FUN-NRC: Paraphrase-Augmented Phrase-Based SMT Systems for NTCIR-10 PatentMT

Atsushi Fujita
Future University Hakodate
http://paraphrasing.org/~fujita/

Marine Carpuat
National Research Council
http://marinecarpuat.weebly.com/
Summary of our systems

- Phrase-based SMT + paraphrases
  - State-of-the-art non-hierarchical system: PortageII @ NRC
    - Almost no language- or domain-specific knowledge
  - Phrase table augmentation
    - Paraphrases in both source & target languages (separately)
    - Comparison of paraphrase collections
    - Aggregation of multiple paths w/ feature engineering
  - Improved performance over a vanilla phrase-based SMT
    - at least BLEU, NIST, and RIBES
Motivation & proposed method
Modern SMT systems: Limitations

Principle

- At source side
  - Unseen expressions will never be translated
  - They are either dropped or retained as is

- At target side
  - Only seen expressions can be generated as hypotheses
  - cf. Language models only ranks the given hypotheses
Expressions that convey the same meaning

Paraphrase: monolingual

Emma burst into tears and he tried to comfort her.

Emma cried, and he tried to console her.

Translation: cross-lingual

Désirez-vous obtenir des conseils pratiques sur le déménagement?

Are you looking for some helpful tips for moving?
Paraphrases

Linguistic expressions in the same language that convey the same meaning

- **Word / word sequence**
  - burst into tears
  - resemble
  - cried
  - look like

- **Clause (simple sentence)**
  - The car collided with the bicycle
  - The car and the bicycle collided

- **Beyond single clause**
  - It was his best suit that John wore to the dance last night.
  - John wore his best suit to the dance last night.
Prior arts in integrating paraphrases to MT

Rewrite input

Augment input
[Onishi+, 10]
[Du+, 10][Jiang+, 11]

Augment translation table
[Callison-Burch, 06]
[Marton+, 09]

Expand training corpus
[Bond+, 08][Nakov+, 11]

Input

[Shirai+, 93][Doi+, 04]
[Xu+, 05]
[Nanjo+, 12]

Expand tuning data
[Madnani+, 07]

Statistical models

Bilingual corpus

Post-edit

Output
Augmentation of translation table

- **Updates from [Callison-burch, 06][Marton+, 09]**
  - Comparison of several paraphrase collections
  - Aggregation of multiple paths (both sides)
    - **Source side (Saug):** translate more phrases
    - **Target side (Taug):** generate more hypotheses
  - Feature engineering for decoding
Key issue: how to realize paraphrases?

- Large-scale knowledge-base is indispensable
  - Handcrafting
  - Automatic paraphrase acquisition (PA)

- Pros. & cons. of prior arts
  - PA from Monolingual non-parallel corpora
    - **Pro.** Large $\rightarrow$ (potentially) high recall
    - **Con.** Only weak evidences $\rightarrow$ low precision
  - PA from Mono/Bi/Multi-lingual parallel corpora
    - **Pro.** Sentence-level equivalence $\rightarrow$ high precision
    - **Con.** Limited availability $\rightarrow$ low recall
**PA from monolingual non-parallel corpora**

- **Distributional Hypothesis [Harris, 68]**
  - Expressions that appear frequently in similar contexts have similar meanings
  - e.g., “Tezgüno” [Pantel+, 02]

- A bottle of **tezgüno** is on the table
- Everyone likes **tezgüno**
- **Tezgüno** makes you drunk
- We make **tezgüno** out of corn

- Similar to wine, cognac, whiskey → alcoholic beverage
- **Con.** Not necessarily equivalent: e.g., antonyms, hypernyms

- resembles
- looks like
- tezgüno
- wine
- decrease
- increase
- liquor
- wine
PA from bilingual parallel corpora

Translations as pivot [Bannard+, 05]
- A more reliable evidence than context
- Obtainable from bilingual parallel corpora
  - i.e., word alignment + phrase extraction

Automatically learned translation table

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>health issue</td>
<td>problème de santé</td>
</tr>
<tr>
<td>health problem</td>
<td>problème de santé</td>
</tr>
<tr>
<td>regional issue</td>
<td>problème régional</td>
</tr>
<tr>
<td>regional problem</td>
<td>problème régional</td>
</tr>
</tbody>
</table>

- Polysemy would generate non-paraphrases
- Con. Parallel corpora << monolingual non-parallel corpora
Paraphrase collections examined

$P_{Seed}$, $P_{Hvst}$, and $P_{OOPH}$

- "health issue" $\Rightarrow$ “problème de santé”
- "health problem" $\Rightarrow$ “problème de santé”
- "look like" $\Rightarrow$ “ressemble”
- "regional issue" $\Rightarrow$ “problème régional”
- "regional problem" $\Rightarrow$ “problème régional”
- "resemble" $\Rightarrow$ “ressemble”

- "health issue" $\Leftrightarrow$ “health problem”
- "look like" $\Leftrightarrow$ “ressemble”
- "regional issue" $\Leftrightarrow$ “regional problem”

- “issue” $\Leftrightarrow$ “problem”; {health, regional, ...}

- "backlog issue" $\Leftrightarrow$ “backlog problem”
- "communal issue" $\Leftrightarrow$ “communal problem”
- "phishing issue" $\Leftrightarrow$ “phishing problem”
- "spatial issue" $\Leftrightarrow$ “spatial problem”

[Fujita+, 12]
Paraphrase collections examined

$P_{\text{Seed}}, P_{\text{Hvst}}, \text{and } P_{\text{OOPH}}$

- **Translation Table**
  - "health issue" ⇒ "problème de santé"
  - "health problem" ⇒ "problème de santé"
  - "look like" ⇒ "ressemble"
  - "regional issue" ⇒ "problème régional"
  - "regional problem" ⇒ "problème régional"
  - "resemble" ⇒ "ressemble"

- **$P_{\text{Seed}}$: Seed Paraphrases**
  - "health issue" ⇔ "health problem"
  - "look like" ⇔ "ressemble"
  - "regional issue" ⇔ "regional problem"

- **$P_{\text{Hvst}}$: Novel Paraphrases**
  - "backlog issue" ⇔ "backlog problem"
  - "communal issue" ⇔ "communal problem"
  - "phishing issue" ⇔ "phishing problem"
  - "spatial issue" ⇔ "spatial problem"

- **$P_{\text{OOPH}}$: unseen phrases ⇒ seen phrases**

---

[Fujita+, 12]
Aggregation of multiple paths (1/2)

- **Source-side augmentation**

  - **Zooming operation**
    - 0.735
  - **Zoom action**
    - 0.709
  - **Zooming**
    - 0.352
  - **Source-side augmentation**
  - **Translation scores**
    - **Forward**
      \[
      p(t | s') = \frac{\sum_{s \in S} \left( p(t | s) \text{Para}(s' \Rightarrow s) \right)}{\sum_{s \in S} \text{Para}(s' \Rightarrow s)}
      \]
    - **Backward**
      \[
      p(s' | t) = \frac{\sum_{s \in S} \left( p(s | t) \text{Para}(s \Rightarrow s') \right)}{\sum_{s \in S} \text{Para}(s \Rightarrow s')}
      \]
Aggregation of multiple paths (2/2)

Target-side augmentation

Translation scores

- **Forward**
  \[
  p(t'|s) = \frac{\sum_{t \in T} \left( p(t|s) \cdot \text{Para}(t \Rightarrow t') \right)}{\sum_{t \in T} \text{Para}(t \Rightarrow t')}
  \]

- **Backward**
  \[
  p(s|t') = \frac{\sum_{t \in T} \left( p(s|t) \cdot \text{Para}(t' \Rightarrow t) \right)}{\sum_{t \in T} \text{Para}(t' \Rightarrow t)}
  \]
## Paraphrase-related Features

<table>
<thead>
<tr>
<th>Features in the translation model</th>
<th>Original</th>
<th>Source-side fabricated</th>
<th>Target-side fabricated</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a1) Forward translation score</td>
<td>Cond.Prob.</td>
<td>[0,1]</td>
<td>[0,1]</td>
</tr>
<tr>
<td>(a2) Backward translation score</td>
<td>Cond.Prob.</td>
<td>[0,1]</td>
<td>[0,1]</td>
</tr>
<tr>
<td>(b1) Obtained from IBM2 alignment</td>
<td>True/False</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>(b2) Obtained from HMM alignment</td>
<td>True/False</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>(b3) Obtained from IBM4 alignment</td>
<td>True/False</td>
<td>False</td>
<td>False</td>
</tr>
<tr>
<td>(c1) Fabricated using Seed</td>
<td>False</td>
<td>True/False</td>
<td>True/False</td>
</tr>
<tr>
<td>(c2) Fabricated using Hvst/OOPH</td>
<td>False</td>
<td>True/False</td>
<td>True/False</td>
</tr>
<tr>
<td>(d1) Unseen in the phrase table</td>
<td>False</td>
<td>True/False</td>
<td>True/False</td>
</tr>
<tr>
<td>(d2) Unseen in the bilingual data</td>
<td>False</td>
<td>True/False</td>
<td>True/False</td>
</tr>
<tr>
<td>(e1) Paraphrase score (Saug/fwd)</td>
<td>1</td>
<td>[0,1]</td>
<td>1</td>
</tr>
<tr>
<td>(e2) Paraphrase score (Saug/bwd)</td>
<td>1</td>
<td>[0,1]</td>
<td>1</td>
</tr>
<tr>
<td>(e3) Paraphrase score (Taug/fwd)</td>
<td>1</td>
<td>1</td>
<td>[0,1]</td>
</tr>
<tr>
<td>(e4) Paraphrase score (Taug/bwd)</td>
<td>1</td>
<td>1</td>
<td>[0,1]</td>
</tr>
</tbody>
</table>
Score of each paraphrase pair (1/2)

**PivProb**: Pivot-based paraphrase probability [Bannard+, 05]

- For $P_{\text{Seed}}$ only

| $s$       | $t$       | $p(s|t)$ | $p(t|s)$ |
|-----------|-----------|----------|----------|
| look like | resemble  | 0.0177   | 0.0061   |
| resemble  | ressemble | 0.0074   | 0.0181   |

- Asymmetric score

$$Para(s_1 \Rightarrow s_2) = p(s_2|s_1) = \sum_{t \in tr(s_1) \cap tr(s_2)} p(s_2|t)p(t|s_1)$$
Score of each paraphrase pair (2/2)

**CosSim**: cosine similarity of “contexts”

- For all of $P_{Seed}$, $P_{Hvst}$, and $P_{OOPH}$
- Contextual similarity in a monolingual corpus
- Adjacent 1- to 4-grams of each token $\rightarrow$ feature vector
  - cf. cheap but noisy features, e.g., bag-of-words
  - cf. accurate but expensive features, e.g., dependency trees

There have been many approaches to compute the similarity between words based on their distribution in a corpus.
Dev & Test
Our base system

- **Portagell 1.0** [National Research Council, 12]
  - A state-of-the-art phrase-based SMT system
    - Reasonably good results at NIST OpenMT 2012 [Foster, 12]
  - Advanced features (cf. Moses)
    - Kneser-Ney translation probability smoothing [Chen+, 11]
    - Hierarchical lexicalized reordering [Cherry+, 12]
    - Lattice-batch-MIRA optimization [Cherry & Foster, 12]
    - etc.
  - User-friendly features
    - Highly tuned libraries for using gigantic models [Germann+, 09]
    - High stability (cf. GIZA++)
    - Fits well to cluster computing environment
Training component models

- Provided data
  - Training bi-text
    - 3.2M sentence pairs
  - Monolingual text
    - Ja: 594M sentences (27.3B words)
    - En: 413M sentences (13.4B words)
  - Data for tuning
    - 2000 sentence pairs

- Component models
  - Language models
    - SRILM-5g
  - Translation models
    - IBM2
    - HMM
    - IBM4
  - Reordering models
    - Lexical model
    - Hierarchical lexical model
  - Paraphrase tables

Parameter tuning
# of learned phrasal equivalent pairs

<table>
<thead>
<tr>
<th># of trans. pairs</th>
<th>Ja → En</th>
<th>En → Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM2</td>
<td>9.1M</td>
<td>9.4M</td>
</tr>
<tr>
<td>HMM</td>
<td>230.6M</td>
<td>234.4M</td>
</tr>
<tr>
<td>IBM4</td>
<td>80.6M</td>
<td>81.8M</td>
</tr>
<tr>
<td>Union</td>
<td>260.4M</td>
<td>264.8M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># of paraphrase pairs</th>
<th>$th_p$</th>
<th>$th_s$</th>
<th>En</th>
<th>Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td>P$_{Seed}$</td>
<td>0</td>
<td>0</td>
<td>7.2M</td>
<td>5.1M</td>
</tr>
<tr>
<td>P$_{Seed}$</td>
<td>0.01</td>
<td>0.1</td>
<td>1.1M</td>
<td>0.8M</td>
</tr>
<tr>
<td>P$_{Hvst}$</td>
<td>0.01</td>
<td>0</td>
<td>272M</td>
<td>143M</td>
</tr>
</tbody>
</table>

$P_{Seed}$ and $P_{Hvst}$ are criteria used for dev&test data driven filtering.

Extraction, filtering, and expansion processes are illustrated.
## Avg. BLEU score over held-out data

On two 2006-2007 dev data (v7, v8)

<table>
<thead>
<tr>
<th>System</th>
<th>Para score</th>
<th>$\text{Ja} \rightarrow \text{En}$</th>
<th></th>
<th>$\text{En} \rightarrow \text{Ja}$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td># of trans. pairs</td>
<td>BLEU</td>
<td># of trans. pairs</td>
<td>BLEU</td>
</tr>
<tr>
<td>Base system</td>
<td>-</td>
<td>18.0M</td>
<td>33.30</td>
<td>15.5M</td>
<td>37.64</td>
</tr>
<tr>
<td>Saug-$P_{\text{Seed}}$</td>
<td>PivProb</td>
<td>27.3M</td>
<td>33.65</td>
<td>+0.35</td>
<td>24.6M</td>
</tr>
<tr>
<td>Saug-$P_{\text{Seed}}$</td>
<td>Cosine</td>
<td>27.3M</td>
<td>33.27</td>
<td>-0.03</td>
<td>24.6M</td>
</tr>
<tr>
<td>Saug-$P_{\text{Hvst}}$</td>
<td>Cosine</td>
<td>23.6M</td>
<td>33.22</td>
<td>-0.08</td>
<td>22.0M</td>
</tr>
<tr>
<td>Saug-$P_{\text{OOPH}}$</td>
<td>Cosine</td>
<td>18.1M</td>
<td>33.72</td>
<td>+0.42</td>
<td>15.6M</td>
</tr>
<tr>
<td>Saug-$P_{\text{Seed}}+P_{\text{Hvst}}$</td>
<td>Cosine</td>
<td>32.8M</td>
<td>32.91</td>
<td>-0.39</td>
<td>30.9M</td>
</tr>
<tr>
<td>Taug-$P_{\text{Seed}}$</td>
<td>PivProb</td>
<td>22.9M</td>
<td>33.34</td>
<td>+0.04</td>
<td>19.6M</td>
</tr>
<tr>
<td>Taug-$P_{\text{Seed}}$</td>
<td>Cosine</td>
<td>22.9M</td>
<td>33.56</td>
<td>+0.26</td>
<td>19.6M</td>
</tr>
<tr>
<td>Taug-$P_{\text{Hvst}}$</td>
<td>Cosine</td>
<td>29.1M</td>
<td>33.43</td>
<td>+0.13</td>
<td>26.8M</td>
</tr>
<tr>
<td>Taug-$P_{\text{OOPH}}$</td>
<td>Cosine</td>
<td>23.4M</td>
<td>33.21</td>
<td>-0.09</td>
<td>21.5M</td>
</tr>
<tr>
<td>Taug-$P_{\text{Seed}}+P_{\text{Hvst}}$</td>
<td>Cosine</td>
<td>33.9M</td>
<td>32.99</td>
<td>-0.31</td>
<td>30.8M</td>
</tr>
</tbody>
</table>
Official results

Human evaluation ($\text{Saug-}P_{OOPH}$)

<table>
<thead>
<tr>
<th></th>
<th>Ja $\rightarrow$ En</th>
<th>En $\rightarrow$ Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>Ranking</td>
</tr>
<tr>
<td>Adequacy</td>
<td>2.89/5.00</td>
<td>10th/18</td>
</tr>
<tr>
<td>Acceptability</td>
<td>0.43/1.00</td>
<td>8th/9</td>
</tr>
</tbody>
</table>

Automatic evaluation

<table>
<thead>
<tr>
<th>System</th>
<th>Ja $\rightarrow$ En</th>
<th>En $\rightarrow$ Ja</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>NIST</td>
</tr>
<tr>
<td>$\text{Saug-}P_{OOPH}$</td>
<td>31.56</td>
<td>8.2507</td>
</tr>
<tr>
<td>$\text{Taug-}P_{\text{Seed}}$</td>
<td>31.65</td>
<td>8.2198</td>
</tr>
<tr>
<td>$\text{*Const-}P_{Tvst}$</td>
<td>30.58</td>
<td>8.1114</td>
</tr>
<tr>
<td>$\text{*Const mixLM}$</td>
<td>30.65</td>
<td>8.1400</td>
</tr>
</tbody>
</table>

*Systems built using only bilingual data.
Implications

- Relatively high **BLEU** and **NIST** scores
  - Useful n-grams (~ phrases) were generated and selected

- Low **RIBES** score and human evaluation score
  - Reordering ability was poor
  - Features of superior systems
    - Structure-aware SMT
    - RBMT adapted to the patent domain
We’ve used 7 for the distortion limit ...

Referring to FIGS. 2 to 8, description will be given to an example of a manufacturing process of the semiconductor storage device which comprises the trench capacitor 120 according to the embodiment.
Relaxation of distortion limit

- Held-out data same as development
  - Obtained significantly higher score
  - Positive impact led by paraphrases was retained

**Graphs:**

- **BLEU (English to Japanese):**
  - Base system
  - Saug-OOPH
  - Taug-Seed

- **RIBES (Japanese to English):**
  - Base system
  - Saug-OOPH
  - Taug-Seed

**Notes:**

- Held-out data same as development
- Obtained significantly higher score
- Positive impact led by paraphrases was retained
Conclusion

- Phrase-based SMT + paraphrases
  - State-of-the-art non-hierarchical system: Portagell @ NRC
    - Almost no language- or domain- specific knowledge
  - Phrase table augmentation
    - Paraphrases in both source & target languages (separately)
    - Comparison of paraphrase collections
    - Aggregation of multiple paths w/ feature engineering
  - Improved performance over a vanilla phrase-based SMT
    - at least BLEU, NIST, and RIBES
Greatest thanks go to

- Supporters of the research program
  - NRC: National Research Council Canada
    - esp. All members in the Portage team
  - FUN: Future University Hakodate
  - JSPS: Japan Society for the Promotion of Science

- PatentMT task organizers