English-Japanese Example-Based Machine Translation Using Abstract Linguistic Representations

Chris Brockett, Takako Aikawa, Anthony Aue, Arul Menezes, Chris Quirk
and Hisami Suzuki
Natural Language Processing Group, Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{chrisbkt,takakoa,anthaue,arulm,chrisq,hisamis}@microsoft.com

Abstract

This presentation describes an example-based English-Japanese machine translation system in which an abstract linguistic representation layer is used to extract and store bilingual translation knowledge, transfer patterns between languages, and generate output strings. Abstraction permits structural neutralizations that facilitate learning of translation examples across languages with radically different surface structure characteristics, and allows MT development to proceed within a largely language-independent NLP architecture. Comparative evaluation indicates that after training in a domain the English-Japanese system is statistically indistinguishable from a non-customized commercially available MT system in the same domain.

Introduction

In the wake of the pioneering work of Nagao (1984), Brown et al. (1990) and Sato and Nagao (1990), Machine Translation (MT) research has increasingly focused on the issue of how to acquire translation knowledge from aligned parallel texts. While much of this research effort has focused on acquisition of correspondences between individual lexical items or between unstructured strings of words, closer attention has begun to be paid to the learning of structured phrasal units: Yamamoto and Matsumoto (2000), for example, describe a method for automatically extracting correspondences between dependency relations in Japanese and English. Similarly, Imamura (2001a, 2001b) seeks to match corresponding Japanese and English phrases containing information about hierarchical structures, including partially completed parses.

Yamamoto and Matsumoto (2000) explicitly assume that dependency relations between words will generally be preserved across languages. However, when languages are as different as Japanese and English with respect to their syntactic and informational structures, grammatical or dependency relations may not always be preserved: the English sentence "the network failed" has quite a different grammatical structure from its Japanese translation equivalent ネットワークに障害が発生した 'a defect arose in the network.' One issue for example-based MT, then, is to capture systematic divergences through generic learning applicable to multiple language pairs.

In this presentation we describe the MSR-MT English-Japanese system, an example-based MT system that learns structured phrase-sized translation units. Unlike the systems discussed in Yamamoto and Matsumoto (2000) and Imamura (2001a, 2001b), MSR-MT places the locus of translation knowledge acquisition at a greater level of abstraction than surface relations, pushing it into a semantically-motivated layer called LOGICAL FORM (LF) (Heidorn 2000; Campbell & Suzuki 2002a, 2002b). Abstraction has the effect of neutralizing (or at least minimizing) differences in word order and syntactic structure, so that mappings between structural relations associated with lexical items can readily be acquired within a general MT architecture.

In Section 1 below, we present an overview of the characteristics of the system, with special reference to English-Japanese MT. Section 2 discusses a class of structures learned through

phrase alignment, Section 3 presents the results of comparative evaluation, and Section 4 some factors that contributed to the evaluation results. Section 5 addresses directions for future work.

1 The MSR-MT System

The MSR-MT English-Japanese system is a hybrid example-based machine translation system that employs handcrafted broad-coverage augmented phrase structure grammars for parsing, and statistical and heuristic techniques to capture translation knowlege and for transfer between languages. The parsers are general purpose: the English parser, for example, forms the core of the grammar checkers used in Microsoft Word (Heidorn 2000). The Japanese grammar utilizes much of the same codebase, but contains language-specific grammar rules and additional features owing to the need for word-breaking in Japanese (Suzuki et al. 2000; Kacmarcik et al. 2000). These parsers are robust in that if the analysis grammar fails to find an appropriate parse, it outputs a best-guess "fitted" parse.

System development is not confined to English-Japanese: MSR-MT is part of a broader natural language processing project involving three Asian languages (Japanese, Chinese, and Korean) and four European languages (English, French, German, and Spanish). Development of the MSR-MT systems proceeds more or less simultaneously across these languages and in multiple directions, including Japanese-English. The Spanish-English version of MSR-MT has been described in Richardson et al. 2001a, Richardson et al 2001b, and the reader is referred to these papers for more information concerning algorithms employed during phrase alignment. A description of the French-Spanish MT system is found in Pinkham & Smets. 2002.

1.1 Training Data

MSR-MT requires that a large corpus of aligned sentences be available as examples for training. For English-Japanese MT, the system currently trains on a corpus of approximately 596,000 pre-aligned sentence pairs. About 274,000 of these are sentence pairs extracted from Microsoft technical documentation that had been professionally translated from English into Japanese. The remaining 322,000 are sentence examples or sentence fragments extracted from electronic versions of student dictionaries.[1]

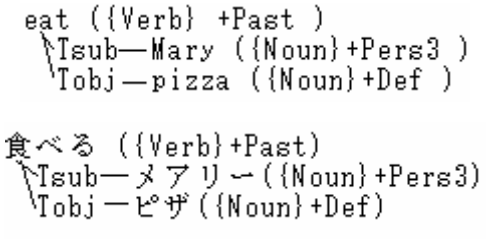

Fig. 1 Canonical English and Japanese Logical Forms

1.2 Logical Form

MSR-MT employs a post-parsing layer of semantic representation called LOGICAL FORM (LF) to handle core components of the translation process, namely acquisition and storage of translation knowledge, transfer between languages, and generation of target output. LF can be viewed as a representation of the various roles played by the content words after neutralizing word order and local morphosyntactic variation (Heidorn 2000; Campbell & Suzuki 2002a; 2002b). These can be seen in the Tsub (Typical Subject) and Tobj (Typical Object) relations in Fig. 1 in the sentence "Mary eats pizza" and its Japanese counterpart. The graphs are simplified for expository purposes.

Although our hypothesis is that equivalent sentences in two languages will tend to resemble each other at LF more than they do in the surface parse, we do not adopt a naïve reductionism that would attempt to make LFs completely identical. In Fig. 2, for example, the LFs of the quantified nouns differ in that the Japanese LF preserves the classifier, yet are similar enough that learning the mapping between the two structures is straightforward. It will be noted that since the LF for each language stores words or morphemes of that language, this level of representation is not in any sense an interlingua.

[1] *Kodansha's Basic English-Japanese Dictionary*, 1999; *Kenkyusha's New College Japanese-English Dictionary*, 4th Edition, 1995 ; and *Kenkyusha's New College English-Japanese Dictionary,* 6th Edition, 1994.

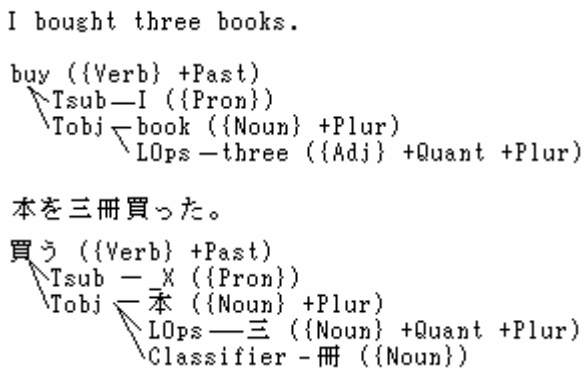

Fig. 2 Cross-Linguistic Variation in Logical Form

1.3 Mapping Logical Forms

In the training phase, MSR-MT learns transfer mappings from the sentence-aligned bilingual corpus. First, the system deploys the general-purpose parsers to analyze the English and Japanese sentence pairs and generate LFs for each sentence. In the next step, an LF alignment algorithm is used to match source language and target language LFs at the sub-sentence level.

The LF alignment algorithm first establishes tentative lexical correspondences between nodes in the source and target LFs on the basis of lexical matching over dictionary information and approximately 31,000 "word associations," that is, lexical mappings extracted from the training corpora using statistical techniques based on mutual information (Moore 2001). From these possible lexical correspondences, the algorithm uses a small grammar of (language-pair-independent) rules to align LF nodes on lexical and structural principles. The aligned LF pairs are then partitioned into smaller aligned LF segments, with individual node mappings captured in a relationship we call "sublinking." Finally, the aligned LF segments are filtered on the basis of frequency, and compiled into a database known as a Mindnet. (See Menezes & Richardson 2001 for a detailed description of this process.)

The Mindnet is a general-purpose database of semantic information (Richardson et al. 1998) that has been repurposed as the primary repository of translation information for MT applications. The process of building the Mindnet is entirely automated; there is no human vetting of candidate entries. At the end of a typical training session, 1,816,520 transfer patterns identified in the training corpus may yield 98,248 final entries in the Mindnet. Only the output of successful parses is considered for inclusion, and each mapping of LF segments must have been encountered twice in the corpus before it is incorporated into the Mindnet.

In the Mindnet, LF segments from the source language are represented as linked to the corresponding LF segment from the target languages. These can be seen in Figs. 3 and 4, discussed below in Section 2.

1.4 Transfer and Generation

At translation time, the broad-coverage source language parser processes the English input sentence, and creates a source-language LF. This LF is then checked against the Mindnet entries.[2] The best matching structures are extracted and stitched together deterministically into a new target-language "transferred LF" that is then submitted to the Japanese system for generation of the output string.

The generation module is language-specific and used for both monolingual generation and MT. In the context of MT, generation takes as input the transferred LF and converts it into a basic syntactic tree. A small set of heuristic rules preprocesses the transferred LF to "nativize" some structural differences, such as pro-drop phenomena in Japanese. A series of core generation rules then applies to the LF tree, transforming it into a Japanese sentence string. Generation rules operate on a single tree only, are application-independent and are developed in a monolingual environment (see Aikawa et al. 2001a, 2001b for further details.) Generation of inflectional morphology is also handled in this component. The generation component has no explicit knowledge of the source language.

2 Acquisition of Complex Structural Mappings

The generalization provided by LF makes it possible for MSR-MT to handle complex structural relations in cases where English and Japanese are systematically divergent. This is

[2] MSR-MT resorts to lexical lookup only when a term is not found in the Mindnet. The handcrafted dictionary is slated for replacement by purely statistically generated data.

Training Data	Translation Output
This URL provides access to public folders. この URL でパブリック フォルダにアクセスできます。	This computer provides access to the internet. このコンピュータでインターネットへアクセスできます。

Table 1. Sample Input and Output

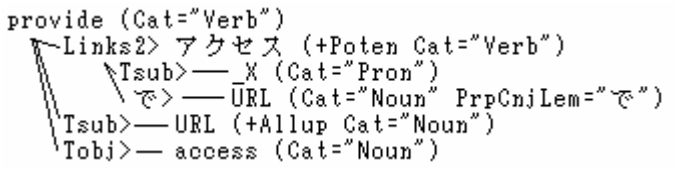

Fig. 3. Part of the Mindnet Entry for "provide"

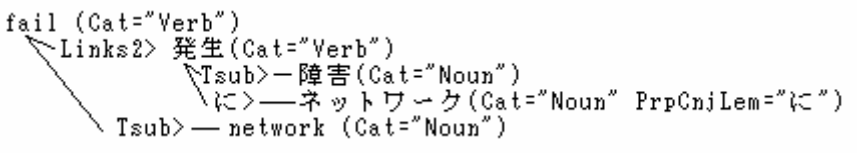

Fig. 4. Part of the Mindnet Entry for "fail"

illustrated by the sample training pair in the lefthand column of Table 1. In Japanese, inanimate nouns tend to be avoided as subjects of transitive verbs; the word "URL", which is subject in the English sentence, thus corresponds to an oblique relation in the Japanese. (The Japanese sentence, although a natural and idiomatic translation of the English, is literally equivalent to "one can access public folders with this URL.")

Nonetheless, mappings turn out to be learnable even where the information is structured so radically differently. Fig. 3 shows the Mindnet entry for "provide," which is result of training on sentence pairs like those in the lefthand column of Table 1. The system learns not only the mapping between the phrase "provide access" and the potential form of アクセス "access", but also the crucial sublinking of the Tsub node of the English sentence and the node headed by で (underspecified for semantic role) in the Japanese. At translation time the system is able to generalize on the basis of the functional roles stored in the Mindnet; it can substitute lexical items to achieve a relatively natural translation of similar sentences such as shown in the right-hand side of Table 1.

Differences of the kind seen in Fig 3 are endemic in our Japanese and English corpora. Fig. 4 shows part of the example Mindnet entry for the English word "fail" referred to in the Introduction, which exhibits another mismatch in grammatical roles somewhat similar to that in observed in Fig. 3. Here again, the lexical matching and generic alignment heuristics have allowed the match to be captured into the Mindnet. Although the techniques employed may have been informed by analysis of language-specific data, they are in principle of general application.

3 Evaluation

In May 2002, we compared output of the MSR-MT English-Japanese system with a commercially available desktop MT system.[3]

[3] Toshiba's *The Honyaku Office* v2.0 desktop MT system was selected for this purpose. *The Honyaku* is a trademark of the Toshiba Corporation. Another desktop system was also considered for evaluation; however, comparative evaluation with that system indicated that the Toshiba system performed marginally, though not significantly, better on our technical documentation.

Evaluation Date	Transfers per Sentence	Nodes Per Transfer
Oct. 2001	5.8	1.6
May 2002	6.7	2.0

Table 2. Number of Transfers and Nodes Transferred per Sentence

Evaluation Date	Word Class	Total	From Mindnet	From Dictionary	Untranslated
Oct. 2001 (250 sentences)	Prepositions	410	17.1%	77.1%	5.9%
	Content Lemmas	2124	88.4%	7.8%	3.9%
May 2002 (520 sentences)	Prepositions	842	61.9%	37.5%	0.6%
	Content Lemmas	4429	95.9%	1.5%	2.6%

Table 3. Sources of Different Word Classes at Transfer

A total of 238 English-Japanese sentence pairs were randomly extracted from held-out software manual data of the same kinds used for training the system. [4] The Japanese sentences, which had been translated by human translators, were taken as reference sentences (and were assumed to be correct translations). The English sentences were then translated by the two MT systems into Japanese for blind evaluation performed by seven outside vendors unfamiliar with either system's characteristics.

No attempt was made to constrain or modify the English input sentences on the basis of length or other characteristics. Both systems provided a translation for each sentence.[5]

For each of the Japanese reference sentences, evaluators were asked to select which translation was closer to the reference sentence. A value of +1 was assigned if the evaluator considered MSR-MT output sentence better and −1 if they considered the comparison system better. If two translated sentences were considered equally good or bad in comparison to the reference, a value of 0 was assigned. On this metric, MSR-MT scored slightly worse than the comparison system rating of −0.015. At a two-way confidence measure of +/−0.16, the difference between the systems is statistically insignificant. By contrast, an earlier evaluation conducted in October 2001 yielded a score of −0.34 vis-à-vis the comparison system.

In addition, the evaluators were asked to rate the translation quality on an absolute scale of 1 through 4, according to the following criteria:

1. *Unacceptable:* Absolutely not comprehensible and/or little or no information transferred accurately.
2. *Possibly Acceptable:* Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately.
3. *Acceptable:* Not perfect, but definitely comprehensible, and with accurate transfer of all important information.
4. *Ideal:* Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred.

On this absolute scale, neither system performed exceptionally well: MSR-MT scored an average 2.25 as opposed to 2.32 for the comparison system. Again, the difference between the two is statistically insignificant. It should be added that the comparison presented here is not ideal, since MSR-MT was trained principally on technical manual sentences,

[4] 250 sentences were originally selected for evaluation; 12 were later discarded when it was discovered by evaluators that the Japanese reference sentences (not the input sentences) were defective owing to the presence of junk characters (*mojibake*) and other deficiencies.

[5] In MSR-MT, Mindnet coverage is sufficiently complete with respect to the domain that an untranslated sentence normally represents a complete failure to parse the input, typically owing to excessive length.

while the comparison system was not specifically tuned to this corpus. Accordingly the results of the evaluation need to be interpreted narrowly, as demonstrating that:

- A viable example-based English-Japanese MT system can be developed that applies general-purpose alignment rules to semantic representations; and
- Given general-purpose grammars, a representation of what the sentence means, and suitable learning techniques, it is possible to achieve in a domain, results analogous with those of a mature commercial product, and within a relatively short time frame.

4 Discussion

It is illustrative to consider some of the factors that contributed to these results. Table 2 shows the number of transfers per sentence and the number of LF nodes per transfer in versions of the system evaluated in October 2001 and May 2002. Not only is the MSR-MT finding more LF segments in the Mindnet, crucially the number of nodes transferred has also grown. An average of two connected nodes are now transferred with each LF segment, indicating that the system is increasingly learning its translation knowledge in terms of complex structures rather than simple lexical correspondences.

It has been our experience that the greater MSR-MT's reliance on the Mindnet, the better the quality of its output. Table 2 shows the sources of selected word classes in the two systems. Over time, reliance on the Mindnet has increased overall, while reliance on dictionary lookup has now diminished to the point where, in the case of content words, it should be possible to discard the handcrafted dictionary altogether and draw exclusively on the contextualized resources of the Mindnet and statistically-generated lexical data. Also striking in Table 2 is the gain shown in preposition handling: a majority of English prepositions are now being transferred only in the context of LF structures found in the Mindnet.

The important observation underlying the gains shown in these tables is that they have primarily been obtained either as the result of LF improvements in English or Japanese (i.e., from better sentence analysis or LF construction), or as a result of generic improvements to the algorithms that map between LF segments (notably better coindexation and improved learning of mappings involving lexical attributes). In the latter case, although certain modifications may have been driven by phenomena observed between Japanese and English, the heuristics apply across all seven languages on which our group is currently working. Adaptation to the case of Japanese-English MT usually takes the form of loosening rather than tightening of constraints.

5 Future Work

Ultimately it is probably desirable that the system's mean absolute score should approach 3 (*Acceptable*) within the training domain: this is a high quality bar that is not attained by off-the-shelf systems. Much of the work will be of a general nature: improving the parses and LF structures of source and target languages will bring automatic benefits to both alignment of structured phrases and runtime translation. For example, efforts are currently underway to redesign LF to better represent scopal properties of quantifiers and negation (Campbell & Suzuki 2002a, 2002b).

Work to improve the quality of alignment and transfer is ongoing within our group. In addition to improvement of alignment itself, we are also exploring techniques to ensure that the transferred LF is consistent with known LFs in the target language, with the eventual goal of obviating the need for heuristic rules used in preprocessing generation. Again, these improvements are likely to be system-wide and generic, and not specific to the English-Japanese case.

Conclusions

Use of abstract semantically-motivated linguistic representations (Logical Form) permits MSR-MT to align, store, and translate sentence patterns reflecting widely varying syntactic and information structures in Japanese and English, and to do so within the framework of a general-purpose NLP architecture applicable to both European languages and Asian languages.

Our experience with English-Japanese example based MT suggests that the problem of MT

among Asian languages may be recast as a problem of implementing a general representation of structured meaning across languages that neutralizes differences where possible, and where this is not possible, readily permits researchers to identify general-purpose techniques of bridging the disparities that are viable across multiple languages.

Acknowledgements

We would like to thank Bill Dolan and Rich Campbell for their comments on a draft of this paper. Our appreciation also goes to the members of the Butler Hill Group for their assistance with conducting evaluations.

References

Aikawa, T., M. Melero, L. Schwartz, and A. Wu. 2001a. Multilingual sentence generation. In *Proceedings of 8th European Workshop on Natural Language Generation*, Toulouse, France.

Aikawa, T., M. Melero, L. Schwartz, and A. Wu. 2001b. Sentence generation for multilingual machine translation. In *Proceedings of the MT Summit VIII*, Santiago de Compostela, Spain.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2): 79-85.

Campbell, R. and H. Suzuki. 2002a. Language-neutral representation of syntactic structure. In *Proceedings of the First International Workshop on Scalable Natural Language Understanding (SCANALU 2002)*, Heidelberg, Germany.

Campbell, R. and H. Suzuki. 2002b. *Language-Neutral Syntax: An Overview.* Microsoft Research Techreport: MSR-TR-2002-76.

Heidorn, G. 2000. Intelligent writing assistance. In R. Dale, H. Moisl and H. Somers (eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text.* Marcel Dekker, New York. pp. 181-207.

Imamura, K. 2001a. Application of translation knowledge acquired by ierarchical phrase alignment. In *Proceedings of TMI*.

Imamura, K. 2001b. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of NLPRS,* Tokyo, Japan, pp 377-384.

Kacmarcik, G., C. Brockett, and H. Suzuki. 2000. Robust segmentation of Japanese text into a lattice for parsing. In *Proceedings of COLING 2000*, Saarbrueken, Germany, pp. 390-396.

Menezes, A. and S. D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 39-46.

Moore, R. C. 2001. Towards a simple and accurate statistical approach to learning translation relationships among words," in *Proceedings, Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter, Association for Computational Linguistics*, Toulouse, France, pp. 79-86.

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn. and R. Bannerji (eds.) *Artificial and Human Intelligence*. Nato Publications. pp. 181-207.

Pinkham, J., M. Corston-Oliver, M. Smets and M. Pettenaro. 2001. Rapid Assembly of a Large-scale French-English MT system. In *Proceedings of the MT Summit VIII*, Santiago de Compostela, Spain.

Pinkham, J., and M. Smets. 2002. Machine translation without a bilingual dictionary. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*. Kyoto, Japan, pp. 146-156.

Richardson, S. D., W. B. Dolan, A. Menezes, and M. Corston-Oliver. 2001. Overcoming the customization bottleneck using example-based MT. In *Proceedings, Workshop on Data-driven Machine Translation, 39th Annual Meeting and 10th Conference of the European Chapter, Association for Computational Linguistics. Toulouse*, France, pp. 9-16.

Richardson, S. D., W. B. Dolan, A. Menezes, and J. Pinkham. 2001. Achieving commercial-quality translation with example-based methods. In *Proceedings of MT Summit VIII*, Santiago De Compostela, Spain, pp. 293-298.

Richardson, S. D., W. B. Dolan, and L. Vanderwende. 1998 MindNet: Acquiring and structuring semantic information from text, *ACL-98.* pp. 1098-1102.

Sato, S. and Nagao M. 1990. Toward memory-based translation. In *Proceedings of COLING 1990*, Helsinki, Finland, pp. 247-252.

Suzuki, H., C. Brockett, and G. Kacmarcik. 2000. Using a broad-coverage parser for word-breaking in Japanese. In *Proceedings of COLING 2000,* Saarbrueken, Germany, pp. 822-827.

Yamamoto K., and Y Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of COLING 2000*, Saarbrueken, Germany, pp. 933-939.