# Extracting Semantic Transfer Rules from Parallel Corpora with SMT Phrase Aligners

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

Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)
Jeju, Korea, July 12 2012

# Outline



NANYANG TECHNOLOGICAL UNIVERSITY

# Outline

Semantic Transfer

# Jaen

- Jaen is a rule-based machine translation system employing semantic transfer rules
- The medium for the semantic transfer is Minimal Recursion Semantics, MRS (Copestake et al., 2005)
- The system consists of two HPSG grammars:
  - JACY parses the Japanese input (Siegel and Bender, 2002)
  - The ERG generates the English output (Flickinger, 2000)
- The third component of the system is the transfer grammar Jaen (Bond et al., 2011):

  **IN** MRS representation produced by the Japanese grammar

  **OUT** MRS representation the English grammar can generate from

# Stochastic Models

- At each step of the translation process, the output is ranked by stochastic models
- Only the 5 top ranked outputs at each step are kept
  - $\Rightarrow$ maximum number of translations: 125 (5x5x5)
- A final reranking using a combined model

NANYANG TECHNOLOGICAL UNIVERSITY

# Architecture

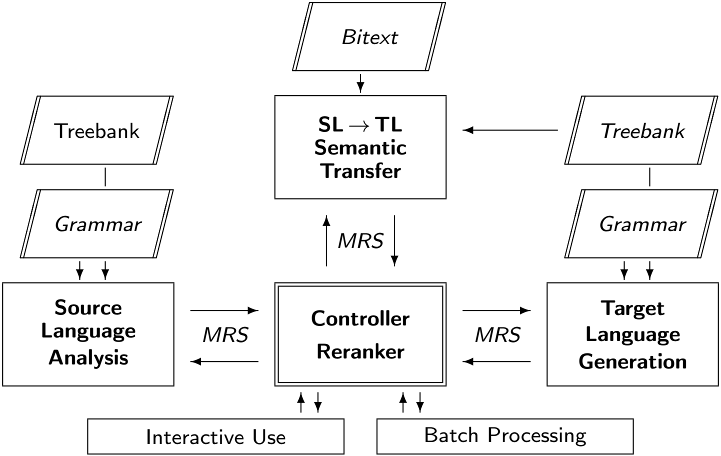


Figure 1: Architecture of the Jaen MT system

# Transfer Rules

- Many transfer rules are simple predicate changing rules:
  - "_hon_n_rel" ⇒ "_book_n_1_rel"
- Other rules are more complex, and may transfer many Japanese relations into many English relations
- In all, there are 61 types of transfer rules

# Most Frequent Rule Types

| Rule type | Hand | Lemma | Pred | Intersect | Union | Total |
|---|---|---|---|---|---|---|
| noun | 64 | 32033 | 31575 | 19100 | 44508 | 44572 |
| n+n_n+n | 0 | 32724 | 18967 | 13494 | 38197 | 38197 |
| n+n_adj+n | 0 | 22777 | 15406 | 10504 | 27679 | 27679 |
| arg12+np_arg12+np | 0 | 9788 | 1774 | 618 | 10944 | 10944 |
| arg1_v | 22 | 8325 | 1031 | 391 | 8965 | 8987 |
| pp_pp | 2 | 146 | 8584 | 19 | 8711 | 8713 |
| adjective | 27 | 4914 | 4034 | 2183 | 6765 | 6792 |
| arg12_v | 50 | 4720 | 1846 | 646 | 5920 | 5970 |
| n_adj+n | 1 | 0 | 4695 | 0 | 4695 | 4696 |
| n+n_n | 0 | 2591 | 3273 | 1831 | 4033 | 4033 |
| n+n+n_n+n | 0 | 3380 | 0 | 0 | 3376 | 3376 |
| n+adj-adj-mtr | 2 | 633 | 2586 | 182 | 3037 | 3039 |
| n_n+n | 1 | 0 | 2229 | 0 | 2229 | 2230 |

Table 1: Most common mtr rule types

NANYANG TECHNOLOGICAL UNIVERSITY

# Handwritten and Automatically Extracted Rules

- The transfer grammar has a core set of 1,415 hand-written transfer rules:
  - function words
  - proper nouns
  - pronouns
  - time expressions
  - spatial expressions
  - the most common open class items
- The rest of the transfer rules (190,356 unique rules) are automatically extracted from parallel corpora

The full system is available from
`http://moin.delph-in.net/LogonTop`
(all components are open source, mainly LGPL and MIT)

# Outline

## Two Methods of Rule Extraction

- Extraction from a Lemmatized Parallel Corpus
- Extraction from a Parallel Corpus of Predicates

# Parallel Corpus

- The parallel corpus we use for rule extraction is a collection of four Japanese English parallel corpora:
  - Tanaka Corpus (2,930,132 words)
  - The Japanese Wordnet Corpus (3,355,984 words)
  - The Japanese Wikipedia corpus (7,949,605 words)
  - The Kyoto University Text Corpus with NICT translations (1,976,071 words)
- Plus the dictionary Edict (3,822,642 words)
- (The word totals include both English and Japanese words)

# Parallel Corpus

- The corpora were divided into into development, test, and training data
- The training data plus the bilingual dictionary was used for rule extraction
- The combined corpus used for rule extraction consists of
  - 9.6 million English words
  - 10.4 million Japanese words

# Procedure 1

## Lemmatizing the Corpus

- We extracted transfer rules directly from the surface lemmas of the parallel text
- The four parallel corpora were tokenized and lemmatized
  - Japanese: the MeCab morphological analyzer
  - English: the Freeling analyzer

## Aligning the Lemmatized Corpus

- We then used MOSES and Anymalign to align the lemmatized parallel corpus

# Procedure 1

## Selection of Alignments

- We selected the alignments that
  - had relatively high probability ($> 0.1$)
  - were known both to the parsing grammar (JACY) and the generating grammar (ERG)

# Procedure 1

## Assigning Semantic Predicates

- The alignments were a mix of one-to-one-or-many and many-to-one-or-many
- For each lemma in each alignment, we listed the possible predicates according to the lexicons of JACY and the ERG
- Many lemmas are ambiguous
  - ⇒ we often ended up with many semantic alignments for each surface alignment
- If a surface alignment contains 3 lemmas with two readings each
  - ⇒ 8 (2x2x2) semantic alignments

NANYANG TECHNOLOGICAL UNIVERSITY

# Procedure 1

## Filtering of Semantic Predicates

- Some lemmas have very rare readings
  - $\Rightarrow$ We parsed the training corpus and made a list of 1-grams of the semantic relations of the highest ranked parses
  - $\Rightarrow$ Predicates with probability $> 0.2$ were considered

# Procedure 1

## Types of Templates

- The semantic alignments were matched against 16 templates
- Seven templates are simple one-to-one mapping templates:
  1. noun $\Rightarrow$ noun
  2. proper noun $\Rightarrow$ proper noun
  3. adjective $\Rightarrow$ adjective
  4. adjective $\Rightarrow$ intransitive verb
  5. intransitive verb $\Rightarrow$ intransitive verb
  6. transitive verb $\Rightarrow$ transitive verb
  7. ditransitive verb $\Rightarrow$ ditransitive verb

# Procedure 1

## Multiword Templates

- Some multiword templates are relatively simple:

8. n+n ⇒ n

(1) 小 テスト-が あっ-た 。
minor test had
*I had a quiz.*

9. arg12+np ⇒ arg12+np_mtr

(2) その 仕事-を 終え-まし-た 。
that job finished
*I finished the job.*

# Procedure 1

## Complex Templates

- Other rules are more complex:

10. n+adj ⇒ adj

(3) 前-の 冬-は 雪-が 多かっ-た 。
previous winter snow much-be
*Previous winter was snowy.*

(4) 雪-の 多い 冬 だっ-た 。
snow much winter was
*It was a snowy winter.*

In all, we extracted 126,964 rules with this method

NANYANG TECHNOLOGICAL UNIVERSITY

# Procedure 1

## Problems with Filtering of Transfer Rules

- We were forced to filter semantic relations that have a low probability in order to avoid translations that do not generalize
  - ⇒ We failed to build rules that should have been built
    - (where an ambiguous lemma has one dominant reading, and one or more less frequent, but plausible, readings)
  - ⇒ We built incorrect rules
    - (where the dominant reading is used, but where a less frequent reading is correct)
- The method is not very precise
  - it is based on simple 1-gram counts
  - we are not considering the context of the individual lemma

NANYANG TECHNOLOGICAL UNIVERSITY

# Procedure 1

## Solution?

- A way to improve the quality of the assignment of the relation to the lemma would be to use a tagger or a parser

## Using the Grammars as Semantic Taggers

- Instead we decided to try a different approach
  - parse the whole parallel training corpus with the parsing grammar and the generation grammar of the MT system
  - produce a parallel corpus of semantic relations instead of lemmas

  $\Rightarrow$ use the linguistic grammars as high-precision semantic taggers

NANYANG TECHNOLOGICAL UNIVERSITY

# Procedure 2

## A Parallel Corpus of Predicates

- The second rule extraction procedure is based on a parallel corpus of semantic representations
- We parsed the training corpus (1,578,602 items)
  - with the parsing grammar (JACY)
  - with the generation grammar (ERG)
  - $\Rightarrow$ a parse with both grammars for 630,082 items
- The grammars employ statistical models trained on treebanks in order to select the most probable analysis
- For our semantic corpus, we used the semantic representation of the highest ranked analysis on either side

NANYANG TECHNOLOGICAL UNIVERSITY

# Procedure 2

```
[ mrs
  LTOP   [h1] h
  INDEX  [e2] e
  RELS   < [ _the_q_rel<0:3>         [ _white_a_1_rel<4:9>     [ _dog_n_1_rel<10:13>
             LBL   [h3] h              LBL   [h7] h              LBL   [h7]
             ARG0  [x5] x              ARG0  [e8] e              ARG0  [x5] ] ,
             RSTR  [h6] h            , ARG1  [x5] ]            ,
             BODY  [h4] h ] ,
           [ _bark_v_1_rel<14:20>
             LBL   [h9] h
             ARG0  [e2]
             ARG1  [x5] ] >
  HCONS  < [ qeq
             HARG  [h6]
             LARG  [h7] ] > ]
```

Figure 2: MRS of *The white dog barks*

# Procedure 2

## Semantic Parallel Corpus

- The resulting parallel corpus of semantic representations had:
  - 4,712,301 relations for Japanese
  - 3,806,316 relations for English
  
  ⇒ a little more than a third of the size of the lemmatized parallel corpus
- The grammars used for parsing are deep linguistic grammars

# Procedure 2

## Using the Grammars for Disambiguation

- Transfer rules extraction from from the semantic parallel corpus is similar to the rule extraction from the lemmatized corpus
- Major difference:
  - the semantic corpus is disambiguated by the grammars

# Procedure 2

## Alignment of Predicates

- The semantic parallel corpus was aligned with MOSES and Anymalign
- We selected the alignments with probability $> 0.01$

# Procedure 2

## Checking Alignments against Rule Templates

- The alignments were checked against 22 rule templates
  - $\Rightarrow$ 112,579 rules
  - (slightly fewer than the number of rules extracted from the lemmatized corpus – 126,964)
- 49,187 of the rules overlap with the rules extracted from the lemmatized corpus
  - $\Rightarrow$ a total number of unique rules of 190,356

# Outline

Experiment and Results

# Three Transfer Grammar Versions

- We made three versions of the transfer grammar:
  - Lemm: the rules extracted from the lemmatized corpus
  - Pred: the rules extracted from the corpus of semantic representations
  - Combined: the union of Lemm and Pred

# Two Additional Transfer Grammar Versions

- We also made two versions of the transfer grammar including only the 15 templates used in both Lemm and Pred:
  - LemmCore
  - PredCore

# Tests

- The five versions of the transfer grammar were tested on the Tanaka Corpus test data (4,500 sentences):

| | Parsing | Transfer | Gener. | Overall | NEVA | Oracle | F1 |
|---|---|---|---|---|---|---|---|
| LemmCore | 79.8% | 46.3% | 56.0% | 20.7% | 18.65 | 22.99 | 19.61 |
| Lemm | 79.8% | 46.6% | 56.0% | 20.8% | 18.65 | 22.99 | 19.69 |
| PredCore | 79.8% | 48.7% | 52.9% | 20.6% | 20.40 | 24.81 | 20.48 |
| Pred | 79.8% | 49.7% | 52.6% | 20.8% | **21.11** | **25.75** | 20.96 |
| Combined | 79.8% | 60.9% | 54.7% | **26.5%** | 19.77 | 24.00 | **22.66** |

Table 2: Evaluation of the Tanaka Corpus Test Data

NANYANG TECHNOLOGICAL UNIVERSITY

# Results

- 644 of the test sentences were translated by all versions of the transfer grammar (Lemm, Pred, and Combined):

| Version | NEVA |
|---|---|
| Lemmatized | 20.44 |
| MRS | **23.55** |
| Lemma + MRS | 23.04 |

Table 3: NEVA scores of intersecting translations

NANYANG TECHNOLOGICAL UNIVERSITY

# Comparison with MOSES, Combined

| | BLEU | METEOR | HUMAN |
|---|---|---|---|
| JaEn First | 16.77 | 28.02 | **58** |
| MOSES | **30.19** | **31.98** | 42 |

Table 4: BLEU Comparison of Jaen loaded with the Combined rules, and MOSES (1194 items)

# Comparison with MOSES, Pred

| | BLEU | METEOR | HUMAN |
|---|---|---|---|
| JaEn | 18.34 | 29.02 | **58** |
| MOSES | **31.37** | **32.14** | 42 |

Table 5: BLEU Comparison of Jaen loaded with the Pred rules, and MOSES (936 items)

# Jaen Errors

- The output of Jaen is mostly grammatical, but it may not always make sense:

| | | |
|---|---|---|
| (5) | Source: | 我々 は 魚 を 生 で 食べる 。 |
| | Ref.: | We eat fish raw. |
| | Moses: | We eat fish raw. |
| | Jaen: | We eat fish in the camcorder. |

- Jaen sometimes gets the arguments wrong:

| | | |
|---|---|---|
| (6) | Source: | 彼 は 大統領 に 選ば れ た 。 |
| | Ref.: | He was elected president. |
| | Moses: | He was elected president. |
| | Jaen: | The president chose him. |

# Moses Errors 1

- The output of Moses is more likely to lack words in the translation:

(7) Source: カーテン が ゆっくり 引か れ た 。
Ref.: The curtains were drawn slowly.
Moses: The curtain was slowly.
Jaen: The curtain was drawn slowly.

- Missing words become extra problematic when a negation is not transferred:

(8) Source: 偏見 は 持つ べき で は ない 。
Ref.: We shouldn't have any prejudice.
Moses: You should have a bias.
Jaen: I shouldn't have prejudice.

# Moses Errors 2

- The output of Moses is more likely to be ungrammatical:

| | | |
|---|---|---|
| (9) | Source: | 私 は 日本 を 深く 愛し て いる 。 |
| | Ref.: | I have a deep love for Japan. |
| | Moses: | I is devoted to Japan. |
| | Jaen: | I am deeply loving Japan. |
| (10) | Source: | 彼女 は タオル を 固く 絞っ た 。 |
| | Ref.: | She wrung the towel dry. |
| | Moses: | She squeezed pressed the towel. |
| | Jaen: | She wrung the towel hard. |

# Outline

Discussion

# Increasing Coverage of Jaen

- In order to get a system with full coverage, Jaen could be used with Moses as a fallback
- This would combine the precision of the rule-based system with the robustness of Moses

# Extending Jaen, by Using More Training Data

- The coverage and the quality of Jaen itself can be extended by using more training data
- Our experience is that this holds even if the training data is from a different domain
- By adding training data, we are incrementally adding rules to the system
- We still build the rules we built before, plus some more rules extracted from the new data
- Learning rules that are not applicable for the translation task does not harm or slow down the system

# Extending Jaen by Adding Templates

- We can also extend the system by adding more transfer templates
- So far, we are using 23 templates
- By adding new templates for multiword expressions, we can increase the precision

# Using Robust Parsing Techniques

- We would also like to get more from the data we have, by making the parser more robust

# Outline

Conclusion

# Conclusion

- Semantic transfer rules can be learned from parallel corpora that have been aligned in SMT phrase tables
- First strategy:
  - lemmatize the parallel corpus and use SMT aligners to create phrase tables of lemmas
  - look up the relations associated with the lemmas

  $\Rightarrow$ 127,000 rules
- Second strategy:
  - parse the parallel corpus

    $\Rightarrow$ a parallel corpus of predicates about a third the size of the full corpus
  - align the parallel corpus of predicates with SMT aligners

  $\Rightarrow$ 113,000 rules

NANYANG TECHNOLOGICAL UNIVERSITY

# Conclusion

- The two rule extraction methods complement each other:
  - About 30% of the sentences translated with one rule set are not translated by the other
  - By merging the two rule sets into one, we increased the coverage of the system to 26.6%
- A human evaluator preferred Jaen's translation to that of Moses for 58 out of a random sample of 100 translations

Francis Bond, Stephan Oepen, Eric Nichols, Dan Flickinger, Erik Velldal, and Petter Haugereid. 2011. Deep open source machine translation. *Machine Translation*, 25(2):87–105. URL http://dx.doi.org/10.1007/s10590-011-9099-4, (Special Issue on Open source Machine Translation).

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28. (Special Issue on Efficient Processing with HPSG).

Melanie Siegel and Emily M. Bender. 2002. Efficient deep processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, pages 1–8. Taipei.