Overcoming the customization bottleneck using example-based MT

Stephen D. Richardson, William B. Dolan, Arul Menezes, Monica Corston-Oliver[†]

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{steveri, billdol, arulm}@microsoft.com

[†]Butler Hill Group
4610 Wallingford Ave. N.
Seattle WA 98103
moco@butlerhill.com

Abstract

We describe MSR-MT, a large-scale hybrid machine translation system under development for several language pairs. This system's ability to acquire its primary translation knowledge automatically by parsing a bilingual corpus of hundreds of thousands of sentence pairs and aligning resulting logical forms demonstrates true promise for overcoming the so-called MT customization bottleneck. Trained on English and Spanish technical prose, a blind evaluation shows that MSR-MT's integration of rule-based parsers, example based processing, and statistical techniques produces translations whose quality exceeds that of uncustomized commercial MT systems in this domain.

1 Introduction

Commercially available machine translation (MT) systems have long been limited in their cost effectiveness and overall utility by the need for domain customization. Such customization typically includes identifying relevant terminology (esp. multi-word collocations), entering this terminology into system lexicons, and making additional tweaks to handle formatting and even some syntactic idiosyncrasies. One of the goals of data-driven MT research has been to overcome this customization bottleneck through automated or semi-automated extraction of translation knowledge from bilingual corpora.

To address this bottleneck, a variety of example based machine translation (EBMT) systems have been created and described in the literature. Some of these employ parsers to produce dependency structures for the sentence pairs in aligned bilingual corpora, which are then aligned to obtain transfer rules or examples (Meyers et al. 2000; Watanabe et al. 2000). Other systems extract and use examples that are represented as linear patterns of varying complexity (Brown 1999; Watanabe and Takeda 1998; Turcato et al. 1999).

For some EBMT systems, substantial collections of examples are also manually crafted or at least reviewed for correctness after being identified automatically (Watanabe et al. 2000; Brown 1999; Franz et al. 2000). The efforts that report accuracy results for fully automatic example extraction (Meyers et al. 2000; Watanabe et al. 2000) do so for very modest amounts of training data (a few thousand sentence pairs). Previous work in this area thus raises the possibility that manual review or crafting is required to obtain example bases of sufficient coverage and accuracy to be truly useful.

Other variations of EBMT systems are hybrids that integrate an EBMT component as one of multiple sources of transfer knowledge (in addition to other transfer rule or knowledge based components) used during translation (Frederking et al. 1994; Takeda et al. 1992).

To our knowledge, commercial quality MT has so far been achieved only through years of effort in creating hand-coded transfer rules. Systems whose primary source of translation knowledge comes from an automatically created example base have not been shown capable of matching or exceeding the quality of commercial systems.

This paper reports on MSR-MT, an MT system that attempts to break the customization bottleneck by exploiting example-based (and some statistical) techniques to automatically acquire its primary translation knowledge from a bilingual corpus of several million words. The system leverages the linguistic generality of existing rule-based parsers to enable broad coverage and to overcome some of the limitations on locality of context characteristic of data-driven approaches. The ability of MSR-MT to adapt automatically to a particular domain, and to produce reasonable translations for that domain, is validated through a blind assessment by human evaluators. The quality of MSR-MT's output in this one domain is shown to exceed the output quality of two highly rated (though not domain-customized) commercially available MT systems.

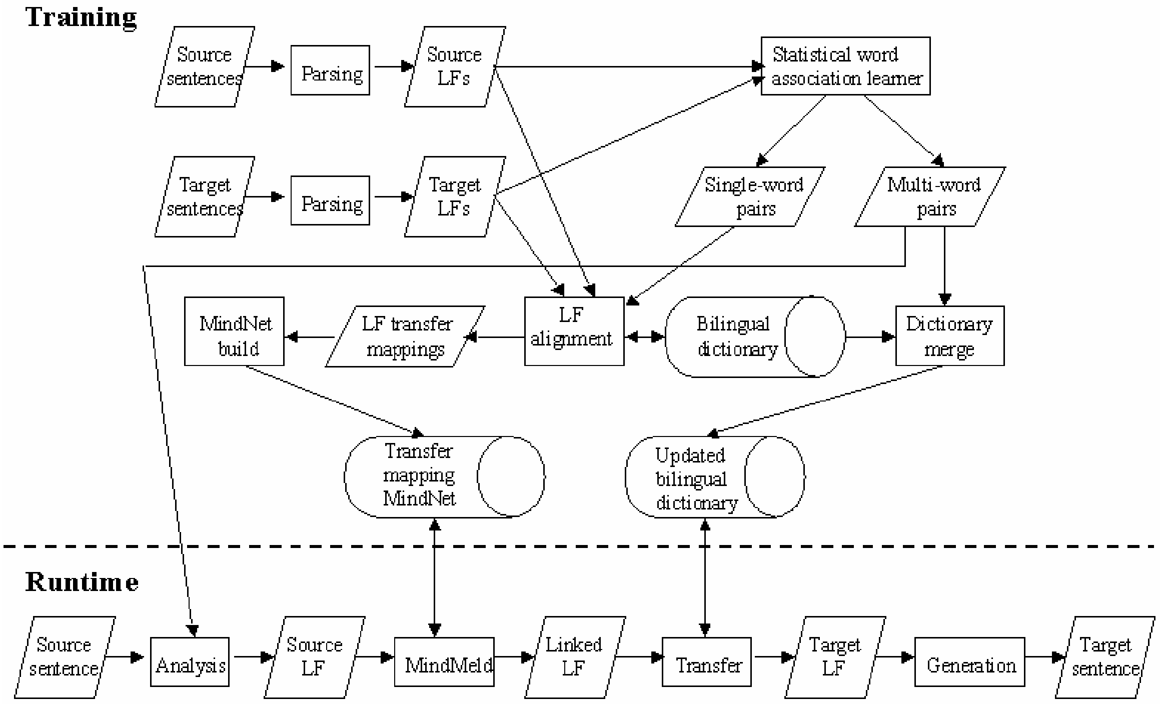

Figure 1. MSR-MT architecture.

We believe that this demonstration is the first in the literature to show that automatic training methods can produce a commercially viable level of translation quality.

2 MSR-MT

MSR-MT is a data-driven hybrid MT system, combining rule-based analysis and generation components with example-based transfer. The automatic alignment procedure used to create the example base relies on the same parser employed during analysis and also makes use of its own small set of rules for determining permissible alignments. Moderately sized bilingual dictionaries, containing only word pairs and their parts of speech, provide translation candidates for the alignment procedure and are also used as a backup source of translations during transfer. Statistical techniques supply additional translation pair candidates for alignment and identify certain multi-word terms for parsing and transfer.

The robust, broad-coverage parsers used by MSR-MT were created originally for monolingual applications and have been used in commercial grammar checkers.[1] These parsers produce a logical form (LF) representation that is compatible across multiple languages (see section 3 below). Parsers now exist for seven languages (English, French, German, Spanish, Chinese, Japanese, and Korean), and active development continues to improve their accuracy and coverage.

[1] Parsers for English, Spanish, French, and German provide linguistic analyses for the grammar checker in Microsoft Word.

Generation components are currently being developed for English, Spanish, Chinese, and Japanese. Given the automated learning techniques used to create MSR-MT transfer components, it should theoretically be possible, provided with appropriate aligned bilingual corpora, to create MT systems for any language pair for which we have the necessary parsing and generation components. In practice, we have thus far created systems that translate into English from all other languages and that translate from English to Spanish, Chinese, and Japanese. We have experimented only preliminarily with Korean and Chinese to Japanese.

Results from our Spanish-English and English-Spanish systems are reported at the end of this paper. The bilingual corpus used to produce these systems comes from Microsoft manuals and help text. The sentence alignment of this corpus is the result of using a commercial translation memory (TM) tool during the translation process.

The architecture of MSR-MT is presented in Figure 1. During the training phase, source and target sentences from the aligned bilingual corpus are parsed to produce corresponding LFs. The normalized word forms resulting from parsing are also fed to a statistical word association learner (described in section 4.1), which outputs learned single word translation pairs as well as a special class of multi-word pairs. The LFs are then aligned with the aid of translations from a bilingual dictionary and the learned single word pairs (section 4.2). Transfer mappings that result from LF alignment, in the form of linked source and target LF segments, are stored in a special repository known as MindNet (section 4.3). Additionally, the learned multi-word pairs are added to the bilingual dictionary for possible backup use during translation and to the main parsing lexicon to improve parse quality in certain cases.

At runtime, MSR-MT's analysis parses source sentences with the same parser used for source text during the training phase (section 5.1). The resulting LFs then undergo a process known as MindMeld, which matches them against the LF transfer mappings stored in MindNet (section 5.2). MindMeld also links segments of source LFs with corresponding target LF segments stored in MindNet. These target LF segments are stitched together into a single target LF during transfer, and any translations for words or phrases not found during MindMeld are searched for in the updated bilingual dictionary and inserted in the target LF (section 5.3). Generation receives the target LF as input, from which it produces a target sentence (section 5.4).

3 Logical form

MSR-MT's broad-coverage parsers produce conventional phrase structure analyses augmented with grammatical relations (Heidorn et al. 2000). Syntactic analyses undergo further processing in order to derive logical forms (LFs), which are graph structures that describe labeled dependencies among content words in the original input. LFs normalize certain syntactic alternations (e.g. active/passive) and resolve both intrasentential anaphora and long-distance dependencies.

MT has proven to be an excellent application for driving the development of our LF representation. The code that builds LFs from syntactic analyses is shared across all seven of the languages under development. This shared architecture greatly simplifies the task of aligning LF segments (section 4.2) from different languages, since superficially distinct constructions in two languages frequently collapse onto similar or identical LF representations. Even when two aligned sentences produce divergent LFs, the alignment and generation components can count on a consistent interpretation of the representational machinery used to build the two. Thus the meaning of the relation *Topic*, for instance, is consistent across all seven languages, although its surface realizations in the various languages vary dramatically.

4 Training MSR-MT

This section describes the two primary mechanisms used by MSR-MT to automatically extract translation mappings from parallel corpora and the repository in which they are stored.

4.1 Statistical learning of single word- and multi-word associations

The software domain that has been our primary research focus contains many words and phrases that are not included in our general-domain lexicons. Identifying translation correspondences between these unknown words and phrases across an aligned dataset can provide crucial lexical anchors for the alignment algorithm described in section 4.2.

In order to identify these associations, source and target text are first parsed, and normalized word forms (lemmas) are extracted. In the multi-word case, English "captoid" processing is exploited to identify sequences of related, capitalized words. Both single word and multi-word associations are iteratively hypothesized and scored by the algorithm under certain constraints until a reliable set of each is obtained.

Over the English/Spanish bilingual corpus used for the present work, 9,563 single word and 4,884 multi-word associations not already known to our system were identified using this method.

Moore (2001) describes this technique in detail, while Pinkham & Corston-Oliver (2001) describes its integration with MSR-MT and investigates its effect on translation quality.

4.2 Logical form alignment

As described in section 2, MSR-MT acquires transfer mappings by aligning pairs of LFs obtained from parsing sentence pairs in a bilingual corpus. The LF alignment algorithm first establishes tentative lexical correspondences between nodes in the source and target LFs using translation pairs from a bilingual lexicon. Our English/Spanish lexicon presently contains 88,500 translation pairs, which are then augmented with single word translations acquired using the statistical method described in section 4.1. After establishing possible correspondences, the algorithm uses a small set of alignment grammar rules to align LF nodes according to both lexical and structural considerations and to create LF transfer mappings. The final step is to filter the mappings based on the frequency of their source and target sides. Menezes & Richardson (2001) provides further details and an evaluation of the LF alignment algorithm.

The English/Spanish bilingual training corpus, consisting largely of Microsoft manuals and help text, averaged 14.1 words per English sentence. A 2.5 million word sample of English data contained almost 40K unique word forms. The data was arbitrarily split in two for use in our Spanish-English and English-Spanish systems. The first sub-corpus contains over 208,000 sentence pairs and the second over 183,000 sentence pairs. Only pairs for which both Spanish and English parsers produce complete, spanning parses and LFs are currently used for alignment. Table 1 provides the number of pairs used and the number of transfer mappings extracted and used in each case.

	Spanish-English	***English-Spanish***
Total sentence pairs	208,730	183,110
Sentence pairs used	161,606	138,280
Transfer mappings extracted	1,208,828	1,001,078
Unique, filtered mappings used	58,314	47,136

Table 1. English/Spanish transfer mappings from LF alignment

4.3 MindNet

The repository into which transfer mappings from LF alignment are stored is known as MindNet. Richardson et al. (1998) describes how MindNet began as a lexical knowledge base containing LF-like structures that were produced automatically from the definitions and example sentences in machine-readable dictionaries. Later, MindNet was generalized, becoming an architecture for a class of repositories that can store and access LFs produced for a variety of expository texts, including but not limited to dictionaries, encyclopedias, and technical manuals.

For MSR-MT, MindNet serves as the optimal example base, specifically designed to store and retrieve the linked source and target LF segments comprising the transfer mappings extracted during LF alignment. As part of daily regression testing for MSR-MT, all the sentence pairs in the combined English/Spanish corpus are parsed, the resulting spanning LFs are aligned, and a separate MindNet for each of the two directed language pairs is built from the LF transfer mappings obtained. These MindNets are about 7MB each in size and take roughly 6.5 hours each to create on a 550 Mhz PC.

5 Running MSR-MT

MSR-MT translates sentences in four processing steps, which were illustrated in Figure 1 and outlined in section 2 above. These steps are detailed using a simple example in the following sections.

5.1 Analysis

The input source sentence is parsed with the same parser used on source text during MSR-MT's training. The parser produces an LF for the sentence, as described in section 3. For the example LF in Figure 2, the Spanish input sentence is Haga clic en el botón de opción. In English, this is literally Make click in the button of option. In fluent, translated English, it is Click the option button.

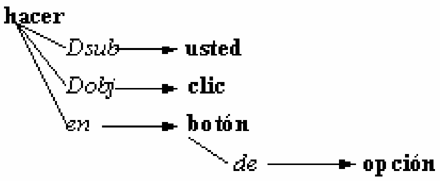

Figure 2. LF produced for *Haga clic en el botón de opción*.

5.2 MindMeld

The source LF produced by analysis is next matched by the MindMeld process to the source LF segments that are part of the transfer mappings stored in MindNet. Multiple transfer mappings may match portions of the source LF. MindMeld attempts to find the best set of matching transfer mappings by first searching for LF segments in MindNet that have matching lemmas, parts of speech, and other feature information. Larger (more specific) mappings are preferred to smaller (more general) mappings. In other words, transfers with context will be matched preferentially, but the system will fall back to the smaller transfers when no matching context is found. Among mappings of equal size, MindMeld prefers higher-frequency mappings. Mappings are also allowed to match overlapping portions of the source LF so long as they do not conflict in any way.

After an optimal set of matching transfer mappings is found, MindMeld creates ***Links*** on nodes in the source LF to copies of the corresponding target LF segments retrieved from the mappings. Figure 3 shows the source LF for the example sentence with additional ***Links*** to target LF segments. Note that ***Links*** for multi-word mappings are represented by linking the root nodes (e.g., **hacer** and **click**) of the corresponding segments, then linking an asterisk (*****) to the other source nodes participating in the multi-word mapping (e.g., **usted** and **clic**). Sublinks between corresponding individual source and target nodes of such a mapping (not shown in the figure) are also created for use during transfer.

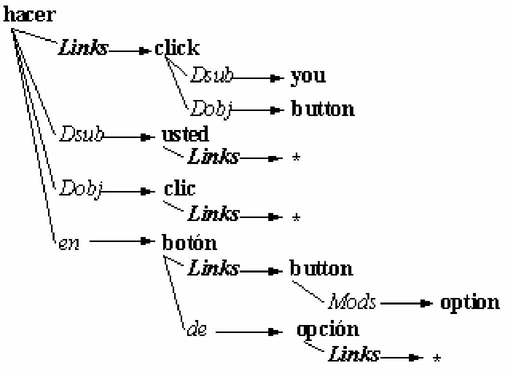

Figure 3. Linked LF for *Haga clic en el botón de opción.*

5.3 Transfer

The responsibility of transfer is to take a linked LF from MindMeld and create a target LF that will be the basis for the target translation. This is accomplished through a top down traversal of the linked LF in which the target LF segments pointed to by ***Links*** on the source LF nodes are stitched together. When stitching together LF segments from possibly complex multi-word mappings, the sublinks set by MindMeld between individual nodes are used to determine correct attachment points for modifiers, etc. Default attachment points are used if needed. Also, a very small set of simple, general, hand-coded transfer rules (currently four for English to/from Spanish) may apply to fill current (and we hope, temporary) gaps in learned transfer mappings.

In cases where no applicable transfer mapping was found during MindMeld, the nodes in the source LF and their relations are simply copied into the target LF. Default (i.e., most commonly occurring) single word translations may still be found in the MindNet for these nodes and inserted in the target LF, but if not, translations are obtained, if possible, from the same bilingual dictionary used during LF alignment.

Figure 4 shows the target LF created by transfer from the linked LF shown in Figure 3.

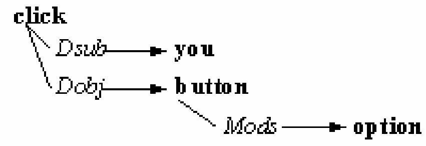

Figure 4. Target LF for *Click the option button.*

5.4 Generation

A rule-based generation component maps from the target LF to the target string (Aikawa et al. 2001). The generation components for the target languages currently handled by MSR-MT are application-independent, having been designed to apply to a range of tasks, including question answering, grammar checking, and translation. In its application to translation, generation has no information about the source language for a given input LF, working exclusively with the information passed to it by the transfer component. It uses this information, in conjunction with a monolingual (target language) dictionary to produce its output. One generic generation component is thus sufficient for each language.

In some cases, transfer produces an unmistakably “non-native” target LF. In order to correct some of the worst of these anomalies, a small set of source-language independent rules is applied prior to generation. The need for such rules reflects deficiencies in our current data-driven learning techniques during transfer.

6 Evaluating MSR-MT

In evaluating progress, we have found no effective alternative to the most obvious solution: periodic, blind human evaluations focused on translations of single sentences. The human raters used for these evaluations work for an independent agency and played no development role building the systems they test. Each language pair under active development is periodically subjected to the evaluation process described in this section.

6.1 Evaluation Methodology

For each evaluation, five to seven evaluators are asked to evaluate the same set of 200 to 250 blind test sentences. For each sentence, raters are presented with a reference sentence in the target language, which is a human translation of the corresponding source sentence. In order to maintain consistency among raters who may have different levels of fluency in the source language, raters are not shown the source sentence. Instead, they are presented with two machine-generated target translations presented in random order: one translation by the system to be evaluated (the experimental system), and another translation by a comparison system (the control system). The order of presentation of sentences is also randomized for each rater in order to eliminate any ordering effect.

Raters are asked to make a three-way choice. For each sentence, raters may choose one of the two automatically translated sentences as the better translation of the (unseen) source sentence, assuming that the reference sentence represents a perfect translation, or, they may indicate that neither of the two is better. Raters are instructed to use their best judgment about the relative importance of fluency/style and accuracy/content preservation. We chose to use this simple three-way scale in order to avoid making any *a priori* judgments about the relative importance of these parameters for subjective judgments of quality. The three-way scale also allows sentences to be rated on the same scale, regardless of whether the differences between output from system 1 and system 2 are substantial or negligible.

The scoring system is similarly simple; each judgment by a rater is represented as 1 (sentence from experimental system judged better), 0 (neither sentence judged better), or -1 (sentence from control system judged better). For each sentence, the score is the mean of all raters' judgments; for each comparison, the score is the mean of the scores of all sentences.

6.2 Evaluation results

Although work on MSR-MT encompasses a number of language pairs, we focus here on the evaluation of just two, Spanish-English and English-Spanish. Training data was held constant for each of these evaluations.

6.2.1 Spanish-English over time

Spanish-English systems	*Mean preference score (7 raters)*	*Sample size*
MSR-MT 9/00 vs. MSR-MT 12/00	0.30 ± 0.09 (at 0.95)	200 sentences
MSR-MT 12/00 vs. MSR-MT 4/01	0.28 ± 0.07 (at 0.99)	250 sentences

This table summarizes two evaluations tracking progress in MSR-MT's Spanish-English (SE) translation quality over a seven month development period. The first evaluation, with seven raters, compared a September 2000 version of the system to a December 2000 version. The second evaluation, carried out by six raters, examined progress between December 2000 and April 2001.

A score of -1 would mean that raters uniformly preferred the control system, while a score of 1 would indicate that all raters preferred the comparison system for all sentences. In each of these evaluations, all raters significantly preferred the comparison, or newer, version of MSR-MT, as reflected in the mean preference scores of 0.30 and 0.28. These numbers confirm that the system made considerable progress over a relatively short time span.

6.2.2 Spanish-English vs. alternative system

Spanish-English systems	*Mean preference score (7 raters)*	*Sample size*
MSR-MT 9/00 vs. Babelfish	-0.23 ± 0.12 (at 0.95)	200 sentences
MSR-MT 12/00 vs. Babelfish	0.11 ± 0.10 (at 0.95)	200 sentences
MSR-MT 4/01 vs. Babelfish	0.32 ± 0.11 (at .99)	250 sentences

This table summarizes our comparison of MSR-MT's Spanish-English (SE) output to the output of Babelfish (http://world.altavista.com/). Three separate evaluations were performed, in order to track MSR-MT's progress over seven months. The first two evaluations involved seven raters, while the third involved six.

The shift in the mean preference score from -0.23 to 0.32 shows clear progress against Babelfish; by the second evaluation, raters very slightly preferred MSR-MT in this domain. By April, all six raters strongly preferred MSR-MT.

6.2.3 English-Spanish vs. alternative system

English-Spanish systems	*Mean preference score (5 raters)*	*Sample size*
MSR-MT 2/01 vs. L&H	0.078 ± 0.13 (at 0.95)	250 sentences
MSR-MT 4/01 vs. L&H	0.19 ± 0.14 (at 0.99)	250 sentences

The evaluations summarized in this table compared February and April 2001 versions of MSR-MT's English-Spanish (ES) output to the output of the Lernout & Hauspie (L&H) ES system (http://officeupdate.lhsl.com/) for 250 source sentences. Five raters participated in the first evaluation, and six in the second.

The mean preference scores show that by April, MSR-MT was strongly preferred over L&H. Interestingly, though, one rater who participated in both evaluations maintained a slight but systematic preference for L&H's translations. Determining which aspects of the translations might have caused this rater to behave differently from the others is a topic for future investigation.

6.3 Discussion

These results document steady progress in the quality of MSR-MT's output over a relatively short time. By April 2001, both the SE and ES versions of the system had surpassed Babelfish in translation quality for this domain. While these versions of MSR-MT are the most fully developed, the other language pairs under development are also progressing rapidly.

In interpreting our results, it is important to keep in mind that MSR-MT has been customized to the test domain, while the Babelfish and Lernout & Hauspie systems have not.[2] This certainly affects our results, and

[2] Babelfish was chosen for these comparisons only after we experimentally compared its output to that of the related Systran system augmented with its computer domain dictionary. Surprisingly, the

means that our comparisons have a certain asymmetry. As our work progresses, we hope to evaluate MSR-MT against a quality bar that is perhaps more meaningful: the output of a commercial system that has been hand-customized for a specific domain.

The asymmetrical nature of our comparison cuts both ways, however. Customization produces better translations, and a system that can be automatically customized has an inherent advantage over one that requires laborious manual customization. Comparing an automatically-customized version of MSR-MT to a commercial system which has undergone years of hand-customization will represent a comparison that is at least as asymmetrical as those we have presented here.

We have another, more concrete, purpose in regularly evaluating our system relative to the output of systems like Babelfish and L&H: these commercial systems serve as (nearly) static benchmarks that allow us to track our own progress without reference to absolute quality.

7 Conclusions and Future Work

This paper has described MSR-MT, an EBMT system that produces MT output whose quality in a specific domain exceeds that of commercial MT systems, thus attacking head-on the customization bottleneck. This work demonstrates that automatic data-driven methods can provide commercial-quality MT.

In future work we hope to demonstrate that MSR-MT can be rapidly adapted to very different semantic domains, and that it can compete in translation quality even with commercial systems that have been hand-customized to a particular domain.

Acknowledgements

We would like to acknowledge the efforts of the MSR NLP group in carrying out this work.

References

Aikawa, T., M. Melero, L. Schwartz, and A. Wu 2001 "Multilingual natural language generation," *Proceedings of 8th European Workshop on Natural Language Generation, Toulouse.*

Brown, R. 1999. "Adding linguistic knowledge to a lexical example-based translation system," *Proceedings of TMI 99*.

Franz, A., K. Horiguchi, L. Duan, D. Ecker, E. Koontz, and K. Uchida 2000. "An integrated architecture for example-based machine translation," *Proceedings of COLING2000*.

Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domashnev, D. Attardo, D. Grannes, and R. Brown 1994. "Integrating translations from multiple sources within the Pangloss Mark III machine translation system," *Proceedings of AMTA94*.

Heidorn, G., K. Jensen, S. Richardson, and A. Viesse 2000. In R. Dale, H. Moisl and H. Somers (eds) *Handbook of Natural Language Processing*. Marcel Dekker Inc.

Meyers, A., M. Kosaka, and R. Grishman. 2000. "Chart-based transfer rule application in machine translation," *Proceedings of COLING98*.

Menezes, A. and S. Richardson 2001. "A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora," *Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001*.

Moore, R. 2001 "Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships Among Words," *Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001*.

Pinkham, J and M. Corston-Oliver 2001 "Adding Domain Specificity to an MT system," *Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001*.

Richardson, S. D., W. Dolan, and L. Vanderwende 1998. "MindNet: Acquiring and Structuring Semantic Information from Text," *Proceedings of COLING-ACL '98*, Montreal.

Takeda, K., N. Uramoto, T. Nasukawa, and T. Tsutsumi 1992. "Shalt 2—a symmetric machine translation system with conceptual transfer," *Proceedings of COLING92*.

Turcato, D., P. McFetridge, F. Popowich, and J. Toole 1999. "A unified example-based and lexicalist approach to machine translation," *Proceedings of TMI 99*.

Watanabe, W. Kurohashi, S. and E. Aramaki 2000. "Finding structural correspondences from bilingual parsed corpus for corpus-based translation," *Proceedings of COLING2000*.

Watanabe, H. and K. Takeda 1998. "A pattern-based machine translation system extended by example-based processing," *Proceedings of COLING98*.

generic SE Babelfish engine produced slightly better translations of our technical data.