Relearning an RBMT System and Pivot Approaches

Marta Berardi, Alina Baranova
Hybrid Machine Translation
SS 2018, 15.05
Can we relearn an RBMT system?

DUGAST, Loïc; SENELLART, Jean; KOEHN, Philipp

*Can we relearn an RBMT system?*

Table of contents

- Systran and Hybrid Machine Translation
- Relearning an RBMT system
- Systran today
SYSTRAN in a nutshell

→ Founded in 1968 out of pioneering research at Georgetown University

→ Goal: creation of software that automatically translates information from one language into another

→ One of the first developers of machine translation software:
  ◆ Cold War, SYSTRAN - US Air Force cooperation
  ◆ Creation of the first translation software from Russian to English
SYSTRAN in a nutshell

➔ Continuous Improvements in MT: Systran R&D
➔ Technology Breakthroughs

- First hybrid translation software solution combining 40+ years of rule-based linguistics and the latest statistical techniques for publishable quality translation
- First Neural Machine translation software producing an automatic translation overachieving the current state of the art, and better than a non-native speaker
SYSTRAN in a nutshell

➔ The SYSTRAN system is traditionally classified as a “rule-based” system
➔ Constant evolution over the years – integrating most of the available productive techniques
➔ Evolution governed by the following principles:
  ➔ Provide a deterministic output
  ➔ Incremental translation quality
SYSTRAN in a nutshell

- Crucial component: linguistic resources
  - 100k – 800k entries
  - Entries = simple or multi-word lexical entries, but also customized disambiguation rules
SYSTRAN Hybrid Technology

→ What is the advantage of combining RMBT with SMT?
→ RB components guarantee:
  ◆ predictable and consistent translations
  ◆ compliance with corporate terminology
  ◆ out-of-domain usability
  ◆ high performance
→ Statistical components learn from existing monolingual and multilingual corpora
  ◆ reduced customization costs
  ◆ improved translation quality within specified domains
SYSTRAN Hybrid Technology

➔ SYSTRAN’s rule-based translation software is the backbone of its hybrid MT
➔ Statistical post-editing
➔ Built-in knowledge (Systran Enterprise Server 7, 2009)
  ◆ Keyword: domain specific terminology
  ◆ Users can add their own linguistic resources such as dictionaries to improve the quality of specific domains or business objectives
  ◆ Built-in or customer specific language models increase disambiguation of the source text
Can we relearn an RBMT system?

➔ 2008, Systran submission at the Workshop on Statistical Machine Translation at ACL

➔ French-English statistical model trained without any human-translation involved:
  ➔ Monolingual corpus translated with Systran rule-based engine to produce parallel corpus
Can we relearn an RBMT system?

- RBMT: manually written rules associated with bilingual dictionaries
- SMT: usage of implicit linguistic information present in translated corpora
- Pioneer experiment: Systran Relearnt, statistical model of the rule-based engine
Can we relearn an RBMT system?

→ Systran relearnt:
  ◆ Phrase-based system
  ◆ Phrase table trained on the rule-based translation of the French Europarl corpus
  ◆ Language model however trained on the real English Europarl data provided for the task
Can we relearn an RBMT system?

➔ Comparison of the performances of three systems:

- Baseline Systran (rule-based): **BLEU 21.27**
- Systran Relearnt (phrase-based, phrase table trained on the rule-based translation of French Europarl corpus): **BLEU 26.57**
- Baseline Moses (phrase-based, phrase table learnt from Europarl parallel data): **BLEU 29.86**
Can we relearn an RBMT system?

Error types

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>Missing Content</td>
</tr>
<tr>
<td>MO</td>
<td>Missing Other</td>
</tr>
<tr>
<td>TCL</td>
<td>Translation Choice (content, lemma)</td>
</tr>
<tr>
<td>TCI</td>
<td>Translation Choice (content, inflection)</td>
</tr>
<tr>
<td>TCO</td>
<td>Translation Choice (other)</td>
</tr>
<tr>
<td>EWC</td>
<td>Extra Word Content</td>
</tr>
<tr>
<td>EWO</td>
<td>Extra Word Other</td>
</tr>
<tr>
<td>UW</td>
<td>Unknown word</td>
</tr>
<tr>
<td>WOS</td>
<td>Word Order, short</td>
</tr>
<tr>
<td>WOL</td>
<td>Word Order, long (distance&gt;=3 words)</td>
</tr>
<tr>
<td>PNC</td>
<td>Punctuation</td>
</tr>
</tbody>
</table>

*Table 4: Short definition of error types*
Can we relearn an RBMT system?

<table>
<thead>
<tr>
<th>System</th>
<th>MC</th>
<th>MO</th>
<th>TCL</th>
<th>TCI</th>
<th>TCO</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTRAN</td>
<td>0.02</td>
<td>0.2</td>
<td>1.11</td>
<td>0.14</td>
<td>0.48</td>
</tr>
<tr>
<td>Releamt</td>
<td>0.22</td>
<td>0.39</td>
<td>0.77</td>
<td>0.22</td>
<td>0.38</td>
</tr>
<tr>
<td>Moses</td>
<td>0.35</td>
<td>0.46</td>
<td>0.63</td>
<td>0.27</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table 5.a: Average number of errors/sentence

<table>
<thead>
<tr>
<th>System</th>
<th>EWC</th>
<th>EW0</th>
<th>UW</th>
<th>WOS</th>
<th>WOL</th>
<th>PNC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTRAN</td>
<td>0</td>
<td>0.72</td>
<td>0.06</td>
<td>0.41</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>Releamt</td>
<td>0.05</td>
<td>0.35</td>
<td>0.09</td>
<td>0.41</td>
<td>0.05</td>
<td>0</td>
</tr>
<tr>
<td>Moses</td>
<td>0.17</td>
<td>0.4</td>
<td>0.12</td>
<td>0.3</td>
<td>0.08</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 5.b: Average number of errors/sentence
Can we relearn an RBMT system?

- Additional non-submitted system:
  - Relearnt system trained using two monolingual corpora (news domain)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size (sentences)</th>
<th>Size (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel FR-EN (Europarl v3)</td>
<td>0.94M</td>
<td>21M</td>
</tr>
<tr>
<td>Monolingual FR (Le Monde 1995)</td>
<td>0.96M</td>
<td>18M</td>
</tr>
<tr>
<td>Monolingual EN (NYT 1995)</td>
<td>3.8M</td>
<td>19M</td>
</tr>
</tbody>
</table>

Table 2: Corpus sizes for the additional model, trained on news domain
Can we relearn an RBMT system?

- Relearnt system trained using two monolingual corpora (news domain)

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU (tuning, nc-dev2007)</th>
<th>BLEU (test, nctest2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYSTRAN</td>
<td>n.a.</td>
<td>21.32</td>
</tr>
<tr>
<td>Relearnt</td>
<td>22.8</td>
<td>23.15</td>
</tr>
<tr>
<td>Baseline</td>
<td>22.7</td>
<td>22.19</td>
</tr>
<tr>
<td>Moses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Results of systems on News task
<table>
<thead>
<tr>
<th><strong>Can we relearn an RBMT system?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SOURCE</strong></td>
</tr>
<tr>
<td><strong>SYSTRAN</strong></td>
</tr>
<tr>
<td><strong>MOSES</strong></td>
</tr>
<tr>
<td><strong>RELEARNT</strong></td>
</tr>
<tr>
<td><strong>REF.</strong></td>
</tr>
</tbody>
</table>
Conclusion

➔ Goal of the experiment: set up a comparison between three different systems with equivalent resources
➔ Intermediate solution between a purely corpus-based statistical system and a rule-based system
➔ Future work: explore the effect of size of the monolingual corpus used for training the translation model
SYSTRAN Today

- Neural Machine Translation (NMT)
- Open-source community
- Drawback: Neural MT engines known to be computationally very expensive
- One-fits-all approach: impossible
- Goal: injecting existing language knowledge in the training process
References


➔ http://www.systransoft.com/
Pivot Language Approach for MT
Table of Contents

➔ Pivot language approach

➔ English — Minangkabau
  ◆ PHMT composite service
  ◆ Experiment
  ◆ Conclusion

➔ Chinese — Spanish
  ◆ Three different methods
  ◆ Using RBMT systems for pivot translation
  ◆ Experiments
  ◆ Conclusion
Pivot Language Approach

➔ Using a third language, for which there exist large source-pivot and pivot-target bilingual corpora

➔ Motivation:
  ◆ Large quantities of data (especially domain-specific) are not available for some language pairs
  ◆ Suitable for low-resource languages
PHMT Composite Service

Google Translate (SL — PL) + Bilingual Dictionary (PL — TL)
PHMT Composite Service

English → Indonesian

SMT: Google Translation (eng - ind)

RBMT: Word-to-Word Translation (ind - min)

Indonesian → Minangkabau
Experiment

➔ The video of English presentation
➔ Played to 165 bachelor students of Informatics, Islamic University of Riau, Indonesia
➔ Results:

<table>
<thead>
<tr>
<th></th>
<th>eng - ind</th>
<th>ind - min</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>3.52</td>
<td>3.05</td>
</tr>
<tr>
<td>adequacy</td>
<td>3.59</td>
<td>3.06</td>
</tr>
</tbody>
</table>

➔ Usability evaluation: mean of 3.71
Conclusion

➔ Fluency and adequacy scores are medium
➔ Still, useful to support multilingual communication
➔ The simplest word-to-word translation should be improved
➔ More entries should be added to the dictionary (only 5,391 at that moment)
WU, Hua; WANG, Haifeng. 
Revisiting pivot language approach for machine translation. 
In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1, August 02-07, 2009, Suntec, Singapore.
Pivot Language Approach: Methods

➔ Triangulation method
➔ Transfer method
➔ Synthetic method
Three Different Methods: Triangulation

- Phrase table multiplication
- \( = \) Multiplying translation probabilities and lexical weights in source-pivot and pivot-target translation models
Three Different Methods: Transfer

- First: translates from the source language to the pivot language using a source-pivot model
- Second: translates from the pivot language to the target language using a pivot-target model
Three Different Methods: Synthetic Method

Two possible methods to obtain a source-target corpora:

A. Obtain target translations for the source sentences in the source-pivot corpus (using pivot-target SMT)
B. Obtain source translations for the target sentences in the pivot-target corpus (using pivot-source SMT)

→ Combine two source-target corpora to a final synthetic corpus
Three Different Methods: Synthetic Method

A. Source Language → Pivot Language → Target Language

B. Source Language → Pivot Language → Target Language
Using RBMT Systems For Pivot Translation

1. Create a synthetic multilingual source-pivot-target corpus

2. Enlarge the size of bilingual training data by:
   ○ Translating monolingual corpora
   ○ Providing alternative translations for bilingual corpora

3. Translate the test set (through pivot into target) and add the translated set to the training data
Experiments
Data

- Performed experiments on spoken language translation
- Translated Chinese to Spanish using English as the pivot language
- Corpora:
  - BTEC (Basic Travel Expression Corpus) Chinese-English (CE): BTEC CE1 and BTEC CE2 (parallel to BTEC ES)
  - BTEC English-Spanish (ES)
  - The HIT olympic CE corpus
  - The Europarl ES corpus
Results: by Using SMT Systems

➔ Using BTEC CE1 and BTEC ES corpora

➔ Synthetic method:
  ◆ Translate the English part of the CE corpus into Spanish (with the ES translation model)
  ◆ Translate the English part of the ES corpus into Chinese (with the ES translation model)
Results: by Using SMT Systems

➔ Three methods achieved comparable quality
➔ The translation selection method is very effective (see “Combination”)
➔ (ASR = automatic speech recognition; CRR = correct recognition result)

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BLEU-Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangulation</td>
<td>33.70/27.46</td>
<td>31.59/25.02</td>
</tr>
<tr>
<td>Transfer</td>
<td>33.52/28.34</td>
<td>31.36/26.20</td>
</tr>
<tr>
<td>Synthetic</td>
<td>34.35/27.21</td>
<td>32.00/26.07</td>
</tr>
<tr>
<td>Combination</td>
<td>38.14/29.32</td>
<td>34.76/27.39</td>
</tr>
</tbody>
</table>

Table 3: CRR/ASR translation results by using SMT systems
Results: by Using Both RBMT and SMT Systems

1. With the RBMT system, the English sentences in the ES corpus are translated into Chinese
   a. **Transfer method:** CE translation quality improved $\Rightarrow$ Spanish translation quality improved as well
   b. **Triangulation method:** add the resulting Chinese-English corpus to the BTEC CE1 $\Rightarrow$ coverage of more phrase pairs
   c. **Synthetic method:** combine the resulting Chinese-Spanish corpus with those produced by the EC and ES SMT systems
Results: by Using Both RBMT and SMT Systems

- Translation quality was greatly improved!
- BLEU scores: at least 5.1 (CRR) and 3.9 (ASR) higher

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BLEU-Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangulation</td>
<td>40.69/31.02</td>
<td>37.99/29.15</td>
</tr>
<tr>
<td>Transfer</td>
<td>42.06/31.72</td>
<td>39.73/29.35</td>
</tr>
<tr>
<td>Synthetic</td>
<td>39.10/29.73</td>
<td>37.26/28.45</td>
</tr>
<tr>
<td>Combination</td>
<td>43.21/33.23</td>
<td>40.58/31.17</td>
</tr>
</tbody>
</table>
Results: by Using Both RBMT and SMT Systems

2. With the CE RBMT system, the size of training data was enlarged by providing alternative English translations for the Chinese part of the CE corpus

Translation quality was further improved!

<table>
<thead>
<tr>
<th>Method</th>
<th>EC RBMT</th>
<th>+ CE RBMT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>BLEU-Fix</td>
</tr>
<tr>
<td>Triangulation</td>
<td>40.69/31.02</td>
<td>37.99/29.15</td>
</tr>
<tr>
<td>Transfer</td>
<td>42.06/31.72</td>
<td>39.73/29.35</td>
</tr>
<tr>
<td>Synthetic</td>
<td>39.10/29.73</td>
<td>37.26/28.45</td>
</tr>
<tr>
<td>Combination</td>
<td>43.21/33.23</td>
<td>40.58/31.17</td>
</tr>
</tbody>
</table>
Results: by Using Both RBMT and SMT Systems

3. With the CE RBMT and ES RBMT systems, the test set was translated through English into Spanish

→ The translated test set was added to the training data
Results: by Using Both RBMT and SMT Systems

→ Translation quality was further improved!

→ **BLEU** scores: at least 2 (CRR) and 1.5 (ASR) higher

<table>
<thead>
<tr>
<th>Method</th>
<th>EC RBMT BLEU</th>
<th>EC RBMT BLEU-Fix</th>
<th>+ CE RBMT BLEU</th>
<th>+ CE RBMT BLEU-Fix</th>
<th>+ Test Set BLEU</th>
<th>+ Test Set BLEU-Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangulation</td>
<td>40.69/31.02</td>
<td>37.99/29.15</td>
<td>41.59/31.43</td>
<td>39.39/29.95</td>
<td>44.71/32.60</td>
<td>42.37/31.14</td>
</tr>
<tr>
<td>Transfer</td>
<td>42.06/31.72</td>
<td>39.73/29.35</td>
<td>43.40/33.05</td>
<td>40.73/30.06</td>
<td>45.91/34.52</td>
<td>42.86/31.92</td>
</tr>
<tr>
<td>Synthetic</td>
<td>39.10/29.73</td>
<td>37.26/28.45</td>
<td>39.90/30.00</td>
<td>37.90/28.66</td>
<td>41.16/31.30</td>
<td>37.99/29.36</td>
</tr>
<tr>
<td>Combination</td>
<td>43.21/33.23</td>
<td>40.58/31.17</td>
<td>45.09/34.10</td>
<td>42.88/31.73</td>
<td>47.06/35.62</td>
<td>44.94/32.99</td>
</tr>
</tbody>
</table>
Results: by Using Both RBMT and SMT Systems

All in all:

- The triangulation and the transfer methods greatly outperformed the synthetic method
- The synthetic method contributes little to the system combination
Results: by Using Monolingual Corpus

➔ With EC and ES RBMT systems, the English part of the HIT Olympics corpus was translated into Chinese and Spanish
➔ The resulting corpus was added to the training data
➔ Translation quality was further improved!

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>BLEU-Fix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangulation</td>
<td>45.64/33.15</td>
<td>42.11/31.11</td>
</tr>
<tr>
<td>Transfer</td>
<td>47.18/34.56</td>
<td>43.61/32.17</td>
</tr>
<tr>
<td>Combination</td>
<td>48.42/36.42</td>
<td>45.42/33.52</td>
</tr>
</tbody>
</table>

Table 5: CRR/ASR translation results by using additional monolingual corpora
Conclusion

→ The triangulation method and the transfer method generally outperform the synthetic method
→ The hybrid method combining RBMT and SMT systems can be used to fill up the data gap between the source-pivot and pivot-target corpora

Additionally:

→ Even if the translation quality of the RBMT system is low, it still greatly improved the translation quality
→ The developed system outperforms the best system for the pivot task in the IWSLT 2008 evaluation campaign
Thanks for your attention!