Experimenting with Phrase-Based Statistical Translation within the IWSLT 2004 Chinese-to-English Shared Translation Task

Philippe Langlais
RALI/DIRO
University of Montreal
Canada
felipe@iro.umontreal.ca

Michael Carl
IAI
Saarbrücken
Germany
carl@iai.uni-sb.de

Oliver Streiter
NUK
National Univ. of Kaohsiung
Taiwan
ostreiter@nuk.edu.tw

IWSLT 2004, Kyoto, Japan
Motivations

• How far can we go in one month of work, starting from (almost) scratch and relying intensively on available packages?

• Interested by the perspective taken by the organizers: validation of existing evaluation methodologies. See also the CESTA project (TECHNOLANGUE):

http://www.technolangue.net/

We participated in:

• The Chinese-to-English track using only the 20K sentences provided
Plan

• Few words on the core engine
• Our phrase-based extractors
• Experiments with phrase-based models (PBMNs)
• Conclusions
The core engine

We used an off-the-shelf decoder: Pharaoh (Koehn, 2004). It requires:

- a flat PBM (*e.g.* small boats ↔ bateau de plaisance 0.82)
  
  details are coming soon

- we used SRILM (Stolcke, 2002) to produce a 3-gram
  
  ngram-count -interpolate -kndiscount1 -kndiscount2 -kndiscount3

- a set of parameters (one for the PBM, one for the language model, one for the length penalty and one for the built-in distortion model)
  
  details are coming soon

Pharaoh is a noisy channel phrase-based statistical engine.
Our phrase-based extractors

We tried two different methods of extraction:

**WABE:** relying on viterbi alignments computed from IBM model 3

We used Giza++ (Och and Ney, 2000) to get them out of an IBM model 3

**SBE:** One capitalizing on redundancies in the training corpus at the sentence level

- WABE = Word-Alignment Based Extractor
- SBE = String-Based Extractor
**WABE: Word-alignment based extractor**

Yet another version of (Koehn et al., 2003; Tillmann, 2003) and others. Basically:

- Considering the intersection of the word links obtained by viterbi alignment in both directions (C-E, E-C)
- (more or less) carefully extending this set of links with links belonging to the union of both sets (C-E,E-C)

Few meta-parameters are controlling the phrases acquired in this way:

**length-ratio**: ratio = 2

**min-max src/tgt length**: min=1, max=8

---

felipe@ rali

IWSLT 2004, Kyoto, Japan
**SBE: String-based extractor**

If two strings are in relation of translation and if part of them also are, then we can induce a specific translation relation between the other parts.

\[
res \leftarrow T = \{(E_i, C_i), i \in [1, |T|]\} \text{ (the training corpus)}
\]

**repeat**

**for all** \(\langle (E_i, C_i), (E_j, C_j) \rangle \in res\) **do**

**if** \(C_j = C_i \alpha\) or \(C_i = C_j \alpha\) **then**

**if** \(E_j = E_i \beta\) or \(E_i = E_j \beta\) **then**

\[
res \leftarrow res \cup (\beta, \alpha)
\]

**until** convergence of \(res\)

54,461 parameters out of 20K sentences
Experiments with PBMs: setting

| corpus   | |          |          | |      |          |          |
|----------|--------|----------|----------|--------|----------|----------|
|          | pair   | Chinese  |          | English|          |          |
|          |        | tokens   | words    | tokens | words    |
| TRAIN    | 20 000 | 182 904  | 7 643    | 188 935| 7 181    |
| TRAIN-A  | 11 884 | 112 000  | 6 456    | 116 343| 6 008    |
| TRAIN-Q  | 8 116  | 70 904   | 4 024    | 72 592 | 3 900    |
| CSTAR    | 506    | 3 515    | 870      | —      | —        |
| TEST     | 500    | 3 794    | 893      | —      | —        |

- the **tokenization** was the one provided, English material was **lowerized**,.
- **punctuation** marks were removed from the translations in accordance to the specifications (s/ \.///g, s/ ?///g, s/ ,//g s/ "//g, s/ \\!///g, s// /g, s/ *// /g)
- source **OOV** appearing in the translations were replaced afterward by the most likely word according to our 3g model (in a left-to-right manner). Uppercased OOV were left unmodified.
**Word-based translation versus PB translation**

<table>
<thead>
<tr>
<th>engine</th>
<th>NIST</th>
<th>BLEU%</th>
<th>mWER</th>
<th>mSER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibm2+3g</td>
<td>5.0726</td>
<td>26.57</td>
<td>60.56</td>
<td>94.47</td>
</tr>
<tr>
<td>Pharaoh</td>
<td>5.5646</td>
<td>26.16</td>
<td>59.70</td>
<td>94.27</td>
</tr>
<tr>
<td>wbm by Pharaoh</td>
<td>4.8417</td>
<td>15.54</td>
<td>64.95</td>
<td>97.63</td>
</tr>
</tbody>
</table>

- *ibm2+3g* is an extension of the decoder described by *(Niessen et al., 1998)*

- *Pharaoh* was run with its default setting; each parameter of the FPBM was scored by relative frequency
**Tuning the decoder**

<table>
<thead>
<tr>
<th>$\lambda_d$</th>
<th>$\lambda_\phi$</th>
<th>$\lambda_w$</th>
<th>$\lambda_{lm}$</th>
<th>NIST</th>
<th>BLEU%</th>
<th>MWER</th>
<th>MSER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5.5646</td>
<td>26.16</td>
<td>59.70</td>
<td>94.27</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1.5</td>
<td>1</td>
<td>6.3470</td>
<td>25.63</td>
<td>58.93</td>
<td>94.27</td>
</tr>
<tr>
<td>.2</td>
<td>.9</td>
<td>-1.5</td>
<td>.8</td>
<td>6.8401</td>
<td>28.44</td>
<td>56.25</td>
<td>94.07</td>
</tr>
</tbody>
</table>

$\lambda_d$, distorsion weight ([0, 1])

$\lambda_\phi$, transfer weight ([0, 1])

$\lambda_w$, word penalty ([-3, 3])

$\lambda_{lm}$, language model weight ([0, 1])

We applied a poor man’s strategy (sampling uniformly the parameter ranges)

→ a relative gain over the default configuration (line 1) of 23%

→ 61% of this gain obtained by tuning only the word penalty parameter
Merging different FPBM

| config | $|p|$ | NIST | BLEU\% | MWER | MSER |
|--------|-------|------|--------|------|------|
| WABE   | 6.8401| 28.44| 56.25  | 94.07|      |
| + WBM  | 7.0766| 31.38| 54.88  | 93.28|      |
| + SBE  | 7.0926| 31.78| 54.56  | 92.69|      |

Merging 2 models was done harshly by:

- copying $p_i(s|t), \forall s$ whenever $t$ has not been seen in one model,
- averaging them in case both $p_1(s|t)$ and $p_2(s|t)$ exist,
- normalizing

$\rightarrow$ a relative gain of 3.7%
The weakness of relative frequency

<table>
<thead>
<tr>
<th>min</th>
<th>max</th>
<th>( \text{model} )</th>
<th>%f1</th>
<th>%f2</th>
<th>%f3+</th>
<th>( %p = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>166 481</td>
<td>90.6</td>
<td>4.9</td>
<td>4.5</td>
<td>74.6</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>153 512</td>
<td>92.7</td>
<td>4.3</td>
<td>3.0</td>
<td>78.5</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>73 369</td>
<td>87.0</td>
<td>7.1</td>
<td>5.9</td>
<td>68.7</td>
</tr>
</tbody>
</table>

- \%f1, \%f2 and \%f3+ stand for the percentage of parameters (pairs of phrases) seen 1, 2 or at least 3 times in the \textsc{train} corpus.

- \( \%p = 1 \) stands for the percentage of parameters that have a relative frequency of 1.
### Scoring phrases with IBM model 1

<table>
<thead>
<tr>
<th>model</th>
<th>NIST</th>
<th>BLEU%</th>
<th>MWER</th>
<th>MSER</th>
</tr>
</thead>
<tbody>
<tr>
<td>relfreq</td>
<td>7.0926</td>
<td>31.78</td>
<td>54.56</td>
<td>92.69</td>
</tr>
<tr>
<td>ibm</td>
<td>7.3067</td>
<td>32.98</td>
<td>53.86</td>
<td>92.49</td>
</tr>
<tr>
<td>relfreq&amp;ibm</td>
<td>7.3118</td>
<td>34.48</td>
<td>52.73</td>
<td>91.90</td>
</tr>
<tr>
<td>relfreq&amp;pn-ibm</td>
<td>7.4219</td>
<td>34.6</td>
<td>53.02</td>
<td>91.70</td>
</tr>
</tbody>
</table>

- baseline model (line 1) = merged FPBM of 306 585 parameters trained by relative frequency.

- rating these parameters by IBM model 1 yields a relative improvement in the NIST score of 3%

- pn-ibm: do not normalize parameters where $|\{s : p(s|t) \exists\}| = 1$ holds
Specific models

<table>
<thead>
<tr>
<th>config</th>
<th>NIST</th>
<th>BLEU%</th>
<th>MWER</th>
<th>MSER</th>
</tr>
</thead>
<tbody>
<tr>
<td>relfreq&amp;ibm</td>
<td>7.3118</td>
<td>34.48</td>
<td>52.73</td>
<td>91.90</td>
</tr>
<tr>
<td>A</td>
<td>7.1862</td>
<td>34.21</td>
<td>53.12</td>
<td>91.18</td>
</tr>
<tr>
<td>Q</td>
<td>6.4995</td>
<td>34.92</td>
<td>52.12</td>
<td>93.00</td>
</tr>
<tr>
<td>specific-lm</td>
<td>7.4702</td>
<td>33.64</td>
<td>53.27</td>
<td>91.90</td>
</tr>
<tr>
<td>A</td>
<td>7.3229</td>
<td>33.66</td>
<td>53.08</td>
<td>90.85</td>
</tr>
<tr>
<td>Q</td>
<td>6.7010</td>
<td>33.58</td>
<td>53.55</td>
<td>93.50</td>
</tr>
</tbody>
</table>

- around 40% of the training sentences were interrogatives ones

⇒ specific language model combined to the general one (specific tuning over 6 parameters)

(we did not observe improvements by modelling specific FPBM's)
Translations we submitted before the deadline

*ibm2+3g* word-based translation engine,

*straight* a WABE FPBM

*merge* the best model obtained by merging word and phrase associations

*QA* the one submitted for manual evaluation

*manual* to measure the usefulness of the automatic translations for human post-editing

**Task:** selecting one translation among the generated ones and enhancing its quality though slight modifications

felipe@ rali

IWSLT 2004, Kyoto, Japan
The manual experiment

- 423 (84.6%) were just selections of one of the automatic translations.

- Out of these 423 translations, 85 (20%) were produced by the word-based engine ($ibm2+3g$).

<table>
<thead>
<tr>
<th>trans1</th>
<th>take a bath for a twin room .</th>
</tr>
</thead>
<tbody>
<tr>
<td>trans2</td>
<td>please take a bath for a double .</td>
</tr>
<tr>
<td>trans3</td>
<td>take a bath of double .</td>
</tr>
<tr>
<td>trans4</td>
<td>take one twin room with bath .</td>
</tr>
<tr>
<td>trans5</td>
<td>have a bath for double .</td>
</tr>
<tr>
<td>trans6</td>
<td>have a twin room with bath , please .</td>
</tr>
<tr>
<td>trans7</td>
<td>have a double room with bath , please .</td>
</tr>
<tr>
<td>manual</td>
<td>please, a twin room with bath .</td>
</tr>
</tbody>
</table>
## Translations we submitted before the deadline

<table>
<thead>
<tr>
<th>config</th>
<th>BLEU%</th>
<th>NIST</th>
<th>GTM</th>
<th>WER</th>
<th>Per</th>
</tr>
</thead>
<tbody>
<tr>
<td>ibm2+3g</td>
<td>27.27</td>
<td>6.55</td>
<td>62.49</td>
<td>58.12</td>
<td>48.82</td>
</tr>
<tr>
<td>straight</td>
<td>30.92</td>
<td>7.52</td>
<td>66.93</td>
<td>56.05</td>
<td>47.90</td>
</tr>
<tr>
<td>merge</td>
<td>35.32</td>
<td>8.00</td>
<td>68.60</td>
<td>51.74</td>
<td>43.86</td>
</tr>
<tr>
<td>QA</td>
<td>33.89</td>
<td>7.85</td>
<td>68.55</td>
<td>53.24</td>
<td>45.14</td>
</tr>
<tr>
<td>manual</td>
<td>36.93</td>
<td>8.13</td>
<td>68.42</td>
<td>49.62</td>
<td>42.53</td>
</tr>
</tbody>
</table>

The ordering of the variants was (almost) consistent with the one observed on the Cstar corpus.
Conclusions

• Is phrase-based translation ≡ Pharaoh(Giza++$^\lambda_g \times$ SRILM$^\lambda_s$) ?
  ↦ at least a decent system can be obtained this way

• Things we tried that did not work better:
  – splitting the training sentences into shorter ones
  – replacing proper names by NAME

• Many factors to be tried:
  – word alignment procedure (Simard and Langlais, 2003)
  – other scoring functions (Zao et al., 2004)

• Not clear whether the best settings we found here would be appropriate for another translation task
References


WABE

Require: $\mathcal{P}, \mathcal{R}, minLength, maxLength, ratio$

Ensure: $res$ contains all the pairs of phrases

1: Initialization
2: $res \leftarrow \{\}$
3: for all $x : 1 \rightarrow |S|$ do $T[x] \leftarrow \{\}$
4: for all $y : 1 \rightarrow |T|$ do $T[y] \leftarrow \{\}$
5:
6: Step 1: $\mathcal{P}$-projection
7: for all $(x,y) \in \mathcal{P}$ do $add(x,y)$
8:
9: Step 2: Extension
10: for $p : 1 \rightarrow 2$ do
11:    repeat
12:    $a \leftarrow \{\}$
13:    for $s : 1 \rightarrow |S|$ do

felipe@ rali

IWSLT 2004, Kyoto, Japan
14: \textbf{for all} $t \in T[s]$ \textbf{do}
15: \hspace{1em} \textbf{if} $p = 2$ \textbf{then}
16: \hspace{2em} \text{neighbor}(x-1,y-1); \text{neighbor}(x+1,y-1);
17: \hspace{2em} \text{neighbor}(x-1,y+1); \text{neighbor}(x+1,y-1);
18: \hspace{1em} \textbf{else}
19: \hspace{2em} \text{neighbor}(x-1,y); \text{neighbor}(x+1,y);
20: \hspace{2em} \text{neighbor}(x,y-1); \text{neighbor}(x,y+1);
21: \hspace{1em} \textbf{for all} $(x, y) \in a$ \textbf{do} \text{add}(x, y)
22: \hspace{1em} \textbf{until} $|a| = 0$
23: 
24: \textit{Step 3: Collect independent boxes}
25: \hspace{1em} $b \leftarrow \{\}$
26: \hspace{1em} \textbf{for} $x : 1 \rightarrow |S|$ \textbf{do}
27: \hspace{2em} $X \leftarrow \{x\}; \ Y \leftarrow \{\}$
28: \hspace{2em} \textbf{repeat}
29: \hspace{3em} $X_m \leftarrow X; \ Y_m \leftarrow Y$
for all $x \in X$ do $Y \leftarrow Y \cup T[x]$

if $Y \neq Y_m$ then

for all $y \in Y$ do $X \leftarrow X \cup T[y]$

until $X = X_m$ and $Y = Y_m$

$b \leftarrow b \cup \left\{ \left( \min\{x : x \in X\}, \max\{x : x \in X\} \right), \left( \min\{y : y \in Y\}, \max\{y : y \in Y\} \right) \right\}$

$x \leftarrow \max\{x : x \in X\} + 1$

Step 4: Combine boxes

for $i : 1 \rightarrow |b|$ do

let $((x_{m_i}, x_{M_i}), (y_{m_i}, y_{M_i})) = b_i$

add($x_{m_i}, x_{M_i}, y_{m_i}, y_{M_i}$)

for $j : i + 1 \rightarrow |b|$ do

let $((x_{m_j}, x_{M_j}), (y_{m_j}, y_{M_j})) = b_j$

if $x_{M_i} + 1 = x_{m_j}$ then

add($x_{m_i}, x_{M_j}, y_{m_i}, y_{M_j}$)