Extracting Transfer Rules for Multiword Expressions from Parallel Corpora

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

Abstract

This paper presents a procedure for extracting transfer rules for multiword expressions from parallel corpora for use in a rule based Japanese-English MT system. We show that adding the multi-word rules improves translation quality and sketch ideas for learning more such rules.

1 Introduction

Because of the great ambiguity of natural language, it is hard to translate from one language to another. To deal with this ambiguity it is common to try to add more context to a word, either in the form of multi-word translation patterns (Ikehara et al., 1991) or by adding more context to the translations in statistical MT systems (Callison-Burch et al., 2005).

In this paper, we present a way to learn large numbers of multi-word translation rules from either dictionaries or parallel text, and show their effectiveness in a semantic–transfer-based Japanese-to-English machine translation system. This research is similar to work such as Nichols et al. (2007). The novelty lies in (i) the fact that we are learning rules from parallel text and (ii) that we are learning much more complex rules.

In Section 2, we outline the semantic transfer machinery and we introduce the DELPH-IN machine translation initiative that provided the resources used in its construction. We describe in more detail how we learn new rules in Section 3, and show their effect in Section 4. We briefly discuss the results and outline future work in Section 5 and, finally, we conclude this paper in Section 6.

2 Semantic transfer

All experiments are carried out using Jaen, a semantic transfer based machine translation system (Bond et al., 2011). The system uses Minimal Recursion Semantics (MRS) as its semantic representation (Copestake et al., 2005). The transfer process takes place in three steps. First, a Japanese string is parsed with the Japanese HPSG grammar, JACY. The grammar produces an MRS with Japanese predicates. Second, the Japanese MRS is transferred into an English MRS. And finally, the English HPSG grammar ERG generates an English string from the English MRS.

At each step of the translation process, stochastic models are used to rank the output. There is a cutoff at 5, so the maximal amount of generated sentences is 125 (5x5x5). The final results are reranked using a combined model (Oepen et al., 2007).

While JACY and the ERG have been developed over many years, less effort has been put into the transfer grammar, and this component is currently the bottleneck of the system. In general, transfer rules are the bottleneck for any system, and there is a long history of trying to expand the number of transfer rules types (Matsuo et al., 1997) and tokens (Yamada et al., 2002).

In order to increase the coverage of the system (the number of words that we can translate) we build rules automatically. We look at strings that have a high probability of being a translation (identified from parallel corpora), and see if they fit a pattern defined in the transfer grammar. A very simple pattern would be that of a noun predicate being transferred as another noun predicate. The transfer rule type for this pattern is given in (1). The type makes

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 92–100,
Portland, Oregon, USA, 23 June 2011.

sure that the LBL and the ARG0 values are kept when the relation is transferred, while the PRED value is left underspecified.[1]

(1)
$$\begin{bmatrix} \textit{noun-mtr} & \\ \text{IN}|\text{RELS} & \left\langle \left[\text{LBL } \boxed{h1}, \text{ ARG0 } \boxed{x1} \right] \right\rangle \\ \text{OUT}|\text{RELS} & \left\langle \left[\text{LBL } \boxed{h1}, \text{ ARG0 } \boxed{x1} \right] \right\rangle \end{bmatrix}$$

The rule for 本 (hon) → *book*, which is a subtype of *noun-mtr*, is given in (2).

(2)
$$\begin{bmatrix} \textit{hon\_book} & \\ \text{IN}|\text{RELS} & \left\langle \left[\text{PRED \_hon\_n\_rel} \right] \right\rangle \\ \text{OUT}|\text{RELS} & \left\langle \left[\text{PRED \_book\_n\_of\_rel} \right] \right\rangle \end{bmatrix}$$

A linguistically more interesting transfer rule is that for *PP* → *Adjective* transfer (see (3)), which takes as input 3 relations (the first for the noun, the second for the postposition, and the third for the quantifier of the noun, all properly linked), and outputs one relation (for the adjective), for example *of an angle* → *angular*, to give an English-to-English example. The output adjective relation is given the same handle, index and external argument as the input postposition, so that the semantic linking with the rest of the MRS is preserved. In this way, modifiers of the PP will modify the Adjective, and so on. The use of this transfer rule is demonstrated in Section 3.1.[2]

(3)
$$\begin{bmatrix} \textit{pp-adj\_mtr} & \\ \text{IN} & \begin{bmatrix} & \left[\text{LBL } \boxed{h1}, \text{ ARG0 } \boxed{x1} \right] \\ \text{RELS} & \left\langle \begin{bmatrix} \text{LBL } \boxed{h0}, \text{ ARG0 } \boxed{e0}, \\ \text{ARG1 } \boxed{ext}, \text{ ARG2 } \boxed{x1} \end{bmatrix} \right\rangle \\ & \left[\text{ARG0 } \boxed{x1}, \text{ RSTR } \boxed{hr} \right] \\ \text{HCONS} & \left\langle \left[\text{HARG } \boxed{hr}, \text{ LARG } \boxed{h1} \right] \right\rangle \end{bmatrix} \\ \text{OUT}|\text{RELS} & \left\langle \begin{bmatrix} \text{LBL } \boxed{h0}, \text{ ARG0 } \boxed{e0}, \\ \text{ARG1 } \boxed{ext} \end{bmatrix} \right\rangle \end{bmatrix}$$

3 Procedure

We are using GIZA++ (Och and Ney, 2003) and Anymalign (Lardilleux and Lepage, 2009) to generate phrase tables from 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 et al. (2010)), the Japanese Wikipedia corpus (7,949,605),[3] and the Kyoto University Text Corpus with NICT translations (1,976,071 words: Uchimoto et al. (2004)). The dictionary is Edict, a Japanese English dictionary (3,822,642 words: Breen (2004)). The word totals include both English and Japanese words.

We divided the corpora into development, test, and training data, and extracted the transfer rules 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 and 10.4 million Japanese words). The Japanese text is tokenized and lemmatized with the MeCab morphological analyzer (Kudo et al., 2004), and the English text is tokenized and lemmatized with the Freeling analyzer (Padró et al., 2010), with MWE, quantities, dates and sentence segmentation turned off.

When applying GIZA++ and Anymalign to the lemmatized parallel corpus they produced phrase tables with 10,812,423 and 5,765,262 entries, respectively, running GIZA++ with the default MOSES settings and Anymalign for approximately 16 hours.

[1] the LBL (label) of the relation is a tag, which can be used to refer to the relation (conventionally written with an *h* for handle). The ARG0 is the index of the relation. Nouns and determiners have referential indices (conventionally written with an *x*), while adjectives and verbs have event indices (written with an *e*).

[2] The HCONS feature has as value a list of *qeq* constraints (equality modulo quantifiers), which function is to express that the label of a relation is equal to a handle in an argument position (without unifying them).

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

We filtered out the entries with an absolute frequency of 1,[4] and which had more than 4 words on the Japanese side or more than 3 words on the English side. This left us with 6,040,771 Moses entries and 3,435,176 Anymalign entries. We then checked against the Jacy lexicon on the Japanese side and the ERG lexicon on the English side to ensure that the source and the target could be parsed/generated by the MT system. Finally, we filtered out entries with a translation probability, P(English|Japanese), of less than 0.1. This gave us 1,376,456 Moses entries and 234,123 Anymalign entries. These were all phrase table entries with a relatively high probability, containing lexical items known both to the parser and the generator.

For each of these phrase table entries, we looked up the lexemes on either side in the Jacy/ERG lexicons, and represented them with the semantic predicate (and their syntactic category).[5] Ambiguous lexemes were represented with a list of predicates. We represented each possible surface rule with a list of all possible semantic predicate rules. So a possible surface rule with two (two times) ambiguous lexical items would give four possible semantic rules, a possible surface rule with three (two times) ambiguous lexical items would give eight possible semantic rules, and so on. A total of 53,960,547 possible semantic rules were created. After filtering out semantic transfer rules containing English predicates of probability less than 0.2 compared to the most frequent predicate associated with the same surface form, this number was reduced to 26,875,672.[6] Each of these rules consists of two ordered lists of semantic predicates (one for Japanese and one for English).

From these possible semantic transfer rules, we extracted transfer rules that fitted nine different patterns. We extracted 81,690 rules from the Moses entries, and 52,344 rules from the Anymalign entries. The total number of rules extracted was 97,478. (36,556 rules overlapped.) Once the rule templates have been selected and the thresholds set, the entire process is automatic.

The distribution of the extracted rules over the nine patterns is shown in Table 1.

In the first three patterns, we would simply see if the predicates had the appropriate '_n_' and '_a_' infixes in them (for nouns and adjectives respectively). 82,651 rules fitted these patterns and were accepted as transfer rules. The last six patterns were slightly more complex, and are described below.

3.1 PP → adjective

Japanese PPs headed by the postposition の *no* "of" often correspond to an adjective in English as illustrated in (4).

(4) a. 小型 の
small.size of
small

b. 音楽 の
music of
musical

In order to extract transfer rules that fit this pattern, we checked for possible semantic rules having two predicates on the Japanese side and one on the English side. The first Japanese predicate would have the infix '_n_' (be a noun), and the second would be '_no_p_rel' (the predicate of the postposition の). The sole English predicate would have the infix '_a_' (be an adjective).

3.2 PP → PP

Japanese PPs headed by the postposition で *de* "with/by/in/on/at" are, given certain NP complements, translated into English PPs headed by the preposition 'by' (meaning 'by means of') where the prepositional object does not have a determiner, as illustrated in (5).

(5) タクシー で
taxi DE
by taxi

By checking for possible semantic transfer rules fitting the pattern *noun* + *de_p_rel* on the Japanese

[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]As shown in (2), predicates reflect the syntactic category of the lexical item by means of an infix, e.g. '_n_' for *noun*.

[6]We used a profile of the English training data from the Tanaka Corpus and the Japanese Wordnet Corpus, parsed with the ERG grammar, to find the probability of each English predicate, given its surface form. For example the word *sleep* is assigned the predicate "_sleep_n_1_rel" 103 times, the predicate "_sleep_v_1_rel" 89 times, and "_sleep_v_in_rel" 2 times. Hence, semantic transfer rules containing the first two are accepted, while rules conataining the last are filtered out.

Input		Output	Moses	Anymalign	Merged rules
noun + noun	→	noun + noun	34,691	23,333	38,529
noun + noun	→	adj + noun	21,129	13,198	23,720
noun + noun	→	noun	11,824	12,864	20,402
PP	→	adj	753	372	1,022
PP	→	PP	131	24	146
verb + NP	→	verb + NP	9,985	1,926	10,256
noun + adj	→	adj	544	243	566
postp + noun + verb	→	verb	1,821	173	1,921
PP + verb	→	verb	812	211	916
Total			81,690	52,344	97,478

Table 1: Transfer rule patterns.

side, and the pattern *by_p_rel* and *noun* on the English side, we created PP to PP transfer rules where, in addition to the predicates stemming from the lexical items, the English determiner was set to the empty determiner (*udef_q_rel*). The resulting transfer rule for (5) is illustrated in (6).

(6)
$$\left[\begin{array}{l} \textit{pp\_pp\_mtr} \\ \text{IN} \quad \left\langle \begin{array}{l} [\text{PRED \_de\_p\_rel}] \\ [\text{PRED udef\_q\_rel}] \\ [\text{PRED \_takushii\_n\_rel}] \end{array} \right\rangle \\ \text{OUT} \quad \left\langle \begin{array}{l} [\text{PRED \_by\_p\_means\_rel}] \\ [\text{PRED udef\_q\_rel}] \\ [\text{PRED \_taxi\_n\_1\_rel}] \end{array} \right\rangle \end{array}\right]$$

With this particular pattern we get transfer rules which prevent us from generating all possible translations of で ('with', 'by', 'on', 'in', or 'at'), and keeps the quantifier unexpressed.

There are many other possible PP→PP patterns, such as 始め に *start in/on/at/to* "in the beginning". We started with one well known idiomatic English type, but should learn many more.

3.3 Verb + NP → Verb + NP

Japanese MWEs fitting the pattern *noun* + *object marker* (を) + *verb* usually are translated into English MWEs fitting one out of three *verb* + *NP* patterns, illustrated in (7). In (7a), the NP has an unexpressed quantifier. The English pattern in these cases will be *verb* + *noun*. In (7b), the NP has an indefinite article. The English pattern will then be *verb* + *_a_q_rel* + *noun*. And in (7c), the NP has definite article. The English pattern will then be *verb* + *_the_q_rel* + *noun*.

(7) a. テニス を し ます
tenisu wo shi masu
tennis ACC do POLITE
play tennis

b. 生計 を 立てる
seikei wo tateru
living ACC stand up
make a living

c. 責め を 負う
seme wo ou
blame ACC bear
take the blame

By adding these rules to the transfer grammar, we avoid generating sentences such as *I play the tennis* and *He took a blame*. In addition, we are able to constrain the translations of the individual words, greatly reducing the transfer search space

3.4 Noun + Adjective → Adjective

Japanese has a multiword expression pattern that is not found in English. In this pattern, *noun* + が (ga) + *adjective* usually correspond to English adjectives, as shown in (8). The pattern is an example of a double subject construction. The Japanese adjective has its subject provided by a noun, but still takes an external subject. Our transfer rule takes this external

subject and links it to the subject of the English adjective.

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

With the new rules, the transfer grammar now correctly translates (9) as *She is very intelligent.* and not *Her head is very good.*, which is the translation produced by the system without the new multiword rules. Notice the fact that the adverb modifying the adjective in Japanese is also modifying the adjective in English.

(9) 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.

Because of the flexibility of the rule based system, we can also parse, translate and generate many variants of this, including those where the adverb comes in the middle of the MWE, or where a different topic marker is used as in (10). We learn the translation equivalences from text n-grams, but then match them to complex patterns, thus taking advantage of the ease of processing of simple text, but still apply them flexibly, with the power of the deep grammar.

(10) 彼女 も 頭 が 大変 いい 。
kanojo mo atama ga taihen yoi .
She FOCUS head NOM very good .
She is also very intelligent.
She is very intelligent also.

3.5 Postp + Noun + Verb → Verb / PP + Verb → Verb

Japanese has two MWE patterns consisting of a postposition, a noun, and a verb, corresponding to a verb in English. The first is associated with the postposition の *no* "of" (see (11)), and the second is associated with the postposition に *ni* "in/on/at/to" (see (12)).

(11) 歴史 の 勉強 を する
rekishi no benkyou wo suru
history of study ACC make
study history

(12) 金魚 に えさ を やる
kingyo ni esa wo yaru
goldfish in/on/at/to feed ACC give
feed the goldfish

In (11), the postposition の *no* "of", the noun 勉強 *benkyou* "study", and the verb する *suru* "make" are translated as *study*, while in (12), the postposition に *ni* "in/on/at/to", the noun えさ *esa* "feed", and the verb やる *yaru* "give" are translated as *feed*. In both MWE patterns, the noun is marked with the object marker を *wo*. The two patterns have different analysis: In (11), which has the *no*-pattern, the postposition attaches to the noun, and the object of the postposition 歴史 *rekishi* "history" functions as a second subject of the verb. In (12), which has the *ni*-pattern, the postposition attaches to the verb, and the object of the postposition金魚 *kingyo* "goldfish" is a part of a PP. Given the different semantic representations assigned to the two MWE patterns, we have created two transfer rule types. We will have a brief look at the transfer rule type for the *no* translation pattern, illustrated in (13).[7]

(13)
$$\left[\begin{array}{l} \textit{p+n+arg12\_arg12\_mtr} \\ \text{IN} \left[\begin{array}{l} \left[\begin{array}{l}\text{LBL } \boxed{h2}, \text{ ARG0 } \textit{event},\\ \text{ARG1 } \boxed{x3}, \text{ ARG2 } \boxed{\mathbf{x2}}\end{array}\right] \\ \text{RELS} \left\langle \left[\text{LBL } \boxed{h2}, \text{ ARG0 } \boxed{x3}\right], \left[\text{ARG0 } \boxed{x3}, \text{ RSTR } \boxed{h3}\right] \right\rangle \\ \left[\begin{array}{l}\text{LBL } \boxed{h1}, \text{ ARG0 } \boxed{e1},\\ \text{ARG1 } \boxed{\mathbf{x1}}, \text{ ARG2 } \boxed{x3}\end{array}\right] \\ \text{HCONS} \left\langle \left[\text{HARG } \boxed{h3}, \text{ LARG } \boxed{h2}\right] \right\rangle \end{array}\right] \\ \text{OUT}|\text{RELS} \left\langle \left[\begin{array}{l}\text{LBL } \boxed{h1}, \text{ ARG0 } \boxed{e1},\\ \text{ARG1 } \boxed{\mathbf{x1}}, \text{ ARG2 } \boxed{\mathbf{x2}}\end{array}\right] \right\rangle \end{array}\right]$$

The input of the *p+n+arg12_arg12_mtr* transfer rule type consists of (i) a postposition relation, (ii) a noun relation, (iii) a quantifier (of the the noun),

[7] The transfer rule type for the *ni* translation pattern (*pp+arg12_arg12_mtr*) is identical to the transfer rule type for the *no* translation pattern except from the linking of the postposition in the input.

and (iv) a verb relation (listed as they appear on the RELS list). The output relation is a verb relation. Notice that the ARG1 of the input verb relation is reentered as ARG1 of the output relation (x1), and the ARG2 of the input postposition relation is reentered as ARG2 of the output relation (x2). The output relation is also given the same LBL and ARG0 value as the input verb relation. In this way, the Japanese MWE is collapsed into one English relation while semantic links to the rest of the semantic representation are maintained.

3.6 Summary

Out of the 26,875,672 possible semantic predicate rules, we extracted 97,478 rules that fitted one of the nine patterns. These rules were then included in the transfer grammar of the MT system.

4 Results

The impact of the MWE transfer rules on the MT system is illustrated in Table 2.

We compare two versions of the system, one with automatically extracted MWE rules and one without. They both have hand-written MWE and single word rules as well as automatically extracted single word rules extracted from Edict by Nichols et al. (2007).

The additional rules in + MWE are those produced in Section 3. The system was tested on held out sections of the Tanaka Corpus (sections 003 to 005). As can be seen from the results, the overall system is still very much a research prototype, the coverage being only just over 20%.

Adding the new rules gave small but consistent increases in both end-to-end coverage (19.3% to 20.1%) and translation quality (17.80% to 18.18%) measured with NEVA (Forsbom, 2003).[8]

When we look only at the 105 sentences whose translations were changed by the new rules the NEVA increased from 17.1% to 21.36%. Investigating the effects on development data, we confirmed that when the new MWE rules hit, they almost always improved the translation. However, there is still a problem of data-sparseness, we are missing instances of rule-types as well as missing many potential rule types.

As an example of the former, we have a pattern for verb+NP → verb+NP, but were unable to learn 慈悲を願う *jihi wo negau* "beg for mercy: lit. ask for compassion". We had one example in the training data, and this was not enough to get over our threshold. As an example of the latter, we do not currently learn any rules for Adverb+Verb→Verb although this is a common pattern.

5 Discussion and Further Work

The transfer rules learned here are based on co-occurrence data from corpora and a Japanese-to-English dictionary. Many of the translations learned are in fact compositional, especially for the compound noun and verb-object patterns. For example, 穴 を 掘る *ana-wo horu* "dig hole" → *dig a whole* would have been translated using existing rules. In this case the advantage of the MWE rule is that it reduces the search space, so the system does not have to consider less likely translations such as *carve the shortages*. More interestingly, 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. Currently, both Jacy and the ERG have idiom processing (based on Copestake et al., 2002), but there are few idiomatic entries in their lexicons. Bilingual data can be a good source for identifying these monolingual idioms, as it makes the non-compositionality explicit. An example of a rule that uses the current idiom machinery is the (hand-built) rule *N-ga chie-wo shiboru* "N squeezes knowledge" → *N racks N's brains*, where the subject is co-indexed with a possessive pronoun modifying the object: *I/You rack my/your brains*. Adding such expressions to the monolingual grammars simplifies the transfer rules and makes the grammars more useful for other tasks.

In this paper we only presented results for nine major multi-word transfer rule types. These were those that appeared often in the training and development data. We can straightforwardly extend this in two ways: by extending the number of rule types

[8] NEVA is an alternative to BLEU that is designed to provide a more meaningful sentence-level score for short references. It is calculated identically to BLEU, but leaving out the log and exponent calculations. We find it correlates highly with BLEU.

Version	Parse coverage	Transfer coverage	Generation coverage	Total coverage	NEVA (%)	F1
− MWE (0 rules)	3614/4500 (80.3%)	1647/3614 (45.6%)	870/1647 (52.8%)	870/4500 (19.3%)	17.80	0.185
+ adj/n (83,217 rules)	3614/4500 (80.3%)	1704/3614 (47.1%)	900/1704 (52.8%)	900/4500 (20.0%)	17.99	0.189
+ PP (1,168 rules)	3614/4500 (80.3%)	1659/3614 (45.9%)	877/1659 (52.9%)	877/4500 (19.5%)	17.88	0.187
+ verb (13,093 rules)	3614/4500 (80.3%)	1688/3614 (46.7%)	885/1688 (52.4%)	885/4500 (19.7%)	17.89	0.186
+ MWE (97,478 rules)	3614/4500 (80.3%)	1729/3614 (47.8%)	906/1729 (52.4%)	906/4500 (20.1%)	18.18	0.190

Table 2: Coverage of the MT system before and after adding the MWE transfer rules.

and by extending the number of rule instances.

Shirai et al. (2001) looked at examples in a 65,500-entry English-Japanese lexicon and estimated that there were at least 80 multi-word Japanese patterns that translated to a single word in English. As we are also going from multi-word to multi-word we expect that there will be even more than this. Currently, adding another pattern is roughly an hour's work (half to make the rule-type in the transfer engine, half to make the rule matcher in the rule builder). To add another 100 patterns is thus 6 weeks work. Almost certainly this can be speeded up by sharing information between the templates. We therefore estimate that we can greatly reduce the sparseness of rule-types with four weeks work.

To improve the coverage of rule instances, we need to look at more data, such as that aligned by Utiyama and Takahashi (2003).

Neither absolute frequency nor estimated translation probability give reliable thresholds for determining whether rules are good or not. Currently we are investigating two solutions. One is feedback cleaning, where we investigate the impact of each new rule and discard those that degrade translation quality, following the general idea of Imamura et al. (2003). The second is the more traditional 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. Ultimately, we would like to extend this approach to crowd source the decisions. There are currently two very successful online collaborative Japanese-English projects (Edict and Tatoeba, producing lexical entries and multilingual examples respectively) which indicates that there is a large pool of interested knowledgeable people.

Finally, we are working in parallel to qualitatively improve the MWE rules in two ways. The first is to extend rules using semantic classes, not just words. This would mean we would need fewer rules, but each rule would be more powerful. Of course, many rules are very idiomatic and should trigger on actual lexemes, but there are many, such as 慈悲を願う *himei wo negau* "beg for mercy" which allow some variation — in this case there are at least three different verbs that are commonly used. At a lower level we need to improve our handling of orthographic variants so that a rule can match on different forms of the same word, rather than requiring several rules. We are working together with the Japanese WordNet to achieve these goals.

The second approach is to learn complex rules directly from the parallel text, in a similar way to (Jellinghaus, 2007) or (Way, 1999). This will be necessary to catch rules that our templates do not include, but it is very easy to over-fit the rules to the translation data. For this reason, we are still constraining rules with templates.

Resource Availability

The MWE expression rules made here and the machine translation system that uses them are available through an open source code repository. Installation details can be found at `http://wiki.delph-in.net/moin/LogonInstallation`. The code to make the rules is undergoing constant revision, when it settles down we intend to also add it to the repository.

6 Conclusion

This paper presented a procedure for extracting transfer rules for multiword expressions from parallel corpora for use in a rule based Japanese-English MT system. We showed that adding the multiword rules improves translation coverage (19.3% to 20.1%) and translation quality (17.8% to 18.2% NEVA). We show how we can further improve by learning even more rules.

Acknowledgments

We would like to thank the members of the LOGON, and DELPH-IN collaborations for their support and encouragement. In addition we would like to thank the developers and maintainers of the other resources we used in our project, especially JMDict, Tatoeba, Anymalign and Moses. This project was supported in part by Nanyang Technological University (through a start-up grant on "Automatically determining meaning by comparing a text to its translation").

References

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*. (Special Issue on Open source Machine Translation, to appear).

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

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *43nd Annual Meeting of the Association for Computational Linguistics: ACL-2005*.

Ann Copestake, Dan Flickinger, Carl J. Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: an introduction. *Research on Language and Computation*, 3(4):281–332. URL `http://lingo.stanford.edu/sag/papers/copestake.pdf`.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag, and Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1941–7. Las Palmas, Canary Islands.

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*.

Satoru Ikehara, Satoshi Shirai, Akio Yokoo, and Hiromi Nakaiwa. 1991. Toward an MT system without pre-editing — effects of new methods in **ALT-J/E** —. In *Third Machine Translation Summit: MT Summit III*, pages 101–106. Washington DC. URL `http://xxx.lanl.gov/abs/cmp-lg/9510008`.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 447–454. Association for Computational Linguistics, Sapporo, Japan. URL `http://www.aclweb.org/anthology/P03-1057`.

Michael Jellinghaus. 2007. *Automatic Acquisition of Semantic Transfer Rules for Machine Translation*. Master's thesis, Universität des Saarlandes.

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.

Yoshihiro Matsuo, Satoshi Shirai, Akio Yokoo, and Satoru Ikehara. 1997. Direct parse tree translation in cooperation with the transfer method. In Daniel Joneas and Harold Somers, editors, *New Methods in Language Processing*, pages 229–238. UCL Press, London.

Eric Nichols, Francis Bond, Darren Scott Appling, and Yuji Matsumoto. 2007. Combining resources for open source machine translation. In *The 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 134–142. Skövde.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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`.

Satoshi Shirai, Kazuhide Yamamoto, and Kazutaka Takao. 2001. Construction of a dictionary to translate japanese phrases into one english word. In *Proceedings of ICCPOL'2001 (19th International Conference on Computer Processing of Oriental Languages*, pages 3–8. Seoul.

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`.

Masao Utiyama and Mayumi Takahashi. 2003. English-Japanese translation alignment data. `http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html`.

Andy Way. 1999. A hybrid architecture for robust MT using LFG-DOP. *Journal of Experimental and Theoretical Artificial Intelligence*, 11. Special Issue on Memory-Based Language Processing.

Setsuo Yamada, Kenji Imamura, and Kazuhide Yamamoto. 2002. Corpus-assisted expansion of manual mt knowledge. In *Ninth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2002*, pages 199–208. Keihanna, Japan.