A Bootstrapped Interlingua-Based SMT Architecture

Manny Rayner, Paula Estrella
Pierrette Bouillon
Outline

- Goals of paper
  - Background
  - Bootstrapping an interlingua-based SMT
  - Experiments
Goals of Paper

- « Relearning Rule-Based MT systems »
  - Usual goal: add robustness
  - E.g. Dugast et al 2008 with SYSTRAN
- Can we do it with a small-vocabulary high-precision system?
  - Our GEAF 2009 paper: it’s not so easy
- Can we do better if we use interlingua in the right way?
Use rule-based MT system to generate training data
Train statistical MT system
Naive approach

(GEAF 2009 paper)

- Naive approach is unimpressive
- If bootstrapped SMT translation different from RBMT translation, usually wrong
- Very poor for English → Japanese
  - Better for English → French
- Tops out quickly, then no improvement
« Relearning Interlingua-Based Machine Translation »

Source representation → RBMT → Interlingua representation → RBMT → Target representation

↑ parsing

Source text

Target text

↑ generation
Relearning Interlingua-Based Machine Translation

Source representation \(\xrightarrow{RBMT}\) Interlingua representation \(\xrightarrow{RBMT}\) Target representation

\[\text{Source text} \xrightarrow{parsing} \text{Source representation} \xrightarrow{RBMT} \text{Interlingua representation} \xrightarrow{RBMT} \text{Target representation} \xrightarrow{\text{generation}} \text{Target text}\]

Source text \(\xrightarrow{SMT}\) ??? \(\xrightarrow{SMT}\) Target text
« Relearning Interlingua-Based Machine Translation »

- Source representation ➔ RBMT ➔ Interlingua representation ➔ RBMT ➔ Target representation
- Source text ➔ parsing ➔ Interlingua text ➔ generation ➔ Target text
- Source text ➔ SMT ➔ Interlingua text ➔ SMT ➔ Target text
Key Questions

- What is «interlingua text»?
- How can we use it to relearn an interlingua-based system as an SMT?
- How well does it work in practice?
Outline

- Goals of paper
  - Background
- Bootstrapping an interlingua-based SMT
- Experiments
MedSLT

- Unidirectional doctor $\rightarrow$ patient spoken translation
- Controlled language, grammar-based
  - Implemented using Regulus platform
- Multi-lingual, interlingua-centred
  - Current prototype: 6 languages, any-to-any
  - English, French, Japanese, Arabic, Catalan, Swedish
- System checks correctness by backtranslating
English MedSLT examples

Where is the pain?
Is the pain in the front of the head?
Do you often get headaches in the morning?
Does bright light give you headaches?
Do you have headaches several times a day?
Does the pain last more than an hour?
Backtranslation

- **Source:** Do you have headaches at night?
- **B/trans:** Do you experience the headaches at night?
- **Target:** Vos maux de tête surviennent-ils la nuit?
- **Target:** Yoru atama wa itamimasu ka?
Think of interlingua as a language
- Define using formal grammar
- Associate text form with representation
- Text form is simplified/telegraphic English

Functions of interlingua grammar
- Allows us to induce an SMT
- Constrains semantic content of input language
- Surface form useful in development/debugging
English sentence

“Does the pain spread to the jaw?”

Interlingua representation

[null=[utterance_type, ynq],
arg1=[symptom, pain],
null=[state, radiate],
null=[tense, present],
to_loc=[body_part, jaw]]

Interlingua Text

“YN-QUESTION pain radiate PRESENT jaw”
Different Forms of Interlingua Gloss

- Current gloss is simplified English
  - Word-order is English-like
- Can have simplified forms of other languages too
  - In particular, Japanese
<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EN</strong></td>
<td>does the pain last for more than one day</td>
</tr>
<tr>
<td><strong>IN/E</strong></td>
<td>YN-QUESTION pain last PRESENT duration more-than one day</td>
</tr>
<tr>
<td><strong>JP</strong></td>
<td>ichinichi sukunakutomo itami wa tsuzukimasu ka</td>
</tr>
<tr>
<td><strong>IN/J</strong></td>
<td>more-than one day duration pain last PRESENT YN-QUESTION</td>
</tr>
</tbody>
</table>
Outline

- Goals of paper
- Background
  - Bootstrapping an interlingua-based SMT
- Experiments
Bootstrapping an interlingua-based SMT

- Randomly generate 1M sentences source data
- Translate using EN-FR and EN-JP RBMT
- Save interlingua in text form
  - Both English (IN/E) and Japanese (IN/J) forms
- Train SMT models using Moses etc
Ways to exploit interlingua text

- Rescoring
  - Do Source \(\rightarrow\) Interlingua in N-best mode
  - Prefer well-formed interlingua text

- Reformulation
  - Split up EN-JP as EN-IN/E + IN/J-JP
  - Use interlingua grammar to do IN/E-IN/J
  - SMT translation only between languages with similar word-orders
Processing pipelines

- (Plain RBMT)
  - Source text \[\rightarrow\] RBMT
  - Target text

- (Plain SMT)
  - Source text \[\rightarrow\] SMT
  - Target text
Processing pipelines

- **SMT + SMT**

  - Source text → SMT → Interlingua text → SMT → Target text

- **SMT + RBMT**

  - Source text → SMT → Interlingua text → RBMT → Target text
Processing pipelines

- SMT + rescoring + SMT

Source text (EN) \(\rightarrow\) SMT \(\rightarrow\) Int. Text (N-best) \(\rightarrow\) Rescore \(\rightarrow\) Int. Text (single) \(\rightarrow\) SMT \(\rightarrow\) Target text (JP)

- SMT + interlingua-reformulation + SMT

Source text \(\rightarrow\) SMT \(\rightarrow\) Int. Text (IN/E) \(\rightarrow\) Reform \(\rightarrow\) Int. Text (IN/J) \(\rightarrow\) SMT \(\rightarrow\) Target text (JP)
Processing pipelines

- Other combinations
  - SMT + rescoring + int-reformulation + SMT
  - SMT + rescoring + RBMT
Outline

- Goals of paper
- Background
- Bootstrapping an interlingua-based SMT
  - Experiments
Experiments

- Evaluate relative performance of different processing pipelines
- Evaluate on held-out part of generated data
  - Measure agreement with RBMT translation
  - GEAF 2009 paper: when SMT and RBMT different, SMT often worse and hardly ever better
- Evaluate best pipelines on real out-of-coverage data
  - Use human judges
# Results on generated data

(Metric: agreement with original RBMT system)

<table>
<thead>
<tr>
<th>Configuration</th>
<th>EN $\rightarrow$ FR</th>
<th>EN $\rightarrow$ JP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain RBMT</td>
<td>(100%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>Plain SMT</td>
<td>65.8%</td>
<td>26.8%</td>
</tr>
<tr>
<td>SMT + SMT</td>
<td>76.6%</td>
<td>10.5%</td>
</tr>
<tr>
<td>SMT + int-reformulation + SMT</td>
<td>---</td>
<td>74.1%</td>
</tr>
<tr>
<td>SMT + int-rescoring + SMT</td>
<td>78.5%</td>
<td>10.8%</td>
</tr>
<tr>
<td>SMT + int-rescore + int-reform + SMT</td>
<td>---</td>
<td>78.5%</td>
</tr>
<tr>
<td>SMT + RBMT</td>
<td>83.5%</td>
<td>81.9%</td>
</tr>
<tr>
<td>SMT + int-rescoring + RBMT</td>
<td>87.0%</td>
<td>87.1%</td>
</tr>
</tbody>
</table>
Results on real data (EN-FR)

(Use best versions: SMT + rescoring + SMT/RBMT)

358 out-of-coverage utterances
245 well-formed interlingua
81 good backtranslation
75/81 SMT + RBMT translations
75/75 good SMT + RBMT translations
81/81 SMT + SMT translations
76/81 good SMT + SMT translations
Results on real data (EN-JP)

(Use best versions: SMT + rescore + reform + SMT/RBMT)

358 out-of-coverage utterances
245 well-formed interlingua
81 good backtranslation
81/81 SMT + RBMT translations
77/81 good SMT + RBMT translations
81/81 SMT + SMT translations
71/81 good SMT + SMT translations
Summary

- Goal: relearn small RBMT system as SMT
- Not trivial if high precision required
- Much better results if we use interlingua
- Key idea: text form of interlingua
  - Use interlingua to reorder SMT output
  - Use interlingua to handle word-order problems
- Good results on EN-FR and EN-JP