2007－NL－178 (12)
2007／3／29

Automatic Transfer Rule Acquisition for Semantic Transfer Based MT

Eric Nichols ♭ Francis Bond ♯ Yuji Matsumoto ♭
♭ Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma Nara 630-0192 Japan
{eric-n,matsu}@is.naist.jp
♯ National Institute of Information and Communications Technology
bond@ieee.org

In this paper, we describe the expansion of a core semantic transfer based Japanese→English machine translation system. We show how transfer rules can be automatically acquired from a bilingual dictionary. We handcraft a small number of transfer rules to increase the system's performance on the BTEC* corpus and present preliminary evaluation of our system on the IWSLT 2006 evaluation data. Finally, we propose a method for using graph isomorphism to acquire transfer rules from parallel corpora.

意味変換翻訳のための自動規則獲得

Eric Nichols ♭ Francis Bond ♯ 松本 裕治 ♭
♭ 奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町8916-5
{eric-n,matsu}@is.naist.jp
♯ 独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
〒 619-0289 京都府相楽郡精華町光台3-5
bond@ieee.org

本稿では、基本的な日英意味変換翻訳システムの拡張について述べる。和英辞 典からの変換ルール獲得の手法を提案する。BTEC*コーパスを対象にシス テムの精度を向上させるために、少数の変換ルールを自動に作成し、 IWSLT2006年の評価データを用いて予備的なシステムの評価を行う。最後に、 グラフ類似性を用いたパラレルコーパスからの変換ルールの自動獲得手法を提案す る。

1 Introduction

While many advances have been made 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 ([14], [15]). Thus, it should come as no surprise that a number of hybrid approaches to MT have been developed. However, currently few high level transfer-based machine translation systems that use automatically acquired translation knowledge. This work aims to fill that gap.

We take the position that machine translation is fundamentally a task of meaning preservation and consider linguistic knowledge essential for solving this problem. Our goal is to produce a machine translation system that represents its translation knowledge in a format that flexible enough that transfer rules can be acquired automatically yet still be understandable to humans. This will allow users to supplement the system's rules with their own linguistic knowledge, making the system customizable to their needs.

We present a semantic transfer based Japanese→English machine translation system. Our system uses a flexible semantic representation produced by high-coverage lexical grammars as its transfer language. This approach gives our system access to the knowledge its needs to generate natural language, while at the same time, the transfer language is sufficiently abstracted away from the syntactic level to eliminate

rules with language-dependant features such as word order. Our system makes it easy to represent alignments on a linguistically-meaningful level.

In this paper, we describe the expansion of the prototype system targeting an existing corpus from the travel domain. We show how rule types originally developed for Norwegian→English were modified to handle the Japanese→English language pair. We propose a method for automatically acquiring translations from aligned sentences by treating semantic analyses as graphs and searching for isometric sub-graphs, and we show how a small-number of hand-crafted rules and rules acquired from a Japanese→English dictionary can bootstrap the process by acting as anchors in the alignment process.

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. We describe the expansion of our prototype system in to target the ATR BTEC Japanese-English corpus and present evaluation against the subset in the IWSLT 2006 testing data in Section 4. Our proposal for acquiring new rules using graph isomorphism is given in Section 5, and, finally, we conclude this paper in Section 6.

2 Related Research

There are some rule-based systems that use automatically acquired transfer rules, but such systems tend to transfer on a shallow level. The morpohological transfer based Spanish-Catalan translation system, interNOSTRUM has been the target of rule-acquisition methods with the macro-based MorphTrans[1] and more recently using statistical alignment templates in [13]. Systems that transfer on the syntactic level include Transfer Driven Machine Translation [7], which also incorporates elements of EBMT, and Yamada et. al's work expanding ALT J/E [17]. Our system's transfer rules use semantic relations that are more generalized than the rules typically employed in these systems.

The phrased-based and hierarchical approaches to statistical machine translation (SMT) are similar to the above systems using automatically acquired transfer rules. Examples include Wu's work with Stochastic Inversion Transduction Grammars [16] and Charniak et. al's syntax-based language models [3], to name a few, but syntactic-based approaches to alignment have been proposed as early as [6] and [8]. These advances make it easier for SMT systems to handle larger translation units, however, we still consider these approaches inefficient to deal with the problems of generating well-formed translation with languages as syntactically different as Japanese and English.

Recently several example-based MT systems have also been developed that use parsers to find examples. Nakazawa et. al used dependency parsers and a bilingual dictionary in [9]. Perhaps the system that bears the most resemblance to our system is LFG-DOP [15], a system that employed Data-Oriented Parsing to obtain alignments from parallel corpora annotated with Lexical Function Grammar parse trees and case-frame information. However, this approach has so far been limited to a small number of languages, and the number of translation "fragments" obtained is large enough to present search space problems.

3 Japanese→English MT with DELPH-IN

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 an target language MRS structure, which is finally used to generate text from a target language HPSG grammar.

JaEn uses the MRS rewrite translation engine from the LOGON[2] Norwegian→English MT system [11] and parsers as well and English and Japanese grammars freely available from the DELPH-IN project[3]. Bond et al. describe the development of the core system in detail and discuss the motivation behind constructing an open-source Japanese→English MT system in [1].

MRS [4] is a semantic formalism representing syntactic relations in a generalized manner. MRS structures are flat, unordered collections of elementary predications (EPs) with handles (h) indicating

[1] http://www.internostrum.com/docum/morphtrans.ps
[2] http://www.emmtee.net
[3] http://www.delph-in.net

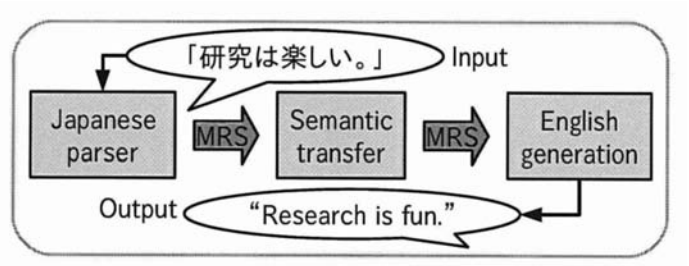

Figure 1: The JaEn machine translation architecture

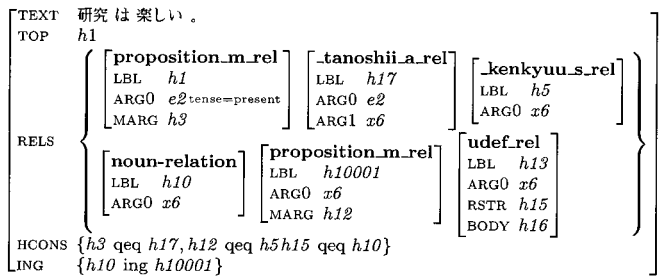

Figure 2: MRS for the sentence 「研究は楽しい。」 *"Research is fun."*

scopal relations, events (e), and entities (x). Figure 2 gives the MRS for the sentence 「研究は楽しい。」 *"Research is fun."*. The sentence is a statement, and the message, **proposition_m_rel**($e2$) indicates this. **tanoshii_a**($e2$,$x6$) 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 [5].

3.1 LOGON Transfer Rules

As illustrated in [10], transfer rules in LINGO take the form of MRS tuples:

```
[CONTEXT:]INPUT[!FILTER]→OUTPUT
```

where `INPUT` is rewritten by `OUTPUT`, 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」 into it "language":

```
gengo-language-mtr :=
[ INPUT.RELS < [ PRED "_gengo_n_1_rel", LBL #h1, ARG0 #x1 ] >,
  OUTPUT.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 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 &
[ INPUT.RELS < [ LBL #h1, ARG0 #x1 ] >,
  OUTPUT.RELS < [ LBL #h1, ARG0 #x1 ] > ].
```

and our example rule can be rewritten as:

```
gengo-language-mtr := noun_mtr &
[ INPUT.RELS < [ PRED "_gengo_n_1_rel" ] >,
  OUTPUT.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.

3.2 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」 *"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**. Thus, we need a special rule type that translates a Japanese verbal noun into an English noun when **noun-relation** is present:

```
vn_mtr := monotonic_mtr &
[ CONTEXT.RELS < [ PRED udef_rel, ARG0 #x0, RSTR #h2 ] >,
  INPUT [ RELS < [ LBL #h0, ARG0 #x0 ],
                 [ PRED proposition_m_rel,  LBL #h4, ARG0 #x0, MARG #h5 ],
                 [ PRED "noun-relation", LBL #h6, ARG0 #x0, ARG1 #h4 ] >,
          HCONS < qeq & [ HARG #h5, LARG #h0 ],
                  qeq & [ HARG #h2, LARG #h6 ] > ],
  OUTPUT [ RELS < [ LBL #h0, ARG0 #x0 ] >,
           HCONS < qeq & [ HARG #h2, LARG #h0 ] > ] ].
```

In short, this rule type removes the **noun-relation** and all semantic relations resulting in the verbal noun's analysis as an event. A rule to translate 「研究」 as the noun *"research"* can now be formulated as:

```
kenkyuu_s_rel-research_n_1_rel-omtr := vn_mtr &
[ INPUT.RELS < [ PRED "_kenkyuu_s_rel" ], ... >,
  OUTPUT.RELS < [ PRED "_research_n_1_rel" ], ... > ].
```

Another interesting case that required an additional rule type is the class of Japanese nouns that translate as adjectives in English. Examples include 「好都合」 as *"favorable"*, and 「独自」 as *"original"*. Here, we have to perform the opposite operation as in `vn_mtr` and convert from an entity to an untensed event:

```
noun_adj_mtr := monotonic_mtr &
[ INPUT.RELS < [ LBL #h, ARG0 #x ] >,
  OUTPUT.RELS < [ LBL #h,
                  ARG0 e & [ TENSE UNTENSED, MOOD INDICATIVE ],
                  ARG1 #x ] > ].
```

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. We also report coverage and BLEU score evaluation of our current system.

Rule type	BTEC* vocab.	Total rules
adjective	490	1,892
intransitive verb	618	1,617
noun	3,940	12,101
transitive verb	1,424	4,446
verbal noun	1,398	4,027

Table 1: Results of automatic transfer rule acquisition from EDICT

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 [12]. 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 approx. 61% 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

Nygaard et al. demonstrated that it is possible to learn transfer rules for some open category lexical items using a bilingual Norwegian→English dictionary in [10]. 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. it 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: (i) a list of HPSG syntactic categories corresponding to the words in the source translation; (ii) a list of HPSG syntactic categories for the target translation words; and (iii) the name of the rule template that can be used to construct a transfer rule. For example, the template mapping (["noun"], ["adjective", "noun"], "n-an") identifies a template that creates a rule to translate a Japanese noun into an English adjective-noun sequence.

We use EDICT[4] , the Japanese→English dictionary created by Jim Breen [2] to automatically acquire transfer rules. EDICT 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.

Transfer rule generation is carried out in the following manner:

1. Look up all words in source language translation in HPSG grammar
   - 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 calling the rule template with lists of MRS relations as its arguments

The results of open category transfer rule acquisition from EDICT are summarized in Table 1. We have also extracted several thousand rules of multiple words in length, but they have not been tested yet.

[4]http://www.csse.monash.edu.au/~jwb/j_edict.html

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'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 occurring handcrafted relations and their 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 119 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 causes our system to generate "I/you/we/he/she/it/they" for every pronoun inserted, so to avoid an explosion in the number of translations, we only allow pronouns to be inserted for the first two argument slots (roughly corresponding to "subject" and "object"). Other advances include the treatment of common modal verbs, and natural generation of determiners for negative clauses. We spent approximately one month on handcrafting transfer rules.

4.4 Evaluation

We evaluated our system on the development set of the IWSLT 2006 evaluation campaign using the rules we acquired and handcrafted as outlined in this section. Evaluation results are summarized in Table 3. We split all translation pairs into individual sentences on our own, yielding a slightly different number of translation sentences than reported in IWSLT 2006's data. Currently, we have increased our system's coverage from a starting point of 1.30% up to 8.04%. In doing so, we are able to translate a large number of sentences with interesting phenomena. Currently, our system's bottleneck is the semantic transfer which succeeds 18.35% of the time in comparison to the over 63% success rate of parsing and generation. We report a Bleu score of 0.265 for the sentences we translate over the entire development set using a 3-gram Bleu score and one reference translation.

5 Rule Acquisition with Graph Isomorphism

In this section, we propose an algorithm to extract transfer rules from aligned MRS structures using sub-graph isomorphism. Sub-graph isomorphism is the task of identifying identical sub-graphs present in different graphs. Graph similarity and isomorphism measures are often used in natural language processing, bioinformatics, and chemistry. Here, we will show how an MRS structure can be converted to a Directed Acyclic Graph (DAG). MRS structures are composed of EPs which are predicates with argument structures. Each of the arguments points to a handle, event, or entity identifier.

Evaluation	# sentences / total	Percentage
Japanese parsing	27,027 / 42,699	63.30%
Semantic transfer	4,959 / 27,027	18.35%
English generation	3,434 / 4,959	69.19%
Overall coverage	3,431 / 42,699	8.04%
Bleu score	**0.265**	

Table 3: Evaluation against IWSLT 2006 development data

1. Convert all EPs into nodes labeled with their relations
2. Convert all handles, events, and entities into nodes labeled with their identifiers
3. For each argument slot in an EP vertex:
   - Create a directed edge from the EP vertex to the vertex corresponding to its argument
   - Label the created edge with the name of the argument slot
4. For each QEQ relation present, create a directed edge labeled "QEQ" from the left handle to the right handle

In order to extract transfer rules, we need to compare the graphs of the MRS structures from parsing sentence-aligned Japanese and English. We propose the following method to find matching areas of the graph that can be turned into transfer rules:

1. Using our existing transfer rules, create anchors linking the Japanese and English MRS DAGs together where ever an existing transfer rule can be applied
2. Once the graphs are connected by anchors, search the graphs for sub-graphs that exhibit isomorphism
   - If these areas do not contain an anchor, they can be extracted as a translation rule if every directed edge coming into the sub-graph has the same label

   This algorithm will identify MRS regions that appear in identical arguments structures but have not been linked yet, because there are no anchors in the sub-graph region. These regions can be transformed into machine translation rules. We are currently carrying out experiments in acquiring transfer rules using this method.

6 Conclusion

We have outlined the expansion of a core Japanese→English semantic transfer based machine translation system. We showed how the semantic representation it uses as a transfer language is flexible and general enough to allow for the easy development of powerful machine translation rules capable of generating fluent translation output while preserving the meaning of the source language adequately. We described the selection of a target corpus, the automatic acquisition of open class transfer rules from a bilingual dictionary, and the handcrafting of transfer rules for the most frequent closed class entities. We also described some of the linguistic problems we encountered in adapting the framework of a Norwegian→English machine translation system to handle the Japanese→English language pair, and gave a preliminary evaluation of our system's capability. Finally, we proposed an algorithm for extracting translation rules from aligned semantic structures and showed how existing transfer rules can be used in the process.

References

[1] F. Bond, S. Oepen, M. Siegel, A. Copestake, and D. Flickinger. Open source machine translation with DELPH-IN. In *Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit*, pages 15–22, Phuket, Thailand, September 2005.

[2] J. Breen. Building an electronic japanese-english dictionary, 1995.

[3] E. Charniak, K. Knight, and K. Yamada. Syntax-based language models for statistical machine translation, 2003.

[4] A. Copestake, D. Flickinger, R. Malouf, S. Riehemann, and I. Sag. Translation using minimal recursion semantics. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, 1995.

[5] D. Flickinger, J. T. Lønning, H. Dyvik, S. Oepen, and F. Bond. SEM-I rational MT. Enriching deep grammars with a semantic interface for scalable machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 165 – 172, Phuket, Thailand, 2005.

[6] R. Grishman and M. Kosaka. Combining rationalist and empiricist approaches to machine translation. In *Proc. of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, pages 263–274, Montreal, Canada, 1992.

[7] K. Imamura, E. Sumita, and Y. Matsumoto. Automatic construction of machine translation knowledge using translation literalness. In *EACL '03: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, pages 155–162, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[8] Y. Matsumoto, H. Ishimoto, and T. Utsuro. Structural matching of parallel texts, 1993.

[9] T. Nakazawa, K. Yu, D. Kawahara, and S. Kurohashi. Example-based Machine Translation based on Deeper NLP. In *Proc. of the International Workshop on Spoken Language Translation*, pages 64–70, Kyoto, Japan, 2006.

[10] L. Nygaard, J. T. Lønning, T. Nordgård, and S. Oepen. Using a bi-lingual dictionary in lexical transfer. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo, Norway, 2006.

[11] S. Oepen, H. Dyvik, J. T. Lønning, E. Velldal, D. Beermann, J. Carroll, D. Flickinger, L. Hellan, J. B. Johannessen, P. Meurer, T. Nordgård, and V. Rosèn. 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, October 2004.

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

[13] F. Sànchez-Martínez and H. Ney. Using alignment templates to infer shallow-transfer machine translation rules. In S. P. Tapio Salakoski, Filip Ginter and T. Pahikkala, editors, *Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, pages 756–767. Springer-Verlag, August 2006. Copyright Springer-Verlag.

[14] K.-Y. Su and J.-S. Chang. Why corpus-based statistics-oriented machine translation. In *Proc. of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT*, pages 249–262, Montreal, Canada, 1992.

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

[16] D. Wu. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *IJCAI*, pages 1328–1337, 1995.

[17] S. Yamada, H. Nakaiwa, K. Ogura, and S. Ikehara. A method of automatically adapting a MT system to different domains. In *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95*, pages 303–310, Leuven, 1995.