Joint Word Alignment and Decipherment Improves Machine Translation

Qing Dou, Ashish Vaswani, and Kevin Knight
10/26/2014
Outline

• What is Decipherment
• Motivation
• Contributions
• Joint Word Alignment and Decipherment
• Deciphering Malagasy
• Conclusions
What is Decipherment?

• Letter Substitution Cipher

plaintext
decipherment is the analysis of documents written in ancient languages
What is Decipherment?

• Letter Substitution Cipher

```
plaintext

decipherment is the analysis of documents written in ancient languages

```

Encryption

```
ciphertext

05 13 19 25 12 14 13 04 02 13 11 16 15
25 22 15 16 14 13 15 17 11 17 08 03 22
25 22 15 09 20 15 05 09 19 07 02 13 11
22 15 06 04 25 16 16 13 11 15 25 11 15
17 11 19 25 13 11 16 15 08 17 11 03 07
17 03 13 22
```
What is Decipherment?

• Letter Substitution Cipher

**Encryption**

plaintext

decipherment is the analysis of documents written in ancient languages

ciphertext

<table>
<thead>
<tr>
<th>a</th>
<th>17</th>
<th>o</th>
<th>09</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>01</td>
<td>p</td>
<td>12</td>
</tr>
<tr>
<td>c</td>
<td>19</td>
<td>q</td>
<td>23</td>
</tr>
<tr>
<td>d</td>
<td>05</td>
<td>r</td>
<td>04</td>
</tr>
<tr>
<td>e</td>
<td>02</td>
<td>s</td>
<td>22</td>
</tr>
<tr>
<td>f</td>
<td>16</td>
<td>t</td>
<td>16</td>
</tr>
<tr>
<td>g</td>
<td>04</td>
<td>u</td>
<td>07</td>
</tr>
<tr>
<td>h</td>
<td>14</td>
<td>v</td>
<td>24</td>
</tr>
<tr>
<td>i</td>
<td>25</td>
<td>w</td>
<td>06</td>
</tr>
<tr>
<td>j</td>
<td>18</td>
<td>x</td>
<td>26</td>
</tr>
<tr>
<td>k</td>
<td>21</td>
<td>y</td>
<td>03</td>
</tr>
<tr>
<td>l</td>
<td>08</td>
<td>z</td>
<td>27</td>
</tr>
<tr>
<td>m</td>
<td>02</td>
<td>_</td>
<td>15</td>
</tr>
<tr>
<td>n</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Decipherment
Substitution Cipher and Translation

• Word Substitution Cipher

plaintext
the head of the german social democratic party …

Encryption

ciphertext
007834 000094 048235 007834 113485 087654 129823 032834 …

Decryption

• Word substitutions also take place in translation
Automatic Decipherment

• A Noisy Channel Model Approach (Knight et al. 2006)

A model of plaintext → Substitute → ciphertext

\[ P(p) \rightarrow P(c|p) \rightarrow c \]
Automatic Decipherment

- A Noisy Channel Model Approach (Knight et al. 2006)

Plaintext unrelated to ciphertext → Substitute

Search $P(c|p)$ to maximize

$$P(c) = \sum_p P(p)P(c|p)$$

EM

FREEZE!

plaintext

$P(p)$

$P(c|p)$

Substitute

ciphertext $c$
Automatic Decipherment

- A Noisy Channel Model Approach (Knight et al. 2006)

Plaintext unrelated to ciphertext

Search $P(c|p)$ to maximize

$$P(c) = \sum_p P(p)P(c|p)$$

\[\text{FREEZE!}\]

$P(p)$

plaintext

Substitute

$P(c|p)$

ciphertext $c$

\[O(N \cdot V^2 \cdot R)\]

- Time Complexity: $O(N \cdot V^2 \cdot R)$

(Forward-backward)

N: Ciphertext length
V: Vocabulary
R: EM iteration
Outline

• What is Decipherment
• Motivation
• Contributions
• Joint Word Alignment and Decipherment
• Deciphering Malagasy
• Conclusions
Motivation

• Decipherment improves machine translation (Dou and Knight 2013)
Motivation

• Decipherment improves machine translation (Dou and Knight 2013)
Motivation

• Decipherment improves machine translation (Dou and Knight 2013)
Motivation

• Joint Alignment and Decipherment?
Contributions

• Proposed a new framework to perform joint word alignment and decipherment

• The joint framework improves both word alignment and machine translation significantly

• Released Malagasy treebank and 15.3 million word Malagasy news data
Outline

• What is Decipherment
• Motivation
• Contributions
• Joint Word Alignment and Decipherment
• Deciphering Malagasy
• Conclusions
Word Alignment

• Word Alignment Model and Objective

Objective:

\[ P(F \mid E) = \sum_a \prod_{j=1}^J d(a) \cdot t(f_j \mid e_{a_j}) \]

distortion probabilities
translation probabilities
Decipherment

• Decipherment Model and Objective

(Dependency based Decipherment  Dou and Knight 2013)

Objective:

$$P(F_{mono}) = \sum_{e} P(e_1, e_2) \prod_{j=1}^{2} t(f_j \mid e_j)$$

LM: Dependency Language Model
(Created from dependency trees)

LM Probabilities (fixed)

translation probabilities
A New Objective

Word Alignment Objective:

\[ P(F \mid E) = \sum_a \prod_{j=1}^J d(a) \cdot t(f_j \mid e_{a_j}) \]

Decipherment Objective:

\[ P(F_{\text{mono}}) = \sum_e P(e_1e_2) \prod_{j=1}^2 t(f_j \mid e_j) \]

Shared Parameters:

\[ t(f \mid e) \]
A New Objective

Word Alignment Objective:

\[ P(F \mid E) = \sum_a \prod_{j=1}^{J} d(a) \cdot t(f_j \mid e_{a_j}) \]

Decipherment Objective:

\[ P(F_{\text{mono}}) = \sum_e P(e_1e_2) \prod_{j=1}^{2} t(f_j \mid e_j) \]

New Objective:

\[
P(\text{JOINT}) = P(F \mid E) + \alpha P(F_{\text{mono}})
\]
Learning Algorithm

• EM

5 iterations of EM on Parallel text only

EM
Parallel
Data
Learning Algorithm

• EM

Collect expected counts for:
\[ t(f | e) \]

Collect expected counts for:
\[ t(f | e) \quad d(a) \]
Learning Algorithm

- **EM**

  - **E Step**
    - Non Parallel Data
    - Parallel Data

```
Collect expected counts for:
\[ t(f | e) \]

Collect expected counts for:
\[ t(f | e) \quad d(a) \]
```

Sum up expected counts
Learning Algorithm

- EM

\[
\begin{align*}
&\text{EM} \\
&\text{Parallel Data} \\
&\text{E Step} \\
&\text{Non Parallel Data} \\
&\text{M Step} \\
&\text{E Step} \\
&\text{Parallel Data} \\
&\text{Update parameters} \\
&t(f \mid e) \quad d(a)
\end{align*}
\]
Learning Algorithm

• EM
E Step

• On Parallel Data
  (Brown et al. 1993, Vogel and Ney 1996)
E Step

• On Parallel Data
  (Brown et al. 1993, Vogel and Ney 1996)

• On Non-parallel Data
  Time complexity: $O(N \cdot V^2 \cdot R)$
  $V$: Vocabulary size  $N$: Ciphertext length

• Not Scalable when  $V \sim 10^5, N \sim 10^7$
E Step

• On Non-parallel Data

Use samples to collect expected counts:

Let $N$ be total number of samples we draw

And $e_1e_2$ be one of them:

$$\text{Expected Count}(f_1,e_1) = \text{Expected Count}(f_2,e_2) = \frac{1}{N} \cdot \text{count}(f_1,f_2)$$
## Word Alignment Experiment

- **Data (Size in tokens)**

<table>
<thead>
<tr>
<th></th>
<th>Spanish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>10.3k</td>
<td>9.9k</td>
</tr>
<tr>
<td>Non Parallel</td>
<td>80 million</td>
<td>400 million</td>
</tr>
<tr>
<td>TreeBank</td>
<td>0.4 million</td>
<td>1.0 million</td>
</tr>
</tbody>
</table>
Decipherment Improves Alignment

Spanish - English

Model 1

HMM

Baseline

Joint

F-Score

Iterations

USC University of Southern California

Information Sciences Institute
Outline

• What is Decipherment
• Motivation
• Contributions
• Joint Word Alignment and Decipherment
• Deciphering Malagasy
• Conclusions
The Malagasy Language

• Is official Language of Madagascar
The Malagasy Language

- Is official Language of Madagascar
The Malagasy Language

- Is official Language of Madagascar
The Malagasy Language

• Is official Language of Madagascar
• Although spoken in African, Malagasy has its root in southeast Asia.
The Malagasy Language

• Is official Language of Madagascar
• Although spoken in African, Malagasy has its root in southeast Asia.
• Has 18 million native speakers
The Malagasy Language

• Is official Language of Madagascar
• Although spoken in African, Malagasy has its root in southeast Asia.
• Has 18 million native speakers
• Is head initial with V-O-S word order. (English: S-V-O)
Malagasy Dependency Parser

• Data

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>120 sentences, 20k tokens</td>
<td>48 sentences, 7k tokens</td>
</tr>
</tbody>
</table>

Spanish parser trained on 400k tokens

• Result on Malagasy

72.4 % directed attachment accuracy
Malagasy Dependency Parser

• More Training Data

  - English
  - Parallel
  - Malagasy

  - English Dependency Tree
  - Manual Project Dependency
  - Malagasy Dependency Tree

  - retrain
  - New Parser

• Result

  Improved to 80.0% from 72.4%
## Malagasy-English MT

- **Data (In tokens)**

<table>
<thead>
<tr>
<th></th>
<th>Malagasy</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parallel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train (GV)</td>
<td>0.9 million</td>
<td>0.8 million</td>
</tr>
<tr>
<td>Tune (GV)</td>
<td>22.2k</td>
<td>20.2k</td>
</tr>
<tr>
<td>Test (GV)</td>
<td>23k</td>
<td>21k</td>
</tr>
<tr>
<td>Test (Web)</td>
<td>2.2k</td>
<td>2.1k</td>
</tr>
<tr>
<td><strong>Non Parallel</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GigaWord</td>
<td>N/A</td>
<td>834 million</td>
</tr>
<tr>
<td>Web</td>
<td>15.3 million</td>
<td>396 million</td>
</tr>
</tbody>
</table>

GV: Global Voices, multilingual international news website
Malagasy-English MT

• Baseline

Phrase-based MT system with Moeses

(Model 3 and Model 4 doesn’t improve BLEU)
Align in 2 directions and used grow-diag-final to extract phrases
Malagasy-English MT

• Joint

Parallel Data

Model1: 5 iterations (parallel only)
5 iterations (Joint)
HMM: 5 iterations (Joint)

Non Parallel Data

Alignment & Decipherment

Translation Model

Align and extract phrases only on one direction P(English|Malagasy)
Malagasy-English MT

• Disjoint
Results on Global Voices

![Bar chart showing BLEU scores for Tune (GV) and Test (GV) with three categories: Baseline, Disjoint, and Joint.](chart.png)
Results on Local News

![Graph showing BLEU scores for Baseline, Disjoint, and Joint methods on test (web) data.]
Conclusion

- Proposed a framework for joint alignment and decipherment
- The joint process improves both alignment and machine translation quality
- Released a mini Malagasy treebank and 15m tokens news data
Thank You!