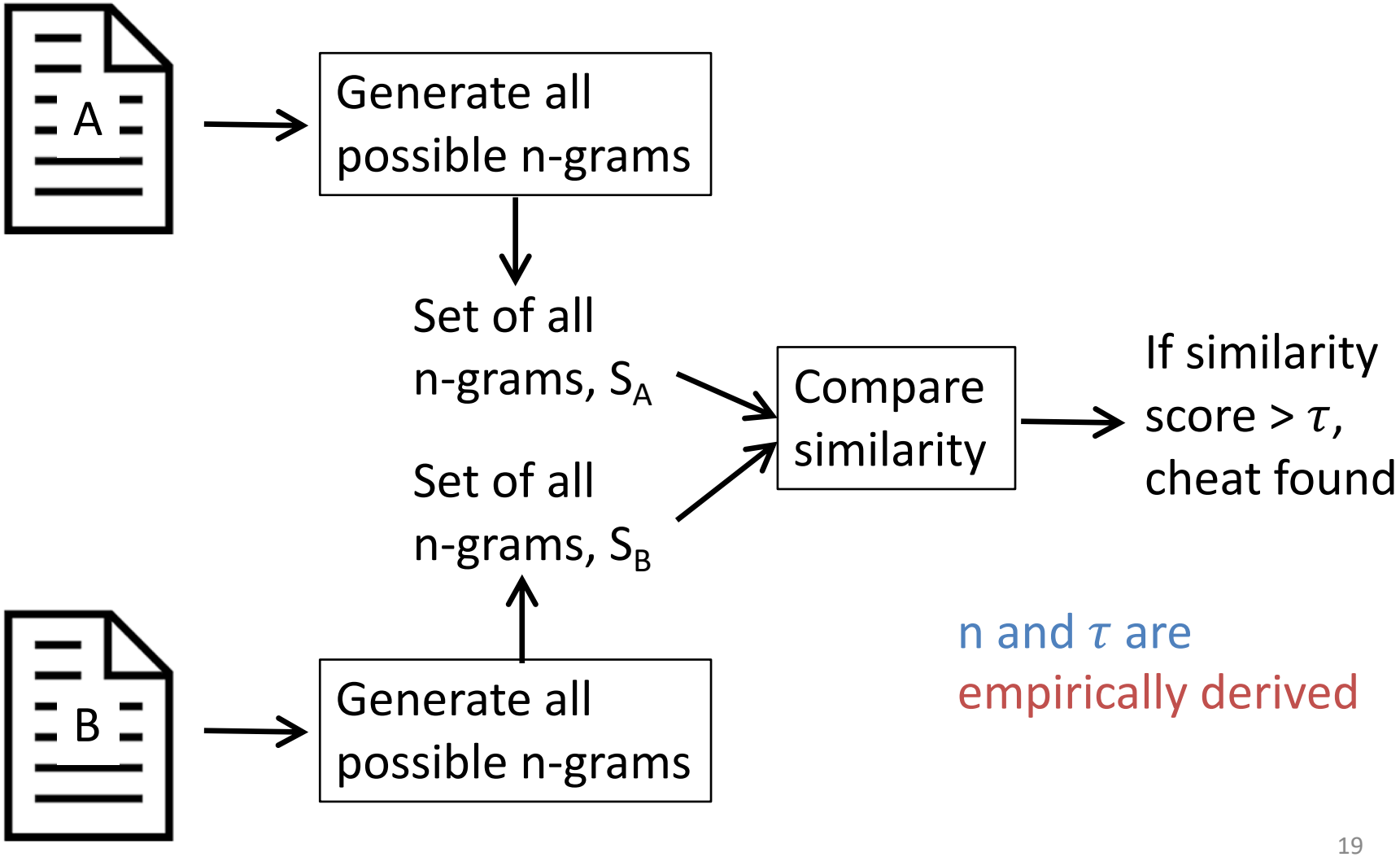
Bigrams:
fat cat
cat can
can sleep
sleep for
for weeks
weeks and
and weeks

Trigrams:
fat cat can
cat can sleep
can sleep for
sleep for weeks
for weeks and
weeks and weeks

Plagiarism Detection Algorithm



Plagiarism Detection Algorithm



n and τ are empirically derived

Measuring Document Similiarity

- Treat each document as a set of overlapping word n-grams
- Compute similarity score based on a set-theoretic notion of containment defined as:

$$C(A,B) = \frac{|S_A \cap S_B|}{|S_B|}$$

- where:
  - A is the source text
  - B is the suspicious text
  - S_A is the set of word n-grams in A
- C(A,B) ranges in [0,1] indicating none to all overlap

Measures percentage of overlap between texts A and B

Time Complexity

- Comparing the full document may be too expensive
  - Let *d* be the length of a document
  - Let *n* be the number of students in a class

Time Complexity

- Comparing the full document may be too expensive
  - Let *d* be the length of a document
  - Let *n* be the number of students in a class
  - Time to strip out n-grams from one doc takes *O(d)*
  - Compute similarity with every potential source takes $O(n^2)$
    - Create set intersection at best *O(d)*

Time Complexity

- Comparing the full document may be too expensive
  - Let *d* be the length of a document
  - Let *n* be the number of students in a class
  - Time to strip out n-grams from one doc takes *O(d)*
  - Compute similarity with every potential source takes $O(n^2)$
    - Create set intersection at best *O(d)*

- Example: ENGL 112
  - n: Class has about 200 students
  - d: 5-page essay is about 1,200 words
  - $n^2*d + 2d = 48{,}001{,}200$ steps
  - If each step took 1 ms, algorithm would take 13 hours to complete

Fingerprinting

- Create a fingerprint of a document to represent text in a compact form
  - Select a subset of n-grams from the text
  - Generate a hash value for each n-gram

Fingerprinting

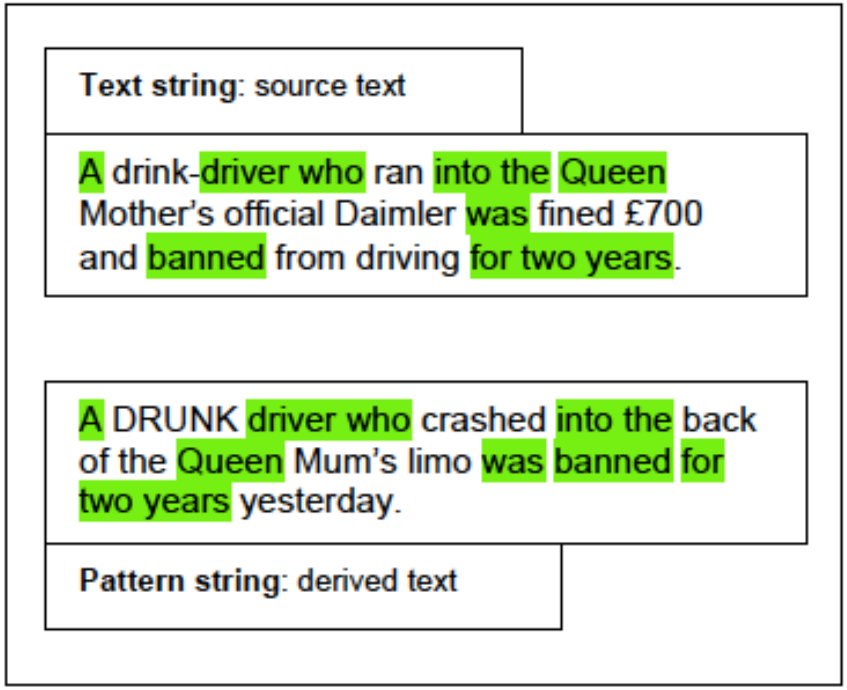
- Create a fingerprint of a document to represent text in a compact form
  - Select a subset of n-grams from the text
  - Generate a hash value for each n-gram

- Common methods:
  - Full fingerprinting: Use all possible n-grams
  - k^{th}-in-sentence: For each sentence, select the n-gram that starts with the k^{th} word

Obstacles

- Data sparseness problem: for large n, n-grams becomes more unlikely
- Doesn't handle domain-specific terminology
- Doesn't detect "smart" plagiarism, e.g.:

Image taken from (Clough 2003)



Alternative Scenario

- **Scenario:** You're in ENGL 112 and are told to write a 10-page essay (approx. 2,500 words). You spend nights writing this essay, and finally get to 9 pages the night before the essay is due. But you fell asleep. Out of pity, your brother writes the last page for you.

- **Problem:** How to detect the last page was written by/taken from someone else?

General Types of Plagiarism Detection

- Extrinsic detection
  - Given a body of source documents and a suspicious document, check to see if pieces of the suspicious doc came from any of the sources
- Intrinsic detection
  - Given one document, see if it is coherent or if pieces of it came from somewhere else
  - Does not use any external sources for comparison

How to tackle this problem?

Exploiting Idiosyncrasies Revisited

- **Solution:** Detect stylistic inconsistency within one text
  - Inconsistency serves as an indicator that multiple authors were involved

- Examples of writing *habits*:
  - Use of particular words (vocabulary richness)
  - Grammatical variations (e.g., punctuation)
  - Average sentence length
  - Frequency of passive voice
  - Distribution of word classes (nouns, adjectives, etc.)
  - Types of errors made

Key Idea in Approach

- **Assumption**: Everyone has a unique set of quantifiable habits

- **Claim**: Variation in the observed occurrences of the habit indicates multiple authorship

The cusum Technique

- cusum is short for cumulative sum
- Given a habit
  - Compute its average value
  - Compare against average sentence length
  - Rate of habit is expected to be consistent
- Analysis using a cumulative sum chart to plot cumulative deviation from the mean
- Provides a cumulative measure of homogeneity

Comparison Steps

- #1: Sentence length baseline
- Let w_r be the number of words in sentence
- Compute average sentence length, $\bar{w}$
- Compute cumulative variance:
  - Variance in length for each sentence is $w_r - \bar{w}$
  - Sum this with sentences preceding it

where:

$$c_i = \sum_{r=1}^{i} (w_r - \bar{w}) \quad \text{or} \quad c_i = \sum_{r=1}^{i} v_r$$

c_i is the cusum value for sentence I

v_r is the variance of the r^{th} sentence

Example

- Paragraph with 4 sentences all of length 5
- $\bar{w} = 5$
- $c_1 = w_r - \bar{w} = 0$
- $c_2 = \sum_{r=1}^{2}(w_r - \bar{w}) = 0 + v_1 = 0$
- $c_3 = \sum_{r=1}^{3}(w_r - \bar{w}) = 0 + v_2 + v_1 = 0$
- $c_4 = \sum_{r=1}^{4}(w_r - \bar{w}) = 0 + v_3 + v_2 + v_1 = 0$

- Text that is entirely consistent will have no cumulative variance

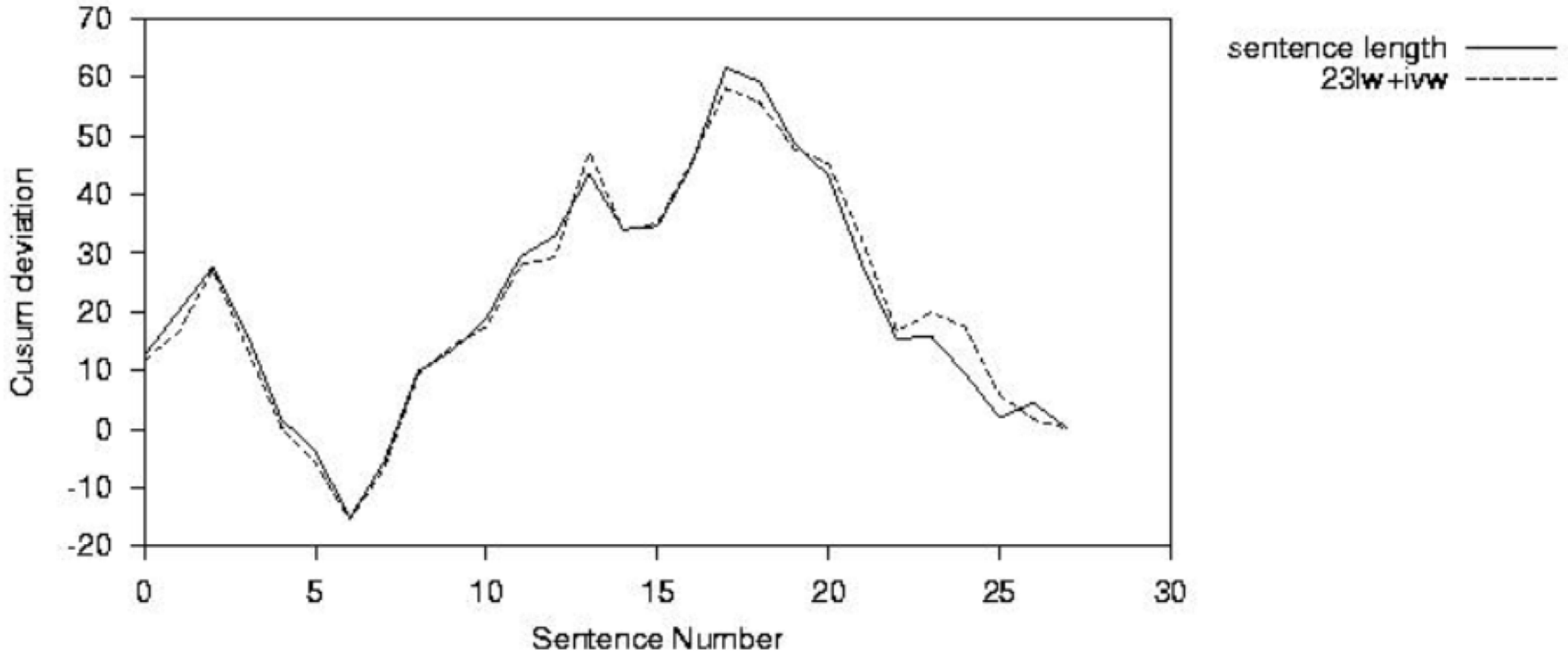
Comparison Steps (cont.)

- #2: Habit computation
- Pick a habit
  - E.g. 2-3 letter words
- Compute average number of habit occurrences per sentence
- Compute cumulative habit variance in the same way

- #3: Plot results of #1 against results of #2

Example Results: One Author

Image taken from Clough (2003):

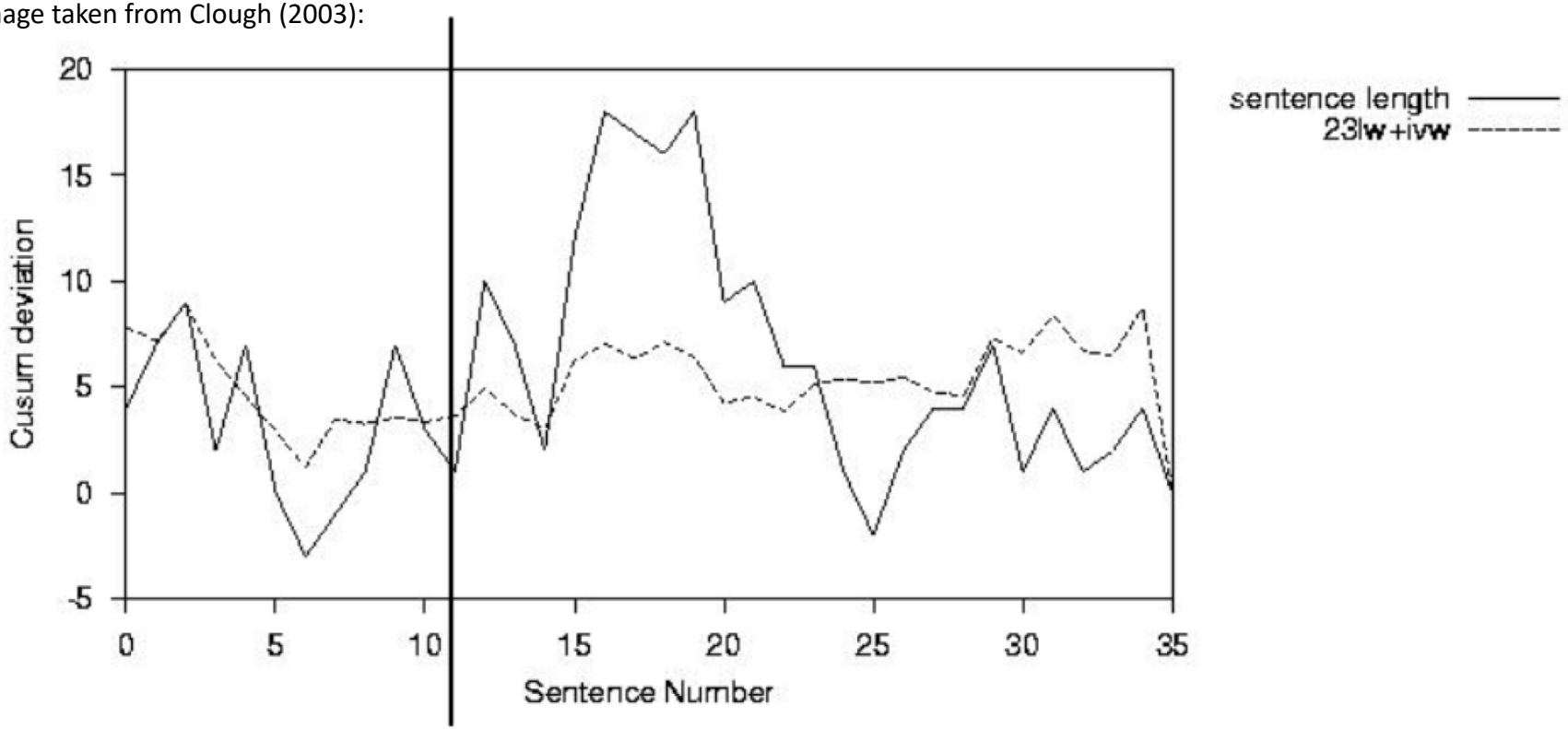


Habit: 2-3 letter words + initial vowel word

Text: BBC news story

Example Results: Multiple Authors

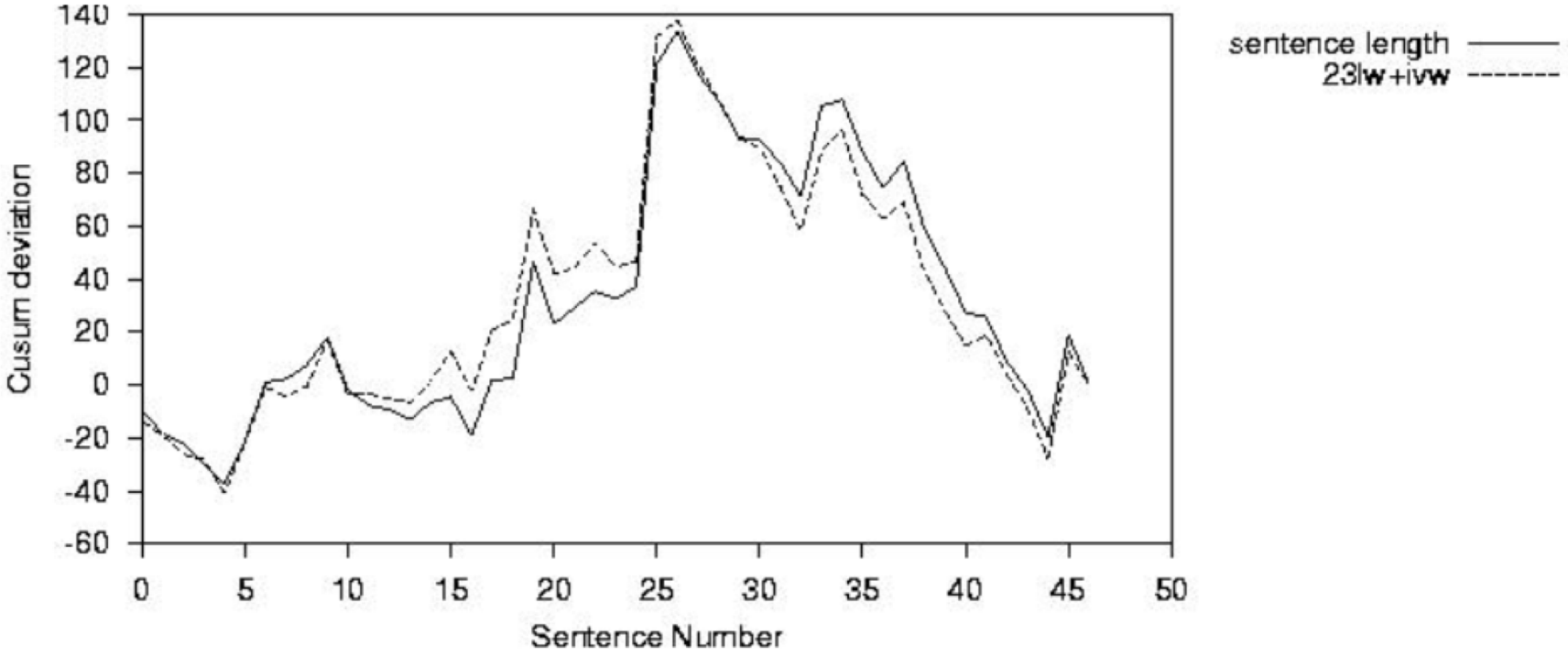
Image taken from Clough (2003):



Text: Two combined news stories on the same topic (from the Sun and Mirror)

Example Results: One Author?

Image taken from Clough (2003):



Text: Chapter 1 of Jane Austin's Northanger Abbey

Lessons Learned

- Choice of habit may not be appropriate
  - Need to try different habits
  - Comparison using multiple habits

- Amount of tolerable deviation must also be empirically derived
  - Is having help from editor/proofreader cheating?

Natural vs. Programming Language

- Natural language
  - Highly complex
  - Lots of redundancies = stylistic variation
  - Open to change (through time)

- Programming language
  - Subset of natural language
  - Highly structured
  - Closed language

Natural vs. Programming Language

- Natural language
  - Highly complex
  - Lots of redundancies = stylistic variation
  - Open to change (through time)

Easy to detect plagiarism

- Programming language
  - Subset of natural language
  - Highly structured
  - Closed language

Much harder to detect

Original source (computer program)

```
private static int partition(Comparable[] a, int lo, int hi)
{
   int i = lo, j = hi+1;
   Comparable v = a[lo];
   while (true)
   {
      while (less(a[++i], v)) if (i == hi) break;
      while (less(v, a[--j])) if (j == lo) break;
      if (i >= j) break;
      exch(a, i, j);
   }
   exch(a, lo, j);
   return j;
}
```

Computer program example 2. Plagiarized code structure with inconsequential changes

```
private static int partition(int[] bob, int left, int right) {
    int x = left;
    int y = right+1;
    for (;;) {
        while (less(bob[left], bob[--y]))
            if (y == left) break;
        while (less(bob[++x], bob[left]))
            if (x == right) break;
        if (x >= y) break;
        swap(bob, y, x);
    }
    swap(bob, y, left);
    return y;
}
```

Example taken from http://www.princeton.edu/pr/pub/integrity/pages/plagiarism/

```
Source: quicksort (int a [ ], int l, int r)
     {
         int v, i, j, t;
         if (r > l)
            {
               v = a [ r ]; i = l-1; j = r;

               for ( ; ; )
                   {
                      while (a [++i ] < v) ;
                      while (a [--j] > v);
                      if (i >= j) break;
                      t = a [i]; a [i] = a [j]; a [j] = t;
                   }
               t = a [i]; a [i] = a [r]; a [r] = t;
               quicksort (a, l, i-1);
               quicksort (a, i+1, r);
            }
     }
```

Student submission:

Equivalent to paraphrasing without citation

```
#define Swap(A,B) { temp=(A); (A)=(B); (B)=A;}

void mysort (const int* data, int x, int y){
   int temp;
   while (y > x){
      int pivot = data[y];
      int i = x-1;
      int j = r;
      while (1){
         while (data [++i] < pivot){/*do nothing*/}
         while (data --j] > pivot){/*do nothing*/}
         if (i >= j) break;
         swap (data [i], data [y];
      }
   swap (data [i], data [j];
   mysort (data, x, i-1);
   x = i+1;
   }
}
```

Example taken from
http://www.upenn.edu/academicintegrity/ai_computercode.html

Key Ideas

- People exhibit idiosyncrasies that make them unique
- Extrinsic and intrinsic detection exploits these idiosyncrasies
- Representation:
  - Word n-grams is a Markov model of words
  - Fingerprinting provides a compact representation of documents
- Algorithm:
  - Set containment to measure similarity between two sets of fingerprints
  - cusum technique analyses variance of habits within single text