Achieving commercial-quality translation with example-based methods

Stephen D. Richardson, William B. Dolan, Arul Menezes, Jessie Pinkham,

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

Abstract

We describe MSR-MT, a large-scale example-based machine translation system under development for several language pairs. Trained on aligned English-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 in this domain exceeds that of uncustomized commercial MT systems.

Keywords

Example-based machine translation, EBMT, Hybrid MT, Example-base learning, Spanish-English MT

1. Introduction

Currently available commercial machine translation (MT) systems rely on hand-coded transfer components which are both difficult and expensive to customize for a particular domain, a fact which has limited their cost-effectiveness and overall utility.

To address this bottleneck, a variety of example based machine translation (EBMT) systems have been created and described in the literature (for a comprehensive overview of work in this area, see Somers 1999). This data-driven approach relies on automated or semi-automated techniques to extract translation knowledge from bilingual corpora.

The field of EBMT has generated interesting experimental results, some of it based on quite large datasets (e.g. Frederking & Brown, 1996). What has so far been lacking, however, is an empirical demonstration that the quality bar set by existing commercial MT systems can be matched or surpassed by an MT system whose primary source of translation knowledge is an automatically-constructed example base.

This paper reports on MSR-MT (Microsoft Research Machine Translation), a translation system that relies on EBMT (and some statistical) techniques to automatically acquire its primary translation knowledge from a bilingual corpus of several million words. MSR-MT 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 quality of MSR-MT's output for the domain to which it has been customized is shown to exceed the output quality of two highly rated (though not fully domain-customized) commercial MT systems.

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. These parsers produce a logical form (LF) representation that is compatible across multiple languages (see section 3 below). Parsers now exist and are under active development for seven languages (English, French, German, Spanish, Chinese, Japanese, and Korean).

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 and a modest bilingual machine readable dictionary, 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.

The bilingual corpus used to produce these systems comes from computer manuals and help text. Sentence alignment is handled by a commercial translation memory (TM) tool.

The architecture of MSR-MT is presented in Figure 1. During training, source and target sentences from the aligned bilingual corpus are parsed to produce LFs. The normalized word forms resulting from parsing are also fed to a statistical word association learner (section 4.1), which outputs learned single word translation pairs as well as multi-word pairs. LFs are then aligned with the aid of translations from a bilingual dictionary and the learned single word pairs (section 4.2). Transfer mappings resulting 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).

At runtime, MSR-MT analyzes source sentences with the same parser used during the training phase (section 5.1). These 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

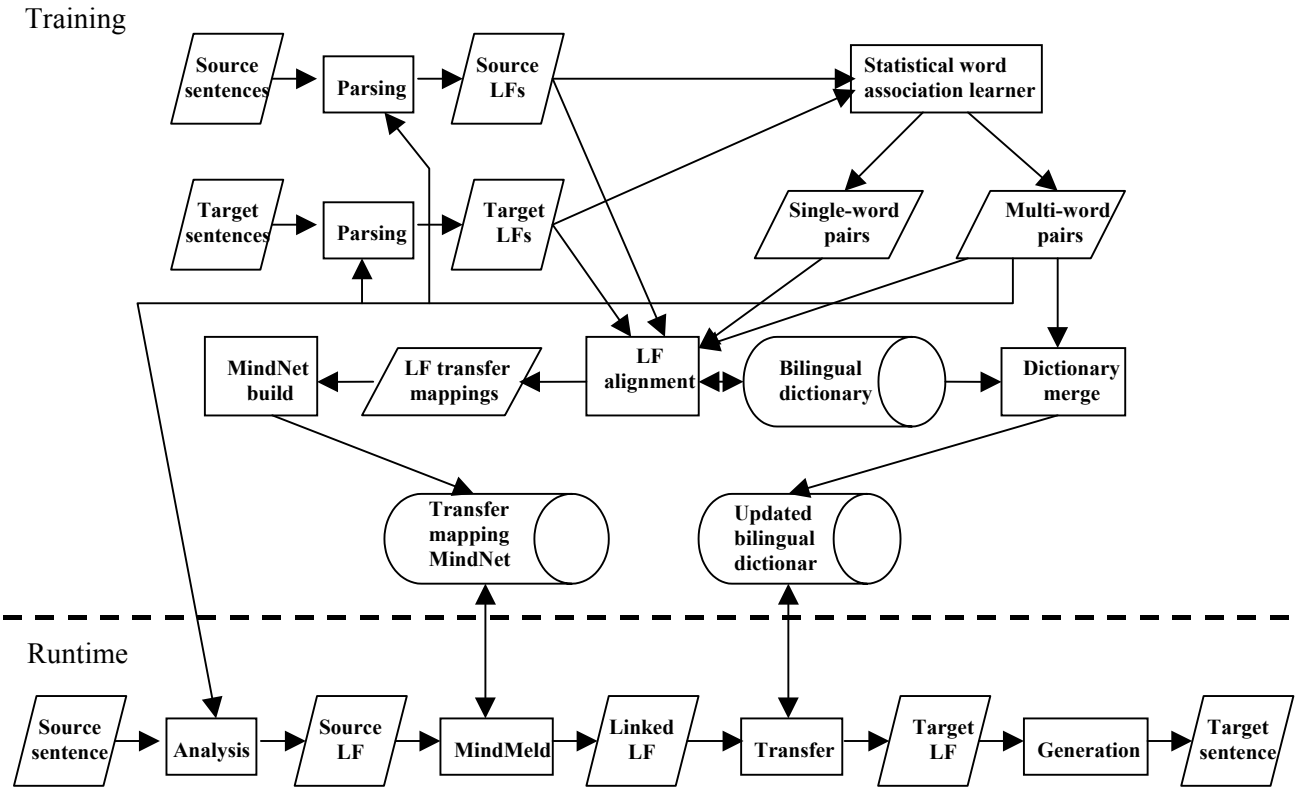

Figure 1. MSR-MT architecture.

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. Syntactic analyses undergo further processing in order to derive logical forms (LFs), which are graph structures that describe labeled dependencies among content words. LFs normalize certain syntactic alternations (e.g. active/passive) and resolve both intrasentential anaphora and long-distance dependencies.

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.

4. Training MSR-MT

This section describes the two primary mechanisms used by MSR-MT to automatically extract translation mappings from parallel corpora.

4.1 Statistical learning of word associations

In order to identify lexical and phrasal translations not contained in our general-domain lexicons, source and target text are first parsed, and normalized word forms (lemmas) are extracted. 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.

Run over our English/Spanish bilingual corpus, this technique produced a total of 9,563 new single word and 4,884 new multi-word associations.

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

4.2 Logical Form Alignment

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- and multi-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 and Richardson (2001) provide further details and an evaluation of the LF alignment algorithm.

The bilingual training corpus to which the alignment algorithm is applied consists largely of Microsoft manual and help text. The portion of the corpus used to train our Spanish-English system contains 208,000 sentence pairs, while the portion used for English-Spanish contains 183,000. English sentences in the entire corpus average 14.1 words and the vocabulary size (number of unique

word tokens) is 41,834, indicating a fairly substantial domain. Only sentence pairs for which both Spanish and English parsers produce complete, spanning parses and LFs are currently used for alignment. Table 1 summarizes the results of processing the corpus.

	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 MindNet's transition from a lexical knowledge base containing information from machine-readable dictionaries to a generalized 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.

5.1 Analysis

An LF is produced for the input source 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* (*Click the option button*, or, literally, *Make click in the button of option*).

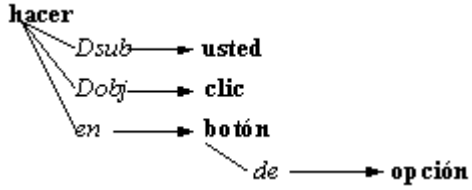

Figure 2. LF produced for *Haga clic en el botón de opción*.

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 searches for 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. Among mappings of equal size, MindMeld prefers higher-frequency mappings. Mappings may also match overlapping portions of the source LF provided 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, as shown in Figure 3. 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.

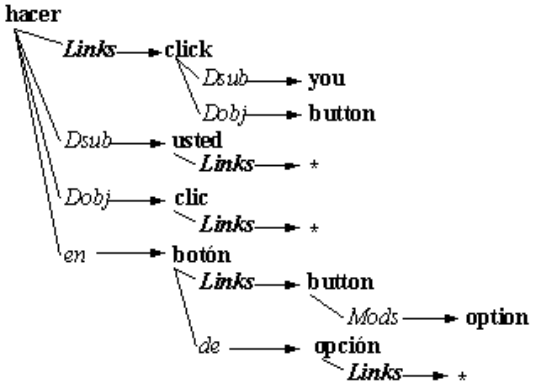

Figure 3. Linked LF for *Haga clic en el botón de opción.*

5.3 Transfer

Transfer takes a linked LF from MindMeld and creates a target LF that will be the basis for the target translation. This involves 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, 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.

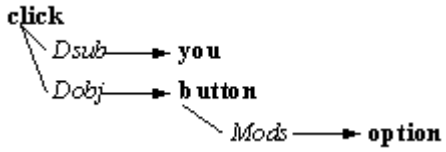

Figure 4. Target LF for *Click the option button.*

5.4 Generation

A rule-based, application-independent generation component maps from the target LF to the target string (Aikawa et al. 2001). The generation component has no information about the source language of input LFs, 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. To lessen this problem, a small set of source-language independent rules applies prior to generation. The need for such rules reflects deficiencies in our current data-driven learning techniques during transfer.

6. Evaluation

In evaluating progress, we have found no effective alternative to the most obvious solution: periodic, blind human evaluations. The human raters used for these evaluations work for an independent agency and played no development role building the systems they test.

6.1 Methodology

For each evaluation, several raters judged the same set of 200-250 sentences. These raters never saw the original source language sentence; instead, they were presented with a human translation in the target language, as well as two machine-generated alternative translations. Their task was to choose between these two alternatives, using the human translation as a reference. "Neither better" was allowed as a third choice. The order in which pairs and sentences were presented was randomized.

Raters were instructed to use their best judgment about the relative importance of fluency/style and accuracy/content preservation.

The scoring system is equally simple; each judgment by a rater was represented as 1 (System A better), 0 (neither better), or -1 (System B 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.

6.1 Results

We focus here on the evaluation of our Spanish-English and English-Spanish systems. The data used in training MSR-MT was held constant for each of these evaluations. Test sentences were not part of the training corpus, and had not been seen by system developers.

Table 2 summarizes an evaluation tracking progress in MSR-MT's Spanish-English translation quality between September 2000 and April 2001.

Spanish-English systems	*Mean preference score (7 raters)*	*Sample size*
MSR-MT 9/00 vs. MSR-MT 12/00	0.30 ± 0.10 (at 0.99)	200 sentences
MSR-MT 12/00 vs. MSR-MT 4/01	0.28 ± 0.07 (at 0.99)	250 sentences

Table 2. Spanish-English over time

A score of -1 would mean that raters uniformly preferred the older system, while a score of 1 would indicate a uniform preference for the newer one. In each of these two evaluations comparing system versions over time, the seven raters significantly preferred the newer version, as reflected in the mean preference scores of 0.30 and 0.28, both of which were significantly greater than 0 at the .99 level. This confirms that the system had made considerable progress over 7 months. The average score across all sentences for all of the seven raters was greater than 0.2, reflecting a strong trend favoring the newer version system.

Table 3 summarizes a comparison of MSR-MT's Spanish-English output to the output of Babelfish (which employs the Systran MT system and is located at http://world.altavista.com/) for a set of 200 source sentences. Three separate evaluations were performed, tracking MSR-MT's progress from September 2000 to April 2001. During this period, the mean preference score shifted from -0.23 to 0.32, showing clear progress against Babelfish. By the second evaluation, raters preferred MSR-MT (the score of 0.11 is significantly greater than 0 at the .95 level). The trend among raters is equally clear: with average scores ranging from 0.035 to 0.215, all seven raters showed at least a slight preference for MSR-MT. By the third evaluation, the aggregate (and each individual rater) showed a strong preference for MSR-MT.

Spanish-English systems	*Mean preference score (7 raters)*	*Sample size*
MSR-MT 9/00 vs. Babelfish	-0.23 ± 0.12 (at 0.99)	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

Table 3. Spanish-English MSR-MT vs. Babelfish

In another comparison, shown in Table 4, we compared February and April 2001 versions of MSR-MT's English-Spanish output to Lernout & Hauspie's equivalent system (http://officeupdate.lhsl.com/) for 250 source sentences. Five raters participated in the first evaluation, and six in the second.

In the first evaluation, where the two systems are statistically tied, the trend among the five raters is less clear. One rater preferred Lernout and Hauspie's output while four preferred MSR-MT.

The mean preference score for the April evaluation shows that MSR-MT was preferred over L&H (significantly greater than 0 at the .99 level). 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.

English-Spanish systems	*Mean preference score (5 or 6 raters)*	*Sample size*
MSR-MT 2/01 vs. L&H	0.078 ± 0.13 (not significant at 0.95)	250 sentences
MSR-MT 4/01 vs. L&H	0.19 ± 0.14 (at 0.99)	250 sentences

Table 4. English-Spanish MSR-MT vs. Lernout & Hauspie

6.2 The role of learned transfer mappings

Table 5 helps clarify the role played during translation by the transfer mappings learned during LF alignment.[1]

In processing a Spanish-English test set of 800 sentences (average length 11.1 words), MSR-MT exploited an average of 4.7 learned mappings per sentence, with the average mapping spanning 1.6 words (the greater the span, the more complex the mapping).

	Mindnet transfer mappings	Spanish-English dictionary.	Same as source
Lemmas	93.4%	4.0%	2.4%
Pronouns	8.3%	56.1%	
Prepositions	32.4%	59.0%	

Table 5. Percentage of learned transfer mappings used during translation

Validating the claim that the automatically-learned mappings are the primary source of translation knowledge in MSR-MT, 93.4% of the content words in this set were translated using these mappings, while just 4% were translated from our general Spanish-English dictionary. Pronouns are not explicitly learned by alignment, but 8.3% of pronoun translations are provided by MindNet anyway; this occurs when the pronoun is part of some larger mapping that is learned by alignment. Prepositions are crucial to translation quality, and in this sample 32.4% of the prepositions in the target translation came from learned mappings, which generally yield better translations than dictionary-derived default translations.

6.3 Discussion

These results document dramatic progress in the development of MSR-MT over a relatively short time. Both the Spanish-English and English-Spanish MSR-MT systems now clearly surpass the comparison commercial systems in translation quality for this domain. While these two language pairs 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] Until we can test the output of our system against a customized version of one of these systems, however, this asymmetry will persist.

In any case, we have a more concrete purpose in regularly evaluating our system relative to the output of systems like Babelfish and L&H: these commercial systems serve as 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 output whose quality in a specific domain exceeds that of commercial MT systems. We believe that this is the first time a system that relies primarily on an automatically created example base has been shown capable of achieving this level of translation quality.

In future work we hope to demonstrate that MSR-MT can be rapidly adapted to very different semantic domains, and we intend to evaluate it against commercial MT systems that have been hand-customized to specific domains.

8. Acknowledgments

We gratefully acknowledge the efforts of the MSR NLP group in carrying out this work, as well as the contributions of the Butler Hill Group in performing the independent evaluations described in section 6.

9. References

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

Frederking, R. and R. Brown (1996). The Pangloss-Lite machine translation system. *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec.

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*, Toulouse, France.

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*, Toulouse, France.

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,* Toulouse, France.

[1] These percentages do not add up to 100% because certain categories, including translations generated by a few simple rules, have been omitted.

[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 generic Spanish-English Babelfish engine produced slightly better translations of our technical data.

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

Somers, H. (1999). Review article: example-based machine translation. *Machine Translation* 14: 113-157.