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

Hirotoshi Taira, Katsuhito 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) (.)

↓ Usually omitted

Jpn: *Φ* *Φ* *shoushou* *ukagai* *tai* *koto ga* *ari masu* .

(some) (ask) (to) (questions) (have)

Zero pronouns (.)

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

No subject, no object

Translate by SMT

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

Reference:

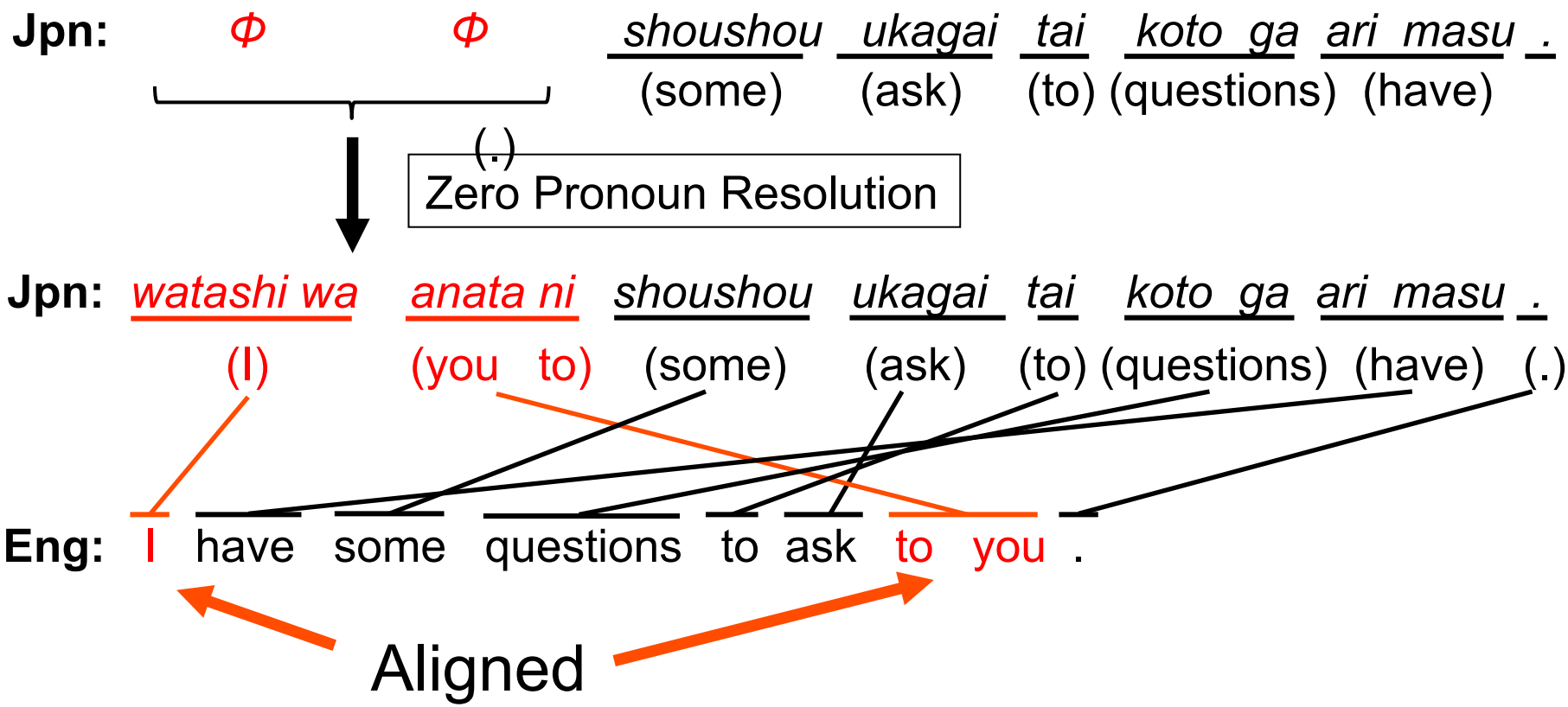
Eng: I have some questions to ask to you .

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

Idea

– Zero pronoun resolution previous to SMT

– Better alignment is expected



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)

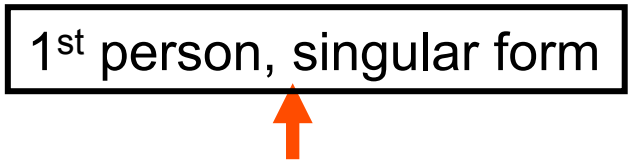
Ellipsis Resolved Data by Human

- 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



(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

Declarative sentence

ano *eiga-wo* *mimashita.*
the movie-OBJ watched

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

Insert ‘watashi-wa’ (I-TOP)

Watashi-wa *ano* *eiga-wo* *mimashita.*
I-TOP the movie-OBJ watched
(= “**I** watched the movie.”)

Question sentence

ano *eiga-wo* *mimashita* *ka* *?*
the movie-OBJ watched QUES ?

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

Insert ‘anata-wa’ (You-TOP)

Anata-wa *ano* *eiga-wo* *mimashita* *ka* *?*
You-TOP the movie-OBJ watched QUES ?
(= “Did **you** watch the movie?”)

Imperative sentence

ano *eiga-wo* *minasai.*
the movie-OBJ watch-IMP
(= “Watch the movie.”)

. 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

original	45.1
baseline	45.4
human	**45.6**

Antecedent F-measure

$$P=|G\cap S|/|S|$$

$$R=|G\cap S|/|G|$$

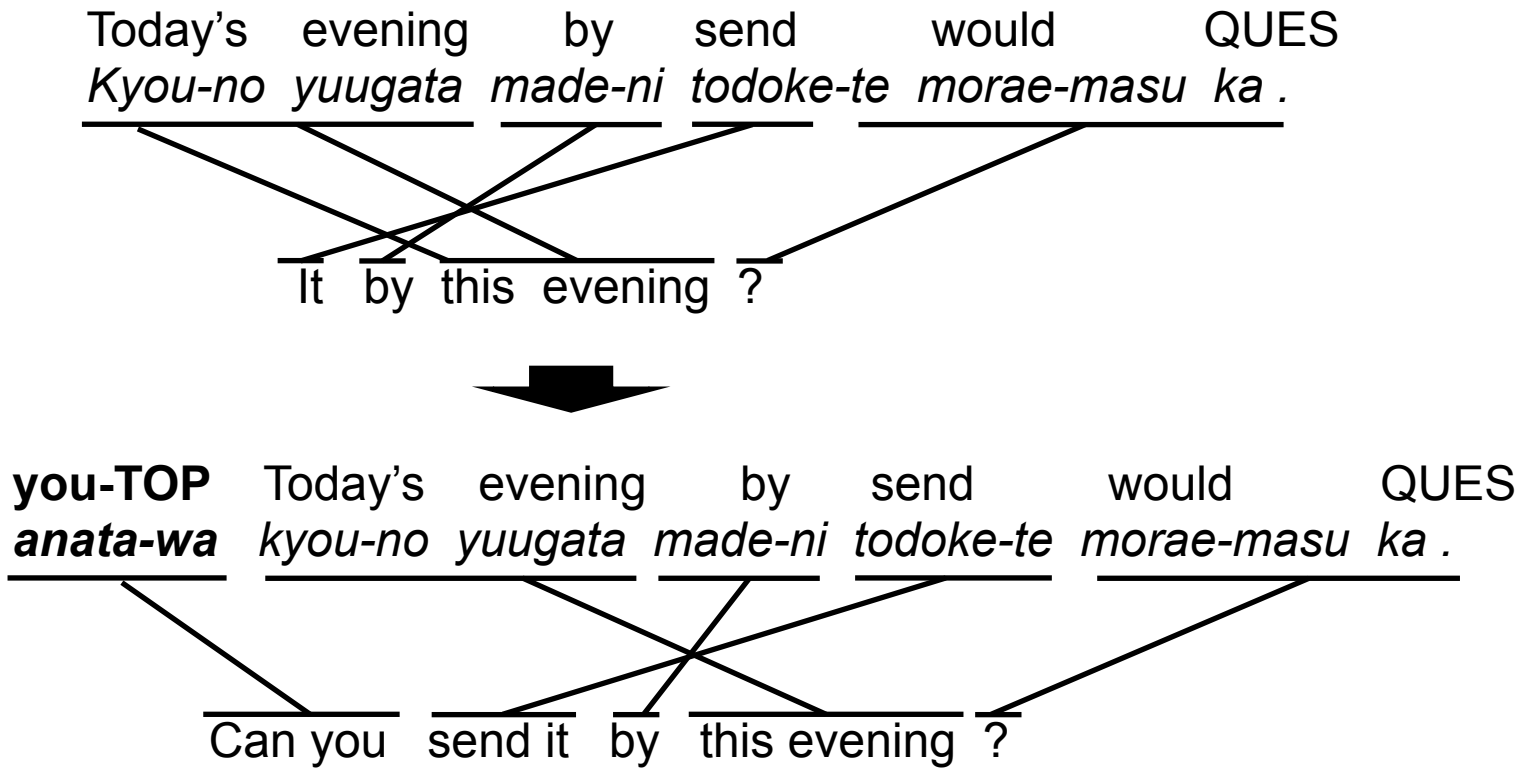
$$F=2PR/(P+R)$$

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

Ant-F	i	my	me	you	your	it
original	53.9	53.3	57.1	55.0	63.1	50.0
baseline	48.9	56.7	57.1	55.9	50.0	45.6
human	**58.4**	**70.5**	**60.3**	**73.7**	**77.5**	**56.5**
#Ref	121	39	32	95	23	51

Effectiveness of Zero Pronoun Resolusion for Decoding

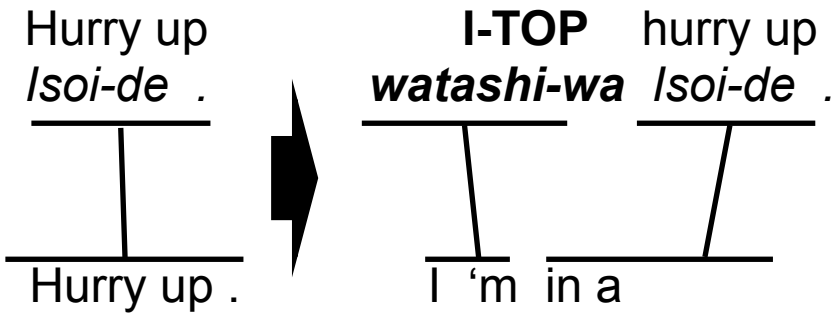
Better case



(Ref) Can you deliver them by this evening?

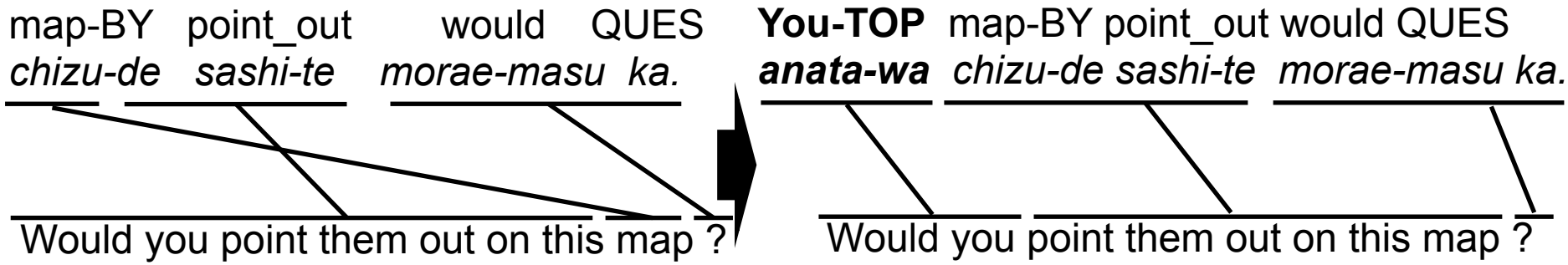
Effectiveness of Zero Pronoun Resolusion for Decoding

Worse case



(Ref) Hurry up.

Not changed



(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