Translation by Pattern Matching

Adam Lopez
University of Edinburgh
Statistical Machine Translation

parallel text +
alignment
Statistical Machine Translation

parallel text + alignment

extract rules
Statistical Machine Translation

parallel text + alignment
Statistical Machine Translation

parallel text + alignment

extract rules

score rules

load rules into memory

decoder
Problem Overview

Solution Framework

Algorithms

Application

Statistical Machine Translation

parallel text + alignment

extract rules

score rules

load rules into memory

number of rules depends on corpus size...

corpus size...

decoder
Statistical Machine Translation

parallel text + alignment

extract rules

score rules

load rules into memory

... and model complexity

decoder
Statistical Machine Translation

parallel text + alignment

extract rules → score rules → filter rules for test set

load filtered rules into memory

decoding algorithm
Baseline Translation Model

- Hierarchical Phrase-based translation (Chiang 2007)
- 1M parallel sentences (27M words)
- GIZA++ alignments (Och & Ney 2003, Koehn et al. 2003)
  - alignments are dense
- Heuristics used to restrict number of extracted rules
- 67M rules, 6.1Gb of data
  - cf. 225M (Zens & Ney 2007), 55M (DeNeefe et al. 2007)
Some Possible Improvements

- 3.5M sentences (2.5M out-of-domain), 100M words
- Discriminatively trained alignments (Ayan & Dorr 2006)
  - Key difference: alignments are sparse
- *Loose* phrase extraction (Ayan & Dorr 2006)
Some Possible Improvements

• 3.5M sentences (2.5M out-of-domain), 100M words
• Discriminatively trained alignments (Ayan & Dorr 2006)
  • Key difference: alignments are *sparse*
• *Loose* phrase extraction (Ayan & Dorr 2006)
Some Possible Improvements

- 3.5M sentences (2.5M out-of-domain), 100M words
- Discriminatively trained alignments (Ayan & Dorr 2006)
  - Key difference: alignments are sparse
- *Loose* phrase extraction (Ayan & Dorr 2006)
Some Possible Improvements

- 3.5M sentences (2.5M out-of-domain), 100M words
- Discriminatively trained alignments (Ayan & Dorr 2006)
  - Key difference: alignments are *sparse*
- *Loose* phrase extraction (Ayan & Dorr 2006)
Some Possible Improvements

- 3.5M sentences (2.5M out-of-domain), 100M words
- Discriminatively trained alignments (Ayan & Dorr 2006)
  - Key difference: alignments are sparse
- Loose phrase extraction (Ayan & Dorr 2006)
Some Possible Improvements

- Rule extraction time: 77 CPU days
  - does not include sorting or scoring!
- Rules counted: 20 billion
  - 2 orders of magnitude larger than state of the art
- Estimated unique rules: 6.6 billion
- Estimated extract file size: 917Gb
- Estimated phrase table size: 600Gb
The Problem

- Current models are bounded by resource limitations.
- We’re already pushing the edge of what’s possible.
- Parallel data aren’t getting any smaller.
- Models aren’t getting any less complex.
The Solution

- Translation by pattern matching.
- Novel pattern matching algorithms.
  - Exploit ideas developed in bioinformatics, IR
- Support for tera-scale translation models.
Idea: Translation by Pattern Matching (Callison-Burch et al. 05, Zhang & Vogel 05)

decoding algorithm

parallel text + alignment in memory

extract and score

sentence-specific rules

pattern matching algorithm
Exact Pattern Matching

Input Pattern: it persuades him and it disheartens him
Exact Pattern Matching

Input Pattern  it persuades him and it disheartens him
=Query Pattern
Pattern Matching for Phrase-Based MT

Input Pattern  it persuades him and it disheartens him
Pattern Matching for Phrase-Based MT

Input Pattern: it persuades him and it disheartens him

Query Patterns:
- it persuades him and it disheartens him
- it persuades him
- it persuades him and it
- it disheartens
- it persuades him
- it persuades him and it
- it disheartens
- it persuades him
- it persuades him and it
- it disheartens
- it persuades him
- it persuades him and it
- it disheartens
- it persuades him
- it persuades him and it
- it disheartens
- it persuades him
Suffix Arrays

it makes him and it mars him , it sets him on and it takes him off .

Text $T$
Suffix Arrays

it makes him and it mars him, it sets him on and it takes him off. #

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text $T$

4  it mars him, it sets him on and it takes him off. #

Suffix 4
Suffix Arrays

it makes him and it mars him, it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

0 it makes him and it mars him, it sets him on and it takes him ...
1 makes him and it mars him, it sets him on and it takes him off. #
2 him and it mars him, it sets him on and it takes him off. #
3 and it mars him, it sets him on and it takes him off. #
4 it mars him, it sets him on and it takes him off. #
5 mars him, it sets him on and it takes him off. #
6 him, it sets him on and it takes him off. #
7 , it sets him on and it takes him off. #
...

Suffix Arrays

it makes him and it mars him, it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

3  and it mars him, it sets him on and it takes him off. #
12  and it takes him off. #
2  him and it mars him, it sets him on and it takes him off. #
15  him off. #
10  him on and it takes him off. #
6  him, it sets him on and it takes him off. #
0 it makes him and it mars him, it sets him on and it takes him ...
4 it mars him, it sets him on and it takes him off. #
...

Suffix Arrays

it makes him and it mars him, it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

3  and it mars him, it sets him on and it takes him off. #
12 and it takes him off. #
2  him and it mars him, it sets him on and it takes him off. #
15 him off. #
10 him on and it takes him off. #
6  him, it sets him on and it takes him off. #
0  it makes him and it mars him, it sets him on and it takes him ...
4  it mars him, it sets him on and it takes him off. #
...
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text T

| 3 | 12 | 2 | 15 | 10 | 6 | 0 | 4 | 8 | 13 | 1 | 5 | 16 | 11 | 9 | 14 | 7 | 17 | 18 |

Suffix Array SA
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text $T$

| 3 | 12 | 2 | 15 | 10 | 6 | 0 | 4 | 8 | 13 | 1 | 5 | 16 | 11 | 9 | 14 | 7 | 17 | 18 |

Suffix Array $SA$

him and it

Query Pattern $w$
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text T

3 12 2 15 10 6 0 4 8 13 1 5 16 11 9 14 7 17 18

Suffix Array SA

him and it

Query Pattern w
Suffix Arrays

it makes him and it mars him. it sets **him on and it takes him off**. #

Text $T$

$$
|   | 3 | 12 | 2 | 15 | 10 | 6 | 0 | 4 | 8 | 13 | 1 | 5 | 16 | 11 | 9 | 14 | 7 | 17 | 18 |
$$

Suffix Array $SA$

him and it

Query Pattern $w$
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text $T$

| 3 | 12 | 2 | 15 | 10 | 6 | 0 | 4 | 8 | 13 | 1 | 5 | 16 | 11 | 9 | 14 | 7 | 17 | 18 |

Suffix Array $SA$

him and it

Query Pattern $w$
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off.

Text $T$

3 12 2 15 10 6 0 4 8 13 1 5 16 11 9 14 7 17 18

Suffix Array $SA$ \( O(|w| \log |T|) \)

him and it

Query Pattern $w$
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #

Text $T$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
</tr>
</tbody>
</table>

Suffix Array $SA$  

$O(|w| \log |T|)$

$O(|w| + \log |T|)$ (Manber & Myers, 93)

Query Pattern $w$

him and it
Suffix Arrays

it makes **him and it mars him** . it sets him on and it takes him off . #

Text $T$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>2</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>13</td>
<td>1</td>
<td>5</td>
<td>16</td>
<td>11</td>
<td>9</td>
<td>14</td>
<td>7</td>
<td>17</td>
<td>18</td>
</tr>
</tbody>
</table>

Suffix Array $SA$ 

$O(|w| \log |T|)$

$O(|w| + \log |T|)$ (Manber & Myers, 93)

$O(|w|)$ (Abouelhoda et al., 04)

Query Pattern $w$
Suffix Arrays

it makes him and it mars him. it sets him on and it takes him off. #

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text $T$

3 12 2 15 10 6 0 4 8 13 1 5 16 11 9 14 7 17 18

Suffix Array $SA$

$O(|w| \log |T|)$

$O(|w| + \log |T|)$ (Manber & Myers, 93)

$O(|w|)$ (Abouelhoda et al., 04)

Query Pattern $w$

on baseline model:

0.009 seconds / sentence
(not including extraction / scoring)
Problem: Phrases with Gaps

- Hierarchical phrase-based translation (Chiang 2005, 2007)

Input: it persuades him and it disheartens him

Source Phrase: it X him
Hierarchical Phrases: Phrases with Gaps

- Hierarchical phrase-based translation (Chiang 2005, 2007)

Input: it persuades him and it disheartens him

Source Phrase: it X him
Hierarchical Phrases: Phrases with Gaps

- Hierarchical phrase-based translation (Chiang 2005, 2007)

Input: it persuades him and it disheartens him

Source Phrase: it X him
Hierarchical Phrases: Phrases with Gaps

- Hierarchical phrase-based translation (Chiang 2005, 2007)

Input: it persuades him and it disheartens him

Source Phrase: it X him
Hierarchical Phrases: Phrases with Gaps

- Hierarchical phrase-based translation (Chiang 2005, 2007)

Input: it persuades him and it disheartens him

Source Phrase: it X and X him
Problem Statement

Given an input sentence, efficiently find all hierarchical phrase-based translation rules for that sentence in the training corpus.
Pattern Matching for Hierarchical PBMT

Input Pattern: it persuades him and it disheartens him
Pattern Matching for Hierarchical PBMT

Input Pattern  it persuades him and it disheartens him

Query Patterns  it
     persuades
     him
     and
     disheartens
     it persuades
     him and
     and it
     it disheartens
     disheartens him

     it persuades him
     persuades him and
     him and it
     and it disheartens
     it disheartens him
     it persuades him and
     persuades him and it
     him and it disheartens
     and it disheartens him
     it persuades him and it
     persuades him and it disheartens
     him and it disheartens him
Pattern Matching for Hierarchical PBMT

Input Pattern
it persuades him and it disheartens him

Query Patterns
it X and
it X it
it X disheartens
it X him
persuades X it
persuades X disheartens
persuades X him
it persuades X it
it persuades X disheartens
it persuades X him
it X and it
it X it disheartens

it X disheartens him
it X and X him
persuades him X disheartens
persuades him X him
persuades X it disheartens
persuades X disheartens him
him and X him
him X disheartens him
it persuades him X disheartens
it persuades him X him
it persuades X it disheartens
it persuades X disheartens him
Pattern Matching for Hierarchical PBMT

Input Pattern

it persuades him and it disheartens him

Query Patterns

it X and it disheartens
it X it disheartens him
persuades him and X him
persuades him X disheartens him
persuades X it disheartens him
it persuades him and X him
it persuades him X disheartens him
it persuades X it disheartens him
it persuades him X disheartens him
it persuades X it disheartens him
it X and it disheartens him
Pattern Matching for Hierarchical PBMT

Input Pattern
it persuades him and it disheartens him

Query Patterns
it X and it disheartens
it X it disheartens him
persuades him and X him
persuades him X disheartens him
persuades X it disheartens him
it persuades him and X him
it persuades him X disheartens him
it persuades X it disheartens him
it X and it disheartens him

This is a variant of approximate pattern matching (Navarro ‘01)
Pattern Matching with Gaps

Query pattern $\alpha$

him X it

<table>
<thead>
<tr>
<th>3</th>
<th>and it mars him, it sets him ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>and it takes him off. #</td>
</tr>
<tr>
<td>2</td>
<td>him and it mars him. it sets ...</td>
</tr>
<tr>
<td>15</td>
<td>him off. #</td>
</tr>
<tr>
<td>10</td>
<td>him on and it takes him off. #</td>
</tr>
<tr>
<td>6</td>
<td>him, it sets him on and it ...</td>
</tr>
<tr>
<td>0</td>
<td>it makes him and it mars ...</td>
</tr>
<tr>
<td>4</td>
<td>it mars him, it sets him on ...</td>
</tr>
<tr>
<td>8</td>
<td>it sets him on and it takes ...</td>
</tr>
<tr>
<td>13</td>
<td>it takes him off. #</td>
</tr>
<tr>
<td>1</td>
<td>makes him and it mars him ...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Pattern Matching with Gaps

Query pattern $\alpha$

him $X$ it

| 3 | and it mars him , it sets him ... |
| 12 | and it takes him off . # |
| 2 | him and it mars him . it sets ... |
| 15 | him off . # |
| 10 | him on and it takes him off . # |
| 6 | him , it sets him on and it ... |
| 0 | it makes him and it mars ... |
| 4 | it mars him , it sets him on ... |
| 8 | it sets him on and it takes ... |
| 13 | it takes him off . # |
| 1 | makes him and it mars him ... |
Pattern Matching with Gaps

Query pattern $\alpha$

him X it

and it mars him, it sets him ...
and it takes him off. #
him and it mars him. it sets ...
him off. #
him on and it takes him off. #
him, it sets him on and it ...
it makes him and it mars ...
it mars him, it sets him on ...
it sets him on and it takes ...
it takes him off. #
makes him and it mars him ...
...
Pattern Matching with Gaps

Query pattern $\alpha$

him X it

Subpatterns $w_i$

him

it

<table>
<thead>
<tr>
<th>3</th>
<th>and it mars him , it sets him ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>and it takes him off . #</td>
</tr>
<tr>
<td>2</td>
<td>him and it mars him . it sets ...</td>
</tr>
<tr>
<td>15</td>
<td>him off . #</td>
</tr>
<tr>
<td>10</td>
<td>him on and it takes him off . #</td>
</tr>
<tr>
<td>6</td>
<td>him , it sets him on and it ...</td>
</tr>
<tr>
<td>0</td>
<td>it makes him and it mars ...</td>
</tr>
<tr>
<td>4</td>
<td>it mars him , it sets him on ...</td>
</tr>
<tr>
<td>8</td>
<td>it sets him on and it takes ...</td>
</tr>
<tr>
<td>13</td>
<td>it takes him off . #</td>
</tr>
<tr>
<td>1</td>
<td>makes him and it mars him ...</td>
</tr>
</tbody>
</table>
# Pattern Matching with Gaps

**Query pattern** $\alpha$

- him
- X
- it

**Subpatterns** $w_i$

- him
- it

<table>
<thead>
<tr>
<th>$i$</th>
<th>$w_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

and it mars him, it sets him ...
and it takes him off. #
him and it mars him, it sets ...
him off. #
him on and it takes him off. #
him, it sets him on and it ...
it makes him and it mars ...
it mars him, it sets him on ...
it sets him on and it takes ...
it takes him off. #
makes him and it mars him ...
Pattern Matching with Gaps

**Query pattern** $\alpha$

<table>
<thead>
<tr>
<th>$\text{him}$</th>
<th>$\text{it}$</th>
<th>$n_i$ Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12</td>
<td>and it mars him , it sets him ...</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>and it takes him off . #</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>him and it mars him . it sets ...</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>him off . #</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>him on and it takes him off . #</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>him , it sets him on and it ...</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>it makes him and it mars ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it mars him , it sets him on ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it sets him on and it takes ...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>it takes him off . #</td>
</tr>
<tr>
<td></td>
<td></td>
<td>makes him and it mars him ...</td>
</tr>
</tbody>
</table>
Problem Overview
Solution Framework
Algorithms
Application

Pattern Matching with Gaps

and it mars him, it sets him ...
and it takes him off. #
him and it mars him. it sets ...
him off. #
him on and it takes him off. #
him, it sets him on and it ...
it makes him and it mars ...
it mars him, it sets him on ...
it sets him on and it takes ...
it takes him off. #
makes him and it mars him ...

...
Pattern Matching with Gaps

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>15</th>
<th>10</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>13</td>
</tr>
</tbody>
</table>
Pattern Matching with Gaps

\[(2, 4)\]
\[(2, 8)\]
\[(2, 13)\]
\[(6, 8)\]
\[(6, 13)\]
\[(10, 13)\]
Pattern Matching with Gaps

RILMS (Rahman et al., 06)
Pattern Matching with Gaps

RILMS (Rahman et al., 06)

linear in number of occurrences of subpatterns: \( O(\sum n_i) \)
Baseline Timing Result

221 seconds per sentence

compare: 0.009 seconds per sentence for contiguous phrases
Complexity Analysis

contiguous

\[
\sum_{w} (|w| + \log |T|)
\]

137 5 27

discontiguous

\[
\sum_{\alpha=w_1 X \ldots X w_I} \sum_{i=1}^{I} (|w_i| + \log |T| + n_i)
\]

2825 3 5 27 82069
Contiguous:
\[
\sum_{w} (|w| + \log |T|)
\]
137 5 27

Discontiguous:
\[
\sum_{\alpha=w_1 X \ldots X w_I} \sum_{i=1}^{I} (|w_i| + \log |T| + n_i)
\]
2825 3 5 27 82069
Exploiting Redundancy

Input Pattern:  it persuades him and it disheartens him

Query Patterns:
- it X and
- it X it
- it X disheartens
- it X him
- persuades X it
- persuades X disheartens
- persuades X him
- it persuades X it
- it persuades X disheartens
- it persuades X him
- it X and it
- it X it disheartens
- it X disheartens him
- it X and X him
- persuades him X disheartens
- persuades him X him
- persuades X it disheartens
- persuades X disheartens him
- him and X him
- him X disheartens him
- it persuades him X disheartens
- it persuades him X him
- it persuades X it disheartens
- it persuades X disheartens him
Exploiting Redundancy

Input Pattern: it persuades him and it disheartens him

Query Patterns:
- it X and it X it
- it X disheartens it X him
- persuades X it persuades X disheartens
- persuades X him it persuades X it
- it persuades X disheartens it X it disheartens
- it X it disheartens it X and it
- it X it disheartens
Exploiting Redundancy

Query Pattern

it persuades X disheartens him
Exploiting Redundancy

Query Pattern: it persuades X disheartens him
Maximal Prefix: it persuades X disheartens

(Zhang & Vogel 2005)
Exploiting Redundancy

Query Pattern: it persuades X disheartens him
Maximal Prefix: it persuades X disheartens
Maximal Suffix: persuades X disheartens him
Prefix Tree with Suffix Links

persuades

it

X

him

X

him

him

him

him
Timing Results

221 seconds/sentence

Baseline
Timing Results

seconds/sentence

Baseline: 221
Prefix Tree: 177
Complexity Analysis

contiguous $\sum_{w} (|w| + \log |T|)$

137 5 27

discontiguous $\sum_{\alpha=w_{1}X\ldots X w_{I}} \sum_{i=1}^{I} (|w_{i}| + \log |T| + n_{i})$

2825 3 5 27 82069
Complexity Analysis

contiguous \[ \sum_{w} (|w| + \log |T|) \]
137 5 27

discontiguous \[ \sum_{\alpha=w_1X...Xw_I} \sum_{i=1}^{I} (|w_i| + \log |T| + n_i) \]
2825 3 5 27 82069
Empirical Analysis

computations (ranked by time)
Distribution of Patterns in Training Data

Pattern types (in descending order of frequency)
Distribution of Patterns in Training Data

Pattern types (in descending order of frequency)
Analysis of Problem

• The expensive computations involve at least one frequent subpattern. There are two cases.
  • A frequent pattern paired with an infrequent pattern
  • Two frequent patterns paired with each other
Frequent $\times$ Infrequent Subpatterns
Frequent × Infrequent Subpatterns
Frequent $\times$ Infrequent Subpatterns
Frequent × Infrequent Subpatterns
Double Binary Search
Baeza-Yates, 04
Double Binary Search

Baeza-Yates, 04

Queryset $Q$

Dataset $D$
Double Binary Search
Baeza-Yates, 04

Queryset \( Q \)  
Dataset \( D \)
Double Binary Search

Baeza-Yates, 04

Queryset $Q$  

Dataset $D$
Double Binary Search
Baeza-Yates, 04

Queryset $Q$  

Dataset $D$
Double Binary Search

Baeza-Yates, 04

Queryset $Q$

Dataset $D$
Double Binary Search

Baeza-Yates, 04
Double Binary Search
Baeza-Yates, 04

Queryset $Q$ \quad Dataset $D$

Upper bound complexity: $|Q| \log |D|$
Obtaining Sorted Sets
Obtaining Sorted Sets

Sort via Stratified Tree
(van Emde Boas et al. 1977)
Obtaining Sorted Sets

Sort via Stratified Tree
(van Emde Boas et al. 1977)

Problem: complexity increases to
\[ O(|Q| \log |D| + (|Q| + |D|) \log \log |T|) \]
Obtaining Sorted Sets

Sort via Stratified Tree
(van Emde Boas et al. 1977)

Problem: complexity increases to
\[ O(|Q| \log |D| + (|Q| + |D|) \log \log |T|) \]

Solution: cache sorted set in prefix tree
Timing Results

- Baseline: 221 seconds/sentence
- Prefix Tree + double binary: 177 seconds/sentence
Timing Results

seconds/sentence

Baseline: 221
Prefix Tree: 177
+ double binary: 174
Obtaining Sorted Sets

Sort via Stratified Tree

Problem: sort complexity is still very high for very frequent patterns
Obtaining Sorted Sets

Solution: precompute the *inverted index* for 1000 most frequent contiguous patterns
Timing Results

- Baseline: 221 seconds/sentence
- Prefix Tree + double binary: 174 seconds/sentence
- Prefix Tree: 177 seconds/sentence
Timing Results

- Baseline: 221 seconds/sentence
- Prefix Tree: 177 seconds/sentence
- Prefix Tree + double binary: 174 seconds/sentence
- Prefix Tree + inverted indices: 44 seconds/sentence
Frequent × Frequent Subpatterns
Frequent × Frequent Subpatterns

Problem:
There is no clever algorithm to solve this problem
Solution: Precomputation

it makes him and it mars him. it sets him on and it takes him off. #

Text
Solution: Precomputation

it makes him and it mars him. it sets him on and it takes him off. #

0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18

text

Most Frequent Patterns

it (4)
him (4)

Precomputed Pattern Matches

it X him
him X it

it X it
him X him
Solution: Precomputation

it makes him and it mars him . it sets him on and it takes him off . #

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Text

<table>
<thead>
<tr>
<th>Most Frequent Patterns</th>
<th>Precomputed Pattern Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>it (4)</td>
<td>it X him</td>
</tr>
<tr>
<td>him (4)</td>
<td>him X it</td>
</tr>
</tbody>
</table>

(0, 2)(0, 6)(13, 15) (2, 4)(2, 8)(10, 13)
(4, 6)(4, 10)(8, 10)(8, 15) (6, 8)(6, 13)

(0, 4)(0, 8) (2, 6)(2, 10)(10, 15)
(4, 8)(4, 13)(8, 13) (6, 10)(6, 15)
Timing Results

seconds / sentence

Baseline 221
Prefix Tree 177
+ double binary 174
+ inverted indices 44
Timing Results

seconds/sentence

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Prefix Tree</th>
<th>+ double binary</th>
<th>+ inverted indices</th>
<th>+ precomp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>221</td>
<td>177</td>
<td>174</td>
<td>44</td>
<td>1</td>
</tr>
</tbody>
</table>
Analysis of Fixed Memory Usage

- Source Text: $|T|$
- Suffix Array: $|T|$
- Alignments: $|T|$
- Target Text: $|T|$
- Total Cost: 4 $|T|$

- For 27M words: about 700M
- including indices for 1000 words: about 2.1 Gb
- for 100 words: 1.1Gb, increases time to 1.6 secs/sent
Longer Spans, Longer Phrases

Maximum Span Length

Maximum Phrase Length
The Tera-Scale Translation Model

- Task: NIST Chinese-English 2005
- Baseline Model: 30.7
- Tera-Scale Model: 32.6
- All modifications contribute to overall score
- With better language model and number translation:
  - Baseline Model: 31.9
  - Tera-Scale Model: 34.5
Open Questions

- Can we improve speed?
- Can we improve memory use? *Compressed self-indexes*?
- Uses for arbitrarily large translation models?
  - Context-sensitive models (Chan et al. 2007, Carpuat & Wu 2007)
  - Factored models (Koehn et al. 2007)
  - Syntax-based model (DeNeefe et al. 2007)
- What other algorithms can we use from bioinformatics?
Thanks

Acknowledgements:
David Chiang, Chris Dyer, Philip Resnik