Automatic Evaluation of Translation Quality for Distant Language Pairs

Stanislav Reichert
University of Tübingen

14th June 2018
Outline

Introduction
  NTCIR Workshops
  NTCIR-10 Patent MT Task
  Japanese-English Translation Results
  English-Japanese Translation Results

From BLEU to RIBES
  Identifying BLEU’s weaknesses

The RIBES metric
  Introducing RIBES
  Refining RIBES
  Evaluating RIBES

Reordering Strategies
  English vs. Japanese
  Pre-ordering models
  Post-ordering models

References
Good translations: I know one when I see one

Figure 1: From Koehn (2009, p. 218)

这个 机场 的 安全 工作 由 以色列 方面 负责。

Israeli officials are responsible for airport security. Israel is in charge of the security at this airport. The security work for this airport is the responsibility of the Israel government. Israeli side was in charge of the security of this airport. Israel is responsible for the airport’s security. Israel is responsible for safety work at this airport. Israel presides over the security of the airport. Israel took charge of the airport security. The safety of this airport is taken charge of by Israel. This airport’s security is the responsibility of the Israeli security officials.
Introduction

NTCIR Workshops
NTCIR-10 Patent MT Task
Japanese-English Translation Results
English-Japanese Translation Results

From BLEU to RIBES
Identifying BLEU’s weaknesses
The RIBES metric
Introducing RIBES
Refining RIBES
Evaluating RIBES

Reordering Strategies
English vs. Japanese
Pre-ordering models
Post-ordering models

References

Good translations: I know one when I see one

Figure 2: From Goto, Chow, et al. (2013)
Reliable automatic evaluation of MT is a desideratum

The NTCIR Workshop is a series of evaluation workshops designed to enhance research in information access technologies (including languages like Japanese and Chinese)

Patents are a challenging domain for MT due to the length and complexity of the patent sentences (JP-EN & EN-JP)
NTCIR-10 Patent MT Task

- NTCIR-8 and NTCIR-9 PatentMT Test Collection
  Japanese-to-English Machine Translation Data package

- Training data for JP-EN and EN-JP subtasks:
  - Approximately 3.2 million patent parallel sentence pairs
  - Monolingual patent corpus in Japanese spanning 13 years

---

## Japanese-English Translation Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Type</th>
<th>Resource</th>
<th>Average score</th>
</tr>
</thead>
<tbody>
<tr>
<td>JAPIO-1</td>
<td>RBMT</td>
<td></td>
<td>3.67</td>
</tr>
<tr>
<td>RBMT1-1</td>
<td>RBMT</td>
<td></td>
<td>3.57</td>
</tr>
<tr>
<td>EIWA-1</td>
<td>HYBRID</td>
<td></td>
<td>3.53</td>
</tr>
<tr>
<td>TORI-1</td>
<td>HYBRID</td>
<td></td>
<td>3.48</td>
</tr>
<tr>
<td>NTITI-1</td>
<td>SMT</td>
<td>✓</td>
<td>3.32</td>
</tr>
<tr>
<td>NTITI-3</td>
<td>SMT</td>
<td>✓</td>
<td>3.26</td>
</tr>
<tr>
<td>RWTH-1</td>
<td>SMT</td>
<td>✓</td>
<td>3.07</td>
</tr>
<tr>
<td>HDU-1</td>
<td>SMT</td>
<td></td>
<td>3.01</td>
</tr>
<tr>
<td>ONLINE1-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.94</td>
</tr>
<tr>
<td>FUN-NRC-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.89</td>
</tr>
<tr>
<td>NTITI-2</td>
<td>SMT</td>
<td>✓</td>
<td>2.87</td>
</tr>
<tr>
<td>BASELINE1-1</td>
<td>SMT</td>
<td></td>
<td>2.81</td>
</tr>
<tr>
<td>KYOTO-1</td>
<td>EBMT</td>
<td></td>
<td>2.74</td>
</tr>
<tr>
<td>BASELINE2-1</td>
<td>SMT</td>
<td></td>
<td>2.68</td>
</tr>
<tr>
<td>OKAPU-1</td>
<td>SMT</td>
<td></td>
<td>2.61</td>
</tr>
<tr>
<td>TRGTK-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.55</td>
</tr>
<tr>
<td>BJTUX-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.25</td>
</tr>
<tr>
<td>ISTIC-1</td>
<td>SMT</td>
<td>✓</td>
<td>1.08</td>
</tr>
</tbody>
</table>

---

**Figure 3:** From Goto, Chow, et al. (2013)
Japanese-English Translation Results

- 4 top systems are RBMT or make use of commercial RBMT systems
- NTITI systems → post-ordering method and pre-ordering methods
- RWTH-1 → phrase-based SMT with a hierarchical phrase reordering model
- The source sentence meaning could be understood (awarded C rank or better) for
  - 55% of the sentences for the best RBMT (JAPIO-1)
  - 38% of the sentences for the best SMT (NTITI-1)
## English-Japanese Translation Results

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Type</th>
<th>Resource</th>
<th>Average score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTITI-2</td>
<td>SMT</td>
<td>✓</td>
<td>3.84</td>
</tr>
<tr>
<td>NTITI-1</td>
<td>SMT</td>
<td>✓</td>
<td>3.81</td>
</tr>
<tr>
<td>JAPIO-1</td>
<td>RBMT</td>
<td>✓</td>
<td>3.53</td>
</tr>
<tr>
<td>RBMT6-1</td>
<td>RBMT</td>
<td>✓</td>
<td>3.47</td>
</tr>
<tr>
<td>EIWA-1</td>
<td>HYBRID</td>
<td>✓</td>
<td>3.42</td>
</tr>
<tr>
<td>ONLINE1-1</td>
<td>SMT</td>
<td>✓</td>
<td>3.38</td>
</tr>
<tr>
<td>BJTUX-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.84</td>
</tr>
<tr>
<td>TSUKU-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.79</td>
</tr>
<tr>
<td>BASELINE1-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.69</td>
</tr>
<tr>
<td>FUN-NRC-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.67</td>
</tr>
<tr>
<td>BASELINE2-1</td>
<td>SMT</td>
<td>✓</td>
<td>2.53</td>
</tr>
<tr>
<td>KYOTO-1</td>
<td>EBMT</td>
<td>✓</td>
<td>2.50</td>
</tr>
<tr>
<td>TRGTK-1</td>
<td>SMT</td>
<td>✓ ✓</td>
<td>2.45</td>
</tr>
<tr>
<td>ISTIC-1</td>
<td>SMT</td>
<td>✓ ✓</td>
<td>2.30</td>
</tr>
</tbody>
</table>

**Figure 4:** From (Goto, Chow, et al., 2013)
EN-JP is a more difficult case for SMT systems

Yet at the NTCIR-9, NTITI-2 (pre-ordering) was significantly better than the best-ranked RBMT system

The source sentence meaning could be understood (awarded C rank or better) for
  ▶ 70% of the translated sentences (vs. 58% for RBMT)
BLEU singing the blues?

- N-gram evaluation metric with modified n-gram precision to deal with variation in phrase order
- Demonstrably correlates with human judgments in many cases
- Can be used for tracking broad, incremental changes to a single system
- However: NIST Machine Translation Evaluation exercise 2005 revealed a major mismatch between the BLEU evaluation and human judgments
BLEU singing the blues?

- Changes in translation quality are reported in opaque BLEU scores, yet actual examples are rarely provided.
- Echizen-ya et al. (2009) show that the popular BLEU and NIST metrics do not work well. Alternatives: IMPACT.
- The global word order is essential for an SMT system working with distant language pairs:
  - Common error: A because of B ⇒ B because of A
  - BLEU disregards global word order.
Disassembling BLEU scores

“Millions of variations on a hypothesis translation [...] receive the same Bleu score” (Callison-Burch, Osborne, & Koehn, 2006, p. 1)

Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

- Appeared calm | when | he was | taken | to the American plane | , | which will | to Miami, Florida .
- which will | he was | , | when | taken | Appeared calm | to the American plane | to Miami , Florida .
- was being led to the | calm as he was | would take | carry him | seemed quite | when | taken
Disassembling BLEU scores

- Equal weighting of all items in the reference sentences
- Words in the hypothesis sentence that did not appear in the reference sentences can be substituted with anything
- BLEU may not be appropriate for comparing systems which employ different translation strategies
The RIBES metric

- Adequacy is more important for distant language pairs (Isozaki, Hirao, et al., 2010)
- NIST, PER, and TER turn out to be inadequate measures
- Solution: rank correlation coefficients compare word ranks in the reference with those in the hypothesis sentences
- RIBES: Rank-based Intuitive Bilingual Evaluation Score
- Defined for each test sentence and averaged score is used for evaluating the entire test corpus (Isozaki et al., 2014)
- Significantly penalizes global word order mistakes
Refining RIBES

- Rank correlation metrics are defined only when there is one-to-one correspondence
- In many cases, Spearman’s $\rho$ and Kendall’s $\tau$ cannot be determined
- Solution:
  - find one-to-one corresponding higher-order n-grams
  - $\rho$ and $\tau$ seem to be better metrics than BLEU
Refining RIBES

- $\rho$ and $\tau$ are normalized (NSR and NKT)
- Precision ($P$), recall, and F-measure to reduce the overestimation of NSR and NKT
- Precision correlates best with adequacy among these three metrics
- New metrics for ($0 \leq \alpha \leq 1$):
  - $\text{NSR}P^{\alpha}$
  - $\text{NKT}P^{\alpha}$
Evaluating RIBES: Results

- NTCIR-7 PATMT JP-EN translation data

<table>
<thead>
<tr>
<th>human judge \ meta-eval</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spm</td>
<td>Prs</td>
</tr>
<tr>
<td>eval\ meta-eval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P$</td>
<td>0.615</td>
<td>0.704</td>
</tr>
<tr>
<td>$R$</td>
<td>0.436</td>
<td>0.669</td>
</tr>
<tr>
<td>$F_\beta=1$</td>
<td>0.525</td>
<td>0.692</td>
</tr>
<tr>
<td>BP</td>
<td>0.000</td>
<td>0.515</td>
</tr>
<tr>
<td>NSR</td>
<td>0.904</td>
<td>0.906</td>
</tr>
<tr>
<td>NSRP$^{1/8}$</td>
<td>0.937</td>
<td>0.905</td>
</tr>
<tr>
<td>NSRP$^{1/4}$</td>
<td><strong>0.947</strong></td>
<td>0.900</td>
</tr>
<tr>
<td>NSRP$^{1/2}$</td>
<td>0.937</td>
<td>0.890</td>
</tr>
<tr>
<td>NSRP$^{1/1}$</td>
<td>0.883</td>
<td>0.872</td>
</tr>
<tr>
<td>NSR $\times$ BP</td>
<td>0.851</td>
<td>0.874</td>
</tr>
<tr>
<td>NKT</td>
<td>0.940</td>
<td>0.922</td>
</tr>
<tr>
<td>NKT$^{1/8}$</td>
<td>0.940</td>
<td>0.913</td>
</tr>
<tr>
<td>NKT$^{1/4}$</td>
<td>0.940</td>
<td>0.904</td>
</tr>
<tr>
<td>NKT$^{1/2}$</td>
<td>0.929</td>
<td>0.890</td>
</tr>
<tr>
<td>NKT$^{1/1}$</td>
<td>0.897</td>
<td>0.869</td>
</tr>
<tr>
<td>NKT $\times$ BP</td>
<td>0.829</td>
<td>0.878</td>
</tr>
</tbody>
</table>

Figure 5: From Isozaki, Hirao, et al. (2010)
Results for new metrics are comparable to conventional methods even for similar language pairs (European)
  ▶ Tested on WMT-07 data (Callison-Burch et al., 2007)
  ▶ With the precision-based penalty, both $\rho$ and $\tau$ outperform conventional methods

With the precision-based penalty, both $\rho$ and $\tau$ outperform conventional methods
English vs. Japanese

- English is a head-initial language (SVO)
- Japanese is a head-final language (SOV). It is also a topic-prominent language
- Scrambling of syntactic chunks called bunsetsu is a frequent phenomenon
- Deterministic rules fail at some sentence constructions
- RIBES performs worse on scrambled sentences
Difficult sentences

(1) Kare wa kōhi wo noma-nai
He TOP coffee OBJ drink-NEG
He doesn’t drink coffee

(2) Kōhi wa noma-nai
Coffee TOP drink-NEG
I don’t drink coffee

(3) Watashi wa jikan ga nai
I TOP time SUB NEG
I don’t have time

(4) Ima wa jikan ga nai
Now TOP time SUB NEG
Now I don’t have time
Reordering Strategies

- Conducting target word selection and reordering jointly: phrase-based SMT, hierarchical phrase-based SMT, syntax-based SMT
- Pre-ordering: reorder the source language sentence into a target language word order. Then, translate the reordered source word sequence using SMT methods
- Post-ordering: reverse of the pre-ordering
Single reordering rule: Head Final English

- Head Finalization: move syntactic heads to the end of the corresponding syntactic constituents (Isozaki, Sudoh, et al., 2010)
  - A HPSG parser ENJU outputs syntactic heads
  - Particle seeds can be generated from the output
- Difficulties:
  - The sentence pair is not an exact one-to-one translation
  - Mistakes in ENJU’s tagging or parsing (esp. VBN/VBD and VBZ/NNS mistakes) or in the GIZA++ automatic word alignment
Rule-based pre-ordering model for JP-EN SMT

- Hoshino et al. (2013) put forth four rules:
  - Modify Japanese dependency trees: every head chunk comes first and its dependent children follow
  - Convert SOV into SVO: place V after the S or the O, if no S. Move V before the rightmost chunk if there is only V
  - Keep the word order of coordinate clauses unchanged and put them into the leftmost slot (punctuation comes into the rightmost slot)
  - For every chunk, swap function and content words → pseudo-prepositional phrases
Rule-based pre-ordering model for JP-EN SMT

Japanese source sentence with predicate-argument analysis:
(dependency and labels)

図2において ガイドバー11と 22の 支持構造も 示す。
In Fig 2 guide bar 11 and 22 for support structure also show.

Rule 1-1 pseudo head-initialization:

図2において 示す。 図2において 支持構造も 22の ガイドバー11と
In Fig 2 show. In Fig 2 support structure also 22 for guide bar 11 and

Rule 1-2 inter-chunk pre-ordering:

図2において 示す。 支持構造も 22の ガイドバー11と
In Fig 2 show. support structure also 22 for guide bar 11 and

Rule 1-3 inter-chunk normalization:

図2において 示す 支持構造も ガイドバー11と 22の。
In Fig 2 show support structure also guide bar 11 and 22 for.

Rule 2 intra-chunk pre-ordering:

において 図2 示す も 支持構造 と ガイドバー11 の 22。
In Fig 2 show also support structure and guide bar 11 for 22.

References

(Komachi et al., 2006):

図2において ガイドバー11と 22の 支持構造も 示す。
In Fig 2 guide bar 11 and 22 for support structures also show.

(Katz-Brown and Collins, 2008):

示す において 図2 も 支持構造 の 22 と 11 ガイドバー。
show In Fig 2 also support structure for 22 and guide bar 11.

Figure 6: From (Hoshino et al., 2013)
Post-ordering framework for JP-EN translation

- Reordering model makes use of syntactic structures (relying on ENJU’s output) (Goto, Utiyama, & Sumita, 2013)
- The model is based on:
  - parsing using probabilistic context free grammar (PCFG)
  - inversion transduction grammar (ITG)
Bibliography I


Bibliography III

