Zero Pronoun Resolution can Improve the Quality of J-E Translation

Hirotoshi Taira, Katsuhiro Sudoh and Masaaki Nagata
NTT Communication Science Labs.
Overview

Jpn-Eng Statistical Translation
– Many ellipses of subjects, objects, and possessive cases (= “Zero Pronouns”) exist in Japanese sentences
– Bad effects to J-E translation quality

This work
– Japanese ellipsis resolution by hand or a baseline method before translation
– Evaluation by “Antecedent F-measure”
– Ellipsis resolution can improve the translation quality
Zero pronoun

– Subjects, objects and possessive cases often drop in Japanese sentences (Nariyama 2003)
– The omitted pronoun is called “zero pronoun”.

Jpn: *watashi wa anata ni shoushou ukagai tai koto ga ari masu* .
(I) (you to) (some) (ask) (to) (questions) (have) (. )

\[\downarrow\]

Usually omitted

Jpn: \(\Phi\) \(\Phi\) *shoushou ukagai tai koto ga ari masu* .
(some) (ask) (to) (questions) (have)

Zero pronouns( \(\Phi\) )
Frequency of Ellipsis in Japanese

– (the National Language Institute for Japanese Language 1995)

Subject Ellipsis in Japanese

Conversation: 74%
Written Texts: 37%
Novel: 20%

– (Ide 2008) Conversation based on Pictures

Ellipsis

Japanese: Subject: 69%, Object: 40%
English: Subject: 15%, Object: 8%
Zero Pronouns Cause Low Quality Translation

Jpn: \textit{shoushou ukagai tai koto ga ari masu .}  
\hspace{2cm} (some)  (ask) (to) (questions) (have) (.)

No subject, no object

Translate by SMT

Eng: You may want to ask a little. (Wrong meaning)

Reference:
Eng: I have some questions to ask to you .

BLEU score is almost not changed, but the translation quality is low.
Idea

– Zero pronoun resolution previous to SMT
– Better alignment is expected

Eng: I have some questions to ask to you.

Jpn: watashi wa anata ni shoushou ukagai tai koto ga ari masu .

Aligned

Zero Pronoun Resolution
Previous Works

– For rule-based translation systems and closed small samples
  - (Yoshimoto, 1988)
  - (Dohsaka, 1990)
  - (Nakaiwa and Yamada, 1997)
  - (Yamamoto et al. 1997)
– For SMT
  - (Furuichi et al. 2011)
    for illustrative sentences in E-J dictionary
    (BLEU evaluation, improved very slightly)
Ellipsis Resolved Data by Human

– BTEC Corpus (Basic Travel Expression Corpus) (Kikui et al., 2003) distributed in IWSLT07
  • Tourism-related Conversation
  • Train: 40k sentences
  • Dev1-3: 1.5k sentences
  • Test: 500 sentences

– Ellipsis Resolved by Human
  • Annotated zero pronouns and the antecedents
  • Based on pronouns in the translated English sentences.
Clues for Ellipsis Resolution

– Spanish to English
  • Spanish allows the ellipsis of subjects
  • The ellipsis leaves traces including the case and the gender on the related sentence

<table>
<thead>
<tr>
<th>1st person, singular form</th>
</tr>
</thead>
</table>

(yo) Tengo algunas preguntas para hacerle a usted.
(I) (have) (some) (questions) (to) (ask) (to) (you).

– Japanese
  • No change of verb form
  • Need to estimate antecedents from contexts (modality, polite expression, previous sentences, etc.)
Baseline System based on Simple Modality Information

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Declarative sentence</td>
<td>Watashi-wa ano eiga-wo mimashita.</td>
<td>I-TOP Watashi-wa ano eiga-wo mimashita. (= “I watched the movie.”)</td>
</tr>
<tr>
<td>Question sentence</td>
<td>Anata-wa ano eiga-wo mimashita ka ?</td>
<td>You-TOP Anata-wa ano eiga-wo mimashita ka ? (= “Did you watch the movie?”)</td>
</tr>
<tr>
<td>Imperative sentence</td>
<td>ano eiga-wo minasai.</td>
<td></td>
</tr>
</tbody>
</table>

No ‘ga’ case (subject) ‘wa’, ‘mo’ (topic)

Insert ‘watashi-wa’ (I-TOP)

Insert ‘anata-wa’ (You-TOP)

NO CHANGE
Experimental Settings

- Phrase based Translation (Moses)
- Tuning: MERT
- Japanese Tokenizer: MeCab
- English Tokenizer: Moses Toolkit
- Decoder: Moses (default settings)
- Zero pronoun resolution for training, dev, and test sets
## Results

### BLEU

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>45.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>45.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>45.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Antecedent F-measure

\[
P = \frac{|G \cap S|}{|S|}
\]

\[
R = \frac{|G \cap S|}{|G|}
\]

\[
F = \frac{2PR}{P+R}
\]

G: the set of the gold standard zero pronouns
S: the set of each pronoun in English translated by decoder the gold standard zero pronouns

<table>
<thead>
<tr>
<th>Ant-F</th>
<th>i</th>
<th>my</th>
<th>me</th>
<th>you</th>
<th>your</th>
<th>it</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>53.9</td>
<td>53.3</td>
<td>57.1</td>
<td>55.0</td>
<td>63.1</td>
<td>50.0</td>
</tr>
<tr>
<td>baseline</td>
<td>48.9</td>
<td>56.7</td>
<td>57.1</td>
<td>55.9</td>
<td>50.0</td>
<td>45.6</td>
</tr>
<tr>
<td>human</td>
<td>58.4</td>
<td>70.5</td>
<td>60.3</td>
<td>73.7</td>
<td>77.5</td>
<td>56.5</td>
</tr>
</tbody>
</table>

#Ref 121 39 32 95 23 51
Effectiveness of Zero Pronoun Resolution for Decoding

Better case

Today's evening by send would QUES
Kyō-no yuugata made-ni todoke-te morae-masu ka.

It by this evening ?

(Ref) Can you deliver them by this evening?

you-TOP anata-wa

Can you send it by this evening ?
Effectiveness of Zero Pronoun Resolution for Decoding

Worse case

Hurry up
Isoi-de .

Hurry up .

I-TOP watashi-wa hurry up Isoi-de .

I’m in a hurry .

Not changed

map-BY point_out would QUES chizu-de sashi-te morae-masu ka.

Would you point them out on this map ?

You-TOP map-BY point_out would QUES anata-wa chizu-de sashi-te morae-masu ka.

Would you point them out on this map ?

(Ref) Would you point one out on this map?
Conclusion

Jpn-Eng Statistical Translation

– Many ellipses of subjects, objects, and possessive cases (= “Zero Pronouns”) exist in Japanese sentences

– Bad effects to J-E translation quality

– Japanese sentence in which ellipsis resolution by hand, improved the score of “Antecedent F-measure”

– Simple baseline system using modality information could not improve the translation quality totally