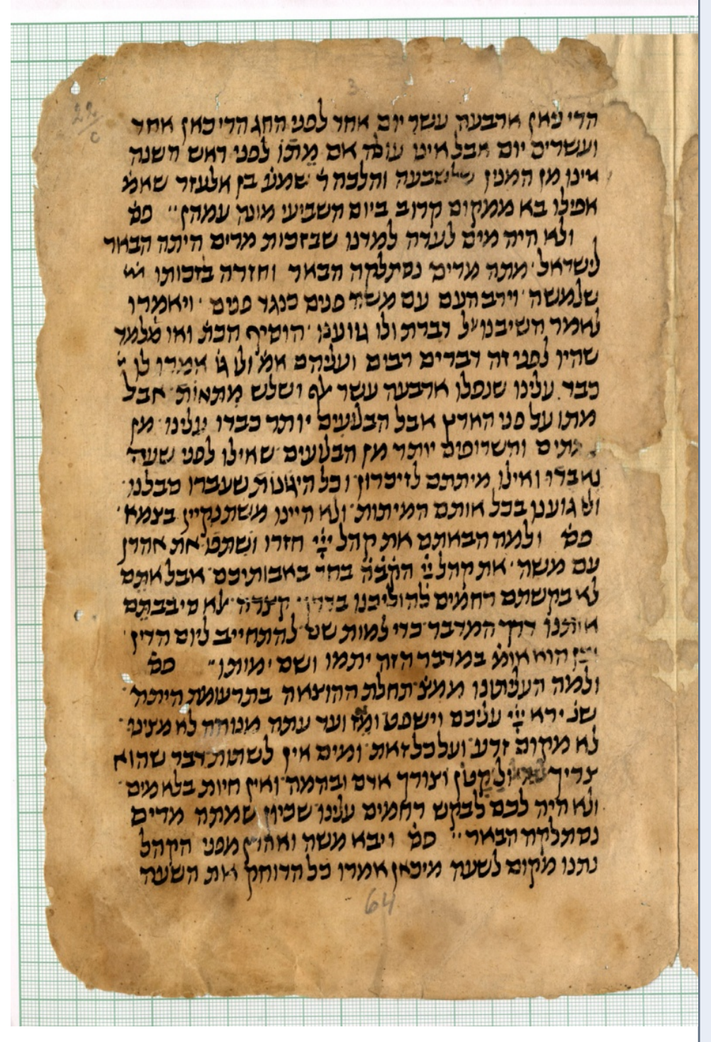
Matching Poorly Recognized Texts to a Corpus

Alex Zhicharevich and Nachum Dershowitz

School of Computer Science, Tel Aviv University

1. Introduction



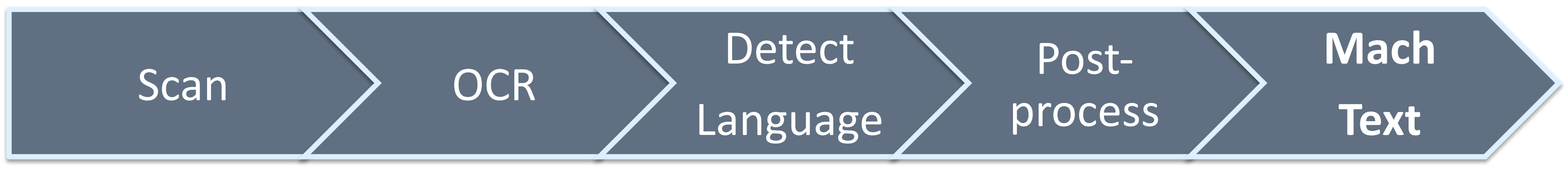
- OCR is an important aspect of digitization of historical texts
- OCR performance on handwritten texts is poor
- Often these are variants of known texts

- So, use approximate search to match bad OCR text with corpus

2. Digitization Process

Given the poor results of OCR on a fragment:

➢ Identify the language using n-gram techniques
➢ Postprocess the OCR result based on a language model
➢ Match the result to a given corpus using an edit-distance measure



Scan

OCR

Detect Language

Post-process

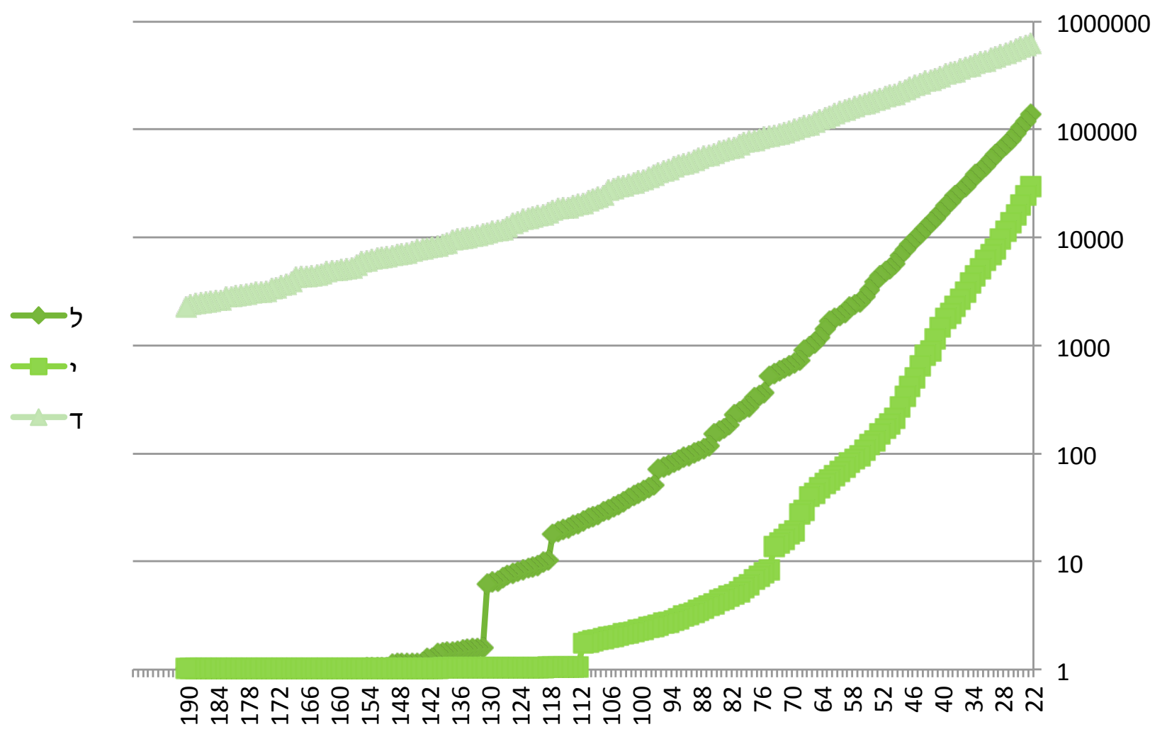
Mach Text

3. OCR Accuracy

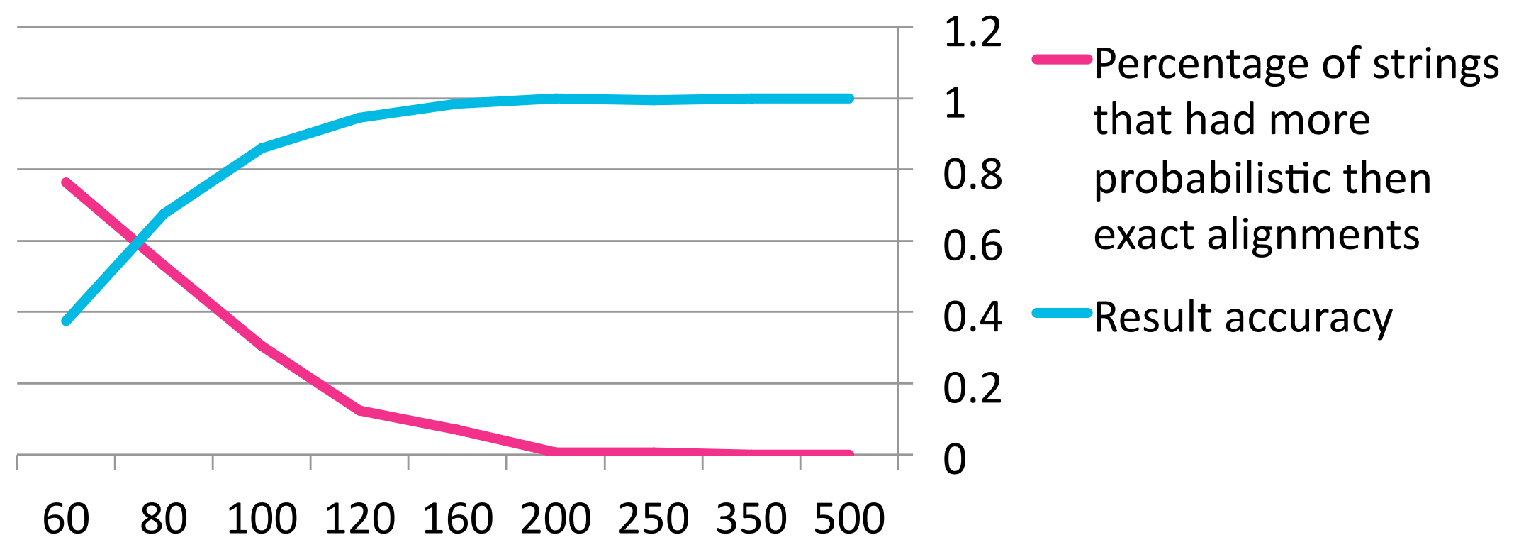
We estimated the desired OCR accuracy using artificially degraded texts

- By assuming OCR perfect performance on a single letter, the needed fragment length was estimated

Fragment size	ל	'	ד
50	5180.67	185.77	208916.5
80	243.278	5.256	72057.76
100	42.796	2.312	33386.34
120	9.268	1.052	15892.2
160	1.032	1.022	4991.454



- Assuming single-letter near-perfect performance, the necessary length increases slightly



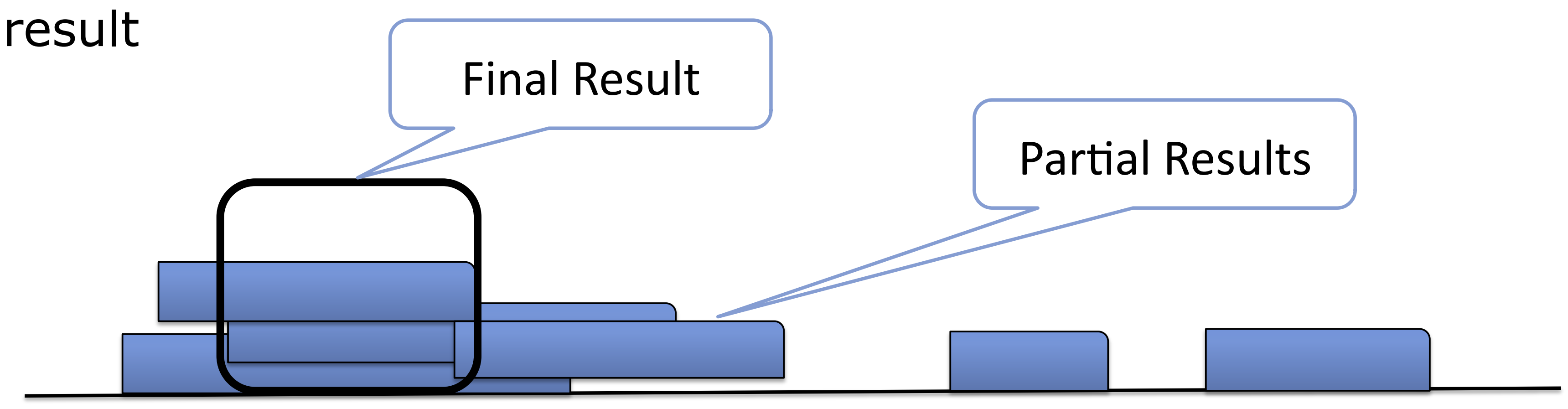
1. In practice, most letters have same recognition rate
2. Insertion and deletion errors dominate

4. Matching Algorithm

- Finding the text fragment with minimal edit distance to a given fragment is a time and space consuming task
- O(mn) time and space complexity, using dynamic programming
- Both corpus and fragment are long enough to use more scalable algorithms

Our Algorithm

▪ Split fragment into short searchable strings
▪ Search for best match using edit-distance dynamic programming
▪ Store results of every search in a list
▪ Use fragment intersection algorithm to find the most probable result



Final Result

Partial Results

5. Experimental Results

Algorithm tested on Genizah fragments with different OCR accuracy

On 60% accurate OCRed texts, found the correct text in the corpus

Below 40%, algorithm had no chance

Original	OCR	Match
שמייםליהוהוהארץנתןלבניאדםלאהמתיםיהללויהולאכלירדידומהואנחנונברךיהמעתהועדעולםהללויהאהבתיכיישמעיהוהאתקוליתחנוניכיהטהאזנוליובימיאקראאפפוניחבלימותומצרישאולמצאוניצרהויגוןאמצאובשםיהוהאקראאנהיהוהמלטהנפשיחנוןיהוהוצדיקואלהינומרחםשמרפתאיםיהוהדלתיולייהושיעשובינפשילמנוחיכייהוהגמלעליכיכיחלצתנפשיממותאתעינימןדמעהאתרגלימדחיאתהלךלפנייהוהבארצותהחייםהאמנתיכיאדבראני	ש$י$$ידוד$רץ$$ן$ב$יאדס$אד$יי$וידולאכליורדידו$ו$$נו$$ךיד$$ד$$לסדלוידאד$ב$יכי$$ידוד$$ו$י$חנו$י$יד$ראזנויו$י$יא$ר$או$י$בלי$$ו$ריש$ו$צאוני$דוגו$א$$א$$שסי$וא$ראא$אידודל$$שי$$וידודו$די$ואדינו$רחסשו$ר$$י$י$ו$ד$י$יי$יידושיעשובי$$י$$נוחיכיכיידוד$$לעיכי$יחל$$$שי$$ו$א$עינ$$ד$$דא$רג$יא$דלדך$יידו$$רד$י$פדא$$$יכי$דברא$יג$יא$דל$	והארץנתןלבניאדםלאהמתיםיהללויהולאכלירדידומהואנחנונברךיהמעתהועדעולםהללויהאהבתיכיישמעיהוהאתקוליתחנוניכיהטהאזנוליובימיאקראאפפוניחבלימותומצרישאולמצאוניצרהויגוןאמצאובשםיהוהאקראאנהיהוהמלטהנפשיחנוןיהוהוצדיקואלהינומרחםשמרפתאיםיהוהדלתיולייהושיעשובינפשילמנוחיכייהוהגמלעליכיכיחלצתנפשיממותאתעינימןדמעהאתרגלימדחיאתהלךלפנייהוהבארצותהחייםהאמנתיכיאדבראני
עניתימאדאניאמרתיבחפזיכלהאדםכזבמהאשיבליהוהכלתגמולוהיעליכוסישועותאשאובשםיהוהאקראנדריליהוהאשלםנגדהנאלכלעמויקרבעינייהוההמותהלחסידיואנהיהוהכיאניעבדךאניעבדךבןאמתךפתחתלמוסרילךאזבחזבחתודהובשםיהוהאקראנדריליהוהאשלםנגדהנאלכלעמובחצרותביתיהוהבתוככיירושלםהללויההללואתיהוהכלגויםשבחוהוכלהאמיםכיגברעלינוחסדוואמתיהוהלעולםהללויההודוליהוהכיטובכילעולם	$$י$ד$נ$א$$$יב$ז$כ$ד$דסכוז$כ$דז$דסכוז$$$ד$שדודכ$$$$ו$ודי$$יכוסישו$ו$אשאובש$יד$וד$א$$$$$רי$י$וד$ש$נ$דד$$לכל$$ויבעי$יידודד$ו$דל$ידיוא$ידודכיא$י$כדךאני$בדךבן$$$ך$$$לו$ריל$$זב$$ב$$ודדו$שסי$ו$א$$ר$$ד$ילידודאשל$$דדלכ$$וב$ו$בי$$דוד$$וככייוש$$ב$$ד$וד$$$ככוש$$ד$$וידדלל$$$י$ודכ$$וידזלל$$$לו$כ$$וש$$$דוכלד$ו$י$כיג$$ליו$$דווא$$יו$ולסד$לוי$ش וכלד$ו$י$כיג$$ליו$$דווא$	יתימאדאניאמרתיבחפזיכלהאדםכזבמהאשיבליהוהכלתגמולוהיעליכוסישועותאשאובשםיהוהאקראנדריליהוהאשלםנגדהנאלכלעמויקרבעינייהוהמותהלחסידיואנהיהוהכיאניעבדךאניעבדךבןאמתךפתחתלמוסרילךאזבחזבחתודהובשםיהוהאקראנדריליהוהאשלמנגדהנאלכלעמובחצרותביתיהוהבתוככיירושלםהללויההללואתיהוהכלגויםשבחוהוכלהאמיםכיגברעלינוחסדוואמתיהוהלעולםהללויההודוליהוהכיטובכילעולםחסדויא
כבשיביאקרבנולחטאתנקבהתמימהיביאנהוסמךאתידועלראשהחטאתושחטאתהלחטאתבמקוםאשרישחטאתהעלהולקחהכהןמדםהחטאתבאצבעוונתןעלקרנתמזבחהעלהואתכלדמהישפךאליסודהמזבחואתכלחלבהיסירכאשריוסר	$$יי$ר$נר$$י$$ביון$$$ע$רשן$$ו$$$$$$$$$ו$$$$$$$$$י$$ן$$ן$$$$$$$ו$$$$$$$$$$$ו$$$$$$$$$$$$$$$$$ז$$ו$$$$$$$שיו$$$	לואחותתובלקיןנעמהויאמרלמךלנשיועדהוצלהשמעןקולינשילמךהאזנהאמרתיכיאישהר

1. String matching algorithms have an limit on accuracy due to edit-distance metrics
    1. Estimate the needed OCR accuracy for edit-distance matching
    2. Preprocess the fragments using statistical language models
2. Refine OCR using language models

www.PosterPresentations.com