Fast Inference in Phrase Extraction Models with Belief Propagation

David Burkett and Dan Klein
UC Berkeley
In the past two years

Constraining Word Alignment Models
In the past two years

Constraining Word Alignment Models
Constraining Word Alignment Models
Constraining Word Alignment Models

In the past two years

过去两年中

In the past two years

过去两年中
Constraining Word Alignment Models

Solution: Fertility constraints
Constraining Word Alignment Models

Solution: Fertility constraints
Constraining Word Alignment Models

Solution: Fertility constraints
Constraining Word Alignment Models

Solution: Fertility constraints

Model class: At most 1-1 matchings
Constraining Word Alignment Models

Solution: Fertility constraints
Model class: At most 1-1 matchings
Problem: Summing is #P-complete!
ITG to the Rescue

(Wu, 1997)
ITG to the Rescue

(Wu, 1997)
ITG to the Rescue

Popular for word alignment

Tree of nested bispans

(Wu, 1997)
ITG to the Rescue

Popular for word alignment

(Zhang & Gildea, 2005)

Tree of nested bispans

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)
- (Haghighi et al., 2009)

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)
- (Haghighi et al., 2009)
- (Burkett et al., 2010)

(Wu, 1997)

Tree of nested bispans
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)
- (Haghighi et al., 2009)
- (Burkett et al., 2010)

Good:

Tree of nested bispans

(Wu, 1997)
ITG to the Rescue

- Popular for word alignment
  - (Zhang & Gildea, 2005)
  - (Cherry & Lin, 2007)
  - (Haghighi et al., 2009)
  - (Burkett et al., 2010)

- Good:
  - Includes fertility constraints

(Wu, 1997)
ITG to the Rescue

- Popular for word alignment
  - (Zhang & Gildea, 2005)
  - (Cherry & Lin, 2007)
  - (Haghighi et al., 2009)
  - (Burkett et al., 2010)

- Good:
  - Includes fertility constraints
  - Polynomial exact inference $O(n^6)$

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)
- (Haghighi et al., 2009)
- (Burkett et al., 2010)

Good:
- Includes fertility constraints
- Polynomial exact inference $O(n^6)$

Bad:

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
  - (Zhang & Gildea, 2005)
  - (Cherry & Lin, 2007)
  - (Haghighi et al., 2009)
  - (Burkett et al., 2010)

Good:
  - Includes fertility constraints
  - Polynomial exact inference $O(n^6)$

Bad:
  - Aggressive nesting requirement

(Wu, 1997)
ITG to the Rescue

Popular for word alignment
- (Zhang & Gildea, 2005)
- (Cherry & Lin, 2007)
- (Haghighi et al., 2009)
- (Burkett et al., 2010)

Good:
- Includes fertility constraints
- Polynomial exact inference $O(n^6)$

Bad:
- Aggressive nesting requirement
- $O(n^6)$ still slow

(Wu, 1997)
Approximate Inference
Approximate Inference

Moral of the talk:
Approximate Inference

Moral of the talk:

Write the model you want, then work out inference
Approximate Inference

Moral of the talk:

Write the model you want, then work out inference

Content of the talk:
Approximate Inference

Moral of the talk:

Write the model you want, then work out inference

Content of the talk:

Better modeling through Belief Propagation
Phrase Extraction
In the past two years

Sentence Pair
### Phrase Extraction

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In</strong></td>
<td><strong>the</strong></td>
<td><strong>past</strong></td>
<td><strong>two years</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sentence Pair**

**Word Alignment**

- **过去**  (past)
- **两年**  (two)
- **中**  (year)
- **in**

- **Phrase Extraction**

- past%
- two%
- year%
- in
In the past two years

Sentence Pair

Word Alignment

Extracted Phrases
Modeling Phrase Extraction

(DeNero & Klein, 2010)
Modeling Phrase Extraction

Input: $s$

(DeNero & Klein, 2010)
Modeling Phrase Extraction

Input: $s$

Output: $(\pi, a)$

(DeNero & Klein, 2010)
Factor Graph Notation
Factor Graph Notation

Variables

\( V_1 \)  \( V_2 \)
\( V_3 \)  \( V_4 \)
Factor Graph Notation

Variables

Factors
Factor Graph Notation

Variables

Factors

\[ P(V) \propto \prod_{F} F(V_F) \]
In the past two years
Phrase Extraction Model: Word Alignments

In the past two years

Word Alignments
In the past two years

\[ a_{1,4} = \text{on} \]
Phrase Extraction Model: Word Alignments

In the past two years

Word Alignments

\[ a_{1,4} = \text{on} \quad a_{1,1} = \text{off} \]
Phrase Extraction Model: Word Alignments

In the past two years

过去 两年 中

Word Alignments

\[ a_{1,4} = 	ext{on} \quad a_{1,1} = 	ext{off} \]
Phrase Extraction Model: Word Alignments

In the past two years

Word Alignments

$$a_{1,4} = \text{on} \quad a_{1,1} = \text{off}$$

$$L_{ij}(a_{ij}) = \left\{ \right.$$
Phrase Extraction Model: Word Alignments

In the past two years

过去 两年 中

Word Alignments

\[ a_{1,4} = \text{on} \quad a_{1,1} = \text{off} \]

\[ L_{ij}(a_{ij}) = \begin{cases} a_{ij} = \text{on} \end{cases} \]
In the past two years

过去两年中

Word Alignments

$$a_{1,4} = \text{on} \quad a_{1,1} = \text{off}$$

$$L_{ij}(a_{ij}) = \begin{cases} 
\exp(w^\top f(i, j, s)) & a_{ij} = \text{on} \\
\end{cases}$$
Phrase Extraction Model: Word Alignments

In the past two years

In the past two years

Word Alignments

\[ a_{1,4} = \text{on} \quad a_{1,1} = \text{off} \]

\[ L_{ij}(a_{ij}) = \begin{cases} 
\exp(w^\top f(i, j, s)) & a_{ij} = \text{on} \\
& a_{ij} = \text{off} 
\end{cases} \]
Phrase Extraction Model: Word Alignments

In the past two years

过去 两年 中

Word Alignments

$$a_{1,4} = \text{on} \quad a_{1,1} = \text{off}$$

$$L_{ij}(a_{ij}) = \begin{cases} 
\exp(w^\top f(i, j, s)) & a_{ij} = \text{on} \\
1 & a_{ij} = \text{off}
\end{cases}$$
Phrase Extraction Model: Word Alignments

In the past two years

Word Alignments

\[ L_{ij}(a_{ij}) = \begin{cases} 
\exp(w^\top f(i, j, s)) & a_{ij} = \text{on} \\
1 & a_{ij} = \text{off}
\end{cases} \]
Phrase Extraction Model: Word Alignments

In the past two years

过去

两年

中
Phrase Extraction Model: Word Alignments

In the past two years

ITG Structural Constraint
Phrase Extraction Model: Word Alignments

In the past two years

过去两年中

$A(a_{1,1}, \ldots, a_{nm})$
In the past two years

\[ A(a_{1,1}, \ldots, a_{nm}) = \begin{cases} 
1 & \text{a is ITG} \\
0 & \text{otherwise}
\end{cases} \]
In the past two years

\[ A(a_{1,1}, \ldots, a_{nm}) = \begin{cases} 
1 & a \text{ is ITG} \\
0 & \text{otherwise} 
\end{cases} \]
In the past two years

Enforce fertility constraints

\[ A(a_{1,1}, \ldots, a_{nm}) = \begin{cases} 
1 & a \text{ is ITG} \\
0 & \text{otherwise} 
\end{cases} \]
In the past two years

- Enforce fertility constraints
- Enable efficient inference

$$A(a_{1,1}, \ldots, a_{nm}) = \begin{cases} 
1 & \text{a is ITG} \\
0 & \text{otherwise}
\end{cases}$$
Phrase Extraction Model: Word Alignments

In the past two years.

Enforce fertility constraints.

Enable efficient inference.

\[ A(a_{1,1}, \ldots, a_{nm}) = \begin{cases} 
1 & \text{a is ITG} \\
0 & \text{otherwise} 
\end{cases} \]
Phrase Extraction Model: Word Alignments

In the past two years

Enforce fertility constraints
Fertility Constraints

\[ U_1^f \rightarrow a_{1,1}, a_{2,1}, \ldots, a_{5,1} \]
Fertility Constraints

\[ U^f_1 \]

\[ U^f_j (a_{1j}, \ldots, a_{nj}) \]
Fertility Constraints

$U_1^f (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}$
Fertility Constraints

\[ U^f_j(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]

\[ k = 3 \]
Fertility Constraints

\[ U_f^f (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]

\[ k = 3 \]
Fertility Constraints

\[ U_1^f \]

\[ U_j^f (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise} 
\end{cases} \]

\[ k = 3 \]

In the past two years
Fertility Constraints

\[ U_{1f} \]

\[ a_{1,1} \quad a_{2,1} \quad \ldots \quad a_{5,1} \]

\[ U_{jf}(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise} 
\end{cases} \]

\[ k = 3 \]
Fertility Constraints

\[ U_f^j (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]

\[ k = 3 \]

In the past two years:

[Diagram showing two sequences with a block of three shaded in the second sequence, suggesting a comparison or identification with the fertility constraints.]
In the past two years

\[ U_j^f(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]

\( k = 3 \)
Fertility Constraints

\[ U^f_1 \]

\[ a_{1,1}, a_{2,1}, \ldots, a_{5,1} \]

\[ U^f_j (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise} 
\end{cases} \]

\[ k = 3 \]

In the past two years

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & 5 \\
\hline
\text{In} & \text{the} & \text{past} & \text{two years} & \text{past} & \text{past}
\end{array}
\]
Fertility Constraints

\[
U_1^f(a_{1,1}, a_{2,1}, \ldots, a_{5,1}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}
\]

\[k = 3\]

In the past two years:  
- In the past two years
- In the past two years
- In the past two years
- In the past two years
- In the past two years
Fertility Constraints

\[ U_1^f \]

- \( a_{1,1} \)
- \( a_{2,1} \)
- \( a_{5,1} \)

\[
U_j^f (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}
\]

\( k = 3 \)

In the past two years:

- The left diagram is valid.
- The right diagram is invalid.
Phrase Extraction Model: Word Alignments

**ITG Structural Constraint**

**Relaxed Constraints**
Phrase Extraction Model: Phrase Pairs

In the past two years

 går

兩年

中

Phrase Pairs
In the past two years

Phrase Pairs
Phrase Extraction Model: Phrase Pairs

In the past two years

过去 两年 中

$$\pi_{0,2,0,2} = \text{off}$$
Phrase Extraction Model: Phrase Pairs

In the past two years

$\pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on}$
Phrase Extraction Model: Phrase Pairs

In the past two years

\[ \pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on} \]
Phrase Extraction Model: Phrase Pairs

\[ \pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on} \]
Phrase Extraction Model: Phrase Pairs

\[ R_{ghk\ell}(\pi_{ghk\ell}) = \begin{cases} \\
\end{cases} \]

\( \pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on} \)
Phrase Extraction Model: Phrase Pairs

\[ R_{ghk\ell} (\pi_{ghk\ell}) = \begin{cases} \pi_{ghk\ell} = \text{on} \end{cases} \]

\[ \pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on} \]

In the past two years

过去

两年

中
Phrase Extraction Model: Phrase Pairs

\[
\pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on}
\]

\[
R_{ghkl}(\pi_{ghkl}) = \begin{cases} 
\exp(w^\top f(g, h, k, l, s)) & \pi_{ghkl} = \text{on} 
\end{cases}
\]
Phrase Extraction Model: Phrase Pairs

In the past two years

\[ \pi_{0,2,0,2} = \text{off} \quad \pi_{3,5,1,3} = \text{on} \]

\[
\begin{align*}
R_{ghk\ell}(\pi_{ghk\ell}) &= \begin{cases} 
\exp(w^\top f(g, h, k, \ell, s)) & \pi_{ghk\ell} = \text{on} \\
1 & \pi_{ghk\ell} = \text{off}
\end{cases}
\end{align*}
\]
In the past two years

过去
中

两年

Model Score

\[
\pi_{0,2,0,2} \quad \text{and} \quad \rho_{0,2,0,2}
\]
Model Score

In the past two years

过去两年中

\[ \pi_{0,2,0,2} \]

\[ R_{0,2,0,2} \]
在过去两年中

\[
\sum_{a_{ij} = \text{on}} w^T f(i, j, s)
\]
Model Score

In the past two years

\[
\sum_{a_{ij} = \text{on}} w^T f(i, j, s)
\]
Model Score

In the past two years

过去两年中

\[ \sum_{a_{ij} = \text{on}} w^{\top} f(i, j, s) + \sum_{\pi_{gkh\ell} = \text{on}} w^{\top} f(g, h, k, \ell, s) \]
Model Score

In the past two years

\[
P(a, \pi | s) \propto \exp \left( \sum_{a_{ij} = \text{on}} w^\top f(i, j, s) + \sum_{\pi_{ghk\ell} = \text{on}} w^\top f(g, h, k, \ell, s) \right)
\]
In the past two years.
Phrase Extraction Model: Word Spans

In the past two years

过去的两年中
In the past two years

 Phrase Extraction Model: Word Spans

过去两年中
Phrase Extraction Model: Word Spans

In the past two years

Word Spans

过去 两 年 中

\( \sigma_1^{f} \)

\( \pi_{0,2,0,2} \)

\( R_{0,2,0,2} \)
Phrase Extraction Model: Word Spans

\[ \sigma_f^1 = [1, 3] \]
Phrase Extraction Model: Word Spans

\[ \sigma_1^f = [1, 3] \]
In the past two years, 

\[ \sigma_1^f = [1, 3] \]
Phrase Extraction Model: Word Spans

In the past two years...

Word Spans

\[ \sigma^f_1 = [1, 3] \]
In the past two years

\[ \sigma_1^f = [1, 3] \]
Full Model: ITG
Full Model: Relaxed
Belief Propagation
Belief Propagation
Belief Propagation
Computing Messages
Computing Messages
Computing Messages
Computing Messages
Computing Messages
Computing Messages
Computing Messages
Computing Messages
Computing Messages

Enumeration is exponential in arity of factor
Computing Messages

Enumeration is exponential in arity of factor
Efficient Propagators

$U_1^f \rightarrow a_{1,1} \rightarrow a_{2,1} \rightarrow \cdots \rightarrow a_{5,1}$
Efficient Propagators

Arity of factor: $O(n)$
Efficient Propagators

Arity of factor: \( O(n) \)  \hspace{1cm} \text{Total assignments:} \hspace{0.5cm} O(2^n)
Efficient Propagators

Arity of factor: $O(n)$

Total assignments: $O(2^n)$

$U^f_j(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}$
Efficient Propagators

Arity of factor: \( O(n) \)  \hspace{2cm} \text{Total assignments:} \quad O(2^n) \\

\[
U^f_j(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}
\]
Efficient Propagators

Arity of factor: $O(n)$

Total assignments: $O(2^n)$

\[ U^f_j (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]
Efficient Propagators

Arity of factor: $O(n)$
Total assignments: $O(2^n)$

$$U^f \left( a_{1,j}, \ldots, a_{n,j} \right) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}$$
Efficient Propagators

Arity of factor: \( O(n) \)  
Total assignments: \( O(2^n) \)

\[
U_f^1(a_{1,1}, \ldots, a_{n,1}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}
\]
Efficient Propagators

$U_1^f$ $a_{1,1}$ $a_{2,1}$ $\cdots$ $a_{5,1}$

Arity of factor: $O(n)$  Total assignments: $O(2^n)$

$U_j^f(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}$
Efficient Propagators

Arity of factor: $O(n)$  
Total assignments: $O(2^n)$

$U^f_j(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases}$
Efficient Propagators

Arity of factor: $O(n)$

Total assignments: $O(2^n)$

$$U^f_j(a_1, \ldots, a_n) = \begin{cases} 1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\ 0 & \text{otherwise} \end{cases}$$
Efficient Propagators

\[ U_1^f \]

\( a_{1,1}, a_{2,1}, \ldots, a_{5,1} \)

**Arity of factor:** \( O(n) \)  
**Total assignments:** \( O(2^n) \)

**Total non-zero assignments:**

\[ U_j^f (a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & \text{if } a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]
Efficient Propagators

\[ U^f_1 \]

\[ a_{1,1} \quad a_{2,1} \quad \cdots \quad a_{5,1} \]

Arity of factor: \( O(n) \)  \quad Total assignments: \( O(2^n) \)

Total non-zero assignments: \( O(kn) \)

\[ U^f_j(a_{1j}, \ldots, a_{nj}) = \begin{cases} 1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\ 0 & \text{otherwise} \end{cases} \]
Efficient Propagators

\[ U_1^f \]

Arity of factor: \( O(n) \)  \quad \text{Total assignments: } \ O(2^n) \n
Total non-zero assignments: \( O(n) \)

\[ U_j^f(a_{1j}, \ldots, a_{nj}) = \begin{cases} 
1 & a_{*j} \text{ have } \leq 1 \text{ block of size } \leq k \\
0 & \text{otherwise}
\end{cases} \]
<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td></td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td></td>
</tr>
<tr>
<td>Factor</td>
<td>How to Sum</td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
</tr>
<tr>
<td>ITG</td>
<td></td>
</tr>
<tr>
<td>Fertility</td>
<td></td>
</tr>
<tr>
<td>Constraints</td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ITG</strong></td>
<td>Use bitext inside-outside</td>
<td></td>
</tr>
<tr>
<td><strong>Fertility Constraints</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td></td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td></td>
</tr>
<tr>
<td>Factor</td>
<td>How to Sum</td>
<td>Cost</td>
<td>Count</td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------------------------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td>1</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td></td>
</tr>
</tbody>
</table>
To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td>1</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
</tr>
</tbody>
</table>
To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
<th>Count</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
<th>Count</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td>1</td>
<td>$O(n^6)$</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
<td></td>
</tr>
</tbody>
</table>
## To ITG or not to ITG

<table>
<thead>
<tr>
<th>Factor</th>
<th>How to Sum</th>
<th>Cost</th>
<th>Count</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITG</td>
<td>Use bitext inside-outside</td>
<td>$O(n^6)$</td>
<td>1</td>
<td>$O(n^6)$</td>
</tr>
<tr>
<td>Fertility Constraints</td>
<td>Exploit sparsity of non-zero assignments</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
<td>$O(n^2)$</td>
</tr>
</tbody>
</table>
Training
Training

Run belief propagation
Training

- Run belief propagation
- Use approximate marginals...
Training

- Run belief propagation

- Use approximate marginals...

\[ P(a_{ij} | s) \]
Training

- Run belief propagation
- Use approximate marginals...

\[ P(a_{ij} \mid s) \quad P(\pi_{ghkl} \mid s) \]
Training

- Run belief propagation

- Use approximate marginals...

\[ P(a_{ij} | s) \quad P(\pi_{ghk\ell} | s) \]

...to compute expected feature counts
Training

- Run belief propagation

- Use approximate marginals...

\[ P(a_{ij} | s) \quad P(\pi_{ghk\ell} | s) \]

...to compute expected feature counts

- Plug into gradient optimizer of choice
Precision/Recall Tradeoff
Precision/Recall Tradeoff

Viterbi Decoding
(\hat{\pi}, \hat{a}) = \arg\max_{(\pi, a)} P(\pi, a | s)
Precision/Recall Tradeoff

Viterbi Decoding

\[ (\hat{\pi}, \hat{a}) = \arg \max_{(\pi, a)} P(\pi, a | s) \]
Precision/Recall Tradeoff

Viterbi Decoding

\[
\hat{(\pi, a)} = \arg\max_{(\pi, a)} P(\pi, a | s)
\]

Marginal Decoding

Phrase Extraction

\[ + \text{Viterbi ITG} \]
Precision/Recall Tradeoff

Viterbi Decoding

\[(\hat{\pi}, \hat{a}) = \arg\max_{(\pi, a)} P(\pi, a | s)\]

Marginal Decoding

\[\hat{\pi} = \{(gh, k\ell) : P(\pi_{ghk\ell} | s) > \tau\}\]

Phrase Extraction

![Graph showing Precision vs Recall with Viterbi ITG marked with a cross.](image)
Precision/Recall Tradeoff

Viterbi Decoding

\((\hat{\pi}, \hat{a}) = \operatorname{argmax}_{(\pi, a)} P(\pi, a|s)\)

Marginal Decoding

\(\hat{\pi} = \{(gh, k\ell) : P(\pi_{ghk\ell}|s) > \tau\}\)

Phrase Extraction

\[
\begin{array}{c|c|c}
\text{Precision} & \text{Recall} \\
\hline
60 & 60 & 60 \\
65 & 65 & 65 \\
70 & 70 & 70 \\
75 & 75 & 75 \\
80 & 80 & 80 \\
85 & 85 & 85 \\
\end{array}
\]

- Viterbi ITG
- BP Relaxed
Viterbi Decoding

\[
(\hat{\pi}, \hat{a}) = \arg\max_{(\pi, a)} P(\pi, a|s)
\]

Marginal Decoding

\[
\hat{\pi} = \{(gh, k\ell) : P(\pi_{ghk\ell}|s) > \tau\}
\]
Precision/Recall Tradeoff

Viterbi Decoding

$$(\hat{\pi}, \hat{a}) = \arg\max_{(\pi, a)} P(\pi, a | s)$$

Marginal Decoding

$$\hat{\pi} = \{(gh, k\ell) : P(\pi_{ghk\ell} | s) > \tau\}$$

Phrase Extraction

- Viterbi ITG
- BP Relaxed
Precision/Recall Tradeoff

**Viterbi Decoding**

\[
(\hat{\pi}, \hat{a}) = \arg\max_{(\pi, a)} P(\pi, a | s)
\]

**Marginal Decoding**

\[
\hat{\pi} = \{(gh, k\ell) : P(\pi_{ghk\ell} | s) > \tau\}
\]
Precision/Recall Tradeoff

**Viterbi Decoding**

\[(\hat{\pi}, \hat{a}) = \text{argmax}_{\pi, a} P(\pi, a | s)\]

**Marginal Decoding**

\[\hat{\pi} = \{ (gh, k\ell) : P(\pi_{ghk\ell} | s) > \tau \}\]

Phrase Extraction

- Viterbi ITG
- BP Relaxed

Graph showing the tradeoff between precision and recall for different values.
Phrase Extraction Results

![Bar chart showing F1 and F5 results for different pipelines]

- **F1**
  - BP ITG: 72.6
  - Viterbi ITG: 71.8
  - BP Relaxed: 71.6

- **F5**
  - BP ITG: 83.5
  - Viterbi ITG: 84.5
  - BP Relaxed: 62.8
Phrase Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>68.4</td>
<td>62.8</td>
</tr>
<tr>
<td>Viterbi ITG</td>
<td>71.6</td>
<td>74.0</td>
</tr>
<tr>
<td>BP ITG</td>
<td>71.8</td>
<td>83.5</td>
</tr>
<tr>
<td>BP Relaxed</td>
<td>72.6</td>
<td>84.5</td>
</tr>
</tbody>
</table>
Phrase Extraction Results

F1  F5
60   65   70   75   80   85

68.4

Pipeline
Phrase Extraction Results

![Bar Chart]

- **F1**: 68.4
- **F5**: 62.8

Legend:
- **Pipeline**
Phrase Extraction Results

- F1: 68.4
- F5: 62.8

Pipeline: 68.4
Viterbi ITG: 62.8
Phrase Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>68.4</td>
<td>62.8</td>
</tr>
<tr>
<td>Viterbi ITG</td>
<td>71.6</td>
<td>62.8</td>
</tr>
</tbody>
</table>

Legend: 
- Pipeline
- Viterbi ITG
Phrase Extraction Results

![Bar Chart]

- **F1**:
  - Pipeline: 68.4
  - Viterbi ITG: 71.6

- **F5**:
  - Pipeline: 62.8
  - Viterbi ITG: 74.0

Legend:
- Pipeline
- Viterbi ITG
Phrase Extraction Results

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>68.4</td>
<td>62.8</td>
</tr>
<tr>
<td>Viterbi ITG</td>
<td>71.6</td>
<td>74.0</td>
</tr>
<tr>
<td>BP ITG</td>
<td>71.8</td>
<td></td>
</tr>
</tbody>
</table>

- Pipeline
- Viterbi ITG
- BP ITG
Phrase Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>Pipeline</th>
<th>Viterbi ITG</th>
<th>BP ITG</th>
<th>BP Relaxed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>68.4</td>
<td>71.6</td>
<td>71.8</td>
<td>62.8</td>
</tr>
<tr>
<td>F5</td>
<td>83.5</td>
<td>74.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phrase Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>68.4</td>
<td>62.8</td>
</tr>
<tr>
<td>Viterbi ITG</td>
<td>71.6</td>
<td>74.0</td>
</tr>
<tr>
<td>BP ITG</td>
<td>71.8</td>
<td></td>
</tr>
<tr>
<td>BP Relaxed</td>
<td>72.6</td>
<td>83.5</td>
</tr>
</tbody>
</table>
Phrase Extraction Results

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pipeline</td>
<td>68.4</td>
<td>62.8</td>
</tr>
<tr>
<td>Viterbi ITG</td>
<td>71.6</td>
<td>74.0</td>
</tr>
<tr>
<td>BP ITG</td>
<td>71.8</td>
<td>83.5</td>
</tr>
<tr>
<td>BP Relaxed</td>
<td>72.6</td>
<td>84.5</td>
</tr>
</tbody>
</table>
Machine Translation Results

- Viterbi ITG
- BP Relaxed
Machine Translation Results

- BLEU (Moses)
  - Viterbi ITG: 33.5
  - BP Relaxed: 33.6
Conclusions
Conclusions

- Modeling phrase extraction improves on pipelined approaches
Conclusions

- Modeling phrase extraction improves on pipelined approaches
- The “right” constraints aren’t always the slowest
Conclusions

- Modeling phrase extraction improves on pipelined approaches
- The “right” constraints aren’t always the slowest
- Marginal decoding allows explicit management of precision/recall tradeoffs
Conclusions

- Modeling phrase extraction improves on pipelined approaches
- The “right” constraints aren’t always the slowest
- Marginal decoding allows explicit management of precision/recall tradeoffs
- Approximate inference techniques let you fit the inference to the model, rather than fitting the model to the inference procedure
Thank You

谢谢

您

！

Thank You