JaEn: Large-Scale Rule Acquisition

Petter Haugereid and Francis Bond
Division of Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore
petterha@ntu.edu.sg, bond@ieee.org

delph-in 2011
Suquamish, 25 June, 2011

Introduction

- We have tried two methods for learning transfer rules for the Japanese-to-English MT system Jaen.
- Partial MRS/Object MRS mismatches
- Learning from
  - Dictionaries
  - Parallel text

Partial MRS/Object MRS mismatches

- Parsed short sentences in the Tanaka corpus and the Japanese Wordnet with Jacy and the ERG. (150,000 sentences)
- Tried to transfer the Japanese MRSs
- Looked at the Partially transferred MRSs and the Object MRSs
- Matched with transfer rule types

$\rightarrow$ 1,110 rules

Procedure — Sources

- We are using Moses (Koehn *et al.*, 2007) and Anymalign (Lardilleux and Lepage, 2009) to generate phrase tables from a collection of four Japanese English parallel corpora and one bilingual dictionary:
  - Tanaka Corpus (2,930,132 words)
  - the Japanese Wordnet Corpus (3,355,984 words)
  - the Japanese Wikipedia corpus (7,949,605)
  - the Kyoto University Text Corpus with NICT translations (1,976,071 words)
  - Edict, a Japanese English dictionary (3,822,642 words)

Procedure — Preparing the training data

- The corpora were divided into development, test, and training data.
- Transfer rules were extracted from the training data.
- The training data of the four corpora together with the Edict dictionary form a parallel corpus of 20 million words.
  - 9.6 million English words
  - 10.4 million Japanese words
- The traning data were tokenized and lemmatized.
  - For Japenese with the MeCab morphological analyzer.
  - For English with the Freeling analyzer.

Procedure — Extracting phrase tables

- We applied GIZA++/Moses and Anymalign to the lemmatized parallel corpus

→ 10,812,423 Moses entries and 5,765,262 Anymalign entries

- We filtered out
  - Entries with an absolute frequency of 1.
  - Entries which had more than 4 words on the Japanese side or more than 3 words on the English side.
  - Entries with lemmas that were not in the lexicons the parser/generator.
  - Entries with a translation probability, P(English|Japanese), of less than 0.1.

→ 2,618,959 entries in total

SMT phrase table entries

Japanese	English	Probability	Source
頭 が 良い	bright	0.473684	Anymalign
頭 が 良い	intelligent	0.2	Moses
頭 が 良い	clever	0.2	Moses

Procedure — Extracting possible semantic rules

- Phrase table entry lemmas were matched with semantic predicates assigned by Jacy/ERG lexicons.
- Each possible surface rule was represented with a list of all possible semantic predicate rules.
  - → A possible surface rule with three (two times) ambiguous lexical items → 2x2x2 = 8 possible semantic rules.
- A total of 46,907,658 possible semantic rules were created.
- Semantic transfer rules containing predicates of probability less than 0.2 were filtered out.
  - → 5,584,604 possible semantic rules.

Procedure — Selecting semantic transfer rules

- The possible semantic transfer rules were matched with nine different patterns/templates.
  - → 29,417 single rules.
  - → 74,847 MWE rules.
  - → 104,264 rule in total.
- Once the rule templates have been selected and the thresholds set, the entire process is automatic.

Extracted single rules

Input		Output	Rules
noun	→	noun	20,207
proper noun	→	proper noun	1,225
adj	→	adj	2,751
intrans verb	→	intrans verb	3,242
trans verb	→	trans verb	1,985
ditrans verb	→	ditrans verb	11
Total			29,417

Table: Transfer rule patterns.

Extracted MWE rules

Input		Output	Rules
noun + noun	→	noun + noun	27,345
noun + noun	→	adj + noun	18,053
noun + noun	→	noun	19,033
noun + adj	→	adj	473
PP	→	adj	856
PP	→	PP	146
verb + NP	→	verb + NP	6,993
postp + noun + verb	→	verb	1,360
PP + verb	→	verb	588
Total			74,847

Table: Transfer rule patterns.

PP → adjective

- Japanese PPs headed by the postposition の *no* "of" often correspond to an adjective in English:

(1) a. 小型　　の
small.size of
small

b. 音楽　の
music of
musical

PP → PP

- Japanese PPs headed by the postposition で *de* "with/by/in/on/at" often translates into English PPs headed by the preposition 'by' where the prepositional object does not have a determiner:

(2) タクシーで
taxi DE
by taxi

PP → PP — Example

(3)
$$\begin{bmatrix} pp\_pp\_mtr & \\ \text{IN} & \left\langle \begin{bmatrix}\text{PRED} & \_de\_p\_rel\end{bmatrix}, \begin{bmatrix}\text{PRED} & \text{udef}\_q\_rel\end{bmatrix}, \begin{bmatrix}\text{PRED} & \_takushii\_n\_rel\end{bmatrix} \right\rangle \\ \text{OUT} & \left\langle \begin{bmatrix}\text{PRED} & \_by\_p\_means\_rel\end{bmatrix}, \begin{bmatrix}\text{PRED} & \text{udef}\_q\_rel\end{bmatrix}, \begin{bmatrix}\text{PRED} & \_taxi\_n\_1\_rel\end{bmatrix} \right\rangle \end{bmatrix}$$

Verb + NP → Verb + NP

- Japanese *noun* + *object marker* (を) + *verb* MWEs usually translates into one out of three English MWEs:

(4) a. テニス を　し ます
tennis ACC do POLITE

play tennis

b. 生計　を　立てる
living ACC stand up

make a living

c. 責め　を　負う
blame ACC bear

take the blame

Noun + Adj → Adj

- The Japanese *noun* + が (ga) + *adjective* MWE pattern is transltated into an adjective in English:

(5)
X ga 背 が 高い
X ga se ga takai
X ga NOM height NOM high
X is tall

Noun + Adj → Adj

- With the new rules, the transfer grammar now correctly translates (6) as *She is very intelligent.* and not *Her head is very good.*
- The adverb modifying the adjective in Japanese is also modifying the adjective in English.

(6)

彼女	は	大変	頭	が	いい	。
kanojo	wa	taihen	atama	ga	yoi	.
She	TOPIC	very	head	NOM	good	.

She is very intelligent.

Passive (and zero pronouns)

この 部屋 は 台所 として 使わ れ て いる 。
Old: You are using this room as the kitchen.
New: This room is used as a kitchen.
Ref: (This room is used as a kitchen.)

彼女 の 名前 は 知ら れ て い ませ ん でし た 。
Old: I was not knowing her name.
New: Her name was not known.
Ref: (Her name was not known.)

- NEVA: 18.48% → 18.83%

Models for transfer and generation

- A trigram model for transfer ranking
- Generation model + 1.4% NEVA

Results

Version	Total coverage	NEVA	F1
2010	18.0%	14.9%	16.3%
2011	27.6%	19.0%	22.5%

Table: Coverage on development data.

Results

- The coverage of the system is 27.6%.
- A human evaluation of the translated test sentences from the last test shows that Jaen performs better than Moses in 53 out of 100 cases.
- The BLEU score of Jaen (24.90) is still below that of Moses (33.63).

Discussion — Compositional rules

- Many of the translations learned are compositional.
- 穴 を 掘る *ana-wo horu* "dig hole" → *dig a whole* would have been translated using existing rules.
- The advantage of the MWE rule is that it reduces the search space

→ The system does not have to consider less likely translations such as *carve the shortages*.

Discussion — Non-compositional rules

- Many of the rules find non-compositional translations, or those where the structure cannot be translated word for word.
- Some of these are also idiomatic in the source and target language.
- One of our long term goals is to move these expressions into the source and target grammars.

Future work — Selecting the good rules

- How do we determine whether rules are good or not.
- Currently we are investigating two solutions:
  - feedback cleaning, where we investigate the impact of each new rule and discard those that degrade translation quality.
  - human-in-the loop: presenting each rule and a series of relevant translation pairs to a human and asking them to judge if it is good or not.

Future work — Qualitative improvement

- We are working in parallel to qualitatively improve the MWE rules in two ways.
  - By extending rules using semantic classes, not just words.
    - → need for fewer rules, but each rule would be more powerful.
  - By learning complex rules directly from the parallel text
    - This will be necessary to catch rules that our templates do not include

Cooperation with Kyung Hee University

- We are participating in Jong-Bok Kim's new project.
- Employ this mechanism for Korean-English and Korean Japanese.
- We would like to share the translation rule types and the translation rule learning infrastructure also to other translation projects.

References

Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., Moran, C., and Birch, A. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Interactive Presentation Sessions*, Prague.

Lardilleux, A. and Lepage, Y. (2009). Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria.