
Discovering Similar Passages Within Large Text Documents

Demetrios Glinos
glinos@eecs.ucf.edu



UNIVERSITY OF CENTRAL FLORIDA

The Problem Domain

- The task is to find one or more passages in one document that are the same or closely similar to passages in another document.
- There can be more than one matching set of passages in a given document pair.
- Corresponding passages may not be in the same order in each document.
- Corresponding passages need not be identical, only similar:
 - Additions or deletions of words and phrases
 - Use of synonyms
 - Alternate grammatical constructions
- Each passage pair, however, presents a text alignment problem.



Application Areas

- Document deduplication
 - **Example:** Recognizing that two documents represent the same content when building a database of medical journal articles or abstracts retrieved from different online sources.
- Textual Entailment Determination
 - **Example:** Recognizing that two sentences mean the same thing despite different grammatical constructions and that can spoof deep parsers.
- Plagiarism Detection
 - **Example:** Recognizing that one document contains substantial passages that have been copied, perhaps modified, from another.



A Simple Example of Cut-and-Paste

- Here, the task is simply to *find* the corresponding passage(s), if any.

The screenshot shows a text alignment tool with two panes. The left pane, titled 'Suspect: original', contains a long paragraph of text about GMAT preparation. The right pane, titled 'Source: original', contains a shorter paragraph of text about GMAT preparation resources. Both panes have tabs for 'Tokens', 'Alignments', and 'Stats'. The suspect text has several lines highlighted in yellow.

Suspect: original

for Atlanta or program MBA MAcc () Online Application Provide transcripts to show a 4-year degree from regionally accredited institution with a minimum 2.5 GPA International applicants must have achieved the minimum required TOEFL score provide evaluation on any international transcripts showing equivalency to a U.S. 9 practice tests, extensive online resources, over 3000 realistic GMAT questions exclusive to Kaplan, and GMAC's with retired GMAT questions Includes access to state-of-the-art virtual classroom and complimentary headset Focuses on advanced content Official Guide 9 practice tests, extensive online resources, over 3000 realistic GMAT questions exclusive to Kaplan, and GMAC's with retired GMAT questions Official Guide Timeline You schedule the timeline with your tutor Usually 6-12 weeks Usually 4-9 weeks, although other options are available Usually 4-9 weeks You control the pace, although suggested time is 4-6 weeks Practice is when you sit down and practice those methods. The best move is to buy one, and only one, commercial test preparation book, complete it, and then practice the methods you've learned on former GMAT exams, such as those found in ETS', plus practice on practice GMAT CATs (computer adaptive tests). You'll need to make a decision on which commercial book you pick - but pick one and stick with it. **You can succeed on the GMAT by preparing in depth for the different question types and being able to take difficult questions and break them down into easier parts that you can quickly solve. Repetition and thorough preparation is a process that rewards those willing to work hard, which means that passing the GMAT is within the reach of virtually anyone willing to invest the time in learning how to handle any question they might face on test day.** Class Time Choose from 15, 25, or 35 tutoring hours Includes free access to GMAT classroom or Advanced course. 12 lessons including diagnostic test, including 3 math-content review sessions 2 hours of one-on-one tutoring 32.5 total lesson hours 9 lessons include diagnostic test 23 total lesson hours 10 lessons 2 hours of one-on-one tutoring 27 total lesson hours 9 lessons 9 lessons include diagnostic test They have over 5,000 GMAT annually students at 65 locations worldwide. Their admissions consultant service has former admissions officers from most of the top 20 business schools. BusinessWeek Over 2000 pages of materials 15 practice tests 42 hours of class instruction Unlimited instructor phone and email tutoring support Instructors have scored above the 99th percentile (760+ GMAT) Classroom Products: Weekend/ Focus course: \$700. 42-hour course \$1400 Online courses: \$750 or \$1100 (live) The GMAT consists of three main parts. The consists of two basic writing tasks -- Analysis of an Issue and Analysis of an Argument. You are allowed 30 minutes to complete each one. Analytical Writing Assessment The contains 37 multiple-choice questions of two question types -- Data Sufficiency and Problem Solving. You are allowed a maximum of 75 minutes to complete the entire section. Quantitative section The contains 41 multiple-choice questions of three question types -- Reading Comprehension, Critical Reasoning, and Sentence Correction. You are allowed 75 minutes to complete this entire section. Verbal

Source: original

Select Meaning By Word
GMAT verbal word list: Improving Vocabulary: English Verb GMAT practice test: Improve Your Vocabulary: Noun GMAT verbal test: Vocabulary Answers: Adjective GMAT test: Vocabulary Test: Noun Vocabulary List GMAT preparation test: Vocabulary Exercise: Verb GMAT test: Vocabulary Exercises: Adjective Verb GMAT verbal preparation: Vocabulary Tests
will help you learn new phrases, idioms, expressions and English grammar structures every single day. And you won't even have to cram any grammar rules or vocabulary words into your head. Instead, you will be absorbing bits and pieces of the English language almost without realizing it. This compact is the only printable English test and flash card collection currently available on the Internet. It contains GMAT vocabulary tests. You will also get GMAT Words as Flash cards and GMAT Vocabulary word list in alphabetical order. This will help learn all the most essential vocabulary words you need if you want to pass the GMAT exam. GMAT Test Package GMAT Test Package 3801.400 unique GMAT test prep system

I want to prepare for the (graduate management admission test) to reach a high score. At English-test.net I can take free interactive questions to increase my GMAT vocabulary and learn the . In addition I can speak to other people who are preparing for the GMAT to share experiences on the MBA GMAT forum. Links to GMAT prep resources: Are you interested in a GMAT test study guide written by ACTUAL Graduate Management Admissions Test experts, who ACTUALLY scored HIGHER than the 99th percentile on the toughest graduate level exams? Our original research into the GMAT reveals specific weaknesses never before discovered that you can exploit to increase your GMAT test score more than you've ever imagined- and it's all available for less than the retail price of the rest of the filler-packed GMAT test prep study guides on the market. is available as an instantly downloadable e-book. It doesn't require any special software- if you can read this web page, and have access to a computer, you have all you need to start using and applying GMAT Secrets in just five minutes. You don't have to wait for anything to come in the mail. Download to your computer immediately! Who Else Wants to Use my Proven Flashcard System to Blow the Lid Off the GMAT Exam? **You can succeed on the GMAT by preparing in depth for the different question types and being able to take difficult questions and break them down into easier parts that you can quickly solve. Repetition and thorough preparation is a process that rewards those willing to work hard, which means that passing the GMAT is within the reach of virtually anyone willing to invest the time in learning how to handle any question they might face on test day.** Our are written in an easy to understand, straightforward style we don't include any more technical jargon than what you need to pass the test. The GMAT Flashcard Secrets system is only available at this web page. Don't decide now if these flashcards are for you. Just get them and try them out. GMAT is a registered trademark of the Graduate Management Admission



How Difficult Can This Be?

- Consider two 5,000-word documents that contain a common passage (i.e., no differences), but we don't know anything about it, not even its length.
- An exhaustive search must test:
 - Every valid length from 1 to 5,000
 - Every shingle of each length in each document
 - Average number of shingles is 2,500
- Result is approx. $(5000)(2500)(2500) =$ over 30 billion passage comparisons.
- This is $O(n^3)$ complexity. If differences are allowed, search is $O(n^4)$.



Our Approach

- Take advantage of the fact that, despite differences, similar passages tend to have aligned concepts.
- We borrow the **Smith-Waterman** dynamic programming algorithm from the bioinformatics community.
- We extend it for large document text similarity applications by specifying:
 - **Recursive descent** – to support discovery of multiple passage pairs
 - **Matrix splicing** – for handling large documents
 - **Chaining** – for connecting passage components
 - **Relaxed similarity measure** – for identifying token matches



A simple (but actual) example

This essay discusses Hamlet's famous soliloquy in relation to the major themes of the play.

(ROOT
(S
(NP (DT This) (NN essay))
(VP (VBZ discusses)
(NP
(NP
(NP (NNP Hamlet) (POS 's))
(JJ famous) (NN soliloquy))
(PP (IN in)
(NP (NN relation))))))
(PP (TO to)
(NP
(NP (DT the) (JJ major) (NNS themes))
(PP (IN of)
(NP (DT the) (NN play))))))
(. .)))

This article discusses the famous Hamlet monologue of the main themes of the game.

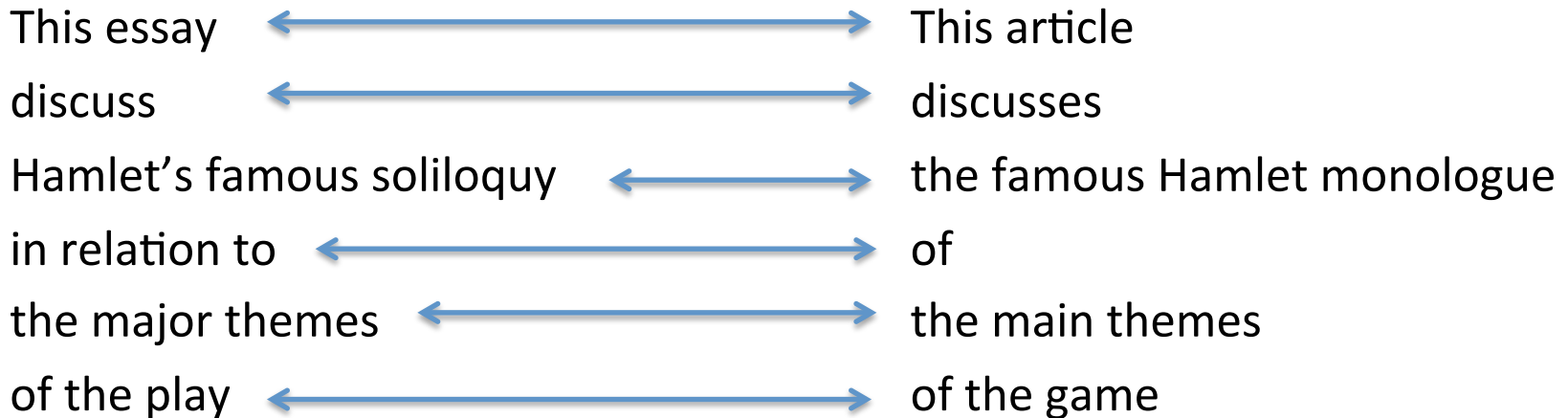
(ROOT
(S
(NP (DT This) (NN article))
(VP (VBZ discusses)
(NP
(NP (DT the) (JJ famous) (NNP Hamlet) (NN monologue))
(PP (IN of)
(NP
(NP (DT the) (JJ main) (NNS themes))
(PP (IN of)
(NP (DT the) (NN game))))))
(. .)))



Concept Alignment

This essay discusses Hamlet 's famous soliloquy in relation to the major themes of the play.

This article discusses the famous Hamlet monologue of the main themes of the game.



The Smith-Waterman Algorithm

- Uses dynamic programming to build a match matrix for the two input documents
- Finds the maximal length alignment
- The algorithm:

$$M(i, j) = \max \begin{cases} M(i-1, j-1) + \text{match}(a_i, b_j) \\ M(i-1, j) + \text{gap} \\ M(i, j-1) + \text{gap} \\ 0 \end{cases}$$

where $\text{match}(a_i, b_j) = +2$, if $a_i = b_j$; and -1 otherwise; and where $\text{gap} = -1$ is the gap penalty.



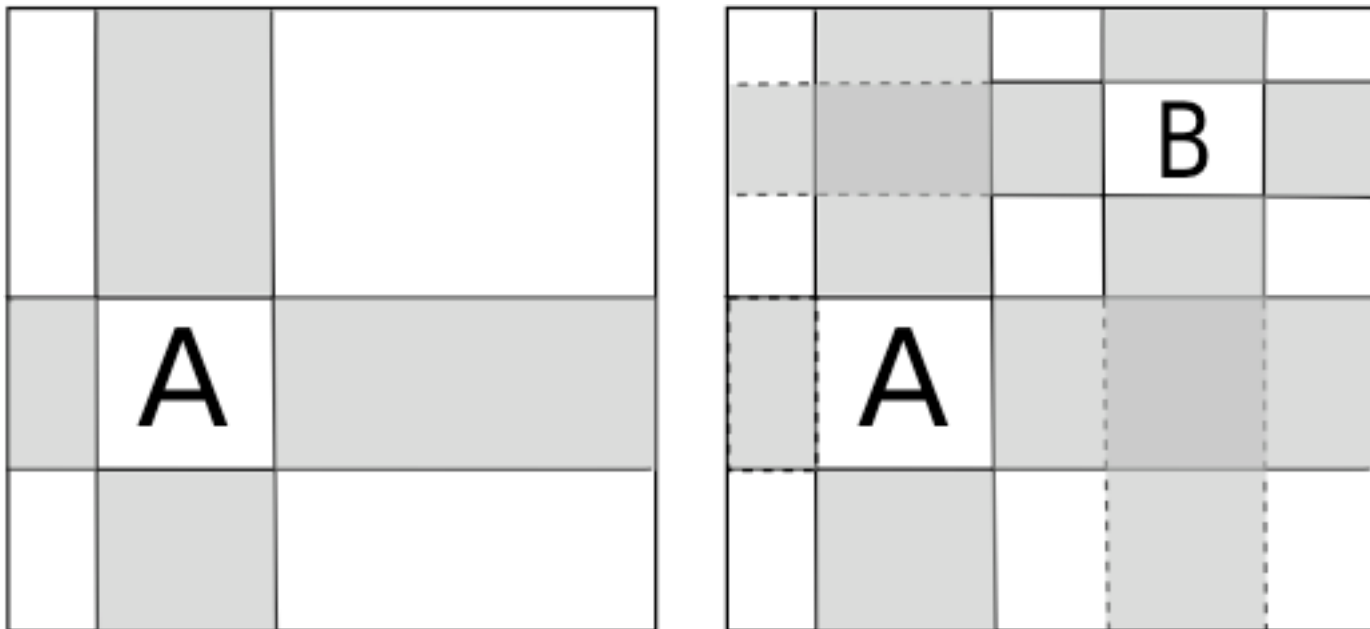
The Match Matrix

	This	essay	discusses	Hamlet	's	famous	soliloquy	in	relation	to	the	major	themes	of	the	play	.	tempus	fugit
This	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
article	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discusses	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
the	0	0	0	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
famous	0	0	0	2	1	0	0	0	0	0	2	1	0	0	2	1	0	0	0
Hamlet	0	0	0	1	1	1	3	2	1	0	1	1	0	0	1	1	0	0	0
monologue	0	0	0	0	3	2	2	2	1	0	0	0	0	0	0	0	0	0	0
of	0	0	0	0	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0
the	0	0	0	0	1	1	1	0	3	2	2	1	0	0	2	1	0	0	0
main	0	0	0	0	0	0	0	0	2	2	1	4	3	2	1	4	3	2	1
themes	0	0	0	0	0	0	0	0	1	1	1	3	3	2	1	3	3	2	2
of	0	0	0	0	0	0	0	0	0	0	2	2	5	4	3	2	2	1	0
the	0	0	0	0	0	0	0	0	2	1	2	1	4	7	6	5	4	3	2
game	0	0	0	0	0	0	0	0	1	1	1	4	3	6	9	8	7	6	5
.	0	0	0	0	0	0	0	0	0	0	3	3	2	5	8	8	7	6	5
carpe	0	0	0	0	0	0	0	0	0	0	2	2	2	4	7	7	10	9	8
diem	0	0	0	0	0	0	0	0	0	0	1	1	1	3	6	6	9	9	8
	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	5	8	8	8

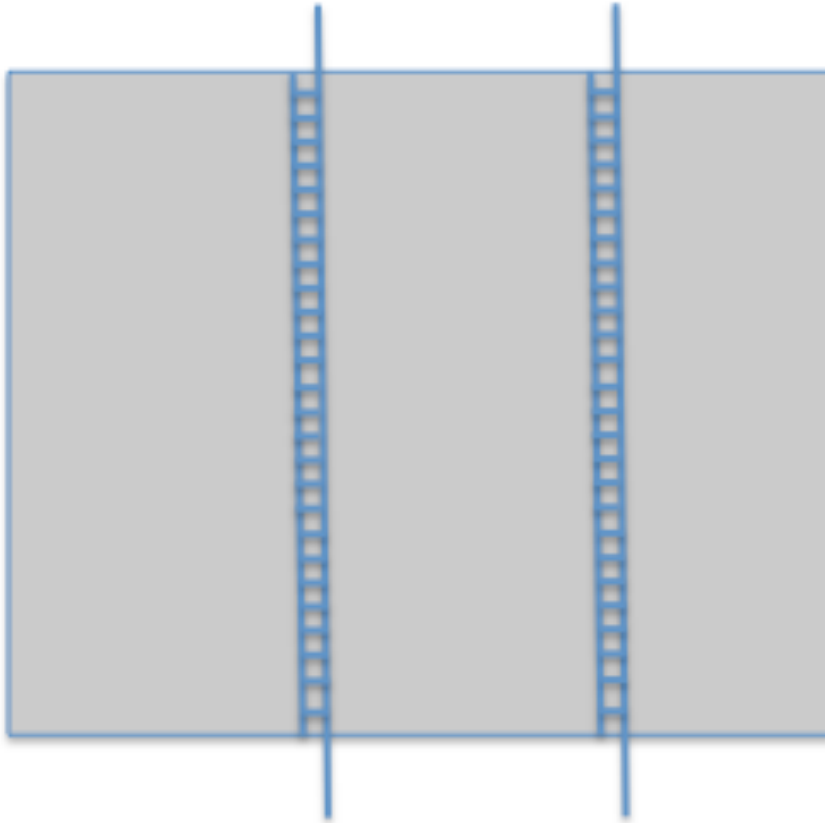


Recursive Descent

- Apply algorithm recursively to unused regions of document space



Matrix Splicing



- Slice to fit segment within available memory
- Column to left of slice preserves state, allowing chains to cross boundaries



Chaining

- Bridge gaps along diagonals if continue on both sides
- Limit 2 gaps bridged per chain

	This	essay	discusses	Hamlet	's	famous	soliloquy	in	relation	to	the	major	themes	of	the	play	.	tempus	fugit
This	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
article	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
discusses	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
the	0	0	0	2	2	1	0	0	0	0	2	1	0	0	2	1	0	0	0
famous	0	0	0	1	1	1	3	2	1	0	1	1	0	0	1	1	0	0	0
Hamlet	0	0	0	0	3	2	2	2	1	0	0	0	0	0	0	0	0	0	0
monologue	0	0	0	0	2	2	1	1	1	0	0	0	0	0	0	0	0	0	0
of	0	0	0	0	1	1	1	0	3	2	2	1	0	0	2	1	0	0	0
the	0	0	0	0	0	0	0	0	2	2	1	4	3	2	1	4	3	2	1
main	0	0	0	0	0	0	0	0	1	1	1	3	3	2	1	3	3	2	2
themes	0	0	0	0	0	0	0	0	0	0	2	2	5	4	3	2	2	1	0
of	0	0	0	0	0	0	0	0	2	1	2	1	1	4	7	6	5	4	3
the	0	0	0	0	0	0	0	0	1	1	1	4	3	3	6	9	8	7	6
game	0	0	0	0	0	0	0	0	0	0	3	3	2	5	8	8	7	6	5
.	0	0	0	0	0	0	0	0	0	0	2	2	2	4	7	7	10	9	8
carpe	0	0	0	0	0	0	0	0	0	0	1	1	1	3	6	6	9	9	8
diem	0	0	0	0	0	0	0	0	0	0	0	0	0	2	5	5	8	8	8



Relaxed Similarity Measure

- Different authors and speakers often use different articles and prepositions when expressing the same concept.
- When testing for matches while building up the match matrix:
 - Equate determiners: *a, an, the*
 - Also equate common prepositions:
of, in, to, for, with, on, at, from, by, about, as, into, like, through, after, over, between, out, against, during, without, before, under, around, among



Test Data

- Although not a perfect match for this algorithm, we chose the 2013 PAN text alignment test corpus, comprising
 - 5,185 document pairs from 3,169 source and 1,826 suspect documents
 - 1,000 pairs each involving *no plagiarism, no obfuscation, random obfuscation*, and *cyclic translation* plagiarism
 - 1,185 pairs involving *summary* plagiarism
 - Source documents:
 - min/mean/max: 104 / 914 / 12,277 words
 - Suspect documents:
 - min/mean/max: 131 / 2,930 / 20,297 words



Aggregate Performance

Target Corpus	PlagDet	Recall	Precision	Granularity	Runtime
No plagiarism	undefined	undefined	0.0000	undefined	1:41.2013
No obfuscation	0.9624	0.9603	0.9644	1.0000	6:47.717
Random obfuscation	0.7958	0.7073	0.9732	1.0413	5:09.562
Cyclic obfuscation	0.8441	0.7506	0.9730	1.0056	6:55.170
Summary obfuscation	0.0984	0.0560	0.9794	1.1099	2:52.770
Overall	0.8404	0.7588	0.9690	1.0177	22:23.481

- Precision uniformly high
- Recall for summary near nil
 - Understandable, since summaries inherently do not preserve order of concepts



Detection Counts

Target Corpus	Document Pairs	Reports	Detections	Cases	Cases Detected
No plagiarism	1,000	5	0	0	0
No obfuscation	1,000	1,160	1,159	1,206	1,159
Random	1,000	1,114	1,109	1,292	1,065
Cyclic	1,000	1,093	1,084	1,308	1,078
Summary	1,185	103	101	236	91
Overall	5,185	3,475	3,453	4,042	3,393

- Low false alarm rate overall
- Manual examination of a number of summary cases detected indicate that the summaries that were detected were largely cut-and-paste excerpts (for which concepts are aligned)



Conclusions and Improvements

- Conclusions
 1. Works well for detecting similar texts whose concepts are more-or-less aligned.
 2. Not well when concepts are not aligned.
 3. Can be a valuable component of a larger system for plagiarism detection (e.g., our entry in PAN 2014, which performed well)

- Improvements
 1. Explicitly include synonymy in similarity determinations
 2. Fine tune treatment of beginnings and endings of chains

