MACHINE TRANSLATION REVIEW

The Periodical
of the
Natural Language Translation Specialist Group
of the
British Computer Society
Issue No. 10
October 1999

The *Machine Translation Review* incorporates the Newsletter of the Natural Language Translation Specialist Group of the British Computer Society and appears twice yearly.

The Review welcomes contributions, articles, book reviews, advertisements, and all items of information relating to the processing and translation of natural language. Contributions and correspondence should be addressed to:

Derek Lewis
The Editor
Machine Translation Review
School of Modern Languages
Queen's Building
University of Exeter
Exeter
EX4 4QH
United Kingdom

Tel: +44 (0)1392 264296
Fax: +44 (0)1392 264377
E-mail: D.R.Lewis@exeter.ac.uk

The *Machine Translation Review* is published by the Natural Language Translation Specialist Group of the British Computer Society. All published items are subject to the usual laws of Copyright and may not be reproduced without the permission of the publishers.

ISSN 1358–8346

Please note: From the next issue in April 2000 this Review will be published electronically and will be available on our web site at the British Computer Society (see page 5). The format will be in HTML, in the same way as some back copies have already been stored electronically, so it will be easy for readers to print copies if they wish. Each section will be separate so readers may print selected parts only.

Some copies will be printed for the Copyright libraries and for purchase at a modest price plus postage and packing for those without electronic access. Members of the Natural Language Translation Specialist Group of the British Computer Society will be advised of each issue.

Contents

Group News and Information

Letter from the Chairman

72 Brattle Wood
Sevenoaks
Kent, TN13 1QU
Tel: 01732 455446
Office: 0171 815 7472
Fax: 0171 815 7550
E-mail: wiggjd@sbu.ac.uk

Plans for the International Machine Translation Conference MT 2000 being organised jointly by our Group and Exeter University to be held at Exeter next year from Monday 20 November to Wednesday 22 November 2000 are going ahead well. A preliminary Call for Papers has been published and an e-mail list has been set up for interested parties to register their interest in being kept informed. You can also obtain up to date information from our web site at the BCS.

Incidentally, if you are interested in keeping in touch with the world of MT the European Association for Machine Translation (EAMT) has now set up an e-mail list at mt-list@eamt.org which can be joined by e-mail to mt-list-request@eamt.org with the word subscribe in the subject line.

As I mentioned last time, the proceedings of the conference at Cranfield in 1994 are now available, and we reproduce here another paper from them by kind permission of the author, Christian Boitet from Joseph Fourier University, Grenoble, to encourage you to buy a copy.

May I remind members yet again, that they do not need to live near London to assist the Committee. We do not have sufficient funds to pay travel expenses for all Committee members to attend meetings, but we still welcome Correspondent members. Correspondent committee members are otherwise treated as full members of the committee and kept advised of all committee business. Anyone interested in helping should contact me or any other committee member.

Our committee still requires a treasurer, although in our case the role is more of an auditor since all our transactions are processed by the BCS. This post does, of course, require some knowledge of accounting, but not much I'm glad to say, and, as mentioned above, it does not need to be someone in the London area. If anybody is interested to know more, please contact me.

Please consider contributing to this Review. We would still welcome more articles, papers and reports on the subject of machine translation and related subjects such as computer assisted language teaching, computer based dictionaries and aspects of multilinguality in computing etc. We would welcome papers from staff and students in linguistics and related disciplines, and from translators and any other users of MT software.

All opinions expressed in this Review are those of the respective writers and are not necessarily shared by the BCS or the Group.'

David Wigg

The Committee

The telephone numbers and e-mail addresses of the Committee are as follows:

David Wigg (Chair)	Tel.: +44 (0)1732 455446 (H)
	Tel.: +44 (0)171 815 7472 (W)
	E-mail: wiggjd@sbu.ac.uk
Monique L'Huillier (Secretary)	Tel.: +44 (0)1276 20488 (H)
	Tel.: +44 (0)1784 443243 (W)
	E-mail: m.l'huillier@rhbnc.ac.uk
Derek Lewis (Editor)	Tel.: +44 (0)1404 814186 (H)
	Tel.: +44 (0)1392 264296 (W)
	E-mail: d.r.lewis@exeter.ac.uk
Douglas Clarke	Tel.: +44 (0)1908 373141
Ian Kelly	Tel.: +44 (0)1276 857599
	E-mail: 100350.3046@compuserve.com
Veronica Lawson	Tel.: +44 (0)171 7359060
	E-mail: veronica@compuserve.com
Roger Harris (Rapporteur)	Tel.: +44 (0)181 800 2903 (H)
	E-mail: rwsh@dircon.co.uk
Correspondent Members:	
Gareth Evans (Minority Languages)	Tel.: +44 (0)1792 481144
	E-mail: g.evans@sihe.ac.uk
Ruslan Mitkov	Tel: +44 (0)1902 322471 (W)
	E-mail: R.Mitkov@wlv.ac.uk

BCS Library

Books kindly donated by members are passed to the BCS library at the IEE, Savoy Place, London, WC2R 0BL, UK (tel: +44 (0)171 240 1871; fax: +44 (0)171 497 3557). Members of the BCS may borrow books from this library either in person or by post. All they have to provide is their membership number. The library is open Monday to Friday, 9.00 am - 5.00 pm.

Website

The website address of the BCS-NLTSG is: http://www.bcs.org.uk/siggroup/sg37.htm

Dialogue-Based MT and Self-explaining Documents as an Alternative to MAHT and MT of Controlled Languages

by

Christian Boitet

GETA, Institut IMAG (UJF and CNRS), GRENOBLE, France
E-mail: Christian.Boitet@imag.fr

Abstract

We argue that, in many situations, Dialogue-Based MT (DBMT) is likely to offer better solutions to translation needs than machine aids to translators or batch MT, even if controlled languages are used. Objections to DBMT have led us to introduce the new concept of the 'self-explaining document', which might be used in monolingual as well as in multilingual contexts, and deeply change our way of understanding important or difficult written material.

1 Introduction

In many situations, documents such as working notes, scientific abstracts, transparencies, calls for proposals, technical documentation, etc., should be translated into several languages, but are not translated, because they are ready at the last moment, and available translators have no time to do the job, or because there are simply no translators to do the job, and of course, in all cases, because no satisfactory MT solution is available.

Our first point is that interactive Dialogue-Based MT systems (DBMT), especially of the kind we are prototyping in the LIDIA project, offer a better hope of solving the problem than machine aids for translators and 'black box' MT, even if controlled languages are used.

Our second point is that the DBMT approach also leads to a new and extremely interesting possibility, that of producing all versions of a document, that is, the source document and all its translations, as 'self-explaining' documents, each consisting of a normal document and its deep or (even better) multilevel disambiguated linguistic representation, augmented by a memory of the original ambiguities and of the disambiguation process.

Finally, we observe that the production of self-explaining documents might also be very useful in monolingual contexts, and perhaps lead to new ways of accessing and using documents of any kind: one could 'click' on any part marked as ambiguous, and get clarifying presentations or paraphrases of it. Thus, an unrestricted self-explaining text would be less ambiguous than a text in a controlled language, which may be unambiguous for a machine but not for a human, and access to texts written in foreign languages would also be facilitated. In this way, authors' true intentions would accompany their productions in other places, times and tongues.

2 Motivations

The idea of DBMT has been proposed and experimented with in various forms during the last 20 years [13, 15, 17, 18, 24, 29, 32, 33, 36-38]. However, it has always been taken for granted that the user should be a specialist, linguist or translator or at least a professional and that consequently the system could and should be specialized. In contrast, we think that DBMT systems should be designed for the general public and should be usable on personal computers. Consequently, the design of the user interface in general, and of the disambiguation dialogues in particular, becomes extremely important.

The main idea of our current concept is that pieces of the text under creation or modification are sent to an analyzer running in the background. If there are ambiguities, be they proper to the source language or relative to the translation into one or several target languages, questions are asked of the author in the source language. The resulting ambiguity-free structures are then sent to transfers and generators into all target languages, producing high quality translations needing no postedition.

During the last few years, we have designed and implemented a mock-up, LIDIA-1 [3-10], to experiment with this concept of DBMT for (non-specialist) individual authors. The mock-up has now only one source language, French, and three target languages, German, English and Russian, but that is only due to the limitations in manpower. This experimentation has led us to various innovations:

- *distributed processing.* The document is created and interactively disambiguated on a middle-range Macintosh, while the various processes of MT proper are performed by a distant server.

- *application to hypertexts*. The documents are in effect hyperdocuments, in the form of HyperCard stacks. Units of translation are HyperCard textual ‘fields’.

- *asynchronous and non-pre-emptive processing.* Units of translation are ‘released’ by the author, and then autonomously travel to the MT server, come back after analysis in a ‘multiple-multilevel-concrete’ form (mmc-structure), announce the presence of ambiguities by letting a button appear next to them, react to the click by engaging in a disambiguation dialogue and then, once disambiguated, travel again autonomously to the MT server to be translated and finally to be inserted in the appropriate field in the target stack.

- *e-mail communication between component processes.* We have switched from a specialized connexion to the use of standard e-mail for all communications between the author workstations and the MT server, which can now be located anywhere in the world.

- *deeper multilevel approach.* We have added a level of ‘interlingual acceptions’ (or word senses) to the classical lexical levels of B. Vauquois’ multilevel transfer approach (‘occurrence’ or wordform, ‘lemma’ or citation form, and ‘lexical unit’ or derivational family).

- *disambiguation strategy.* We have developed a generator of disambiguation dialogues which non-specialists can easily understand and which does not rely on too sophisticated linguistic processing, so that the disambiguator can run in real time on the author’s personal computer.

- *control by reverse translation.* On demand, the system translates back from the target uma-structures (unambiguous, multilevel, abstract), providing a feed-back through a paraphrase in the source language of the translation.

- *homogeneity of knowledge sources.* In the current state of the implementation, this concerns only the lexical knowledge: both the lexical disambiguation messages and the MT dictionaries are obtained from the same multilingual lexical data base, PARAX [2], itself implemented in HyperCard.

An interesting possibility offered by our distributed technique is to build DBMT applications by using heterogeneous components. For example, the source text could be written (in French) in Paris, sent for analysis to our server in Grenoble, disambiguated interactively in Paris, and then sent to our server to produce translations in English, German and Russian, and to a server in Japan to produce a Japanese translation. This would only require appropriate 'filters' (format transducers) between intermediate structures, and agreements with server operators.

3 Self-explaining documents

In the course of our experimentation, we have (again) observed that translation introduces ambiguities which are not present in the source text. *Traduttore, traditore…* It also happens that all disambiguated analyses of a sentence produce the same translation, which is as ambiguous as the original. One example was the translation from French into Russian of the famous sentence 'The man sees the girl in the park with a telescope'.

Then, goes the objection, what is the use of disambiguating the source text if ambiguities reappear in the translation(s), or even worse if new ones are created? Would it not be better to try and produce translations which preserve the ambiguities, and dispense with interactive disambiguation altogether?

Unfortunately, the experience of human translation shows that ambiguities can be *exactly* preserved only in some cases, and that to do it purposefully is quite difficult and often leads to unnatural ways of expression in the translated text. It is also quite clear that the 'transferable' ambiguities vary with the target language. Finally, although some texts may be intentionally ambiguous, especially in poetry and politics, we take it that the vast majority of ambiguities are not intentional, but are due to the intrinsic nature of natural languages. Of course, some authors write more clearly than others, but all authors write unambiguously in any programming language, unambiguous by construction, and ambiguously in any natural language, ambiguous by nature!

This has led us to the idea of *self-explaining documents:* if the target documents are accompanied by their (unambiguous) linguistic structure, with the indications of potentially ambiguous parts, and if the reader in the target language may obtain a clarification of unclear parts in a user-friendly way, the objection disappears. As human users are notably not very sensitive to ambiguities, however, we should find a way to warn the reader that the target text is ambiguous.

In a multilingual DBMT setting, there is a very simple solution to this task. The system simply analyzes the target text with the analyzer of the target language and gets the corresponding mmc-structures. It then runs the disambiguation dialogue on the target side in automatic 'mute' mode, that is by having the system itself answer each question so that the accompanying structure is contained in the selected subset at each point and memorizes questions and answers. It is then possible to show the presence of ambiguities by any convenient means, such as by creating buttons on which the reader may click to obtain the clarification *which would have been given by the author himself, were the text to have been written in the target language!* To simplify this process, the accompanying structure should then be unambiguous and ‘concrete’.

Let us clarify what we call ‘concrete’ and ‘abstract’ linguistic structures. A ‘concrete’ representation of a text is such that the corresponding text can be recovered from it by using a standard traversal algorithm and simple morphological and graphematical generation rules. Familiar examples are textbook constituent structures and dependency structures (with left-to-right traversal of the leaves or infix traversal of all nodes). Otherwise, we say that the representation is ‘abstract’. Note that the information contained in both kinds of structures (on labels and other more or less complex annotations) may be of the same linguistic ‘depth’: there may be ‘deep’ concrete structures and ‘surface’ abstract structures, in this sense, although the opposite is of course more frequent.

Take for example the sentence: *‘The customers were not given their money back by the cashier but by the waiter.’* A ‘multilevel’ head-driven concrete structure could be:

```
S[type=assertive, time=past, aspect=perfective, tense=c-past, voice=passive…]
 (NP[semrel=dest, logrel=arg2, synfunc=subj, sem=human, num=plur…]
    (Art[lex=‘the’, semrel=deict, synfunc=det, number=plur, deter=definite…]
    Noun[lex=‘customer’, synfunc=head, sem=human, number=plur…])
  aux[lex=‘be’, tense=pret, pers=3, number=plur…]
  neg[lex=‘not’]
  vrb[lex=‘give’, synfunc=head, voice=passive, tense=ppart, vbpart=‘back’…]
  NP[semrel=patient, logrel=arg1, synfunc=obj1, number=sing…]
    (adjposs[lex=‘his’, semrel=poss, synfunc=det, number=plur, deter=definite…]
    Noun[lex=‘money’, synfunc=head, number=sing…])
  vbpart[lex=‘back’]
  NP[semrel=agent, logrel=arg0, synfunc=agcomp, number=sing, neg=not-but…]
    (prep[lex=‘by’, synfunc=reg]
    art[lex=‘the’, semrel=deict, synfunc=det, number=sing, deter=definite…]
    Noun[lex=‘cashier’, synfunc=head, sem=human, number=sing, neg=not…]
    NP[semrel=id, logrel=arg0, synfunc=coord, number=sing…]
       (conj[lex=‘but’, synfunc=reg]
       prep[lex=‘by’, synfunc=reg]
       art[lex=‘the’, semrel=deict, synfunc=det, number=sing, deter=definite…]
       Noun[lex=‘waiter’, synfunc=head, sem=human, number=sing…]))
  punct[lex=‘.’])
```

Syntactic categories have been used here as main labels, with phrases (syntagmas) in capitals and preterminals in small letters. Acronyms should be self-explaining.

In an abstract structure, some lexical information would be ‘featurized’, and order could be normalised, leading to:

```
S[type=assertive, time=past, aspect=perfective, tense=c-past, voice=passive…]
  (vrb[lex='give'.'back', synfunc=head, voice=passive, tense=c-past…]
  NP[semrel=agent, logrel=arg0, synfunc=agcomp, num=sing, neg=not-but…]
    (neg[lex='not']
     Noun[lex='cashier', synfunc=head, sem=human, number=sing, deter=definite…]
     NP[semrel=id, logrel=arg0, synfunc=coord, num=sing…]
       (conj[lex='but', synfunc=reg]
        Noun[lex='waiter', synfunc=head, sem=human, number=sing, deter=definite…]))
  NP[semrel=patient, logrel=arg1, synfunc=obj1, num=sing…]
    (adjposs[lex='his', semrel=poss, synfunc=det, number=plur, deter=definite…]
     Noun[lex='money', synfunc=head, number=sing…])
  NP[semrel=dest, logrel=arg2, synfunc=subj, sem=human, number=plur…]
    (Noun[lex='customer', synfunc=head, sem=human, number=plur, deter=definite…]))
```

Abstract representations of utterances are far superior to concrete representations as input and output structures of transfers in semantic transfer MT or as ‘lexical-conceptual structures’ [23] in interlingual MT, especially between distant languages. But their relation to the corresponding utterances is not as clear, a natural consequence of abstraction. That ‘remoteness’ is even more apparent with other types of structures, such as conceptual graphs, logical formulae or interlingual representations *à la KBMT*-89 [27]. By contrast, concrete structures are clearly more adequate for interactive disambiguation. They are also superior for a variety of future applications. For example, no text processor today is able to replace ‘give-back’ by ‘return’ in the preceding example, not to speak of changing the modality, the tense, the voice or the discourse type (say, from direct to indirect or affirmative to negative). Self-explaining documents would make that possible.

Here is a functional diagram (figure 22 - 1) of the processes we have discussed above. In gma-structures (generating, multilevel and abstract), non-interlingual linguistic levels are underspecified and, if present, are used only as reflections of corresponding surface levels in the source language and are recomputed in the first generation phase, which we call ‘paraphrase choice’.

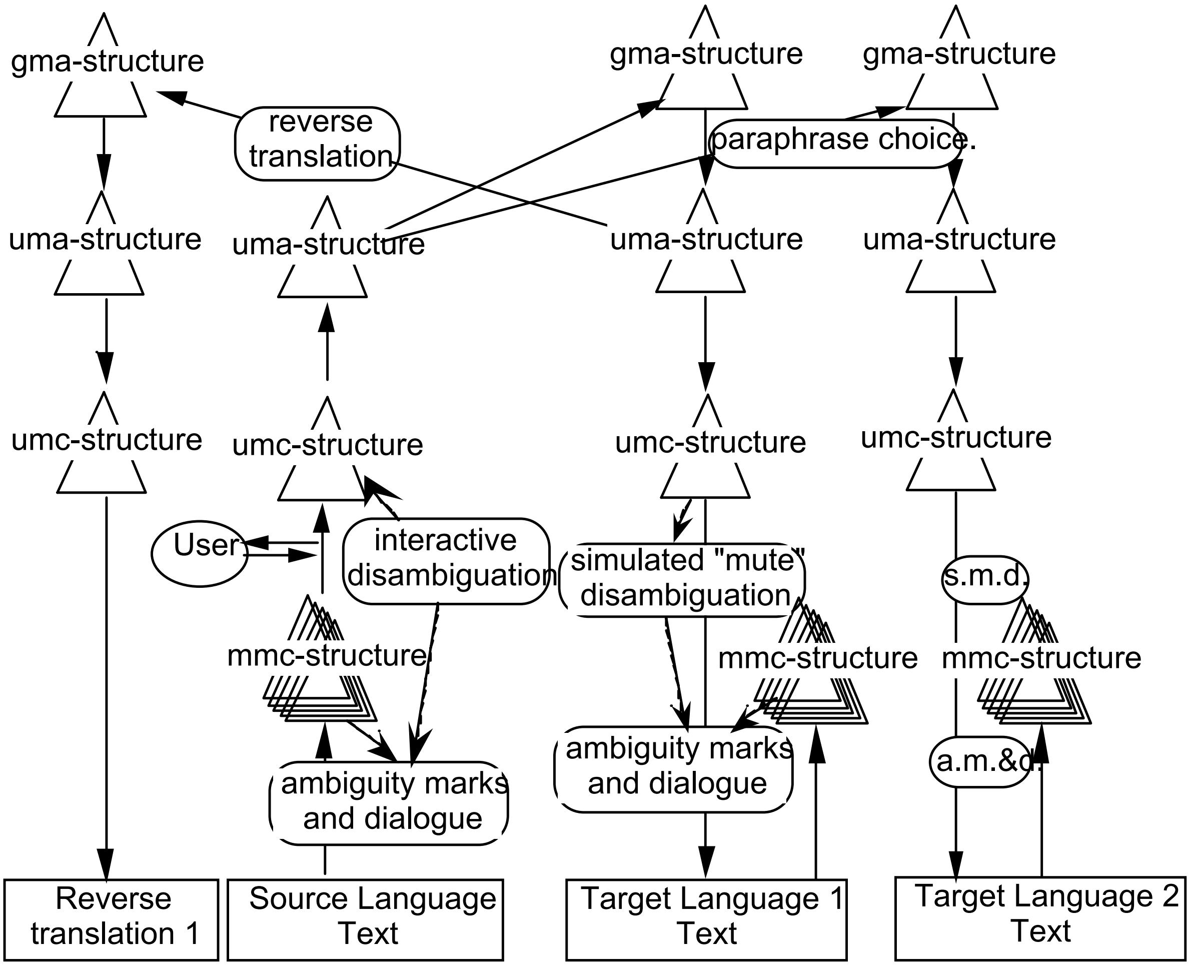

Figure 22 – 1

4 Alternatives

Is DBMT really a better approach than other alternatives? Our answer is a definite yes, because:

- very often, these alternatives are not really feasible,
- the results can be intrinsically better, if presented as self-explaining documents.

First*, machine aids for translators* [1, 21, 25, 26, 34] are usable only if there are available and affordable translators and if they have enough time to do the job. But, in many situations, there are simply no such translators, especially if translations are required in several languages. For example, multinational firms, banks, etc., have many uncovered translation needs. In Europe, scientists and engineers are engaged in many projects where communication is hampered by the language barrier.

Even if competent translators are available, the delays are often such that translation is impossible. That is for example the case in European institutions, which are theoretically required to issue all important documents in all official languages but are unable to do so, although they employ more than 1,200 full-time translators and translate more than 1.2 million pages a year. But the final versions of these documents are too often ready at the last moment. Here, it would make more sense to analyze and disambiguate their parts as soon as they are ready and to translate them at the last moment.

Another possibility, often advocated, is to write in *controlled languages* designed to be unambiguous and use ‘black box’ MT. This can be very successful in restricted situations as in the case of the TITUS system of the Institut Textile de France. But it is very difficult to force people to write in a controlled language. It proved for instance impossible to adapt the TITUS system [14] to the context of the CDST (Center for Scientific and Technical Documentation of CNRS). Another weak point of controlled languages is that they are difficult to design and very task- and domain-specific.

Finally, controlled languages are unambiguous for the analyzer designed to process them, but not for humans. While this may be convenient in the context of man-machine communication, it may be counterproductive, or even dangerous, in the context of human communication. If, for example, ‘to replace (a mechanical part)’ is intended to mean only ‘to replace by a new thing’ (‘remplacer’), and not ‘to put back in place’ (‘replacer’), a mechanic may well understand the second, unintended meaning, and put back in place an airplane part which should be replaced by a new one, leading to an accident.

If the concept of self-explaining document may be made to work in broad contexts, texts could be written without restrictions stronger than the usual ones which concern style and terminology and at the same time be in effect less ambiguous than texts written in controlled languages. Even if translation is not an issue, then the availability of *‘text explainers’* might be a major advance in document processing technology.

5 Perspectives

Full-scale, general DBMT systems ‘for everybody’ would require extremely large grammatical and lexical knowledge bases. To cover a whole language, a lexical data base should contain of the order of 3 million terms, corresponding to 4 to 5 million monolingual acceptions, and perhaps to twice as many interlingual acceptions for systems designed to handle 10 to 20 languages.

The development costs are staggering and probably out of reach if conventional lingware engineering techniques are used. For instance, it has cost EDR (Tokyo) about 1,200 man-years to develop 300K terms in 2 languages (200K terminological and 100K general), with the associated 640K interlingual concepts (200K terminological and 2*300K general, minus 60K common). At least 100 times more (1.2 million man-years!) would be needed for 3 million terms in 20 languages.

This is why we advocate a step by step approach, and the development of new groupware techniques for developing very large lexical data bases.

First, text explainers could be developed for several languages, in specific domains and situations. Here, the figures would be 10 to 30K terms, 100 to 300 times less than those mentioned above.

Second, the (monolingual) analyzers and dictionaries developed for text explainers might be reused for building DBMT systems for the same restricted domains and situations. Transfers and generators would not be too costly to develop. The lexical work would consist in integrating all monolingual lexical data bases in a unique multilingual data-base, thereby refining each monolingual data-base according to the obtained set of interlingual acceptions [30].

In a third step, general text explainers and DBMT systems could then be developed, by progressively merging and extending specific systems. This ‘divide-and-conquer’ strategy might break down the cost and make the whole enterprise feasible in the long run.

Even if the cost were not a problem, developing extremely large lexical data bases and keeping them up to date would be impossible using only professional teams of lexicographers. The quantity is too huge, changes and innovations are too fast, and only specialists can be competent in specific domains. We consider it as a major challenge to develop *groupware lexical data base development techniques* which might be based on the *contribution of lexical information by users* of existing text explainers or DBMT systems.

A distributed architecture would be very advantageous for that purpose, because lexical information created by the users on their personal computers might transparently and automatically be sent to the servers. For example, authors might add new senses to existing terms, add new terms, and propose translations in the languages they know. Professional teams would then process and refine this ‘raw lexical material’ on server sites. This idea is not so far-fetched, and very similar to that used in Eurolang Optimizer™ [1], a distributed environment designed for professional translators.

References

[1] Blanc É., Sérasset G. and Tchéou F. (1994) *Designing an Acception-Based Multilingual Lexical Data Base under HyperCard: PARAX.* Research Report, GETA, IMAG (UJF and CNRS), Aug. 1994

[2] Blanchon H. (1992) *A Solution to the Problem of Interactive Disambiguation.* Proc. COLING-92, Nantes, 23-28 July 1992, C. Boitet, ed., vol. 4/4, pp. 1233-1238

[3] Blanchon H. (1994) *Perspectives of DBMT for monolingual authors on the basis of LIDIA-1, an implemented mockup.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, vol. 1/2, pp. 115—119

[4] Boitet C. (1989) *Motivation and Architecture of the LIDIA Project.* Proc. MTS-II (MT Summit), Munich, 16-18 août 1989, pp. 50—54

[5] Boitet C. (1990) *Towards Personal MT : on some aspects of the LIDIA project.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 30-35

[6] Boitet C. (1993) *La TAO comme technologie scientifique : le cas de la TA fondée sur le dialogue.* In ‘Études et Recherches en Traductique’, A. Clas and P. Bouillon, ed., Presses de l’Université de Montréal, Montréal, pp. 109—148

[7] Boitet C. (1994) *Dialogue-Based Machine Translation and Sub-Languages.* Proc. ICLA-94, Penang, Malaysia, 26-28 July 1994, USM

[8] Boitet C. and Blanchon H. (1993) *Dialogue-Based MT for Monolingual Authors and the LIDIA project.* Proc. NLPRS’93 (Natural Language Processing Rim Symposium, Fukuoka, 6-7/12/93, H. Nomura, ed., Kyushu Institute of Technology, pp. 208—222

[9] Boitet C. and Blanchon H. (1994) *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup.* Machine Translation. (to appear)

[10] Brown R. D. (1989) *Augmentation.* Machine Translation, 4, pp. 1299-1347

[11] Brown R. D. and Nirenburg S. (1990) *Human-Computer Interaction for Semantic Disambiguation.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 42-47

[12] Chandler B., Holden N., Horsfall H., Pollard E. and McGee Wood M. (1987) *N-tran Final Report.* Alvey Project, 87/9, CCL/UMIST, Manchester

[13] Ducrot J.-M. (1988) *Le système TITUS IV.* In 'Traduction Assistée par Ordinateur. Actes du séminaire international sur la TAO et dossiers complémentaires', A. Abbou, ed., Observatoire des Industries de la Langue (OFIL), Paris, mars 1988, pp. 55—71

[14] Huang X. M. (1990) *A Machine Translation System for the Target Language Inexpert.* Proc. COLING-90, Helsinki, 20-25 Aug. 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 364-367

[15] Hutchins W. J. (1986) *Machine Translation : Past, Present, Future.* Ellis Horwood, John Wiley and Sons, Chichester, England, 382 p.

[16] Kay M. (1973) *The MIND system.* In 'Courant Computer Science Symposium 8: Natural Language Processing', R. Rustin, ed., Algorithmics Press, Inc., New York, pp. 155-188

[17] Kay M. (1980) *The Proper Place of Men and Machines in Language Translation.* Research Report, CSL-80-11, Xerox, Palo Alto Research Center, Oct. 1980

[18] Kittredge R. (1983) *Sublanguage — Specific Computer Aids to Translation — a survey of the most promising application areas.* Contract n° 2-5273, Université de Montréal et Bureau des Traductions, mars 1983, 95 p.

[19] Kittredge R. (1986) *Analyzing Language in Restricted Domains.* In 'Sublanguage Description and Processing', R. Grishman and R. Kittredge, ed., Lawrence Erlbaum, Hillsdale, New-Jersey

[20] Langé J.-M. (1994) *Systèmes d'aide à la traduction : un point de vue industriel.* Proc. TALN-94, journées du PRC-CHM, Marseille, 7—8 avril 1994, Univ. de Marseille

[21] Lehrberger J. and Bourbeau L. (1988) *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation.* John Benjamins, 240 p.

[22] Levin L. and Nirenburg S. (1994) *The Correct Place of Lexical Semantics in Interlingual MT.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, vol. 1/2, pp. 349—355

[23] Maruyama H., Watanabe H. and Ogino S. (1990) *An Interactive Japanese Parser for Machine Translation.* Proc. COLING-90, Helsinki, 20-25 août 1990, H. Karlgren, ed., ACL, vol. 2/3, pp. 257-262

[24] Melby A. K. (1981) *Translators and Machines - Can they cooperate ?* META, **26**/1, pp. 23-34

[25] Melby A. K. (1982) *Multi-Level Translation Aids in a Distributed System.* Proc. COLING-82, Prague, 5-10 juillet 1982, vol. 1/2, pp. 215-220

[26] Nirenburg S. (1989) *Knowledge-based Machine Translation.* Machine Translation, 4, pp. 5-24

[27] Nyberg E. H. and Mitamura T. (1992) *The KANT system: Fast, Accurate, High-Quality Translation in Practical Domains.* Proc. COLING-92, Nantes, 23-28 July 92, C. Boitet, ed., ACL, vol. 3/4, pp. 1069—1073

[28] Sadler V. (1989) *Working with analogical semantics: Disambiguation technics in DLT.* T. Witkam, ed., Distributed Language Translation (BSO/Research), Floris Publications, Dordrecht, Holland, 256 p.

[29] Sérasset G. (1994) *Interlingual Lexical Organisation for Multilingual Lexical Databases.* Proc. 15th International Conference on Computational Linguistics, COLING-94, Kyoto, Japan, 5-9 Aug. 1994, 6 p.

[30] Sérasset G. (1994) *An Interlingual Lexical Organization Based on Acceptions.* Proc. ICLA-94, Penang, Malaysia, 26-28 July 1994, USM, 12 p.

[31] Somers H. L., Tsujii J.-I. and Jones D. (1990) *Machine Translation without a source text.* Proc. COLING-90, Helsinki, 20-25 Aug. 1990, H. Karlgren, ed., ACL, vol. 3/3, pp. 271-276

[32] Tomita M. (1986) *Sentence Disambiguation by asking.* Computers and Translation, **1**/1, pp. 39-51

[33] Tong L. C. (1987) *The Engineering of a Translator Workstation.* Computers and Translation, 2/4, pp. 263—273

[34] Vauquois B. (1988) *BERNARD VAUQUOIS et la TAO, vingt-cinq ans de Traduction Automatique, ANALECTES. BERNARD VAUQUOIS and MT, twenty-five years of MT.* C. Boitet, ed., Ass. Champollion and GETA, Grenoble, 700 p.

[35] Wehrli E. (1992) *The IPS System.* Proc. COLING-92, Nantes, 23-28 July 1992, C. Boitet, ed., vol. 3/4, pp. 870-874

[36] Whitelock P. J., Wood M. M., Chandler B. J., Holden N. and Horsfall H. J. (1986) *Strategies for Interactive Machine Translation: the experience and implications of the UMIST Japanese project.* Proc. COLING-86, Bonn, 25-29 août 1986, IKS, pp. 25-29

[38] Wood M. M. G. and Chandler B. (1988) *Machine Translation For Monolinguals.* Proc. COLING-88, Budapest, 22-27 Aug. 1988, pp. 760—763

Report on Commercial Machine Translation in a Manufacturing Industry Domain

by

Orie Fukutomi

School of Humanities, Languages and Social Sciences
University of Wolverhampton, UK
E-mail: n9327544@wlv.ac.uk

Abstract

The aim of this paper is to provide a report on an experiment using a Commercial Machine Translation (CMT) software in a manufacturing company in the UK, with particular reference to Japanese/English Machine translation. It presents the main difficulties involved in the translation of industrial documents from Japanese to English and discusses how the productivity and quality of translation can be improved through the use of commercial Machine Translation (MT) software. They are proposed from a translator's point of view in a manufacturing factory. The survey focuses on a manufacturing organisation which does not have the resources needed to develop their own MT system. The globalisation of the Japanese manufacturing industry makes it necessary for the translation of manuals and other documents to be as rapid as possible. In this paper, linguistic features of both English and Japanese are discussed from the evaluation experiment in order to make up writing rules for members of staff at Makita Manufacturing Europe. It also discusses the viewpoint of British engineers when translated manuals are read. The result of the experiment is also examined in terms of whether it reflects the theoretical viewpoint or not.

1 Outline Structure

The disappointment expressed about translations generated by MT has not stopped a substantial number of people from using it for serious documentation. Although it is still impossible for MT to process natural language completely, the level achieved is significant. Yet, not many users are informed about the properties or strengths of MT. In this paper, one Commercial Machine Translation software package, Honyaku Adapter II (HAII) from NEC, has been chosen in order to demonstrate how it can be utilised in a medium-sized company, Makita Manufacturing Europe Ltd. (MME), where there is no translation department.

The author chose test sets from a Japanese organisation, Japan Electronic Industry Development Association (JEIDA) [Isahara, 95]. However, this experiment focuses only on the manufacturing domain, where the language can be a sublanguage. As can be seen from the example below, test sets are not particularly relevant to the evaluation of sentences from the manufacturing domain.

Thus, instead of using actual sentences from JEIDA's test sets for the evaluation, the author used the test sets' categories for a CMT to translate actual texts from MME. In this way, the result of the evaluation offers immediate feedback to manufacturing companies that require translation services. The experiment procedure was as follows. Firstly, more than five hundred Japanese sentences were chosen from the typical types of document held by Makita Manufacturing Europe Ltd.(MME), and then these were translated using the Commercial Machine Translation software, HAII. Secondly, an evaluation was carried out using criteria from JEIDA's test sets. In this paper, linguistic features of both English and Japanese are discussed from the evaluation experiment in order to make up writing rules for members of staff at Makita Manufacturing Europe. Accordingly, recommendations are made concerning how a company seeking globalization could resolve communication problems by the employment of Commercial Machine Translation.

2 General Background of Commercial Machine Translation Products (Japanese/English) on the Market.

The availability of CMT for non-professional translators has increased tremendously over the last five years. Ten years ago, MT systems in Japan, like MT systems in other countries, were only run on large scale main frame computers or on UNIX workstation. The cost of the system was in the region of £20k and the market being targeted consisted of professional translators and translation agencies [Johnson 97]. Hence target customers were organisations where the volume of translation required was high, so the MT system is paid off.

Now, many MT packages cost under 9,800 yen (approximately 100 US dollars) in the market. This is an incredibly low price compared to the price of MT software packages being sold in other countries. For example, SYSTRAN sells English to Japanese and Japanese to English MT software for 995 US dollars, and the same product is sold at 128,000 yen (approximately 1,000 US dollars) on the Japanese market. According to Ian Hutchins the quality of affordable MT in the European market is not as high as Japanese counterparts [Hutchins, 99].

Owing to the fact that any customer can pick up a CMT at any software shop in Japan, a user's guide is the most valuable information source for users, partly because it is also the only source of information. However, very few user's guides give a thorough explanation about the weaknesses of CMT and information about how to make the most of it For instance, it asks users to cut long sentences into shorter sentences and to add subjects and objects if they are omitted. This may sound helpful, but unless an explanation is provided as to what the subject and object are in a sentence, this type of guidance is not practical at all.

If a user is an MT researcher, such as an author who understands the weaknesses and strengths of the MT, this single guidance page may suffice. Yet most users are not MT specialists. They require more information and an explanation in order to comprehend MT capabilities. The developer should remember that users of MT are not necessarily well educated or well informed in the field of linguistics. The lack of quality in user's manuals creates disappointment among CMT users who have high expectations the product when purchasing. This report will enhance the use of MT in the manufacturing industry.

3 HAII dictionary and Creating a User's Dictionary

Creating a user's dictionary is a vital part of the customisation of MT. The dictionary of HAII stores 100,000 words. A user's dictionary can be created in order to improve the quality of translation. NEC, a developer of HAII, regularly updates the dictionary for newly coined words and offers it to customers over the internet. Also, user's dictionaries are shared and the company makes them available to other users on the internet. The company is hoping to offer a facility which enables their users to share the user's dictionary from the MT products of their competitors [Kamei et al 97]. Yet it is not sufficient and the current situation requires the building of a user's dictionary.

It is easy to say that you can create your own dictionary and that you will then receive a high quality translation. However, creating a user's dictionary is not as simple as it may look. First of all the creator must possess knowledge about basic Japanese and English grammar in order to build a useful dictionary.

For example, if the user wants to enter a Japanese verb, he/she ought to know the type of verb: There are eight types of verbs in Japanese depending on how they conjugate. Therefore, it is essential for the user to acquire basic Japanese grammar as well as English grammar. It is vital for him/her to know a part of speech in both Japanese and English, in order to produce a user's dictionary. So how can this be expected of a user who does not know much about English grammar? Probably he/she would claim that this is the reason why he/she bought this MT package. If they do, they would not need to use MT. Guidance is therefore needed for a user to create a user's dictionary. Also it has to be pointed out that native Japanese engineers do not necessarily have sufficient knowledge of Japanese grammar.

Furthermore the survey shows that the use of a sublanguage within a company is so significant that it is not straightforward or realistic to impose the same terminology among different organizations [Nikkei 28 May 99].

4 Industrial Background of Japanese Companies

This section briefly explains the background of the globalization of the Japanese manufacturing industry, and also the background of the Makita Manufacturing Europe (MME) where the author conducted the MT evaluation experiment.

MME started its operation in the early 1990's in Telford. It is a power tool manufacturer for professionals, and they decided to produce power tools for the European market. They brought machines and equipment from Japan to set up production lines, and training has been carried out using manuals from Japan. When the necessity for translation of these documents was identified, they decided to hire an in-house translator. There are more than 25 Japanese manufacturing companies like MME in Telford.

In most cases the Japanese companies try to use English for their communication. Yet it is not easy for Japanese employees to compose documents in English due to the fact that communication skills were not widely taught in the Japanese education system until recently.

5 Evaluation of Performance

Since its development MT evaluation has been discussed extensively, yet there is not one standard evaluation method due to the fact that MT evaluation has different requirements for the different MT stakeholders such as translators, information consumers, managers, researchers and novice users. On top of this, sometimes there is more than one translation example for a particular sentence from the source text [White 98, Jones 98]. It is possible to carry out evaluation of a CMT either from a user's point of view or a developer's point of view. There are more than a few research groups which have conducted research on MT evaluation in Japan [Nagao 96, Ikehara et al 94, Tomita 92, Ikehara et al 92] as well as in other countries [White and Taylor 98, Povlsen et al 98, Bech 97]. This paper does not discuss the evaluation itself but focuses on linguistic features identified in the evaluation experiment.

In the following sections linguistic features of the original text from Makita Japan are analysed in order to make up writing rules for members of staff at Makita Manufacturing Europe.

5.1 Result of the Experiment from HAII and MME Language

This section presents the investigation on the CMT experiment done at Makita Manufacturing Europe (MME). The author worked for the company as an in-house translator from October 1996 to November 1997.

5.1.1 MME vocabulary word selection

Each company may use different terminology when they mean the same object or act. MME is no exception. There are a few words that are used at MME which cannot be found in a technical dictionary. Therefore the author needed to input the MME vocabulary into the dictionary of Honyaku Adapter II (HAII). For example, as MME manufactures power tools, a few electrical tests are required, and 'taiatsu shiken' is one of them. The word is found in a technical dictionary as 'endurance test', however, MME uses the English terminology, 'withstanding test' as in the following table.

Table 1:

Japanese (MJ document)	Dictionary (English translation)	MME Term (English)
Taiatsu shiken	endurance test	Withstanding test

Moreover, engineers at Makita Japan (MJ) sometimes write manuals with regional dialect. MT cannot process some expressions that are only used in a certain area in Japan.

[n: noun p: particle v: verb]

e.g.	**Sagyosha**	**ga**	**boranai**	
	n	p	v	
	operators		not	wait for

	You	**ni**	**ki**	**wo**	**tsukenu**
	p	p	n.	p	v
			to		care take

(Translation by a human translator: Care should be taken for not letting operators wait for the material.)

The second example is always used in an operation manual. It is grammatically wrong in Japanese because it uses the wrong particle 'ni' (meaning 'at'), but it should use 'wo' (object marker) instead.

e.g.	**Hako**	**na**	**narabe**
	n	p	v
	box	in	line up

(Place the carton on the line.)

The last word ‘narabe’ also has a missing element. It misses out the ending ‘ru’. ‘Narabe’ is the stem of the verb ‘naraberu’. If the sentence has to be rewritten in formal Japanese, it will be ‘Hako wo naraberu’.

We can easily change the definition of a word in a dictionary of HAII. Therefore, the author changed the definition of a word when necessary. For example, HAII translated the word ‘okyaku san’ into ‘a guest’. This word can be translated into a few Japanese words, but ‘a customer’ is more appropriate for an MME document.

When HAII comes across with a new word, it leaves it out. Consequently the translated sentence includes odd Japanese word in Japanese characters. Since Makita uses special terminology, whenever this case arises, it has to be post-edited and the word registered in the user’s dictionary. Here are some examples:

Table 2:

Japanese (MJ document)	Translated by HAII	MME Term (English)
Shijisho	Indication book	Instruction manual
Shimetsuke torque	Fastening-up torque	Tightening torque

5.1.2 (1) Long Sentences

Inputting short sentences for MT is a typical instruction written in the user’s guide when we buy an MT system. However, HAII could manage to translate rather long sentences without any pre-editing as in the following example.

[n: noun p: particle v: verb adj: adjective adv: adverb]

e.g.	**‘Akafuda-sakusen**	**to wa**
	n	p p
	red card strategy	as for

akai	**fuda**	**wo**	**tsukatte**
adj	n	p	v
red	tag		attach

kojo	**ni**	**habikotteiru**
n	p	v
factory	in	be rampant

aka	**wo**
n	p
dirt	

dare	**ni**	**demo**
n	p	p
everyone	to	even

wakaru	**yo**	**ni**	**suru**
v	p	p	v
understand	in	order	to do

seiri	**no**	**yarikata**	**desu**
n	p	n	p
sorting out	of	method	it is

HAII produced the following translation for the above Japanese sentence in a document. 'The red card strategy is a way of the arrangement to make anyone understand the dirt which is being rampant at the factory using the red card.'

Although the translation was not very natural, it is understandable by British engineers and team leaders in a factory. It has to be noted that there are always about ten Japanese members of staff at MME, and they can explain if British members are not sure of the meaning of the translation. However, the experiment shows that a sentence which does not have a subject and object and has too many embedding phrases is not to be translated satisfactorily.

5.1.2 (2) Changes Have to Be Given in Original Sentences

Since Japanese engineers never do a formal technical writing course they often enter sentences which are ambiguous. Also, the fact that a subject and object can be omitted in a Japanese sentence allows them to write more ambiguous sentences.

For example, there are many sentences without subjects and objects in Makita document, which need pre-edition in order to be processed by MT, as in sentence 87:

e.g.

Kojo	**dewa**	**okakusan**	**ni**
n	p p	n	p
factory	in	customers	by

yorokonde	**itadekeru**	**shinamono**
v	p	n
feel happy	(polite ending)	products

wo	**tukuru**	**tameni,**
p	v	p p
	make	in order to

mainichi	**isshoukenmei**	**desu**
n	adv.	p
everyday	hard	

The above sentence omits the subject of the sentence, watashi tachi (we).

Therefore, the translated sentence generated by HAII did not convey the correct meaning. After adding the subject of the sentence, the MT produced the acceptable translation as follows:

'At the factory, to make the product which it is possible for the customer to be glad about, we are strenuous every day.'

5.1.2.(3) Imperative Sentences

Instruction manuals are full of imperative sentences. '-Suru koto' is the phrase used as a command form. 'Koto' is a nominalizer attached after a declarative statement. However, 'koto' also has a different meaning, 'thing' in Japanese. HAII translated 'koto' into the latter meaning, therefore the pre-editor should have rewritten the whole sentence using the imperative form. (Te-form of verb + kudasai)

5.1.2 (4) Phrasal Verbs and Idiomatic Expression

In case of translating idiomatic expressions, HAII has a disadvantage. For example, in sentence 11, HAII broke up the idiomatic expression into individual words and translated each word separately. Therefore, the sentence is not meaningful and needs to be post-edited. For example:
e.g.

ato	**wo**	**tatanai**
n	p	v
back	not	cut off

HAII translation is '<u>doesn't cut off the back</u>'. HAII could not recognise the phrasal verb 'ato wo tatanai' (never stop), so it translated word by word.

5.1.2(5) Problems of Double Subjects

In most languages, a simple sentence has only one subject (nominative case), whereas in Japanese, many adjectives and some verbs can dominate two surface subject cases within a simple sentence. This is called double-subject construction which confuses MT when analysing sentences [Oku 96] .

6 Suggestions in Document Preparation (Writing Rules)

The quality of input sentences affects the quality of translation, so staff at MME should be provided with adequate information on this subject. As we know, natural language is dynamic and flexible, and ambiguity is inevitable. We must emphasise this point strongly to CMT users, and educate them how to utilise CMT in their organisations. CMT is too good to be ignored or abandoned, though it does not have human intelligence. Customers will be disappointed when a product fails to meet expectation. Therefore, adequate PR is needed for penetrating the market in order to enhance MT user awareness. The following are basic writing rules for MME [see Appendix].

7 Summary of the Report

The result of the study can be utilised for enhancing communication within a small to medium sized organisation such as MME. As the globalization of Japanese manufacturers becomes common, problems caused by lack of communication can seriously affect management. MT systems are certainly not magic wands which resolve all the communication problems occurring at MME, but they will without doubt help enhance communication between British and Japanese staff if we all know how to utilise them. They will help decrease the time and cost of rough translation. The point is how we can make the most of the existing system.

Japanese managers tend to think that MT is for Japanese staff only. However, it is also a useful tool for British staff. Viewpoints of British staff have rarely been discussed by researchers. British employees working at Japanese factories are used to listening to English spoken by Japanese staff. Translated documents can be understood even if they are not flawless sentences. As long as British employees see English texts, they are more motivated. If there is a substantial amount of Japanese documents left without translation in the company, British employees feel they are not in a team. They suspect that any hidden information is in a Japanese document. The problem of this psychological barrier between Japanese and British employees within the Japanese manufacturing companies overseas is often identified and it should not be ignored if the productivity of the factory is to be enhanced. Therefore, it is also crucial for the management of any Japanese manufacturing company to make sure that British engineers can also have access to an MT system when they introduce it.

The introduction of the new system may look exciting and promising in a company, but we must make sure that we use it as a part of our work. The author recommends that MME should appoint one person who is responsible for MT and keep updating the system constantly. He/she must report regularly on how it is utilised within the company. In this way, MT systems can find their place in a company. Otherwise, the novelty will soon wear off and no one will bother to switch it on.

References

Hutchins, W. J. and Somers,.H. L. *An Introduction to Machine Translation*, Academic Press, UK, 1992

Ikehara, S., Shirai, S., and Ogura, K. *Criteria for evaluating the lunguistic quality of Japanese to English machine translations*, Journal for Japanese Society for Artificial Intelligence. Vol.9 1994

Isahara, H. *JEIDA:s Test-Sets for Quality Evaluation of MT Systems - Technical Evaluation from the Developer's Point of View* --, Proc. Of MT Summit V, Luxembourg, July 19 - 22, 1995

Johnson, I., *Personal Translation Applications*, Proc. Of ASLIB 97 Translating and the Computer 19, November 13 and 14, 1997, London

Jones, I., *A translation service to maximize quality and efficiency*, Proc. Of ASLIB 98 Translating and the Computer 20, November 12 and 13, 1997, London

Kamei, S., Itoh, E., Fujii, M., Hirai, T., Saitoh, Y., Takahashi, M., Hiyama, T., Muraki, K., *Sharable Formats and Their Supporting Environments for Exchanging User Dictionaries Among Different MT Systems as a Part of AAMT Activities*, Proc. Machine Translation Summit VI, pp132-141, 1997, San Diego, USA

Tatakau Nihongo, (Fighting Japanese), p.1, 28 May 1999, Nikkei Shinbun Newspaper, Tokyo, Japan

Nagao, M.(Editor and author), Sato, S., Kurohashi, S., Tsunoda, T., Shizen Gengo Shori (Natural Language Processing) in Japanese 1996 Iwanami Shoten Publishing Company, Tokyo, Japan

Oku, M., *Analysing Japanese Double-Subject Construction having an Adjective Predicate*, COLING 96 Proc. Pp. 865-870 1996]

Povlsen, C, Underwood, N., Music, B., Neville, A. *Evaluating Text-type Suitability for Machine Translation a Case Study on an English-Danish MT System* Proc. Of the First

International Conference on Language Resources and Evaluation, Granada pp.27-31, 28-30 May, 1998, Spain,

Shirai, S., Yokoi, A., Okuyama, N., Kawamura, M., Ikehara, S., *The Support System for end users to create dictionary of English-Japanese Combining Pattern*, Proc. Of the Fourth Annual Meeting of the Association for Naural Language Processing in Japan pp.568-571, Kyushu, Japan, March 23-26,1998

White, John, *MT Evaluation*, Tutorial given at MT Summit VI, San Diego: Association for Machine Translation in the Americas, 29 Oct.-1 Nov. 1997, Sandiago, USA,

APPENDIX [Rules for Users of HAII]

Here are basic rules for members of staff at MME to make most of the commercial machine translation system, HAII.

1. Input a grammatical sentence.

(Do not omit the subject or object of a sentence, or do not input a fragment of a sentence.)

2. Input the correct kanji.

(A person who uses HAII in the company should check whether he/she is inputting the correct kanji or not.)

3. Avoid using a phrasal verb.

4. Use a hiragana or kanji character, and avoid using katakana for nouns which are not borrowed words from foreign languages.

(HAII sometimes does not recognise a word written in katakana character.)

5. Avoid using a comma, but use a word ‘to’ (and).

6. Avoid using kanji for the verb ‘okonau’ because it tends to mistranslate.

7. Avoid a sentence with embedded clauses.

(Make it simple and split up into more than one sentence.)

8. Specify singular or plural if necessary.

(HAII treats a noun as a singular form.)

9. Avoid using ‘koto’ or dictionary form when making an imperative sentence.

(Although it is a very typical way of making up an imperative form in Japanese, HAII has difficulty in translating these types of imperative sentences.)

10. Use a sentence from example translation database of HAII when translating the opening and closing part of a business letter.

(Set phrases from typical Japanese business letters should be translated as a whole sentence, but not word by word.)

11. Avoid using a word which has more than one meaning.

(It is better to use the consistent terminology, so review the definition of well-used words within the company.)

12. Input Makita terminology.

(Most terminology is not in the HAII dictionary, therefore it is necessary to input them in the dictionary.)

13. Remember that HAII dictionary only carries one definition for one word.

(It takes a while until the system is tuned. Since each entry of a lexicon from HAII dictionary cannot store more than one meaning, it is advisable to use an entry in only a single domain such as a translating operation manual.)

Book Review

Gregory Grefenstette (ed) (1998) *Cross-Language Information Retrieval*, Boston/Dordrecht/ London: Kluwer Academic Publishers, ISBN 0 7923 8122 X, I-VII and 182 pp, £78.25, hardback

Cross Language Information Retrieval (CLIR) addresses the growing need to access large volumes of data across language boundaries. The typical requirement is for the user to input a free form query, usually a brief description of a topic, into a search or retrieval engine which returns a list, in ranked order, of documents or web pages that are relevant to the topic. The search engine matches the terms in the query to indexed terms, usually keywords previously derived from the target documents. Unlike monolingual information retrieval, CLIR requires query terms in one language to be matched to indexed terms in another. Matching can be done by bilingual dictionary lookup, full machine translation, or by applying statistical methods. A query's success is measured in terms of recall (how many potentially relevant target documents are found) and precision (what proportion of documents found are relevant). Issues in CLIR are how to translate query terms into index terms, how to eliminate alternative translations (e.g. to decide that French 'traitement' in a query means 'treatment' and not 'salary'), and how to rank or weight translation alternatives that are retained (e.g. how to order the French terms 'aventure', 'business', 'affaire', and 'liaison' as relevant translations of English 'affair'). Grefenstette provides a lucid and useful overview of the field and the problems. The volume brings together a number of experiments and projects in CLIR.

Mark Davies (New Mexico State University) describes Recuerdo, a Spanish retrieval engine which reduces translation ambiguities by scanning indexes for parallel texts; it also uses either a bilingual dictionary or direct equivalents from a parallel corpus in order to compare results for queries on parallel texts. Lisa Ballesteros and Bruce Croft (University of Massachusetts) use a 'local feedback' technique which automatically enhances a query by adding extra terms to it both before and after translation; such terms can be derived from documents known to be relevant to the query. Christian Fluhr at al (DIST/SMTI, France) outline the EMIR (European Multilingual Information Retrieval) and ESPRIT projects. They found that using SYSTRAN to machine translate queries and to access material from various multilingual databases produced less relevant results than a method referred to as 'multilingual reformulation' (the mechanics of which are only hinted at).

An interesting technique is Latent Semantic Indexing (LSI), described by Michael Littman et al (Brown University) and, most clearly, by David Evans et al (Carnegie Mellon University). LSI involves creating matrices of documents and the terms they contain and 'fitting' related documents into a reduced matrix space. This effectively allows queries to be mapped onto a common semantic representation of the documents.

Eugenio Picchi and Carol Peters (Pisa) report on a procedure to create links between translation equivalents in an Italian-English parallel corpus. The links are used to construct parallel linguistic contexts in real-time for any term or combination of terms that is being searched for in either language. Their interest is primarily lexicographic but they plan to apply the same procedure to comparable corpora, i.e. to texts which are not translations of each other but which share the same domain.

Kiyoshi Yamabana et al (NEC, Japan) address the issue of how to disambiguate between alternative translations of query terms. Their DMAX (double maximise) method looks at co-occurrence frequencies between both source language words and target language words in order to arrive at the most probable translation. The statistical data for the decision are derived, not from the translation texts but independently from monolingual corpora in each language. An interactive user interface allows the user to influence the selection of terms during the matching process.

Denis Gachot et al (SYSTRAN) describe the SYSTRAN NLP browser, a prototype tool which collects parsing information derived from a text or corpus previously translated with SYSTRAN. The user enters queries into the browser in either a structured or free form and receives grammatical and lexical information about the source text and/or its translation. The retrieved output from a query including the phrase 'big rockets' may be, for instance, a sentence containing 'giant rocket' which is semantically ranked above 'military rocket'.

David Hull (Xerox Research Centre, Grenoble) describes an implementation of a weighted Boolean model for Spanish-English CLIR. Users construct Boolean-type queries, weighting each term in the query, which is then translated by an on-line dictionary before being applied to the database. Comparisons with the performance of unweighted free-form queries ('vector space' models) proved encouraging.

Two contributions consider the evaluation of CLIR systems. In order to by-pass the time-consuming and expensive process of assembling a standard collection of documents and of user queries against which the performance of an CLIR system is manually assessed, Páriac Sheridan et al (ETH Zurich) propose a method based on retrieving 'seed documents'. This involves identifying a unique document in a database (the 'seed document') and, for a number of queries, measuring how fast it is retrieved. The authors have also assembled a large database of multilingual news documents for testing purposes. By storing the (fairly short) documents in a structured form tagged with descriptor codes (e.g. for topic, country and area), the test suite is easily expanded while remaining consistent for the purposes of testing. Douglas Ouard and Bonne Dorr (University of Maryland) describe an evaluation methodology which appears to apply LSI techniques in order to filter and rank incoming documents designed for testing CLIR systems.

The volume provides the reader an excellent overview of several projects in CLIR. It is well supported with references and is intended as a secondary text for researchers and practitioners. It highlights the need for a good, general tutorial introduction to the field.

Derek Lewis, University of Exeter

Conferences and Workshops

The following is a list of recent (i.e. since the last edition of the MTR) and forthcoming conferences and workshops. Telephone numbers and e-mail addresses are given where known (please check area telephone codes).

23–27 August 1999
TKE99: 5th International Conference on Terminology and Knowledge Engineering
Innsbruck, Austria
http://gtw-org.uibk.ac.at/tke.html

1–3 September 1999
NTCIR/IREX Joint Workshop
KKR Hotel Tokyo, Tokyo, Japan
http://www.rd.nacsis.ac.jp/~ntcadm/workshop/joint/

7–10 September 1999 |
34th Colloquium of Linguistics: Linguistics on the Way into the New Millennium
University of Mainz, Germany
Tel: +49 7274 508457, fax: +49 7274 508429
http://www.fask.uni-mainz.de/lk/

5–7 November 1999
4th TELRI European Seminar: Text Corpora and Multilingual Lexicography
Bratislava, Slovakia
Tel: +49 621 1581427, fax: +49 621 1581415
http://www.telri.de

10–11 November 1999
ASLIB: Translating and the Computer 21
1 Great George Street, London, SW1
Tel: +44 (0)20 7903 0032, fax: +44 (0)20 7903 0011, e-mail: barbara.hobbs@aslib.co.uk
http://www.aslib.co.uk

3–4 December 1999
NLULP99: 6th International Workshop on Natural Language Understanding and Logic Programming
Las Cruces, New Mexico, USA
http://www.lim.univ-mrs.fr/NLULP99

4 December 1999
CALL99: Improving student performance in language learning through ICT
University of Warwick, Coventry
Tel: +44 (0)171 379 5101 ext 240, fax: +44 (0)171 379 5082, e-mail:
confs.direct@cilt.org.uk

10 December 1999
CLIN99: Computational Linguistics in the Netherlands
Utrecht University
E-mail: Paola Monachesi <clin99@let.uu.nl>

9–11 March 2000
JADT 2000: 5th International Conference on the Statistical Analysis of Textual Data
École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Tel : +41 21 693 27 35, fax : +41 21 693 52 25, e-mail: jadt2000@lia.di.epfl.ch
http://liawww.epfl.ch/jadt2000

22–24 March 2000
ACIDCA2000: Corpora and NLP
Monastir, Tunisia
Fax: +44 216 4 296 229, e-mail: R.Mitkov@wlv.ac.uk

23–25 March 2000
UNTELE 2000: 3rd Conference on the Use of New Technologies in Foreign Language Teaching. Virtual Environments and Language Learning
E-mail: untele@utc.fr, http://www.utc.fr/~untele

6–7 April 2000
3rd Annual CLUK Research Colloquium
University of Brighton and University of Sussex
Carole.Tiberius@itri.brighton.ac.uk
http://www.cogs.susx.ac.uk/lab/nlp/cluk/

29–30 April 2000
CLAW 2000: 3rd International Workshop on Controlled Language Applications
Seattle, Washington, USA
http://www.up.univ-mrs.fr/~veronis/claw2000

29 April–3 May 2000
ANLP-NAACL2000: Language Technology Joint Conference on Applied Natural Language Processing (North American Chapter of the Association for Computational Linguistics)
Seattle, Washington, USA
http://www.gte.com/anlp-naacl2000

31 May–2 June 2000
LREC 2000: The European Language Resources Association (ELRA)
Athens, Greece
http://www.icp.grenet.fr/ELRA/lrec2000.html

21–25 July 2000
ALLC/ACH 2000: Association for Literary and Linguistic Computing Association for Computers and the Humanities Joint International Conference
University of Glasgow, Scotland, UK.
E-mail: J.Anderson@hatii.arts.gla.ac.uk

31 July – 4 August 2000
COLING 2000: 18th International Conference on Computational Linguistics
Saarbruecken (conference), Luxembourg (workshops), Nancy (tutorials)
E-mail: uszkoreit@dfki.de, or: kay@parc.xerox.com
http://www.coling.org/call.html

23–25 August 1999
TMI99: 8th International Conference on Theoretical and Methodological Issues in Machine Translation
Chester, UK
E-mail: Arturo Trujillo (iat@ccl.umist.ac.uk), Harold Somers (harold@ccl.umist.ac.uk)
http://www.ccl.umist.ac.uk/events/tmi99/

13–17 September 1999
MT Summit VII: MT in the Great Translation Era
Singapore
E-mail: secret-4@tokyo.intergroup.co.jp, or: vicky@krdl.org.sg
http://www.jeida.or.jp/aamt/mts99.html

5–7 November 1999
4th TELRI (Trans-European Language Resources Infrastructure) European Seminar: Text Corpora and Multilingual Lexicography
Bratislava, Slovakia
Tel: +49 621 1581 427, 6-13, fax: +49 621 1581 415
http://www.telri.de

5–7 November 1999
NLPRS99: 5th Natural Language Processing Pacific Rim Symposium
Beijing, China
Tel: +82 42 869 3565, fax: +82 42 867 3565, e-mail: nlprs99@korterm.kaist.ac.kr
http://korterm.kaist.ac.kr/~nlprs99

3–4 December 1999
NLULP99: 6th International Workshop on Natural Language Understanding and Logic Programming
Las Cruces, New Mexico, USA
http://www.lim.univ-mrs.fr/NLULP99/

20–22 December 1999
IWPT99: 6th International Workshop on Parsing Technologies
Trento, Italy
Tel: +31 13 466 3060, fax: +31 13 466 3110, e-mail: Harry.Bunt@kub.nl
http://wwwseti.cs.utwente.nl/Docs/parlevink/sigparse/

6–18 August 2000
ESSLLI2000: 12th European Summer School in Logic, Language and Information
Birmingham, UK
Tel: +44 (161) 275 6170, fax: +44 (161) 275 6204, e-mail: franconi@cs.man.ac.uk

http://www.folli.uva.nl/Esslli/2000/esslli-2000.html
8–12 August 2000
EURALEX 2000: 9th EURALEX International Congress
Stuttgart, Germany
Fax: +49 711 121 1366, e-mail: elx2000@ims.uni-stuttgart.de
http://www.ims.uni-stuttgart.de/euralex

31 August – 2 September 2000
EUROCALL 2000: Innovative Language Learning in the Third Millennium
University of Abertay, Dundee, Scotland, UK
Fax: +44 (0)1482 473816, e-mail: eurocall@hull.ac.uk

13–16 September 2000
TSD 2000: 3rd International Workshop on Text, Speech and Dialogue
Brno, Czech Republic,
E-mail: tsd2000@fi.muni.cz, http://www.springer.de/comp/lncs/authors.html

14–16 September 2000
CercleS International Conference: New Challenges for Language Centre Management
Antwerp, Belgium
E-mail: Julie Venner <J.C.Venner@selc.hull.ac.uk>

3–6 October 2000
ACL2000: 38th Annual Meeting of the Association for Computational Linguistics
Hong Kong, China
E-mail: acl2k@cis.udel.edu
http://www.mri.mq.edu.au/conf/acl99/

20–22 November 2000
MT 2000: Machine Translation, Multilingualism and the Millennium
University of Exeter, UK
Tel/fax: +44 (0)1392 264296, e-mail: D.R.Lewis@exeter.ac.uk
http://www.bcs.org.uk/siggroup/sg37.htm

MEMBERSHIP: CHANGE OF ADDRESS

If you change your address, please advise us on this form, or a copy, and send it to the following (this form can also be used to join the Group):

Mr. J.D.Wigg
BCS-NLTSG
72 Brattle Wood
Sevenoaks, Kent TN13 1QU
U.K.

Date: ....../....../......

Name: ....
Address: ....
....
Postal Code: .... Country: ....
E-mail: .... Tel.No: ....
Fax.No: ....

Note for non-members of the BCS: your name and address will be recorded on the central computer records of the British Computer Society.

Questionnaire

We would like to know more about you and your interests and would be pleased if you would complete as much of the following questionnaire as you wish (please delete any unwanted words).

1. a. I am mainly interested in the computing/linguistic/user/all aspects of MT.
   b. What is/was your professional subject? ....
   c. What is your native language? ....
   d. What other languages are you interested in? ....
   e. Which computer languages (if any) have you used? ....

2. What information in this Review or any previous Review have you found:
   a. interesting? Date ....
   ....
   ....
   b. useful (i.e. some action was taken on it)? Date ....
   ....
   ....

3. Is there anything else you would like to hear about or think we should publish in the *MT Review*?
   ....
   ....
   ....
   ....

4. Would you be interested in contributing to the Group by,

   a. Reviewing MT books and/or MT/multilingual software
   b. Researching/listing/reviewing public domain MT and MNLP software ....
   c. Designing/writing/reviewing MT/MNLP application software ....
   d. Designing/writing/reviewing general purpose (non-application specific) MNLP .... procedures/functions for use in MT and MNLP programming ....
   e. Any other suggestions? ....
   ....
   ....
   ....

Thank you for your time and assistance.