Combining Resources for Open Source Machine Translation

Eric Nichols,[♯] **Francis Bond,**[♮] **Darren Scott Appling,**[♭] **Yuji Matsumoto**[♯]
[♯] Graduate School of Information Science, Nara Institute of Science and Technology
{eric-n,matsu}@is.naist.jp
[♮]National Institute of Information and Communications Technology
bond@ieee.org
[♭] College of Computing, Georgia Institute of Technology
darren.scott.appling@gatech.edu

Abstract

In this paper, we present a Japanese→English machine translation system that combines rule-based and statistical translation. Our system is unique in that all of its components are freely available as open source software. We describe the development of the rule-based translation engine including transfer rule acquisition from an open bilingual dictionary. We also show how translations from both translation engines are combined through a simple ranking mechanism and compare their outputs.

1 Introduction

While there have been many advances in the field of machine translation, it is widely acknowledged that current systems do not yet produce satisfactory results. At the same time, many researchers also recognize that no single paradigm solves all of the problems necessary to achieve high coverage while maintaining fluency and accuracy in translation (Way, 1999). It is our position that translation is a problem of meaning preservation, and that deep NLP is essential in meeting goals of high quality translation.

Our ultimate aim is to have a robust, high quality and easily extensible Japanese↔English machine translation system. Current stochastic MT systems are both robust and of high quality, but only for those domains and language pairs where there is a large amount of existing parallel text. Changing the type of the text to be translated causes the quality to drop off dramatically (Paul, 2006). Quality is proportional to the log of the amount of training data (Och, 2005), which makes it hard to quickly extend a system. Rule-based systems can also produce high quality in a limited domain (Oepen et al., 2004). Further, it is relatively easy to tweak rule-based systems by the use of user dictionaries (Sukehiro et al., 2001), although these changes are limited in scope.

Our approach to producing a robust, high quality system is to concentrate on translation quality and system extensibility, without worrying so much about coverage. We are able to do this because of the availability of a robust open source statistical machine translation systems (Koehn et al., 2007). As long as we can produce a system that produces good translations for those sentences it can translate, we can fall back on the SMT system for sentences that it cannot translate.

This leaves the problem of how to build a system that is both high quality and easily extensible. To gain high quality, we accept the brittleness of a rule-based semantic transfer system. In particular, by using a precise grammar in generation we ensure that the output is (almost always) grammatical. Rule types are hand-made. As far as possible we share types with the Norwegian→English system developed in the LOGON project (Oepen et al., 2004). To make the system (relatively) easily extensible, we construct transfer rules instances from a plain bilingual dictionary. As far as possible, we aim to concentrate our rule building efforts on closed-class words, and then fill in the open class transfer rules by automatic conversion

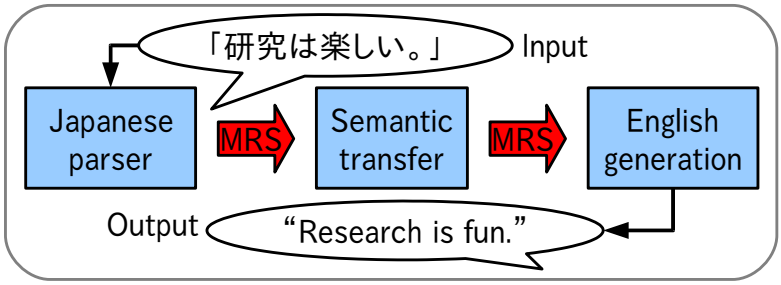

Figure 1: The Jaen machine translation architecture

of the bilingual lexicon. Finally, in future work, we will learn extra rules from aligned corpora.

In order to make this possible, we are working with an existing large scale collaborative Japanese-multilingual dictionary project (JMdict: Breen, 2004).

This paper is organized as follows. In Section 2, we present related research. In Section 3, we outline the development of our core system, and we introduce the DELPH-IN machine translation initiative that provided the resources used in its construction. In Section 4 we describe the expansion of our prototype system to target the Japanese-English section of the ATR Basic Travel Expression Corpus (BTEC*). In Section 5 we outline its integration with the Moses statistical machine translation system, and we compare translation results of these two systems in Section 6. We briefly discuss future work in Section 7, and, finally, we conclude this paper in Section 8.

2 Related Research

Recently, several large open source machine translation projects have been started. Section 3.1 describes the LOGON system, which provides many of the components for our Japanese→English system, Here, we will discuss two other large systems: OpenTrad and OpenLogos.

OpenTrad is a Spanish open source translation initiative consiting of a general MT framework and two engines (Armentano-Oller et al., 2005). The engines are Apertium, a shallow transfer system used for Castillan Spanish↔Catalan, Galician, and Portuguese, with other languages recently added, including English and French. There is also a structural transfer system used for Castillan Spanish↔Basque. Both systems share components (tokeniser, deformatter, reformatter, etc.) and are released under the GPL.

OpenLogos[1] is a 30 year-old commercial transfer system (Scott, 2003) that has recently been released as open source. It can translate from German or English into a number of languages including French, Italian, Spanish, and Portuguese. The system is released under a dual license (commercial/GPL).

Our project is much smaller than either of these, still being closer to its research roots.

3 Japanese→English RBMT with DELPH-IN

The first version of this system is described in detail in Bond et al. (2005). The architecture of our Japanese→English system (hereafter referred to as “Jaen”) is semantic transfer via rewrite rules, as shown in Figure 1. The source text is parsed using an HPSG grammar for the source language, and a semantic analysis in the form of Minimal Recursion Semantics (MRS) is produced. That semantic structure is rewritten using transfer rules into a target-language MRS structure, which is finally used to generate text from a target-language HPSG grammar.

Statistical models are used at various stages in the process. There are seperate models for analyses, transfer and generation, combined as described in Oepen et al. (2007). At each stage we prune the search space, only passing n different results (5 by default) to the next stage.

Although we mainly discuss Jaen in this paper, we have also built a reverse system, Enja, using the same components.

[1] http://logos-os.dfki.de/

3.1 System Components

The grammars and processing systems we use are all being developed within the DELPH-IN [2] project (Deep Linguistic Processing with HPSG Initiative) and are available for download. The lexicon is from an unconnected project (JMdict [3]).

3.1.1 Processing Engines

Jaen uses the LKB (Copestake, 2002) for both parsing and generation. The entire source is released under a very open license, essentially the same as the MIT License. The transfer engine is the MRS rewrite translation engine from the LOGON [4] Norwegian→English MT (Oepen et al., 2004), which is integrated with the LKB.

3.1.2 Grammars

We use HPSG-based grammars of Japanese and English, also from the DELPH-IN project (JACY; Siegel (2000) and the English Resource Grammar (ERG; Flickinger (2000)). Both grammars were originally developed within the Verb*mobil* machine translation effort, but over the past few years have been used for a variety of tasks, including automatic email response and extracting onotlogies from machine readable dictionaries.

The grammars are being developed by seperate groups of researchers, but both are part of the Matrix multilingual grammar engineering effort (Bender et al., 2002). The Matrix consists of a skeleton of grammatical and lexical types, combined with a system of semantic representation known as Minimal Recursion Semantics. The Matrix constitutes a formal backbone for a large scale grammar of, in principle, any language. New grammar resources (e.g., for Italian and Norwegian) were built using the Matrix as a 'starter-kit for grammar writing'. Three existing grammars (English, German, and Japanese) were adapted to the Matrix restrictions.

Other linguistic resources that are available as part of the DELPH-IN open-source repository include a broad-coverage grammar for German and a set of 'emerging' grammars for French, Korean, Modern Greek, Norwegian, Spanish, Swedish, and Portuguese.

[2] http://www.delph-in.net

[3] http://www.csse.monash.edu.au/∼jwb/j_jmdict.html

[4] http://www.emmtee.net

3.1.3 Lexicon

We use JMDict, the Japanese→Multilingual dictionary created by Jim Breen (Breen, 2004) to automatically acquire transfer rules. JMDict has approximately 110,000 main entries, with an additional 12,000 entries for computing and communications technology, and dictionary of over 350,000 proper names. The dictionary is primarily used by non-native speakers of Japanese as an aid to read Japanese. It is widely used, and is increasing in size at the rate of almost 1,000 entries a month (Bond and Breen, 2007).

Because the end users of the dictionary are people, the translations are often more informative than the most common translation equivalents. For example, 医者 *isha* "doctor" is translated as "medical doctor", and フランス語 *furansugo* "French" "French language", in oder to disambiguate them from "Doctor [of Philosophy]" and "French [National]" respectively. These are both correct translations, but they are not necessarily ideal for an MT system: in context, the meaning is normally clear and a translation of just "doctor" or "French" would be preferable.

3.2 Transfer Formalism

MRS (Copestake et al., 2005) is a precise, but underspecified, language-specific semantic representation. MRS structures are flat, unordered collections of elementary predications (EPs) with handles (h) indicating scopal relations, events (e), and entities (x). Figure 2 gives the MRS for the sentence "Research is fun." The sentence is a statement, and the message, `proposition_m_rel(e2)` indicates this. `tanoshii_a_rel(e2,x6)` is an event, and takes `kenkyuu_s_rel(x6)` as its subject. `noun-relation(x6)` nominalizes `kenkyuu_s_rel(x6)`, which is normally an event, turning it into an entity. MRS provides several features that make it attractive as a transfer language, such as uniform representation of pronouns, specifiers, temporal expressions, and the like over grammars. More details can be found in Flickinger et al. (2005).

3.3 Transfer Rules

As illustrated in Oepen et al. (2004), transfer rules take the form of MRS tuples:

```
研究 が 楽しい
 [ LTOP: h1
   INDEX: e2 [ e TENSE: PRES
                MOOD: INDICATIVE
                PROG: - PERF: - ]
   RELS: <
   [ PRED  proposition_m_rel
     LBL: h1
     ARG0: e2
     MARG: h3 ]
   [ PRED  "_kenkyuu_s_rel"
     LBL: h4
     ARG0: x5
     ARG1: u7
     ARG2: u6 ]
   [ PRED  "noun-relation"
     LBL: h8
     ARG0: x5
     ARG1: h9 ]
   [ PRED  proposition_m_rel
     LBL: h9
     ARG0: x5
     MARG: h10 ]
   [ PRED  udef_rel
     LBL: h11
     ARG0: x5
     RSTR: h12
     BODY: h13 ]
   [ PRED  "_tanoshii_a_rel"
     LBL: h14
     ARG0: e2
     ARG1: x5 ] >
   HCONS: < h3 qeq h14, h10 qeq h4,
            h12 qeq h8 > ]
```

Figure 2: MRS for 研究 が 楽しい *research is fun* "kenkyuu ga tanoshii"

```
[CONTEXT:]IN[!FILTER]->OUT
```

where IN(PUT) is rewritten by OUT(PUT), and the optional CONTEXT specifies relations that must be present for the rule to match, and conversely, FILTER specifies relations whose presence blocks a rule from matching. Consider the following transfer rule to translate 言語 *gengo* into "language":

```
gengo-language-mtr :=
[ IN.RELS < [ PRED"_gengo_n_1_rel",
                LBL #h1, ARG0 #x1 ] >,
  OUT.RELS < [ PRED"_language_n_1_rel",
                 LBL #h1, ARG0 #x1 ] > ].
```

This rule rewrites any instance of `gengo_n_1_rel` with `language_n_1_rel`. `#h1` and `#x1` indicate that the LBL and ARG0 arguments of the MRS produced must be preserved. While this may seem like a fairly easy to understand rule, we must repeat the constraint on LBL and ARG0 every time we write a rule to translate nouns. In order to avoid such redundancy in rule writing, LOGON allows the user to specify rule types that can encapsulate common patterns in rules. The above rule can be generalized to cover nouns:

```
noun_mtr := monotonic_mtr &
[IN.RELS < [ LBL #h1, ARG0 #x1 ] >,
 OUT.RELS < [ LBL #h1, ARG0 #x1 ] > ].
```

and our example rule can be rewritten as:

```
gengo-language-mtr := noun_mtr &
[ IN.RELS < [PRED "_gengo_n_1_rel" ] >,
  OUT.RELS <[PRED "_language_n_1_rel"]>].
```

The LOGON system contains a rich definition of rule types - many of which were immediately applicable to Jaen. Jaen inherited from LOGON rule types for open category lexical items such as common nouns, adjectives, and intransitive & transitive verbs. In addition, LOGON contains a number of rule types to specify rules for quantifiers, particles, and conjunctions, providing much of the framework needed to develop Jaen.

3.4 Rule Types Unique to Jaen

Here, we briefly describe a few rule types that were developed to handle linguistic phenomena unique to Japanese→English translation. In Figure 2, we see an example of the Japanese verbal noun, 研究 *kenkyuu* "research" being used as a noun. In Jaen, Japanese verbal nouns are analyzed as events, and they produce messages accordingly. When it is being used as a noun, `kenkyuu_s_rel` is wrapped with the relation noun-relation. We handle these constructions with a special rule that nominalizes the verbal noun by removing its event and the associated message and replacing them with and entity when it appears as a noun:

```
vn-n_jf := monotonic_mtr &
[ CONTEXT.RELS < [ PRED "ja:udef_rel",
                   ARG0 #x0 ] >,
  IN [RELS <[PRED "ja:noun-relation",
          LBL #h6, ARG0 #x0, ARG1 #hp],
         [PRED "ja:proposition_m_rel",
          LBL #hp, ARG0 #ep, MARG #h5 ],
         [PRED #pred, LBL #h0, ARG0 #ep ]>,
      HCONS < qeq & [HARG #h5, LARG #h0 ]>],
  OUT [RELS <[PRED #pred, LBL #h6,
             ARG0 #x0 ]>,
       HCONS < > ] ].
```

In short, this rule type removes the noun-relation and all semantic relations resulting in the

verbal noun's analysis as an event. This change makes it possible to treat verbal nouns identically to regular nouns in the rest of our transfer rules, eliminating the need to create multi-word transfer rules that have to distinguish between nouns and verbal nouns. This simplifies rule development significantly. Thus, a rule to translate 研究 as the noun "research" can now be created using the standard noun template:

```
kenkyuu_s-research_n-omtr := noun_mtr &
[IN.RELS <[PRED "_kenkyuu_s_rel"]>,
 OUT.RELS<[PRED "_research_n_1_rel"]>].
```

4 Expansion of the Core Jaen System

In this section, we describe the process in which the core Jaen system was expanded by targeting a Japanese→English corpus, and using open category transfer rules acquired from a bilingual dictionary to guide the manual development of a small number of transfer rules for the highest occurring closed class rules.

4.1 Targeting the ATR BTEC* Corpus

As development and testing data, we are currently using the ATR Basic Travel Expression Corpus as made available in the IWSLT 2006 evaluation campaign (Paul, 2006). As is indicated in its name, the BTEC* corpus consists of short spoken sentences taken from the travel domain. We selected it because is it a commonly used development set, making our results immediately comparable to a number of different systems, and because our Japanese HPSG parser can successfully analyze approximately 65% of its sentences, providing us with a good base for development. The BTEC* data supplied in the ITWSLT 2006 evaluation campaign consists of almost 40,000 aligned sentence pairs. Sentences average 10.0 words in length for Japanese and 9.2 words in length for English. There are 11,407 unique Japanese tokens and 7,225 unique English tokens.

4.2 Acquiring Open Category Transfer Rules from Bilingual Dictionaries

Nygård et al. (2006) demonstrated that it is possible to learn transfer rules for some open category lexical items using a bilingual Norwegian→English dictionary. They succeeded in acquiring over 6,000 rules for adjectives, nouns, and various combinations thereof. Their method entailed looking up the semantic relations corresponding to words in a translation pair, and matching the results using simple pattern matching to identify compatible rule types.

Our approach is an effort to generalize this approach by using rule templates to generate transfer rules from input source and target MRS structures. Template mappings are used to identify translation pairs where there is a compatible rule type that can be used to create a transfer rule. A template mapping is a tuple consisting of:

- a list of HPSG syntactic categories corresponding to the words in the source translation
- a list of HPSG syntactic categories for the target translation words; and
- the name of the rule template that can be used to construct a transfer rule

Consider the following template mapping:

```
T([noun], [adjective, noun], n-adj+n)
```

This template mapping above identifies a template that creates a rule to translate a Japanese noun into an English adjective-noun sequence.

Transfer rule generation is carried out in the following manner:

1. Look up each word from source-language translation in HPSG lexicon
   - Retrieve syntactic categories and MRS relations
   - Enumerate every possible combination for words with multiple entries
   - Refactor results into separate lists of syntactic categories and MRS relations
2. Repeat 1. for all words in target-language translation
3. Map template mappings onto source and target syntactic categories
   - Translations that match indicate existence of compatible rule template
4. Create a transfer rule by combining the rule template and lists of source and target MRS relations

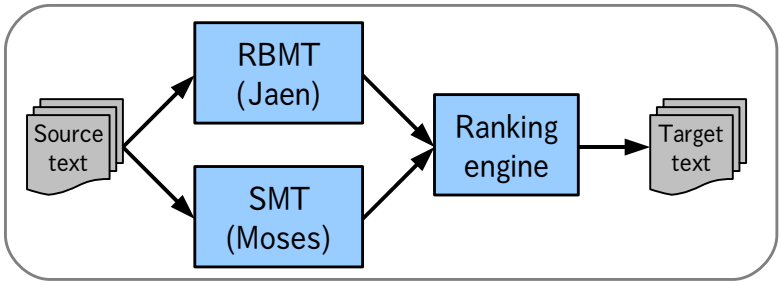

Figure 3: The combined Jaen and Moses system

Using this algorithm we can extract rules from any list of word pairs and have created rules from the EDR[5] Electronic Dictionary, Wikipedia[6] article links, and GIZA++ (Och and Ney, 2003) word alignments from the IWSLT 2006 training data. Our primary source of rules, however, is JMDict. The results of open category transfer rule acquisition from JMDict are summarized in Table 1.

4.2.1 Enhancing the Bilingual Dictionary

The resource bottleneck is a well know problem for machine translation systems. As part of our strategy to overcome it, we are consciously avoiding the creation of specialty lexicons. Instead we are reusing and contributing to an existing dictionary.

JMDict, is an online multilingual Japanese dictionary with a large user base. Users are free to edit and contribute to JMDict, assuring that errors in the lexicon are identified and corrected, and that it can be easily expanded. In order to increase the quality and coverage of JMDict and encourage other users to submit, we make our changes to the dictionary available to the community. In some cases, this means enhancing the descriptive power of JMDict's entries.

We have enhanced the JMdict lexicon in two ways (Bond and Breen, 2007). The first is an explicit distinction between transfer equivalents and explanations:

(1) 点 [てん] ...
```
<gloss g_type="equ">spot</gloss>
<gloss g_type="exp">counter for
goods or items</gloss>
```

The second is to explictly separate disjunctive entries:

(2) 田地 [でんち; でんじ]
```
<gloss>farmland</gloss>
<gloss>rice field or paddy</gloss>
→
<gloss>rice field</gloss>
<gloss>rice paddy</gloss>
```

These two extensions make it possible to produce transfer rules only for those entries which are true translations.

4.3 Handcrafting Closed Category Transfer Rules

In order to decide which semantic relations to write transfer rules for by hand, we used the automatically acquired translation rules in the above section and attempted to translate sentences from the BTEC* corpus. Whenever a relation failed to transfer, the system would be unable to generate a translation, and an error message was produced. We counted the relations and identified the most frequently occurring closed class relations as candidates for handcrafting a transfer rule. There are currently a total of 195 handcrafted rules in our system. A list of the 10 most common untranslatable relations and glosses of the translations we created are given in Table 2.

In handcrafting transfer rules for our system, we also encountered several linguistic problems that needed to be solved in order to achieve high-quality translation results, the most interesting of which was pronoun generation in English. Since our Japanese semantic analyses indicate when arguments of a predicate have been omitted, we came up with a small set of rules that checks what restrictions, if any, are placed on the omitted arguments, and we replace them with underspecified English pronouns, since the nature of the omitted argument is unknown. This leads to over-generation of pronouns, which can cause a com-

[5]http://www2.nict.go.jp/r/r312/EDR/
[6]http://www.wikipedia.org

binatorial explosion in the number of translations for sentences with multiple ellipsed pronouns. To avoid this problem, we only allow pronouns to be inserted for the first two argument slots (roughly corresponding to "subject" and "object").

Other advances made include the treatment of common modal verbs, and natural generation of determiners for negative clauses. We have spent approximately three man months on handcrafting transfer rules.

5 Combining RBMT and SMT

Our end goal is to produce a high-quality, robust machine translation system. To do so, we combine our rule based system with that of an open source statistical machine translation system as shown in Figure 3. The output of the two systems are combined, and a ranking component selects the best possible output. Our current ranking mechanism is a simple cascaded model — we select the RBMT system's output whenever possible, falling back to the SMT system otherwise.

For the fall-back system we use Moses (Koehn et al., 2007), an open source statistical machine translation system that is the result of collaboration at the 2006 John Hopkins University Workshop on Machine Translation. The main component is a beam-search decoder, but it also includes a suite of scripts that, when used together with GIZA++ and SRILM (extensible language modeling toolkit, 2002), make it possible to learn factored phrase-based translation models and carry out end-to-end translation.

We followed the instructions for creating a basic phrase-based factorless system on the Moses homepage[7]. This gave us a system that is comparable to several of that participants in the IWSLT 2006 evaluation.

6 Evaluation

We tracked our coverage on the training set of the IWSLT 2006 evaluation campaign using the rules we acquired and handcrafted as outlined in Section 4.3. Evaluation results are summarized in Table 4. We split all translation pairs into individual sentences by tokenizing on sentence ending punctuation such as "." and "?" yielding a slightly different number of translation sentences than reported in IWSLT 2006's data.

Currently, we have increased our system's coverage tenfold from a starting point of 1.3% up to 13%. In doing so, we are able to translate a large number of sentences with interesting phenomena. Our system's bottleneck is semantic transfer which succeeds over 33% of the time in comparison to the over 65% success rate of parsing and near 60% of generation.

While our currently level of coverage with Jaen makes a quantitative comparison with Moses uninformative, we give a qualitative comparison of the two systems in Figure 3. This small selection of sample translations illustrates the strengths and weaknesses of each of the systems.

As seen in translations 1, 2, and 8, both systems are capable of exactly reproducing the reference for some sentences. Our rule-based system does a better job at preserving structure in translations 4, 5, and 7. Sometimes Moses will omit words entirely; missing the modifier of "hotel" in 4 and the direct object of "see" in 5. While Jaen does not produce perfect translations in these translations, it can be argued that it preserves more of the meaning content of the source sentence.

On the other hand, Jaen often translates quite literally, with the odd-sounding "front money government" being a word-for-word rendering of the Japanese with some slight ambiguity in translating the word corresponding to "government." Sometimes this literal translation can work out well, as in translation 3, where the phrase "this vicinity" is produced in place of the SMT system and reference's use of "here".

Both Jaen and Moses can leave a Japanese word in the translation in-tact. In translation 6, an alignment was not produced for 腹部 *stomach* "fukubu", and it was left untranslated. In translation 2, there is a transliteration of the word 日本 *Japan* "nihon" that is a result of Japanese proper nouns storing transliterations of themselves in their MRS structures. This information is accessible by the English grammar during generation, and, thus "Nihon" is produced.

We feel that the strengths and weaknesses of these two translation systems complement each other; Jaen does a better job at preserving the structure of sentence, where Moses is more ca-

[7] http://www.statmt.org/wmt07/baseline.html

pable at picking up idiomatic, non-compositional translations. Combining their outputs allows us to select the best output possible.

7 Future Work

In addition to the constant work on improving the quality of the system by expanding the inventory of rules, and providing feedback to the component grammars, we are working learning rules from examples. The basic idea is to parse both the source and target and language sentences, then transfer the source and attempt to align the (possibly partial) translation with the parse of the reference translation. Aligned MRS structures can be learned as rules.

A similar approach has been taken by Jellinghaus (2007). The main differences are that they only align very similar sentences; always start the alignment from the root (the handle of the MRS); and directly align the source and target MRSes.

Another area we are working to improve is the translation ranking component of our system combiner. The current method relies on Jaen's statistical models to select the best translation, however, our current models often produce unsatisfiable results. We are exploring methods of directly applying Moses' statistical models to rank system output regardless of its origin.

8 Conclusion

We presented a Japanese→English machine translation system that contains both rule-based and statistical translation engines. All of the components in our system are open source, and excluding the BTEC* data, the resources used in our system are also freely available.

The rule-based translation engine of our system uses a rich semantic representation as a transfer language, allowing the development of powerful transfer rules that produce high-quality translations. By targeting an appropriate corpus for development, automatically acquiring rules from bilingual dictionary, and hand-crafting transfer rules to handle the most common linguistic phenomenon, we were able to greatly extend the RBMT engine's coverage.

The statistical machine translation engine provides a robust fallback for sentences the rule-based system cannot cover. A simple ranking mechanism makes it possible to immediately combine the results of our two translation engine; a better ranking model could help improve overall quality even further.

Comparison of the rule-based and statistical engines showed that their strengths and weaknesses complement each other well. We are optimistic in the potential our combined system has for generating robust and high-quality translations.

Acknowledgments

We would like to thank the members of the LOGON, Hinoki and DELPH-IN projects, especially Stephan Oepen, for their support and encouragement. In addition we would like to thank the developers of the other resources we used in our project, especially JMDict and Moses.

References

Carme Armentano-Oller, Antonio M. Corbí-Bellot, Mikel L. Forcada, Mireia Ginestí-Rosell, Boyan Bonev, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, and Felipe Sánchez-Martínez. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *Open-Source Machine Translation: Workshop at MT Summit X*, pages 23–30. Phuket.

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14. Taipei, Taiwan.

Francis Bond and James Breen. 2007. Semi-automatic refinement of the JMdict/EDICT Japanese-English dictionary. In *13th Annual Meeting of The Association for Natural Language Processing*, pages 364–367. Kyoto.

Francis Bond, Stephan Oepen, Melanie Siegel, Ann Copestake, and Dan Flickinger. 2005. Open source machine translation with DELPH-IN. In *Open-Source Machine Translation:*

Workshop at MT Summit X, pages 15–22. Phuket.

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

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

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.

SRILM An extensible language modeling toolkit. 2002. Andreas stolcke. In *International Conference on Spoken Language Processing*, volume 2, pages 901–904. Denver.

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).

Dan Flickinger, Jan Tore Lønning, Helge Dyvik, Stephan Oepen, and Francis Bond. 2005. SEM-I rational MT: Enriching deep grammars with a semantic interface for scalable machine translation. In *Machine Translation Summit X*, pages 165–172. Phuket.

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

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 Demo and Poster Sessions*, pages 177–180. Prague. URL `http://www.statmt.org/moses/`.

Lars Nygård, Jan Tore Lønning, Torbjørn Nordgård, and Stephan Oepen. 2006. Using a bi-lingual dictionary in lexical transfer. In *EAMT-2006*, pages 233–238. Oslo.

Franz Josef Och. 2005. *Statistical Machine Translation: Foundations and Recent Advances*. MT Summit, Phuket.

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

Stephan Oepen, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosèn. 2004. Som å kapp-ete med trollet? Towards MRS-based Norwegian–English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD.

Stephan Oepen, Erik Velldal, Jan Tore Løning, Paul Meurer, and Victoria Rosen. 2007. Towards hybrid quality-oriented machine translation —On linguistics and probabilities in MT—. In *Eleventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-2007*. Skövde. (this volume).

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–15. Kyoto, Japan.

Bernard (Bud) Scott. 2003. The Logos model: An historical perspective. *Machine Translation*, 18:1–72.

Melanie Siegel. 2000. HPSG analysis of Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 265 – 280. Springer, Berlin, Germany.

Tatsuya Sukehiro, Mihoko Kitamura, and Toshiki Murata. 2001. Collaborative translation environment 'Yakushite.Net'. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium: NLPRS-2001*, pages 769–770. Tokyo.

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.

Rule type	BTEC* vocabulary	Total rules	Examples
Adj→Verb	98	250	不安→to worry
Verb→Adj	239	268	有り得る→likely
Adj+Noun→Adj+Noun	478	527	渋い ワイン→white wine
Intransitive Verb	1,273	2,519	現れる→to appear
Noun→Adj.+Noun	2,262	2,787	悪玉→bad character
Adj, Adverb	2,660	3,023	青い→green
Noun+Noun→Noun	2,945	3,135	アイディア 商品→novelty
Noun→Noun+Noun	2,100	3,588	甘党→sweet tooth
Noun+Noun→Adj+Noun	3,974	4,482	暗黒 物質→dark matter
Transitive Verb	3,299	5,344	選ぶ→ to choose
Noun+Noun→Noun+Noun	5,303	7,909	操り 芝居→puppet show
Noun	14,489	16,242	字→character
Total	39,120	50,074	

Table 1: Results of automatic transfer rule acquisition from JMDict

Frequency	Semantic relation	Translation
25,927	"_ni_p_rel"	に → in, to, into
25,056	"cop_id_rel"	だ, です → to be
22,976	"_no_p_rel"	XのY → X Y, X's Y, Y of X
10,375	"_de_p_rel"	で → in, on, at, with
9,696	"rareru_rel"	∼られる → passive
9,528	"neg_v_rel"	∼ない → negation
8,848	"_exist_v_rel"	ある → to be, to have
7,627	"_kono_q_rel"	この → this
4,173	"tai_rel"	∼たい → to want to
3,588	"_hour_n_rel"	時 → time, hour

Table 2: Most frequently occuring source language relations and their hand-crafted translations

	Jaen	Moses	Reference
1	Are Japanese dogs big?	It is a big dog in Japan?	Are Japanese dogs big?
2	Where is there a Nihon embassy?	Where is the Japanese Embassy?	Where is the Japanese Embassy?
3	Is there a hotel in this vicinity?	Is there a hotel near here?	Is there a hotel around here?
4	A center hotel.	The hotel.	The Center Hotel.
5	Did you see criminals?	Did you see the?	Did you see who did it?
6	Abdomens hurt.	腹部 aches.	I have a stomach ache.
7	Please do an allergy check.	I am allergic to check, please.	I'd like to have an allergy test, please.
8	Is it a front money government?	Do I need to pay in advance?	Do I need to pay in advance?

Table 3: Sample translations from Jaen and Moses systems

IWSLT 2006 Training data results				
Parsing	28,175	/	42,699	65.98%
Transfer	9,355	/	28,175	33.20%
Generation	5,523	/	9,355	59.04%
Overall	**5,523**	/	**42,699**	**12.93**%

Table 4: Coverage for Jaen on the IWSLT 2006 traning data