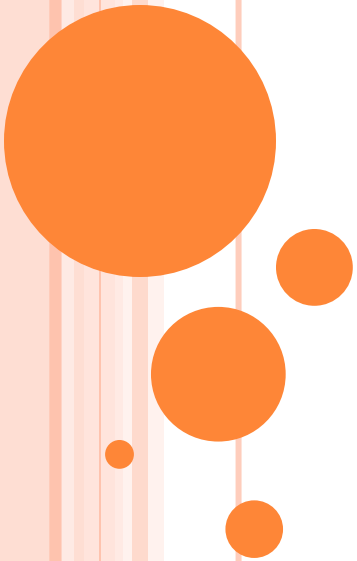

Verb subgategorisation in English-Japanese Machine Translation

Alexis Kauffmann

Contents

- Introduction and related research about subcategorisation lexicons
- Context: Improvement of an English-Japanese LBMT system for better complex sentence translation
- Purpose and method for subcategorisation translation
- Application: Modifying LATL databases:
  - -decription
  - -difficult points
  - - experiment
- Results
- Conclusion
- Next Steps of my research

Indtroduction and related research about subcategorisation lexicons

- As you already know (Sasano, 夏勉強会 2010)
- Famous databases: FrameNet, Propbank, SUCAT, 格フラーム
- LATL databases (Nerima, Werhli,Scherrer, 2009): -handmade monolingual and bilingual lexical databases. For example, 11544 English verb subcats, 7079 French ones, 6843 German ones.

  - transitivity generated bilingual databases

  - Japanese lexical databases

ords enabled
onnées 'infatec' ouverte ! (utilisateur: alexis)

Bilingual

Source Language: English

Target Language: Japanese

go

Search

Search

go : V [111048081] - [NP _ S]
go : V [111048581] - [NP _]
go : V [111048668] - [NP _ FP]
go (into) : V [111048692] - [NP _ PP]
go (from) (to) : V [111049200] - [NP _ PP PP]
go (with) : V [111054031] - [NP _ PP]
go (against) : V [111057199] - [NP _ PP]

Info

行く (から) (まで) : V [611219792] - [NP _ PP P
なる : V [611161400] - [NP _]
行く : V [611207272] - [NP _]
行く (に) : V [611219793] - [NP _ PP]
足を運ぶ : V [611167307] - [NP _]
出かける : V [611174359] - [NP _]
作動する : V [611192433] - [NP _]

Info

Remove

Argument Correspondence

Target Language Correspondence

Source Language Correspondence	1	2	3	No Arg
1	●	○	○	○
2	○	●	○	○
3	○	○	●	○

111049200

611219792

Features

Modify

Preference: 6

Translation Context: Standard

Description

Swap

Update & Swap

Update

Introduction and related research about subcategorisation lexicons

- LATL databases contain data for every category of word. For verbs: syntactic subcat and semantic case frame and other syntactic and semantic information about the verb, and possibly about each argument.
- Quality and drawbacks of LATL verb databases
  - syntactically very correct (hand-made)
  - semantically, could do better,
  - could be bigger.

Context: Improvement of an English-Japanese LBMT system for better complex sentence translation

- A need to translate better complex structures
- Work on modality translation
- Work on infinitives, gerundives and object sentence translation

  => need for subcategorisation data about verbal/sentential objects
- Also a need for less mistakes in nominal argument postpositional particles handling

  => need for global subcategorisation frame data

Purpose and method for subcategorisation translation

- Purpose: Handling subcategorisation translation:

  -to know Japanese verb subcategorisation frames

  -to choose the right case frame and to transfer the arguments correctly
- Method: Using Kawahara sensei's Japanese subcategorisation frame data

Application: Modifying LATL databases

- LATL Japanese monolingual Data

- Taken from a Japanese dictionary file:

- 34700 Verb entries. One entry per subgategorisation frame.

- Transitive/intransitive only

=> problems in bilingual correspondences

- About 50 entries entered or corrected by hand

Application: Modifying LATL databases: description

- LATL grammatical functions:
  - Subject
  - Direct Object
  - Indirect/Prepositional Object
    (+ preposition type)
  - Sentencial/verbal
    (+ verb tense information)
  - Adjectival predicate

Application: Modifying LATL databases : description

- Kawahara sensei's data Japanese particles:
  - が
  - を
  - が2
  - に
  - と
  - から、まで、より、へ

Application: Modifying LATL databases : difficult points

- In order to make the right bilingual correspondences, matching of corresponding subcat frames

  ex: がを ---> S O

  が から まで ---> S PO(from) PO(to)

- Problem1

  がに ---> S PO or O S

Application: Modifying LATL databases : difficult points

- In order to make the right bilingual correspondences, matching of corresponding subcat frames

  ex: がを ---> S O

  が から まで ---> S PO(from) PO(to)

- Problem1

  がに ---> S PO or O S

  Answer: with additional case structure data, showing which one is the agent. The agent is more likely to be the subject.

Application: Modifying LATL databases : difficult points

- Problem2 : selection of the right correspondence when several possible ones

- Problem3 : how to predict asymmetrical correspondences?

  ex: 見る(がを) ---> to look (S PO(at))

Application: Modifying LATL databases : difficult points

- Problem2 : selection of the right correspondence when several possible ones

  Answer: additional data indicating number of occurences can help.

- Problem3 : how to predict asymmetrical correspondences?

  ex: 見る(がを) ---> to look (S PO(at))

Application: Modifying LATL databases : difficult points

- Problem2 : selection of the right correspondence when several possible ones

  Answer: additional data indicating number of occurences can help.

- Problem3 : how to predict asymmetrical correspondences?

  ex: 見る(がを) ---> to look (S PO(at))

  Partial Answer: generation of an asymetrical correspondence only when no possible symetrical correspondence exist.

Application: Modifying LATL databases : experiment

- Creation of the Japanese verb entries for LATL monolingual database, formatting Kawahara sensei's data.
- Insertion of the part of the data that was not already there.
- Knowing LATL bilingual database verb correspondences (without considering subcategorisation), creation of the possible bilingual correspondences, with symetric subcat most often, and asymetric subcat sometimes
- Insertion of absent bilingual correspondences
- Lowering correspondence scores of unprobable existing bilingual entries

Results

- About 5000 monolingual verb subcat entries inserted.
- About 2500 were already there.
- About 8000 bilingual correspondences entered and many old correspondences bilingual almost deleted.

Conclusion

- Japanese subcategorisation knowledge much better in the database
- Need to test the results with the MT system
- Need to improve by hand subcat for verbs with 3 arguments or more, verbal objects and adjectival objects.

9)Next Steps of my research

- Finishing improving subcat data
- Finishing programming translation of modality and verbal/sentential objects

- Improving collocation/multi-word expressions translation