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

Abstract

This paper presents two procedures for extracting transfer rules from parallel corpora for use in a rule-based Japanese-English MT system. First a "shallow" method where the parallel corpus is lemmatized before it is aligned by a phrase aligner, and then a "deep" method where the parallel corpus is parsed by deep parsers before the resulting predicates are aligned by phrase aligners. In both procedures, the phrase tables produced by the phrase aligners are used to extract semantic transfer rules. The procedures were employed on a 10 million word Japanese English parallel corpus and 190,000 semantic transfer rules were extracted.

1 Introduction

Just like syntactic and semantic information finds its way into SMT models and contribute to improved quality of SMT systems, rule-based systems benefit from the inclusion of statistical models, typically in order to rank the output of the components involved. In this paper, we present another way of improving RBMT systems with the help of SMT tools. The basic idea is to learn transfer rules from parallel texts: first creating alignments of predicates with the help of SMT phrase aligners and then extracting semantic transfer rules from these. We discuss two procedures for creating the alignments. In the first procedure the parallel corpus is lemmatized before it is aligned with two SMT phrase aligners. Then the aligned lemmas are mapped to predicates with the help of the lexicons of the parsing grammar and the generating grammar. Finally, the transfer rules are extracted from the aligned predicates. In the second procedure, the parallel corpus is initially parsed by the parsing grammar and the generating grammar. The grammars produce semantic representations, which are represented as strings of predicates. This gives us a parallel corpus of predicates, about a third of the size of the original corpus, which we feed the phrase aligners. The resulting phrase tables with aligned predicates are finally used for extraction of semantic transfer rules.

The two procedures complement each other. The first procedure is more robust and thus learns from more examples although the resulting rules are less reliable. Here we extract 127,000 semantic transfer rules. With the second procedure, which is more accurate but less robust, we extract 113,000 semantic transfer rules. The union of the procedures gives a total of 190,000 unique rules for the Japanese English MT system Jaen.

2 Semantic Transfer

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 the two HPSG grammars: JACY, which is used for the parsing of the Japanese input (Siegel and Bender, 2002) and the ERG, used for the generation of the English output (Flickinger, 2000). The third component of the system is the transfer grammar, which transfers the MRS representation produced by the Japanese grammar into an MRS representation that the English grammar can generate from: Jaen (Bond et al., 2011).

At each step of the translation process, the output

Proceedings of SSST-6, Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 67–75,
Jeju, Republic of Korea, 12 July 2012.

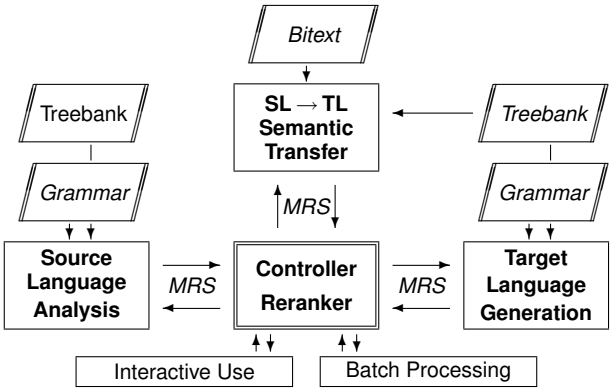

Figure 1: Architecture of the Jaen MT system.

is ranked by stochastic models. In the default configuration, only the 5 top ranked outputs at each step are kept, so the maximum number of translations is 125 (5x5x5). There is also a final reranking using a combined model (Oepen et al., 2007).

The architecture of the MT system is illustrated in Figure 1, where the contribution of the transfer rule extraction from parallel corpora is depicted by the arrow going from Bitext to Semantic Transfer.

Most of the rules in the transfer grammar are simple predicate changing rules, like the rule for mapping the predicate "_hon_n_rel" onto the predicate "_book_v_1_rel". Other rules are more complex, and transfers many Japanese relations into many English relations. In all, there are 61 types of transfer rules, the most frequent being the rules for nouns translated into nouns (44,572), noun noun compounds translated into noun noun compounds (38,197), and noun noun compounds translated into adjective plus noun (27,679). 31 transfer rule types have less than 10 instances. The most common rule types are given in Table 1.[1]

[1]Some of the rule types are extracted by only one extraction method. This holds for the types *n_adj+n_mtr*, *n+n+n_n+n_mtr*, *n+n_n_mtr*, *pp+np_np+pp_mtr*, and *arg1+pp_arg1+pp_mtr*, *adj_pp_mtr*, and *preposition_mtr*. The lemmatized extraction method extracts rules for triple compounds *n+n+n_n+n*. This is currently not done with the semantic extraction method, since a template for a triple compound would include 8 relations (each noun also has a quantifier and there are two compound relations in between), and the number of input relations are currently limited to 5 (but can be increased). The rest of the templates are new, and they have so far only been successfully integrated with the semantic extraction method.

The transfer grammar has a core set of 1,415 hand-written transfer rules, covering function words, proper nouns, pronouns, time expressions, spatial expressions, and 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` (different components have different licenses, all are open source, mainly LGPL and MIT).

3 Two methods of rule extraction

The parallel corpus we use for rule extraction is a collection of four Japanese English parallel corpora and one bilingual dictionary. The corpora are the Tanaka Corpus (2,930,132 words: Tanaka, 2001), the Japanese Wordnet Corpus (3,355,984 words: Bond, Isahara, Uchimoto, Kuribayashi, and Kanzaki, 2010), the Japanese Wikipedia corpus (7,949,605 words),[2] and the Kyoto University Text Corpus with NICT translations (1,976,071 words: Uchimoto et al., 2004). The dictionary is Edict (3,822,642 words: Breen, 2004). The word totals include both English and Japanese words.

The corpora were divided into into development, test, and training data. The training data from the four corpora plus the bilingual dictionary was used for rule extraction. The combined corpus used for rule extraction consists of 9.6 million English words and 10.4 million Japanese words (20 million words in total).

3.1 Extraction from a lemmatized parallel corpus

In the first rule extraction procedure we extracted transfer rules directly from the surface lemmas of the parallel text. The four parallel corpora were tokenized and lemmatized, for Japanese with the MeCab morphological analyzer (Kudo et al., 2004), and for English with the Freeling analyzer (Padró et al., 2010), with MWE, quantities, dates and sentence segmentation turned off. (The bilingual dictionary was not tokenized and lemmatized, since the entries in the dictionary are lemmas).

[2]The Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles: `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`.

Rule type	Hand	Lemma	Pred	Intersect	Union	Total
noun_mtr	64	32,033	31,575	19,100	44,508	44,572
n+n_n+n_mtr	0	32,724	18,967	13,494	38,197	38,197
n+n_adj+n_mtr	0	22,777	15,406	10,504	27,679	27,679
arg12+np_arg12+np_mtr	0	9,788	1,774	618	10,944	10,944
arg1_v_mtr	22	8,325	1,031	391	8,965	8,987
pp_pp_mtr	2	146	8,584	19	8,711	8,713
adjective_mtr	27	4,914	4,034	2,183	6,765	6,792
arg12_v_mtr	50	4,720	1,846	646	5,920	5,970
n_adj+n_mtr	1	-	4,695	-	4,695	4,696
n+n_n_mtr	0	2,591	3,273	1,831	4,033	4,033
n+n+n_n+n_mtr	0	3,380	-	-	3,376	3,376
n+adj-adj-mtr	2	633	2,586	182	3,037	3,039
n_n+n_mtr	1	-	2,229	-	2,229	2,230
pp-adj_mtr	27	1,008	971	1	1,978	2,005
p+n+arg12_arg12_mtr	1	1,796	101	35	1,862	1,863
pp+np_np+pp_mtr	0	-	1,516	-	1,516	1,516
pp+arg12_arg12_mtr	0	852	62	26	888	888
arg1+pp_arg1+pp_mtr	1	-	296	-	296	297
monotonic_mtr	139	-	-	-	-	139
adj_pp_mtr	0	-	112	-	112	112
preposition_mtr	53	-	34	-	34	87
arg123_v_mtr	3	30	14	8	36	39

Table 1: Most common mtr rule types. The numbers in the Hand column show the number of hand-written rules for each type. The numbers in the Lemma column, show the number of rules extracted from the lemmatized parallel corpus. The numbers in the Pred column show the number of rules extracted from the semantic parallel corpus. The Intersect column, shows the number of intersecting rules of Lemma and Pred, and the Union column show the number of distinct rules of Lemma and Pred.

We then used MOSES (Koehn et al., 2007) and Anymalign (Lardilleux and Lepage, 2009) to align the lemmatized parallel corpus. We got two phrase tables with 10,812,423 and 5,765,262 entries, respectively. MOSES was run with the default settings, and Anymalign ran for approximately 16 hours.

We selected the entries that had (i) a translation probability, P(English|Japanese) of more than 0.1,[3] (ii) an absolute frequency of more than 1,[4] (iii) fewer than 5 lemmas on the Japanese side and fewer than 4 lemmas on the English side,[5] and (iv) lexical entries for all lemmas in Jacy for Japanese and the ERG for English. This gave us 2,183,700 Moses entries and 435,259 Anymalign entries, all phrase table entries with a relatively high probability, containing lexical items known both to the parser and the generator.

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 the parsing grammar (Jacy) and the generating grammar (ERG). Since many lemmas are ambiguous, we often ended up with many semantic alignments for each surface alignment. If a surface alignment contained 3 lemmas with two readings each, we would get 8 (2x2x2) semantic alignments. However, some of the seman-

[3]This number is set based on a manual inspection of the transfer rules produced. The output for each transfer rule template is inspected, and for some of the templates, in particular the multi-word expression templates, the threshold is set higher.

[4]The absolute frequency number can, according to Adrien Lardilleux (p.c.), be thought of as a confidence score. The larger, the more accurate and reliable the translation probabilities. 1 is the lowest score.

[5]These numbers are based on the maximal number of lemmas needed for the template matching on either side.

tic relations associated with a lemma had very rare readings. In order to filter out semantic alignments with such rare readings, we parsed the training corpus and made a list of 1-grams of the semantic relations in the highest ranked output. Only the relations that could be linked to a lemma with a probability of more than 0.2 were considered in the semantic alignment. The semantic alignments were matched against 16 templates. Six of the templates are simple one-to-one mapping templates:

1.	noun	⇒	noun
2.	adjective	⇒	adjective
3.	adjective	⇒	intransitive verb
4.	intransitive verb	⇒	intransitive verb
5.	transitive verb	⇒	transitive verb
6.	ditransitive verb	⇒	ditransitive verb

The rest of the templates have more than one lemma on the Japanese side and one or more lemmas on the English side. In all, we extracted 126,964 rules with this method. Some of these are relatively simple, such as 7 which takes a noun compound and translates it into a single noun, or 8 which takes a VP and translates it into a VP (without checking for compositionality, if it is a common pattern we will make a rule for it).

7. n+n ⇒ n

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

8. arg12+np ⇒ arg12+np_mtr

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

Other examples, such as 9 are more complex, here the rule takes a Japanese noun-adjective combination and translates it to an adjective, with the external argument in Japanese (the so-called second subject) linked to the subject of the English adjective. Even though we are applying the templates to learn rules to lemma n-grams, in the translation system these rules apply to the semantic representation, so they can apply to a wide variety of syntactic variations (we give an example of a relative clause below).

9. 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.

Given the ambiguity of the lemmas used for the extraction of transfer rules, we were forced to filter semantic relations that have a low probability in order to avoid translations that do not generalize. One consequence of this is that we were not building rules that should have been built in cases where an ambiguous lemma has one dominant reading, and one or more less frequent, but plausible, readings. Another consequence is that we were building rules where the dominant reading is used, but where a less frequent reading is correct. The method is not very precise since it is based on simple 1-gram counts, and we are not considering the context of the individual lemma. A way to improve the quality of the assignment of the relation to the lemma would be to use a tagger or a parser. However, instead of going down that path, we decided to parse the whole parallel training corpus with the parsing grammar and the generation grammar of the MT system and produce a parallel corpus of semantic relations instead of lemmas. In this way, we use the linguistic grammars as high-precision semantic taggers.

3.2 Extraction from a parallel corpus of predicates

The second rule extraction procedure is based on a parallel corpus of semantic representations, rather than lemmatized sentences. We parsed the training corpus (1,578,602 items) with the parsing grammar (Jacy) and the generation grammar (ERG) of the MT system, and got 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.

The semantic representation produced by the ERG for the sentence *The white dog barks* is given in Figure 2. The relations in the MRSs are represented in the order they appear in the analysis.[6] In the semantic parallel corpus we kept the predicates, e.g. *_the_q_rel*, *_white_a_1_rel*, and so on, but we did not keep the information about linking. For verbs, we attached information about the valency. Verbs that were analyzed as intransitive, like *bark* in Figure 2, were represented with a suffix *1x*, where *1* indicates argument 1 and *x* indicates a referential index: *_bark_v_1_rel@1x*. If a verb was analyzed as being transitive or ditransitive, this would be reflected in the suffix: *_give_v_1_rel@1x2x3x*. The item corresponding to *The white dog barks* in the semantic corpus would be *_the_q_rel _white_a_1_rel _dog_n_1_rel _bark_v_1_rel@1x*.

The resulting parallel corpus of semantic representations consists of 4,712,301 relations for Japanese and 3,806,316 relations for English. This means that the size of the semantic parallel corpus is a little more than a third of the lemmatized parallel corpus. The grammars used for parsing are deep linguistic grammars, and they do not always perform very well on out of domain data, like for example the Japanese Wikipedia corpus. One way to increase the coverage of the grammars would be to include robustness rules. This would decrease the reliability of the assignment of semantic relations, but still be more reliable than simply using 1-grams to assign the relation.

The procedure for extracting semantic transfer rules from the semantic parallel corpus is similar to the procedure for extraction from the lemmatized corpus. The major difference is that the semantic corpus is disambiguated by the grammars.

As with the lemmatized corpus, the semantic parallel corpus was aligned with MOSES and Anymalign. They produced 4,830,000 and 4,095,744 alignments respectively. Alignments with more than 5 relations on either side and with a probability of less than 0.01 were filtered out.[7] This left us with 4,898,366 alignments, which were checked against 22 rule templates.[8] This produced 112,579 rules, which is 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, which gives us a total number of unique rules of 190,356. The distribution of the rules is shown in Table 1.

Some of the more complex transfer rules types like *p+n+arg12_arg12_mtr* and *pp+arg12_arg12_mtr* were extracted in far greater numbers from the lemmatized corpus than from the corpus of semantic representations. This is partially due to the fact that the method involving the lemmatized corpus is more robust, which means that the alignments are done on 3 times as much data as the method involving the corpus of semantic predicates. Another reason is that the number of items that need to be aligned to match these kinds of multi-word templates is larger when the rules are extracted from the corpus of semantic representations. (For example, a noun relation always has a quantifier binding it, even if there is no particular word expressing the quantifier.) Since the number of items to be aligned is bigger, the chance of getting an alignment with a high probability that matches the template becomes smaller.

One of the transfer rule templates (*pp_pp_mtr*) generates many more rules with the method involving the semantic predicates than the method involving lemmas. This is because we restricted the rule to only one preposition pair (*_de_p_rel* ↔ *_by_p_means_rel*) with the lemmatized corpus method, while all preposition pairs are accepted with the semantic predicate method since the confidence in the output of this method is higher.

4 Experiment and Results

In order to compare the methods for rule extraction, we made three versions of the transfer grammar, one including only the rules extracted from the lemma-

[6] Each predicate has the character span of the corresponding word(s) attached.

[7] A manual inspection of the rules produced by the template matching showed that most of the rules produced for several of the templates were good, even with a probability as low as 0.01. For some of the templates, the threshold was set higher.

[8] The reason why the number of rule templates is higher with this extraction method, is that the confidence in the results is higher. This holds in particular for many-to-one rules, were the quality of the rules extracted with from the lemmatized corpus is quite low.

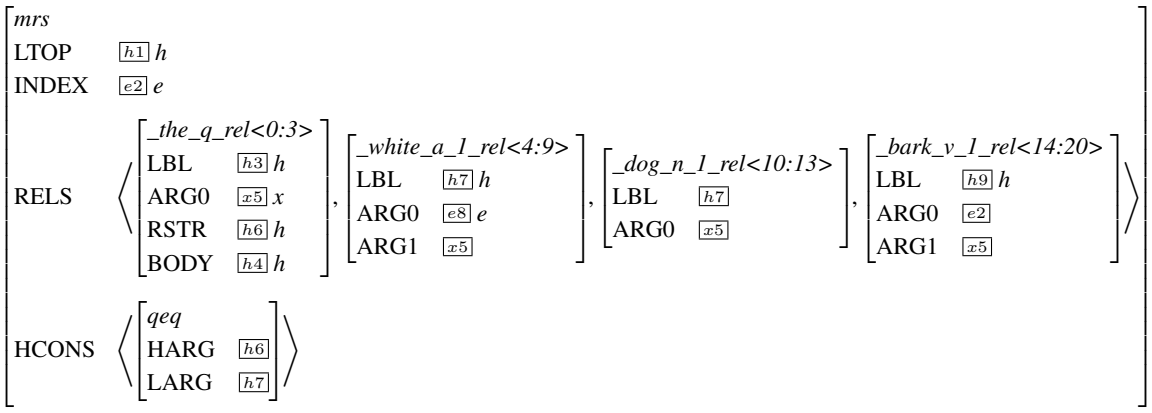

Figure 2: MRS of *The white dog barks*

tized corpus (Lemm), one including only the rules extracted from the corpus of semantic representations (Pred), and one including the union of the two (Combined). In the Combined grammar, the Lemm rules with a probability lower than 0.4 were filtered out if the input relation(s) are already translated by either handwritten rules or Pred rules since the confidence in the Lemm rules is lower.

Since the two methods for rule extraction involve different sets of templates, we also made two versions of the transfer grammar including only the 15 templates used in both Lemm and Pred. These were named LemmCore and PredCore.

The five versions of the transfer grammar were tested on sections 003, 004, and 005 of the Tanaka Corpus (4,500 test sentences), and the results are shown in Table 2. The table shows how the versions of Jaen performs with regard to parsing (constant), transfer, generation, and overall coverage. It also shows the NEVA[9] scores of the highest ranked translated sentences (NEVA), and the highest NEVA score of the 5 highest ranked translations (Oracle). The F1 is calculated based on the overall coverage and the NEVA.

The coverage of Lemm and Pred is the same; 20.8%, but Pred gets a higher NEVA score than Lemm (21.11 vs. 18.65), and the F1 score is one percent higher. When the Lemm and Pred rules are combined in Combined, the coverage is increased by almost 6%. This increase is due to the fact that the Lemm and Pred rule sets are relatively complimentary. Although the use of the Lemm and Pred transfer grammars gives the same coverage (20.8%), only 648 (14.4%) of the test sentences are translated by both systems. The NEVA score of Combined is between that of Lemm and Pred while the F1 score beats both Lemm and Pred.

When comparing the core versions of Lemm and Pred, LemmCore and PredCore, we see the same trend, namely that coverage is about the same and the NEVA score is higher when the Pred rules are used.

644 of the test sentences were translated by all versions of the transfer grammar (Lemm, Pred, and Combined). Table 3 shows how the different versions of Jaen perform on these sentences. The results show that the quality of the transfer rules extracted from the MRS parallel corpus is higher than the quality of the transfer rules based on the lemmatized parallel corpus. It also shows that there is a small decrease of quality when the rules from the lemmatized parallel corpus are added to the rules from the MRS corpus.

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

Table 3: NEVA scores of intersecting translations

The two best-performing versions of JaEn, Pred and Combined, were compared to MOSES (see Table 4 and Table 5). The BLEU scores were calculated with `multi-bleu.perl`, and the METEOR

[9]NEVA (N-gram EVAluation: Forsbom (2003)) is a modified version of BLEU.

	Parsing	Transfer	Generation	Overall	NEVA	Oracle	F1
LemmCore	3590/4500	1661/3590	930/1661	930/4500	18.65	22.99	19.61
	79.8%	46.3%	56.0%	20.7%			
Lemm	3590/4500	1674/3590	938/1674	938/4500	18.65	22.99	19.69
	79.8%	46.6%	56.0%	20.8%			
PredCore	3590/4500	1748/3590	925/1748	925/4500	20.40	24.81	20.48
	79.8%	48.7%	52.9%	20.6%			
Pred	3590/4500	1782/3589	937/1782	937/4500	**21.11**	**25.75**	20.96
	79.8%	49.7%	52.6%	20.8%			
Combined	3590/4500	2184/3589	1194/2184	1194/4500	19.77	24.00	**22.66**
	79.8%	60.9%	54.7%	**26.5**%			

Table 2: Evaluation of the Tanaka Corpus Test Data

scores were calculated with `meteor-1.3.jar` using default settings.[10] The human score is a direct comparison, an evaluator[11] was given the Japanese source, a reference translation and the output from the two systems, randomly presented as A or B. They then indicated which they preferred, or if the quality was the same (in which case each system gets 0.5). All the translations, including the reference translations, were tokenized and lower-cased. In both comparisons, MOSES gets better BLEU and METEOR scores, while the Jaen translation is preferred by the human evaluator in 58 out of 100 cases.

	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)

	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)

The two systems make different kinds of mistakes. The output of Jaen is mostly grammatical, but it may not always make sense. An example of a nonsense translation from Jaen is given in (5).[12]

(5) S: 我々は魚を生で食べる。
R: We eat fish raw.
M: We eat fish raw.
J: We eat fish in the camcorder.

Jaen sometimes gets the arguments wrong:

(6) S: 彼は大統領に選ばれた。
R: He was elected president.
M: He was elected president.
J: The president chose him.

The output of Moses on the other hand is more likely to lack words in the translation, and it is also more likely to be ungrammatical. A translation with a missing word is shown in (7).

(7) S: カーテンがゆっくり引かれた。
R: The curtains were drawn slowly.
M: The curtain was slowly.
J: The curtain was drawn slowly.

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

(8) S: 偏見は持つべきではない。
R: We shouldn't have any prejudice.
M: You should have a bias.
J: I shouldn't have prejudice.

Sometimes the Moses output is lacking so many words that it is impossible to follow the meaning:

[10] The METEOR evaluation metric differs from BLEU in that it does not only give a score for exact match, but it also gives partial scores for stem, synonym, and paraphrase matches.

[11] A Japanese lecturer at NTU, trilingual in English, Japanese and Korean, not involved in the development of this system, but with experience in Japanese/Korean MT research.

[12] The examples below are taken from the development data of the Tanaka Corpus. 'S' stands for 'Source', 'R' stands for 'Reference translation', 'M' stands for 'Moses translation,' and 'J' stands for 'Jaen translation.'

(9) S: 脳が私達の活動を支配している。
R: Our brains control our activities.
M: The brain to us.
J: The brain is controlling our activities.

Also the output of Moses is more likely to be ungrammatical, as illustrated in (10) and (11).

(10) S: 私は日本を深く愛している。
R: I have a deep love for Japan.
M: I is devoted to Japan.
J: I am deeply loving Japan.

(11) S: 彼女はタオルを固く絞った。
R: She wrung the towel dry.
M: She squeezed pressed the towel.
J: She wrung the towel hard.

5 Discussion

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. 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. Jaen has a rule pre-selection program which, before each translation task selects the applicable rules. When the system does a batch translation of 1,500 sentences, the program selects about 15,000 of the 190,000 automatically extracted rules, and only these will be loaded. Rules that have been learned but are not applicable are not used.[13]

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

The predicate alignments produced from the parallel corpus of predicates are relatively precise since the predicates are assigned by the grammars. This allows us to extract transfer rules from alignments that are given a low probability (down to 0.01) by the aligner.

We would also like to get more from the data we have, by making the parser more robust. Two approaches that have been shown to work with other grammars is making more use of morphological information (Adolphs et al., 2008) or adding robustness rules (Cramer and Zhang, 2010).

6 Conclusion

We have shown how semantic transfer rules can be learned from parallel corpora that have been aligned in SMT phrase tables. We employed two strategies. The first strategy was to lemmatize the parallel corpus and use SMT aligners to create phrase tables of lemmas. We then looked up the relations associated with the lemmas using the lexicons of the parser and generator. This gave us a phrase table of aligned relations. We were able to extract 127,000 rules by matching the aligned relations with 16 semantic transfer rule templates.

The second strategy was to parse the parallel corpus with the parsing grammar and the generating grammar of the MT system. This gave us a parallel corpus of predicates, which, because of lack of coverage of the grammars, was about a third the size of the full corpus. The parallel corpus of predicates was aligned with SMT aligners, and we got a second phrase table of aligned relations. We extracted 113,000 rules by matching the alignments against 22 rule templates. These transfer rules produced the same number of translation as the rules produced with the first strategy (20.8%), but they proved to be more precise.

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.

References

Peter Adolphs, Stephan Oepen, Ulrich Callmeier, Berthold Crysmann, Dan Flickinger, and Bernd Kiefer. 2008. Some fine points of hybrid natural language parsing. In European Language Re-

[13] The pre-selection program speeds up the system by a factor of three.

sources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1380–1387. Marrakech, Morocco.

Francis Bond, Hitoshi Isahara, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2010. Japanese WordNet 1.0. In *16th Annual Meeting of the Association for Natural Language Processing*, pages A5–3. Tokyo.

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).

James W. Breen. 2004. JMDict: a Japanese-multilingual dictionary. In *Coling 2004 Workshop on Multilingual Linguistic Resources*, pages 71–78. Geneva.

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.

Bart Cramer and Yi Zhang. 2010. Constraining robust constructions for broad-coverage parsing with precision grammars. In *Proceedings of COLING-2010*, pages 223–231. Beijing.

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).

Eva Forsbom. 2003. Training a super model look-alike: Featuring edit distance, n-gram occurrence, and one reference translation. In *In Proceedings of the Workshop on Machine Translation Evaluation. Towards Systemizing MT Evaluation*.

Philipp Koehn, Wade Shen, Marcello Federico, Nicola Bertoldi, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Ondrej Bojar, Richard Zens, Alexandra Constantin, Evan Herbst, Christine Moran, and Alexandra Birch. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Interactive Presentation Sessions*. Prague. URL `http://www.statmt.org/moses/`.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 230–237. Association for Computational Linguistics, Barcelona, Spain.

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

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, and Victoria Rosen. 2007. Towards hybrid quality-oriented machine translation. on linguistics and probabilities in MT. In *11th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2007*, pages 144–153.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. La Valletta. (`http://nlp.lsi.upc.edu/freeling`.

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.

Yasuhito Tanaka. 2001. Compilation of a multilingual parallel corpus. In *Proceedings of PACLING 2001*, pages 265–268. Kyushu. (`http://www.colips.org/afnlp/archives/pacling2001/pdf/tanaka.pdf`).

Kiyotaka Uchimoto, Yujie Zhang, Kiyoshi Sudo, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2004. Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. In Gilles Sérasset, editor, *COLING 2004 Multilingual Linguistic Resources*, pages 57–64. COLING, Geneva, Switzerland. URL `http://acl.ldc.upenn.edu/W/W04/W04-2208.bib`.