A research perspective on how to democratize machine translation and translation aids aiming at high quality final output

Christian BOITET

GETA, CLIPS, IMAG-campus, F-38041 Grenoble cedex 9, France

Christian.Boitet@imag.fr

Abstract

Machine Translation (MT) systems and Translation Aids (TA) aiming at cost-effective high quality final translation are not yet usable by small firms, departments and individuals, and handle only a few languages and language pairs. This is due to a variety of reasons, some of them not frequently mentioned. But commercial, technical and cultural reasons make it mandatory to find ways to democratize MT and TA. This goal could be attained by: (1) giving users, free of charge, TA client tools and server resources in exchange for the permission to store and refine on the server linguistic resources produced while using TA; (2) establishing a synergy between MT and TA, in particular by using them jointly in translation projects where translators codevelop the lexical resources specific to MT; (3) renouncing the illusion of fully automatic general purpose high quality MT (FAHQMT) and go for semi-automaticity (SAHQMT), where user participation, made possible by recent technical network-oriented advances, is used to solve ambiguities otherwise computationnally unsolvable due to the impossibility, untractability or cost of accessing the necessary knowledge; (4) adopting a hybrid (symbolic & numerical) and "pivot" approach for MT, where pivot lexemes are UNL or UNL inspired English-oriented denotations of (sets of) interlingual acceptions or word/term senses, and the rest of the representation of utterances is either fully abstract and interlingual as in UNL, or, less ambitiously but more realistically, obtained by adding to an abstract English multilevel structure features underspecified in English but essential for other languages, including minority languages.

Keywords

Machine Translation, Translation Aids, MT/TA synergy, Memory-Based TA, Generalized Lexical Contribution, mixed MT approaches, SAHQMT

Introduction

Why speak about democratizing Machine Translation (MT) and Translation Aids (TA) when so many MT systems and computerized lexicons are available off the shelf at cheap prices? Because what is on the market is adequate for "MT for watchers", or "informative MT" (see [8] for this terminology), or "dictionary aids", but inadequate when it comes to quality MT output or really efficient TA, necessarily based on large translation memories.

More precisely, Machine Translation (MT) systems and Translation Aids (TA) aiming at cost-effective high quality final translation are not yet usable by small firms, departments and individuals, and handle only a few languages and language pairs.

This is due to a variety of reasons, some of them not frequently mentioned, which are briefly analyzed in section 1, in which we also detail reasons for democratizing MT and TA. The following sections (2—4) explain in more detail the methods we propose to combine to attain this goal.

1 Why quality MT and TA are only for niches and rich users

1.1 Machine Translation (MT)

Quality MT is almost always equated with fully automatic high quality MT (FAHQMT). But the last 50 years of R&D in FAHQMT have amply demonstrated that it is only possible in restricted typologies of texts (domain, grammatical constructions & semantic interpretations), and cost-efficient if the volume is very large (between 5 and 10 million words)[1].

A model of the situation could be the (tentative) formula: « Coverage * Quality = K » for a certain K which maximum depends on the MT technology used and which real value is determined by the level of "elbow grease" (human sweat) invested in the linguistic knowledge and practical know-how encoded in the system, as well as by the suitability of the task to this suboptimization approach [23] of "MT for translators". For example, according to J. Chandioux, weather bulletins are suitable for MT

[1] Think of the difficulty to translate "MT and TA" by "TA et MT" in French — because, possibly, "Machine Translation" = "Traduction Automatique" and "Translation Aids" = "Machines pour Traducteurs" !

[15] (and not really for TA), but weather alerts are not (and can be handled by TA).

As a result, the primary users of FAHQMT are almost always specialized posteditors, competent in both languages and in the domain, and employed by large and rich organizations.

A further obstacle to the wider availability of quality MT is the proprietary policy of MT vendors, who allow users to modify the lingware only by adding items to user dictionaries and modifying the list and priorities of the dictionaries.

This is too bad, because further significant improvement could be obtained if their clients, companies or organizations, could employ specialists to improve the details of all dictionaries and to tailor the rest of the lingware (grammar, heuristics, semantic restrictions and interpretations) to the typology at hand. Another advantage is that clients could then keep their texts and terminology confidential.

This would be a better and practical marketing strategy, as exemplified by the case of expert systems, which have been deployed successfully at a large number of sites and are served by in-house specialists.

1.2 Translation Aids (TA)

To be cost-effective, quality TA must be based on large translation memories and used by professionals working on large translation tasks presenting a high repetition rate. A typical case is that of successive versions of technical documents.

However, that is not enough. To match an input fragment on a translation memory, the current TA systems depend on the specific format of the memory. If the documents to be translated come in a large variety of formats and the translation delay is very short, memory-based TA (MBTA) systems become unusable.

For that reason, Hewlett Packard Grenoble localization center, which routinely translates and prints a 200 pages document from English into about 30 languages in 2 weeks, has abandoned the idea to use MBTA after having tried all commercial products.

Another obstacle to the use of MBTA is the complexity of the products and the pricing policies. A young freelance professional translator simply cannot buy the full versions of current quality TA products (including the aligner, the task scheduler and the terminology extractor). Even if s/he could, installing the product may be nightmarish[2]. Running it can be as bad. To find a cheap, simple and powerful TA product is still a dream.

Simple dictionary tools are not as effective as MBTA, but can be quite useful. Unfortunately, they are also quite expensive, and despite that never complete and up-to-date.

It was a surprise to learn from a young professional translator running a 3 person company that the cheapest and most useful translation aid for them was Systran, the result of which they do not postedit, but only consult as a quick, context-sensitive dictionary aid, often more up-to-date than heavy on-line terminological data bases. Rough translation cannot be cost-efficiently revised as quality raw translation, but it is cheap, and useful in other ways.

As a result, the primary users of quality MBTA products are again almost always specialists, translators and posteditors, competent in both languages and in the domain, and employed by large and rich organizations.

1.3 Why "democratize" quality MT & TA?

One could claim that there is no need to change that situation. But commercial, technical and cultural reasons make it mandatory to find ways to democratize MT and TA.

First, the need for producing large quantities of high-quality translations in many professional contexts is dramatically increasing, due to internationalization and global communication facilities. But there are not enough professional translation offices, so that individual professional translators, retired translators and bilingual professional and perhaps also occasional translators, have to take care of a possibly large part of the task. Also, the SMEs do more and more export, but are neither niches nor rich users.

Second, there are many situations where individuals or small groups really need to write in their own language and get quality translations in English and possibly other languages. This is the case of almost all scientists for which English is a second language, but who find it very difficult or impossible to write directly in English efficiently. Another example is the writing of proposals to answer European calls for projects: while the calls are translated by the translation services of the EU, the answers may theoretically be written in any official language of the EU. In practice, however, they are not translated into the language(s) of the reviewers, so that people feel obliged to translate them in English, which, considering the allotted time and the size of the proposals, puts non English native speakers at a decided disadvantage.

Third, multilingual private contexts are also quickly increasing with the Internet revolution. Classrooms are paired between schools of various countries, families send their chidren abroad to learn a foreign language and want to communicate with the host parents, etc.

Fourth, the number of languages to handle is on the rise, for economical, cultural and political reasons. It is not possible any more to limit MT and TA to the 4 or 5 bigger languages of the EU plus Russian, Japanese and Chinese. The public does not want to get the documentation of videorecorders in English, or so badly translated in their language that they become ununderstandable and sometimes dangerous. All languages, be they rich or

[2] That is true even if professional computer scientists are available. We have some bad memories of installing Eurolang Optimizer and the mandatory SQL server on a Windows-NT PC. Offering a set of tools like Xerox Multilingual Suite also calls for the assistance of a system programmer. Integration and simplicity are still on demand.

poor, heavy or light in terms of current industrial translation load, large or small wrt the number of their speakers, simply have to be given access to modern technology.

1.4 Outline of how to democratize MT and TA

This goal could be attained by:

1. promoting a "generalized lexical contribution" by exchanging free tools with resources created by users,
2. developing a synergy between MT and TA, not only by including calls to MT in TA products, but by co-developing parts of the MT and TA linguistic resources,
3. extending the set of MT primary users to the authors of texts, by involving them in interactive disambiguation in a semi-automatic approach, the only one realistic for getting high-quality outputs from arbitrary inputs,
4. adopting a hybrid (symbolic & numerical) and "pivot" approach for MT, the "pivot" lexemes being denoted by English inspired strings for ease of development for dozens of "light" languages in addition to the few "heavy" languages already rich in tools and resources.

The scenario of Method 1 (Generalized Lexical Contribution) is the following: as for Netscape, the client software is free. Anyone wanting to translate a document (which may be in a variety of possible formats such as Word, Interleaf, PageMaker, Excel, Eudora…) first downloads and starts the client TA software (cTA). Then, s/he sends the document to the server TA software (sTA), alongside with useful information (source format and language, target languages, domain and class of document, known similar documents if any…). The sTA filters it into a special fTA format used by all TA components and preprocesses it by: (a) lemmatizing it to retrieve words and phrases with their equivalents in the target language(s), (b) retrieving exact and approximate matches from the translation memory, and (c) enriching the fTA file with the information retrieved.

The sTA then sends back the preprocessed fTA file to the user, by e-mail or push. The user translates it, using the bilingual or multilingual editor and dictionary manager included in the cTA. Using the dictionary manager is more attractive if, as in Eurolang Optimizer, the cTA contains the lemmatizer, so that a word or phrase entered in the dictionary becomes immediately available for the rest of the document. The editor automatically keeps the 2 or more versions of each fragment (paragraph, sentence, phrase in a bulleted list) aligned.

When the job is finished, the user sends the translated fTA file back to the server, to get the target version(s) in the desired format[3]. The sTA applies adequate filters to the fTA file to do that, and sends the resulting final translation(s) in final format(s) back to the user. It also extracts from the fTA file the dictionary modifications done by the user, and prepares them to be revised by the manager of the lexical resources at the server site. It also puts the aligned fragments found in the fTA in the format of the translation memory, so that the manager of the textual resources can screen them before deciding to include them in the translation memory or not.

As no full documents, but only terms and fragments, are kept on the server, there should be no problems of confidentiality or copyright. The user would explicitly have to waive the intellectual property rights on the lexical information added or modified before being allowed to run the cTA.

[3] It may be the input format or another one, e.g., for English-Japanese, Word as input and EgWord or IchiTarou as output.

2 Synergy between MT and TA

2.1 Motivations

Some MBTA products call an MT system when they find no match for a sentence in their memory (Eurolang Optimizer+Logos, TM2+LMT). But this poses problems of lexical incoherence between the TA and MT parts, of quality, and of feasability.

Making TA and MT translation proposals lexically coherent implies, at least, that the TA dictionaries are put in the MT user dictionaries. That is the first degree of the MT/TA synergy.

To really raise the quality of MT translation proposals, it also seems necessary that translators using TA codevelop parts of the MT lexical resources, by entering some detailed syntactic and semantic information not used by TA, and that a team of professional MT developers work to specialize the rest of the MT system (grammars, heuristics…) to the utterances for which MBTA does not work well, hoping that they constitute a sublanguage suitable for MT.

Is that feasible? To specialize a quality MT system costs about 30—50 man.years, while this cost can be amortized on perhaps only 30% to 40% of the input. We outline a possible solution below.

2.2 Specializing an MT system while translating with TA

The classical way to develop and deploy quality MT has been as follows. A large (and rich) organization has to produce quality translations on a large scale into one or more language. People in charge hear about MT, and make preliminary studies which show the potential benefits of using quality MT.

They understand that it is not really possible to acquire an MT system off the shelf and adapt it to the task, because the intrinsic quality limits of the technology of such products are too low. Spending hundreds of hours on tuning user dictionaries will not improve the quality to the required level.

However, they find that no adequate quality MT system exists. An operation is then started to develop such a system. In the mean time, translations are produced the old way. When the MT system is ready, it is deployed to replace the human raw translation phase.

In the case of METEO (Canadian Meteorological Center), initial development took 1 year to about 6 researchers. Packaging and improving the system to a really cost-efficient quality level cost about the same (3 years to 2 developers). But the system can only handle weather bulletins, not weather situations or alerts.

In the case of Caterpillar [28], the developement effort started about 1990. 5 years and 5 M$ later, the MT system was ready for translating into French and Spanish, but the 8 or 9 other target languages were still under way.

As a matter of fact, there are now very few potential users of quality MT ready to invest heavily and to wait several years to see the first returns on their investment. They know that using MBTA (translation memories) can give good returns very quickly (after the first 800 or 1,000 pages have been processed). The "entry ticket" for quality MT has become more expensive since MBTA are available. They also know that the quality MT system delivered is likely to be adequate for only a part of the translation task, so that they will have to install TA, and probably MBTA, to complement MT.

We propose to inverse these 2 steps. First, install a MBTA system and begin to translate. In parallel, develop the MT system, aiming specifically at the parts of the texts not well handled by translation memory based techniques. To develop it, use some software and lingware specialists, half of them from the MT developer team and half from the client. Use also the translators as codevelopers of a lexical data base which will be the common source of the TA and MT dictionaries. Finance the development of the MT system by the cost reduction obtained after the MT system is deployed and coupled to the TA system.

Take the example of a translation task of 60,000 standard pages of 250 words to translate over 3 years. This represents about 80,000 hours of work without TA. With a MBTA system, supposing we get 20% exact matches and 40% approximate matches, productivity is increased by 2 (40,000 hours), because the 20% exact matches cost nothing in human time, and the 40% approximate matches cost only the postedition time, 1/4 of the previous time, or 10%. The last 40% (no match) cost the same as before.

We will compare two situations. In the first, a team of 10 translators use the MBTA and work 100% of their time on the translation job. In the second, the team has the same size, but is composed differently.

During the first year, it is composed of 4 software and lingware developers, and of 6 translators. The translators spend 70% of their time on the translation job, thereby enriching the bilingual or multilingual term memory, and the other 30% on indexing the MR-oriented syntactic and semantic properties of these terms in the lexical data base.

After 12 months, the 6 translators will have spent about 7,000 hours on translation and postedition, producing 10,760 pages, and it is not unreasonable to suppose that the first operational version of the MT system will have been developed, starting from an existing core version. At that time, replace the 2 MT developers coming from the MT vendor by 2 translators. Let the 2 client developers take care of the MT software and lingware, and let the 8 translators continue to translate at 70% and index at 30% (or less if the rate of vocabulary increase decreases). After 30 months, the TA/MT combination (with 6 then 8 translators) will just have beaten the pure TA approach with 10 translators.

We summarize this in the following 2 tables, where we suppose that the MT system is called only on the 40% giving no match. It is however possible to call it on the parts giving approximate matches, and that its translation proposals, which rely on full linguistic analysis, are better than the memory proposals. In such as case, returns could be obtained in less than 30 months (2.5 years).

	h/p v %	**Hyp.1**	**Hyp. 2**
Pages		**10,000**	**60,000**
Raw Human Transl.	1 h	10,000	60,000
Revision	0.33 h	3,333	20,000
HT (hours)	**1.33 h**	**13,333**	**80,000**
Exact corresp.	20%	2,000	12,000
Approx. corresp.	40%	4,000	24,000
Others	40%	4,000	24,000
Raw Human Transl.	1 h	4,000	24,000
Revision	0.33 h	2,667	16,000
TA (hours)	**0.67 h**	**6,667**	**40,000**
Raw Human Transl.	1 h	0	0
Revision	0.33 h	2,667	16,000
TA/MT (hours)	**0.27 h**	**2,667**	**16,000**

Months elapsed	142.33 h	**12**	**30**
Pages/translator	214	0	6,405
(10 transl.+ TA)	**2,135**	**25,620**	**64,050**
Pages by TA	149	1,793	0
6 translators at 70%	897	10,760	0
Pages by TA/MT	374	0	6,725
8 translators at 70%	2,989	0	53,802
Pages by TA/MT		**10,760**	**64,562**

In this scenario, MT vendors do not lose, on the contrary. They codevelop the first specialized version, and avoid from then on to support any obligation of results. That will be the responsibility of the client developer teams. At the same time, MT vendors can sell training, annex tools, and maintenace of the purely software components (development environment, specialized languages…).

2.3 Integrated lexical architecture (TA + MT)

Translators are usually not lexicographers. To let them index MT-oriented properties in the dictionaries, one should limit the sophistication of the linguistic description, or at least organize the description of lexical properties at several levels. That has been successfully done by Sharp for its DUET system: the basic system contains relatively few semantic features, but it is possible to add arbitrarily many domain-related codes in the dictionaries.

These codes are used indirectly by the grammars (through operations on the names of the attributes having these codes as values), so that the grammars can remain the same while the codes are changed.

To ensure lexical coherency between MT and TA components, the best solution seems to centralize lexical information in a unique multilingual lexical data base (MLDB), from which MT and TA active dictionaries are extracted, either by a periodic global compilation step, or, more attractively, in an incremental, on-demand way.

This idea of multi-application MLDB has been stressed by most recent projects on multilingual dictionaries. However, it has been implemented only partially. MT-oriented MLDB have been built to be independent of a particular MT system[4], but no MLDB integrating the terms and informations necessary for MT (general terms as well as terminology, morpho-syntactic categories, predicative frames, syntactico-semantic valencies, semantic features, derivations, word sense identifiers…) and those useful for TA (definitions, examples of use…).

Building such data bases supposes a very open software organization [11, 30], which is delicate but possible to implement with current techniques. Another reason why they donot really exist yet is perhaps that they could only be built in contexts where TA and MT techniques would be tightly integrated, as in the above scenario.

Note again that, for such a scenario to succeed, the MT provider should adopt an open, not proprietary, policy. If that is impossible for private companies, publicly funded MT groups should play that role.

3 Interaction with authors in a SAFQMT approach

3.1 Limits of automatic disambiguation

One very important aspect of quality MT democratization is to build very high quality practical multitarget translation systems usable by individuals. For text translation, very high quality means that, as in the case of the METEO system [15], about 3 to 5% editing operations have to be performed on the output to reach perfection, or, equivalently, that revising a standard page of 250 words to reach professional quality takes less than one minute (instead of 20 for a good raw translation of a technical page, produced by a qualified professional). In METEO, this is possible only because the translation inputs, weather bulletins, are extremely well suited to the heuristic "sub-optimization approach" to MT.

For the applications we have in mind, automatic disambiguation alone is not going to reach that kind of quality level in any foreseeable future. In the case of METEO, automatic disambiguation has indeed permitted to reach a quality level of 95—97%. But this remains an isolated example, and no comparable applications have been found, despite intense research by the CITI at Montréal.

[4] For example, BDTAO built by B'VITAL/SITE for the Ariane/aéro/F-E project, or the lexical data base of the JICST MU/Majestic system in Tokyo.

We have said that METEO handles only a very restricted type of documents, the weather bulletins. Trying to adapt it to the apparently very similar texts of weather alerts necessitated a huge increase in the dictionary size and degraded quality far too much, and J. Chandioux eventually developed a TA environment for that second kind of texts[5].

The best that can be obtained in the case of technical manuals seems to be a revision time of 15 to 10 minutes per page, or, in our numerical approximation, something like a 55 to 70% "quality". A more intuitive grading would be "just good enough" to "quite good". Trying to apply MT for watchers or MT for revisors techniques to very varied texts always leads to such a poor quality that revision is far too costly, or even impossible (the revisors prefer to translate again from scratch).

3.2 Complementarity of automatic and interactive disambiguation

3.2.1 Objective and time limits for interactive disambiguation

When we speak of using interactive disambiguation, people often reject the idea, assuming we would like to use interactive disambiguation only, and the number of questions would be tremendously high. It may be useful to clarify these two points.

First, what can it possibly mean to « use only interactive disambiguation » in an NLP system? Obviously, that the system does not solve any ambiguity at all. But that is never the case. For example, even the most primitive system, having to handle « time flies », with categories (N|V) (N|V), will not admit the sequence (V V). What is meant, then, is that using interactive disambiguation would necessarily lead the systems designer to adopt a lazy strategy and not to solve many ambiguities which "should be" easy to solve automatically.

A possible answer to that is to say that the systems designers should try to establish a kind of hierarchy between ambiguity types according to the difficulty of their automatic solution, and solve only the easy and moderately difficult ones (see [12, 13] for such "ambiguity labelling"). One should stop trying to disambiguate automatically when the results are not reliable or when the efforts are disproportionate with the results: get 90% of the job done with 10% of the effort.

The point concerning the number of disambiguating questions is also interesting. First, note that any system which uses only 100% reliable disambiguation techniques is bound to produce a number of interpretations exponential in the length N of the considered utterance, that is, $O(2^{KN})$ for some K. There are at least two reasons for that: if the words with lexical content represent a proportion P of the words and have in average M distinct meanings for the same morphosyntactic class, we get $2^{PN \log M}$ interpretations, without taking mutual information into account — but this may well be the part in lexical disambiguation we do not want to tackle automa-

[5] Personal communication, June 1996.

tically because it is too difficult. Similarly, as natural languages in general are intrinsically ambiguous, an utterance of length N has an exponential number of "skeleton" bracketings.

Suppose, then, that the number of interpretations of our generic utterance is of the form $A.2^{KN}$. Suppose further that answering a disambiguation question excludes, in average, a proportion Q of the remaining interpretations. If the questions have 2 choices, Q=1/2 in average. If the questions have more choices, say C in average, Q≈(C-1)/C in average. Taking the worst case, Q=1/2, we see that the number of disambiguation questions can be approximated by $\boxed{\text{Nb\_questions} \approx KN \log A}$.

This means that, in all cases where we want to use interactive disambiguation, the number of questions will be a linear function of the number of words.

What we should ask (or be asked), then, is the following: « For a text of N words, how big can be the number of questions per word? ». As questions can be short or long, and as the correct choice may be preselected heuristically in some proportion of the questions, a better formulation is: *« How much time per word can be spent in interactive disambiguation? »*.

3.2.2 Feasibility of interactive text disambiguation

Suppose we want to translate a typical scientific paper of 6,000 words (12 pages in Word, or 24 standard pages of 250 words), it will take a professional about 24 hours to produce a good raw translation, and 8 hours to revise for very high quality. If we translate our own prose, knowing our terminology well, we are not likely to produce a very good result under 12 hours for the first draft and 4 hours for the revision. Our final quality will not be that of a professional, but enough for submission to a journal, or a conference.

Is it possible that an MT system asks the user to answer disambiguation questions during 12 hours, equivalent to what is necessary for producing the draft? Perhaps yes, especially if the translation system is multitarget, because the economy would be 12 hours for each target language, starting with the second. But answering questions may be far more boring than producing a translation draft, which in itself is often quite tedious! Let us suppose, then, that we do not want the user to answer questions for more than 2/3 of that time. For 6,000 words, this means that we could take up to 8 hours for disambiguation questions. Equivalently, we arrive at 20 mn per standard page. It is not unreasonable to say that a user can answer about 10 questions per minute, especially, if the correct choice is preselected more often than not. In terms of number of questions, this means about 20 questions per sentence of 25 words, or about one question per content word (about 15 here) and 5 questions for the rest (attachment, aspect, modality…).

We never tried to write the "dummiest possible" analyzer and to see how many interpretations it would deliver. However, we know by the previous experience of CETA, where the first step of structural analysis was to use a CFG in Chomsky Normal Form, that such binary grammars often lead to the production of many "parasite" ambiguities, that is, of structures which would be the same had they be produced by a "flatter" grammar. With the current trend of going backwards 30 years and favor binary rules, we do not really think it will be possible to disambiguate between all structures with only 5 questions per sentence of 25 words, because 5 questions disambiguate between 32 structures in average. We rather think something like 15 questions (32K structures) would be realistic. In such a case, interactive disambiguation can not be used alone: it is necessary to add some degree of automatic disambiguation to reduce sets of equivalent structures to singletons.

However, our previous experiments in the framework of the LIDIA project, where the all-path analyzer does not produce parasite ambiguities, as well as reports from MicroSoft on their large coverage analyzer for English have led us to believe that a reasonable analyzer performing only 100% sure and relatively easy automatic disambiguation can indeed deliver an ambiguous result which can be disambiguated in about 20 mn per standard page, or 2 mn (≈20 questions) per typical sentence of 25 words. As interactive disambiguation should be done incrementally, when the user feels like it, and not imposed on him/her, it may be better to give the final numbers for sentences, not pages.

3.3 Example scenario

Authors of documents, possibly not knowing the target language, can become primary users of quality MT only if the MT system relies on a user-friendly interactive disambiguation step, and if interactive disambiguation is accepted by the authors, or even better, made interesting. What me mean is that the user should be allowed to ask questions about the disambiguating questions, and to navigate from the current application to "discover" related information.

Suppose that the task is to send the minutes of a meeting by e-mail, in the languages determined by the personal profiles of the addressees associated with their e-mail nicknames. One addressee may want to read such e-mails in the original language if it is in some list, in his language, and in English. Another one may prefer to get it only in his/her language, etc.

After having written the minutes, say in English, the secretary of the meeting runs some (spelling, grammar, style) checkers, and sends the message as usual. The e-mail server processes it, and establishes the list of all target languages. If no translation is required, the message is sent as usual. Otherwise, the e-mail server filters the message into some appropriate format and sends it to an analysis server.

The result is then sent to a disambiguation server which will detect, for each utterance, all ambiguity patterns, and prepare the elements of the associated disambiguation dialogues (support of the ambiguity, rephrasing associated with each possibility, etc.). The ambiguities may concern the source language or the passage into one or more target languages.

The e-mail message is then sent back to the author with a message indicating the presence of ambiguities and an attached file in a special format (fID). When the author clicks on the corresponding icon in this e-mail, interactive disambiguation begins. The interactive disambiguator can run on the user machine, or on the disambiguation server (through a Web browser, or a simple http applet).

Suppose the meeting was about maritime exports/imports with Asia and the following sentence appears: « Our captain brought back blue bowls and plates. » A first question could be:

◎ blue bowls and blue plates
O plates and blue bowls

and the second:

◎ marine captain
O airforce captain
O artillery captain
O infantry captain
O cavalry captain
O …

If the author asks why that question is asked, the disambiguator, knowing that s/he is interested in German, would answer that these senses correspond to different translations in German, and give them. From there, the author could also follow some links to get further explanations (e.g., about "Rittmeister" for "cavalry captain").

After interactive disambiguation has been (totally or partially) performed by the author, the modified fID file is sent back to the e-mail server, which processes it again to send it to translation servers (usual transfers & generators, or UNL deconverters). It then assembles the results into appropriate e-mails and sends them to their addressees, either within a normal e-mail message, or as an attached file to avoid loss of information, in particular for languages using complex writing systems.

4 Hybrid (symbolic & numerical) and "pivot" approach for MT

Our last suggestion for democratizing quality MT has to do with the internal organization of the MT system itself.

4.1 Hybrid symbolic & numerical MT systems

An important point in future large coverage quality MT systems is that they should be adaptable to users and tasks. For this, the best approach seems to combine symbolic and numerical techniques.

Purely symbolic, knowledge-intensive methods have given very good results on restricted tasks, but can not be scaled up and porting such a specialized quality MT system to a another restricted task is costly.

Other methods, purely statistical, or "example-based", have been proposed around 1984, 15 years ago. However, no quality system has resulted. According to an old joke, MT in the USSR was MT without machines and without translations. At IBM, Jelinek used also to say that, each time he fired a linguist, his speech recognizer improved. Unfortunately for him and the purely statistical approach, this has not come true for MT, be it quality MT or informative MT: in a famous DARPA experiment, his system did worse than Systran on fragments from the Hansard corpus (minutes of the Canadian Parliament debates, in English and French), although it has been trained on it and Systran had neve tackled it before.

In other words, quality MT systems must have a symbolic, knowledge-intensive backbone. To make them more "continuous", or adaptable, or personalizable, the best way seems to add a measure of numerical techniques. We studiously avoid more precise terms like "statistic", "fuzzy", etc., because the precise techniques may vary from component to component.

To personalize the lexical data base, it may be enough to handle it as a large Hopfield neural net, that is, a graph with terms and senses on the nodes, attractive arcs between terms and related terms and senses, and repulsive arcs between exclusive senses of the same term. Weights on nodes represent the importance or preference of terms and senses, and weights on arcs attraction or repulsion.

That is the technique adopted at Microsoft labs, where B. Dolan's "lexical priming" technique shows that, without any heavy statistical learning, very good results can be obtained. General terms have 12-15 senses, but, if a sentence is presented to the system, no more than 2-3 senses per word appear as likely in the context. Tuning the weights can be done incrementally by using feed-back from users choices.

The case of the grammars is more delicate. Probabilistic or weighted extended context-free devices seem to be the most robust, but tuning the weights is not as straightforward as for the lexicon.

4.2 Hybrid pivot approach

Finally, we advocate to adopt a "pivot" approach for MT, where pivot lexemes are UNL or UNL inspired English-oriented denotations of (sets of) interlingual acceptions or word/term senses, and where the rest of the representation of utterances is either fully abstract and interlingual as in UNL, or, less ambitiously but more realistically, obtained by adding to an abstract English multilevel structure features underspecified in English but essential for other languages, including minority languages.

The term "hybrid pivot" was coined by Shaumjan, and used at CETA before 1970 to denote representations in which the lexemes are abstract, but language-dependent ("lexical units", or derivational families), while all other attributes and relations are interlingual.

Here, we use "hybrid pivot" in another acception. To handle a large variety of languages, it seems necessary to use interlingual lexemes (IL). We can not hope that the IL correspond to fully disambiguated senses. As the UW (universal words) of the UNL project, the ILs have to represent sets of senses.

If the system is to be "democratic" with respect to all languages, it is also reasonable to adopt the UNL strategy of denoting the ILs by strings using English terms, sim-

ply because English is the de facto lingua franca of modern science, and we can expect every lingware developer in the world to know it well enough to understand the meaning(s) of an IL and relate it to words and terms in his/her language.

Another aspect of "hybridicity" is that the pivot representation should be multilevel. We do not want to say that it should countain levels of interpretation relative to the surface expression in some particular languages. For instance, nothing like "English impersonal passive" should appear in a pivot structure.

The idea here is that there is a minimal common core of interlingual attributes and relations which every enconverter should produce in the pivot structures (entity/predicate, semantic relation/argument position…), a kind of "intersection over languages", and also a "maximal envelope", to be expressed only through interlingual attributes, of features which are universally understood, but underspecified in some language and absolutely necessary in another (aspect, modality, sex, quantity or abstract number, social level, etc.).

Using such hybrid pivot representations, it becomes possible to "grade" the resulting quality. If "enconversion" into the pivot is done automatically, only the minimal core will be available, and ambiguities will have been solved automatically (by heuristics which we know are nothing more than educated guesses, often wrong in more than 30% of the cases). If full interactive disambiguation is used, the source language ambiguities will be solved, and the underspecified features will have been made precise for the benefit of deconversion into all other languages. With partial interactive disambiguation, quality would be somewhere in the middle.

Conclusion

Quality MT and TA are currently usable only in niches and by the rich (people and languages). They must be democratized and become usable by small firms, young individual translators, bilinguals performing occasional translations, and even by individuals incompetent in foreign languages. We have proposed to combine four methods to reach that goal.

Although, from the research perspective, we are confident that they are necessary, and then, once implemented, they will indeed lead to the stated goal, some support is necessary from public-oriented organizations, because changing pricing policies, product architectures, and proprietary attitudes, usually meets with strong resistance. Also, we would welcome opportunities to convince MT and TA vendors that these changes would benefit them, not only in the long term, but quite probably very soon.

Finally, democratizing MT and TA would be quite good from the research point of view, not only because it would give a lot of job opportunities for NLP specialists, but also because it would produce a very large quantity of data (aligned terms, phrases and sentences) on which new studies and experimentations could be performed, in a variety of languages.

Acknowledgements

All thanks to the program and organization committees of MTS-VII, and in particular to Pr. J.U. Tsujii, who triggered the theme of this communication.

References

[1] **Blanc É., Sérasset G. & Tchéou F. (1994)** *Designing an Acception-Based Multilingual Lexical Data Base under HyperCard: PARAX.* Research Report, GETA, IMAG (UJF & CNRS), Aug. 1994, 10 p.

[2] **Blanchon H. (1992)** *A Solution to the Problem of Interactive Disambiguation.* Proc. COLING-92, Nantes, 23-28 July 1992, 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, 5-9 Aug. 1994, vol. 1/2, pp. 115—119.

[4] **Boitet C. (1988)** *Hybrid Pivots using m-structures for multilingual Transfer-Based MT Systems.* Jap. Inst. of Electr., Inf. & Comm. Eng., Jap. Inst. of Electr., Inf. & Comm. Eng., June 1988, NLC88-3, pp. 17—22.

[5] **Boitet C. (1988)** *PROs and CONs of the pivot and transfer approaches in multilingual Machine Translation.* Proc. Int. Conf. on "New directions in Machine Translation", 18–19 August 1988, Foris Publications, pp. 93—108.

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

[7] **Boitet C. (1994)** *Dialogue-Based MT and self-explaining documents as an alternative to MAHT and MT of controlled languages.* Proc. Machine Translation 10 Years On, 11-14 Nov. 1994, Cranfield University Press, pp. 22.1—9.

[8] **Boitet C. (1995)** *Factors for success (and failure) in Machine Translation — some lessons of the first 50 years of R&D.* Proc. MTS-V (Fifth Machine Translation Summit), 11—13 July 1995, CEE, 17 p.

[9] **Boitet C. (1997)** *GETA's MT methodology and its current development towards personal networking communication and speech translation in the context of the UNL and C-STAR projects.* Proc. PACLING-97, Ohme, 2-5 September 1997, Meisei University, pp. 23-57. (invited communication)

[10] **Boitet C. & Blanchon H. (1994)** *Multilingual Dialogue-Based MT for monolingual authors: the LIDIA project and a first mockup.* Machine Translation, **9**/2, pp. 99—132.

[11] **Boitet C. & Nédobejkine N. (1986)** *Toward integrated dictionaries for M(a)T: motivations and linguistic organization.* Proc. COLING-86, Aug. 1986, IKS, pp. 423—428.

[12] **Boitet C. & Tomokiyo M. (1995)** *Ambiguities & ambiguity labelling: towards ambiguity data bases.* Proc. RANLP'95 (Recent Advances in NLP), Tzigov Chark, 14—16 September 1995, pp. 13—26. (also published in a book)

[13] **Boitet C. & Tomokiyo M. (1996)** *Theory and Practice of Ambiguity Labelling with a View to Interactive Disambiguation in Text and Speech MT.* Proc. COLING-96, Copenhagen, 4—5 Aug. 1996, CST, Univ. of Copenhagen, vol. 1/2, pp. 119-124.

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

[15] **Chandioux J. (1988)** *10 ans de METEO (MD).* 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. 169—173.

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

[17] **Ducrot J.-M. (1982)** *TITUS IV.* In "Information research in Europe. Proc. of the EURIM 5 conf. (Versailles)", P. J. Taylor, ed., ASLIB, London.

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

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

[20] **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.

[21] **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.

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

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

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

[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] **Melby A. K., Smith M. R. & Peterson J. (1980)** *ITS : An Interactive Translation System.* Proc. COLING-80, Tokyo, 30/9-4/10/80, pp. 424—429.

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

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

[29] **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.

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

[31] **Somers H. L., Tsujii J.-I. & Jones D. (1990)** *Machine Translation without a source text.* Proc. COLING-90, 20-25 Aug. 1990, 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. T. (1986)** *English-Malay Translation System: A Laboratory Prototype.* Proc. COLING-86, Aug. 1988, IKS, pp. 639—642.

[34] **Tsujii J.-I. (1987)** *What is pivot?* Proc. MTS-I (MT Summit), Hakone, pp. p. 121.

[35] **Uchida H. (1989)** *ATLAS.* Proc. MTS-II (MT Summit), Munich, 16-18 août 1989, pp. 152-157.

[36] **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 & GETA, Grenoble, 700 p.

[37] **Vauquois B. & Boitet C. (1985)** *Automated translation at Grenoble University.* Computational Linguistics, **11**/1, January-March 85, pp. 28—36.

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

[39] **Whitelock P. J., Wood M. M., Chandler B. J., Holden N. & 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.

[40] **Whitkam T. (1987)** *Interlingual MT – an industrial initiative.* Proc. MTS-I (Machine Translation Summit), Hakone, pp. 135—140.

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

[42] **Zaharin Y. (1986)** *Strategies and heuristics in the analysis of a natural language in Machine Translation.* Proc. COLING-86, Bonn, Aug. 1986, pp. 136—139.

-o-o-o-o-o-o-o-o-o-o-o-

Contents

-o-o-o-o-o-o-o-o-o-o-