From

$n$-gram-based
to

CRF-based
Translation Models

Lavergne - Crego - Allauzen - Yvon

LIMSI
I CAN HAS SIMPLER TRAINING?
A principio erat
verbum: et verbum erat apud 
Deum: et Deus
erat verbum. Hoc
erat in principio
Deum. Omnia per ipsi
sunt: et sive ipsa factum
sub: quod factum est.
1-gram-based noisy channel MT

Bi-text

Monolingual text

Word alignments & translation model & reordering model

Language model

Decoder

\[ e^* = \arg\max_e p(f|e)p(e) \]
**phrased-based discriminative MT**

- **Bi-text**
  - Monolingual text
  - Language model
  - Phrase translation model + reordering model
- **Tuning bi-text**
  - MERT
- **Decoder**
  - $e^* = \arg\max_e \lambda_{TM} \log p(f|e) + \lambda_{LM} \log p(e) + \sum_i \lambda_i f_i(f, e)$
Reordering for $n$-gram models

Training from word-aligned POS-tagged bi-text

Input search graph

The use of long tuples impoverishes the probability estimates of the translation model, as longer tuples appear less often in training than the smaller ones (data sparseness problem). Therefore, language pairs with significant differences in word order may suffer from poor probability estimates.

The motivation for extending the input graph is double: first, the aim to improve translation quality is met by the ability of reordering following the patterns as explained previously. Second, the reordering decision is more informed since it is taken during decoding using all the SMT models.

The extension procedure is outlined as follows: starting from the monotonic graph, new arcs, covering the source words in the desired word order, are added to the graph by means of covering (translating) some uncovered source word. However, this also results in many arcs as there are words present in the pattern. The unfolding technique makes use of the word alignments. It can be decomposed into two main steps: first, the words in one side are grouped when linked to the same word (or group) in the other side. The procedure loops grouping words in both sides until no new groups are obtained.
n-gram-based discriminative MT

Bi-text

Monolingual text

Language model

Tuning bi-text

n-gram translation model

Reordered word graph + Monotone decoder

\[ e^* = \arg\max_e \lambda_{TM} \log p(\tilde{f}, \tilde{e}) + \lambda_{LM} \log p(e) + \sum_i \lambda_i f_i(f, e) \]
CRF-based MT

\[ \tilde{f} : \{ \text{la} \} \{ \text{voilée} \} \{ \text{femme} \} \{ \text{demanda} \} \{ \text{de nouveau} \} \]

CRF translation model

\[ P(y^L_1|x^L_1) = \frac{1}{Z(x^L_1; \theta)} \exp(\theta^T G(x^L_1, y^L_1)) \]

FST: - segment & reorder
- monotone decoder

\[ \tilde{e}^* = \arg \max_{\tilde{e}} P(\tilde{e}|f) \]

\[ \approx \arg \max_{\tilde{f} \in \mathcal{L}(f), \tilde{e}} P(\tilde{e}, |\tilde{f}, f) P(\tilde{f}|f) \]
n-gram generative story

reordering (POS)

n-gram translation

segmentation
$$P(e|f) = \sum_{\tilde{f}, \tilde{e} \mid \phi(\tilde{e}) = e} P(\tilde{e}, \tilde{f}|f)$$
$$= \sum_{\tilde{f}, \tilde{e} \mid \phi(\tilde{e}) = e} P(\tilde{e}, |\tilde{f}, f)P(\tilde{f}|f)$$
$$= \sum_{\tilde{f}, \tilde{e} \mid \phi(\tilde{e}) = e} P(\tilde{e}, |\tilde{f})P(\tilde{f}|f)$$
CRF Features

- translation features \([\text{trs}]\)
- +context features \([\text{ctx}]\)
- +suffix/prefix features \([\text{ix}]\)
- +segmentation (length) features \([\text{seg}]\)
- +distortion (\(\Delta+\text{lex}\)) features \([\text{ord}]\)
- target bigram features \([\text{trg}]\)
- +LM (non-local)
Negative results

<table>
<thead>
<tr>
<th></th>
<th>dev</th>
<th>test</th>
<th># feat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moses (3g)</td>
<td>21.2</td>
<td>20.5</td>
<td></td>
</tr>
<tr>
<td>n-gram (2g,3g)</td>
<td>20.6</td>
<td>20.2</td>
<td>755K</td>
</tr>
<tr>
<td>n-gram (3g,3g)</td>
<td>21.5</td>
<td>21.2</td>
<td>755K</td>
</tr>
<tr>
<td>CRF (trs,trg)</td>
<td>-</td>
<td>18.3</td>
<td>660K</td>
</tr>
<tr>
<td>CRF +ctx</td>
<td>-</td>
<td>18.8</td>
<td>1.5M</td>
</tr>
<tr>
<td>CRF +ix,ord,seg</td>
<td>-</td>
<td>19.1</td>
<td>1.5M</td>
</tr>
<tr>
<td>CRF +ix,ord,seg+3g</td>
<td>-</td>
<td>19.1</td>
<td>1.5M</td>
</tr>
</tbody>
</table>

+Language model does not help

Weaknesses:
- scoring of reordering & segmentation
- target LM
Oracle segmentation / reordering

<table>
<thead>
<tr>
<th></th>
<th>dev</th>
<th>test</th>
<th># feat.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>decoding with optimal segmentation/reordering</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRF (trs,trg)</td>
<td>23.8</td>
<td>25.1</td>
<td>660K</td>
</tr>
<tr>
<td>CRF +ctx</td>
<td>24.1</td>
<td>25.4</td>
<td>1.5M</td>
</tr>
<tr>
<td>CRF +ix,ord,seg</td>
<td>24.3</td>
<td>25.6</td>
<td>1.5M</td>
</tr>
</tbody>
</table>

| **decoding with optimal reordering** |     |      |         |
| n-gram (2g,3g)          | 20.6| 24.1 | 755K    |
| n-gram (3g,3g)          | 21.5| 25.2 | 755K    |
| CRF trs,trg            | -   | 22.8 | 660K    |
| CRF +ctx               | -   | 23.1 | 1.5M    |
| CRF +ix,ord,seg        | -   | 23.5 | 1.5M    |

| **regular decoding** |     |      |         |
| Moses (3g)            | 21.2| 20.5 |         |
| n-gram (2g,3g)        | 20.6| 20.2 | 755K    |
| n-gram (3g,3g)        | 21.5| 21.2 | 755K    |
| CRF (trs,trg)         | -   | 18.3 | 660K    |
| CRF +ctx              | -   | 18.8 | 1.5M    |
| CRF +ix,ord,seg       | -   | 19.1 | 1.5M    |
| CRF +ix,ord,seg+3g    | -   | 19.1 | 1.5M    |