Keynote Address

Some notes on the state of the art:

Where are we now in MT: what works and what doesn't? And the role of MT as an international collaborative activity

Yorick Wilks

University of Sheffield, UK

Abstract

The paper examines briefly the impact of the "statistical turn" in machine translation (MT) R&D in the last decade, and particularly the way in which it has made large scale language resources (lexicons, text corpora etc.) more important than ever before and reinforced the role of evaluation in the development of the field. But resources mean, almost by definition, co-operation between groups and, in the case of MT, specifically co-operation between language groups and states.

The paper then considers what alternatives there are now for MT R&D. One is to continue with interlingual methods of translation, even though those are not normally thought of as close to statistical methods. The reason is that statistical methods, taken alone, have almost certainly reached a ceiling in terms of the proportion of sentences and linguistic phenomena they can translate successfully. Interlingual methods remain popular within large electronics companies in Japan, and in a large US Government funded project (PANGLOSS).

The question then discussed is what role there can be for interlinguas and interlingual methods in co-operation in MT across linguistic and national boundaries. The paper then turns to evaluation and asks whether, across national and continental boundaries, it can become a co-operative or a "hegemonic" enterprise. Finally the paper turns to resources themselves and asks why co-operation on resources is proving so hard, even though there are bright spots of real co-operation.

1 Introduction: The debate over the "statistical MT" hypothesis

In the last ten years, empiricism has struck computational linguistics in general and MT in particular, where by empiricism I mean a move to methods based on large scale language data, usually corpora of texts, sometimes including dictionary texts, available on computers, rather than on a priori linguistic theories and rules. One of the most striking examples was the purely statistical approach to machine translation at the IBM Watson Research Laboratories which made use of the very large Canadian English/French parliamentary corpus (Brown et al., 1988). The results were striking: with virtually none of the conventional sources of linguistic knowledge (lexicons, syntax, semantics, etc.), the system produced figures of between 50 and 65% of sentences correctly translated, depending on the relationship of the training to the experimental corpus. Although the result was astonishing to many, more detailed critiques (e.g. Wilks, 1994) have pointed out that the figure has remained static if only pure statistical methods are used, that some linguistic phenomena are seemingly resistant to this approach, that the system, CANDIDE, has never actually beaten SYSTRAN in a direct competition of unseen texts from areas different from the training corpus, and that the economics of corpus availability and production are probably against any commercial and general development of CANDIDE for new languages.

All that is now in the past, and we can ask what the effect of the IBM work has been on MT and computational linguistics in general. One could say the alternatives are the following:

* Going on with theoretical linguistic development, which one could deem “linguistics as chemistry”, in search of the correct and devastating formula.

* Machine-aided translation, which supplements computational lacunae by having a human in the translation loop, and has been much used in commercial systems;

* Keep on hacking in the hope that, like SYSTRAN, a system can grow to an acceptable level of performance, perhaps by blending the best of statistical and symbolic components.

There are systems, still under development, in both commercial environments and research laboratories that have adopted all these latter day strategies, sometimes more than one at once. One could also argue that all those strategies agree on most of the following morals that can and have been drawn from where we are now for future MT systems.

2 Future MT systems

* Unaided statistical methods will probably not be enough for any viable system, commercial or otherwise, since they do not lead to a system that beats SYSTRAN, which is available for a large range of languages.

* One should be more sceptical than ever about a system that works on some data, because all MT systems work to some degree, whatever their assumptions: word-for-word MT as much as pure statistical MT. Coverage is as much a criterion as quality of translation.

* There are proven bags of tricks in MT, as Bar Hillel always argued (1960) and no amount of theoretical research is going to diminish their importance.

* Symbolic and statistical methods can be combined, and that seems to be where most MT research is at the moment.

* Interlingual methods remain popular, in spite of the above, at least in Japan and the US.

* Evaluation continues to drive MT, and helps keep old systems alive. The last ARPA evaluations showed SYSTRAN still very much in the game, but with small commercial upstarts beating the research systems, and much closer to the old, established, and more expensive ones than the latter find comfortable.

* Thanks to IBM, resource driven systems are here to stay, at least for the foreseeable future and Big-Data-Small-Program may still be a good ideal, from SYSTRAN to IBM. Here one can take for contrast theoretically motivated systems like EUROTRA (Johnson et al. 1985).

Let us now turn to some issues at the junction of resources, evaluation and interlinguas.

3 Modalities of international cross-language co-operation

Co-operation is now crucial to MT because resource creation demands it, and resources are now considered crucial to MT by all except those still firmly committed to formal linguistic methods, and who have therefore effectively withdrawn from empirical and evaluation-driven MT.

Obvious types of co-operation are:

* between monolingual groups within states (usually monolingual)

* between monolingual groups within the (multilingual) EU

* between groups or state organisations within blocs (US, EU, Japan), where one of those blocs is monolingual, one multilingual, and one (The US) with aspects of both.

The next question is: what should be the basis of that co-operation if it is across languages and cultures (e.g. in writing the analysis, generation and transfer modules of a conventionally structured MT system)?

Should it be on the basis of:

* each partner doing what they do best (as opposed to everyone doing and redoing everything)?

* each partner doing their own language (as opposed to "I'll help you with yours")?

* each partner doing their own interlinguas (as opposed to "I'll believe more in mine if you can use it too")?

* each partner doing their own evaluation of their own modules (as opposed to "I'll evaluate yours and you mine")?

But, historically not all insight is from inside a language: one has only to think of the early keyboards for Chinese, which came from the West, and the fact that Jespersen, a Dane, produced the first full descriptive linguistic grammar for English. The recent morpholympics competition was, I think, won by a Finnish analyser of German which beat all the groups from Germany.

Genuine co-operation, on the other hand, can include offers such as the free availability of JUMAN, the Japanese segmenter from Kyoto University, which is of the "I'll help you do my language" type, and which is quite different from "I'll do mine and you do yours", an attitude which drastically limits possible forms of co-operation. On the other hand, the new Finnish constraint parser for English (Karlsson, 1990) is "I'll help you do yours". If one doubts the need for this kind of thing, I can cite from personal experience the project at CRL-NMSU which built a Spanish lexicon from an English one largely because we could not find a Spanish machine-readable lexicon at all.

Consider, as part of this issue, the problem of the mutual perceptions of Japanese and English speakers: each group sees their own language as mysterious and hard to specify by rules. The proof of this, for English speakers, is that vast numbers of foreigners speak English but find it so hard to get the language right, as opposed to communicate adequately with it. Yet, and as a way of reaching the same conclusion from the opposite evidence, the Japanese sometimes infer, from the fact that so few foreigners speak Japanese at all, let alone perfectly, that they cannot. One imagines that this attitude will soon change, as foreigners speaking Japanese, at least adequately, become commonplace. This situation creates a paradox for speakers of English because it is so widely used; with the result that native speakers often implicitly divide the language into two forms: where one is the "International English" which they understand but cannot speak.

A side-effect of the IBM statistical methods for MT was that they showed the surprising degree to which you do not have to understand ANYTHING of the language you are processing. Most workers in the language industries find this conclusion intuitively unacceptable, even if they do

not subscribe to what one might call the "meaning and knowledge" analysis still popular within many Japanese systems, as it used to be for English during the "artificial intelligence" period in the 1970's. Its basis in both languages was what is usually called paucity of structural information, or some such phrase, which opposes the two languages to, say, Spanish or German, whose speakers tend to believe their language rule governed. Most commentators on recent MT developments contrast as radically opposed the IBM statistical methods to those earlier AI methods explored in the US. But that contrast can disguise the closeness of Meaning-Knowledge systems to statistical systems: both rest on quantifiable notions of information or knowledge. AI systems for MT like "preference semantics" (Wilks, 1977) can be seen as quantitative systems that, at the time, lacked the empirical data, since provided by more recent approaches like (Grishman and Sterling, 1989).

Systems that emphasise the core role of verb meaning (all those going back to Fillmore and case in AI and computational linguistics generally, and beyond him to the verb centred tradition of classical logic) have to deal, in the end, with the vacuity of much verb meaning ("Kakeru" in Japanese or "Make" in English are classic examples) and the reliance for understanding their use on the types of things you can do with, say, keys and locks, or scrolls and branches (in the case of Kakeru). Similar situations for English arise when only the object (bed, versus book, versus point etc.) of the verb give any content at all to the meaning of "make" when used with them.

Perhaps, as with DO, BE, HAVE, in English, those verbs are almost entirely redundant and the verb name is no more than a pointer to constrain abnormal uses: you could delete such verbs from a text and still guess rightly what was going on; or at least you could with Kakeru if you could distinguish open and close (a lock with a key) from the wider context available. One could put this in symbolic terms as "nouns prefer verbs as well as vice versa", but that is no more, in the case of the vacuous verbs above, than recapitulating the basics of information theory, in that these verbs carry little or no information. Text statistics, of the IBM type, reflect this and so should our analysis.

My point here is that, with these phenomena, symbolic and statistical analyses are saying the same thing in different ways, though the symbolic tradition inherits various prejudices (like the structural primacy of verbs in English), where statistical methods are simply unprejudiced.

4 The relationship to MT evaluation

Certain issues to do with MT evaluation follow from the discussion of the last section, particularly in connection with international co-operation in MT, particularly projects that require modules of a single system to be built in different countries, as is standard in EU R&D. Let us consider module interfaces (which may or may not be considered as interlinguas, which raise other, special, issues):

How can you evaluate an international/intermodule project properly?

The EU MT project EUROTRA (Johnson et al. 1985) was designed on the assumption that national/language groups built modules for their own language(s) and the system was held together by a strong structure of software design and, above all, agreed interfaces. But how could one assign blame for error (if any) inside an overall project designed like this after a bad evaluation of overall performance. In fact no serious evaluations of that project based on quantitative assessment of output were ever done, but that is beside the point for this abstract discussion.

EUROTRA was not, in its final form an interlingual system, but imagine a two module interlingual system. Some have certainly written about the possibility of evaluating the modules:

Source Language--> INTERLINGUA and INTERLINGUA ----> Target Language

separately. But could this method for assignment of error be of more than internal team interest if this were an international co-operative project? Or, more precisely, for a given bad translation, how could one know for certain which of those modules was at fault, if each chose, chauvinistically, to blame the other? Clearly, that would only be possible if they had a clear way of deciding for a given sentence what was its correct interlingual representation. If he could do that it would be clear whether or not the first module produced that representation: if it did, the error must be in module two, and if not it would be in module one.

Although not interlingua based, the EUROTRA groups had to agree on module interfaces that are, in effect, interlinguas in the sense of this discussion; it was just there was more than one of them, because there were more than two modules required for a translation. In any case the groups there shared similar language-family assumptions so the interface was not too hard to define. But could Japanese and English speakers agree on a joint interlingua without an indefinite number of arbitrary decisions, such as what are the base meanings of kakeru?

One possible way out of the problem of agreeing on an interlingua between two very different languages, and assuming one did not take the "third way" out of selecting another existing language as an interlingua: might it be possible to define two interlinguas (one J-orientated; one E-orientated) and use both, perhaps comparing translations achieved by the two routes from source to target? That would at least have the virtue of having to have an interlingua based only on one of the two languages and which might therefore not be comprehensible to the other team.

But we will always have the residual problem, rarely mentioned, that one cannot program the module Source-->INTERLINGUA unless one is a "native speaker" of that interlingua (i.e. a native speaker of the language on which it is based), but then the other team will not be able to program the module INTERLINGUA-->Target. A moment's reflection should show that the "two directions" solution is not a solution at all, because both teams can only program one module for each route, so there is no translation produced. In practice, this would just become a blame shifting mechanism: "Our part was fine, so the problem must be in your generation!".

Suppose we retain the earlier assumption that everyone does analysis and generation of their own native language, and see what the possible models would be if we did have both a J-based interlingua (JINT) and an E-based one (EINT):

i. J source---> (J group)---> JINT-------> (E group)---> E target

ii. R R EINT R R

iii. E source---> (E group)---->EINT------>(J group)---> J target

iv. R R JINT R R

The question we raised above was whether, say, an English-speaking group could do task (iv). It is crucial to recall at this point that some Japanese-speaking groups do perform tasks like (ii): the NEC MT group has used an English-like interlingua, and the EDR lexical group in Tokyo has certainly produced large numbers of codings in an E-based interlingua for Japanese word sense, which is effectively task (ii) without any generation to follow.

The solution may then be that we should learn enough of each other's languages to use each other's interlinguas, and then compare the effectiveness of the routes above. And we would probably want to add a safety clause that the evaluation of any module into or out of an interlingua based on language X should be done by the speakers of language Y.

If there are also to be rules going between the interlinguas we shall have what some Japanese groups are calling semantic transfer. Whatever that is, it is quite distinct from syntactic transfer, which is right or wrong and capable of extraction from data, as in the work of Matsumoto and colleagues (e.g. Utsuro et al. 1994). This relativist notion of an interlingua, explicitly dependent on actual natural languages, is one quite separate from the classical notion, of the sort once advocated by Schank (1973) where there could not be more than one interlingua, almost by definition. The tradition being explored in this paper (cf. Wilks et al. 1995) is that if interlinguas in fact have characteristics of natural languages, then the relativist tradition may be the only one with a future.

5 Relativism and interlinguas in MT

I would suggest that one can no longer continue to say, as many still do with straight faces, that items in an interlingua look like words but are in fact "just labels". This ignores the degree to which they are used as a language along with assumptions brought in from languages. They always look like languages, like particular languages, as we saw above, so maybe they are languages.

Remember Ogden's Basic English (Ogden 1942): a reduced primitive language of some thousand words, about the size of the inventory of head notions in a thesaurus like Roget, and about half the size of the LDOCE defining vocabulary (Procter 1978). The words of basic English were also highly ambiguous because of the small size of the set, as is the LDOCE defining vocabulary, a task Guo set out to rectify by a handtagging of the LDOCE defining vocabulary, to produce what he called Mini-LDOCE (Guo 1992). Interlingual items are ambiguous in exactly the same way, though this fact is rarely discussed or tackled. It did surface briefly during discussion at a Pennsylvania seminar on the EDR dictionary, when EDR colleagues explained how hard they sometimes found it to understand the EINT structures they had created in the conceptual part of EDR, and this was in part because the EINT words have senses they did not know. This may be a paradoxical advantage, as I shall discuss in a moment.

If this point of view has merit, then many empirical possibilities arise immediately: one would be to adapt to this task some of the systems for producing and checking controlled languages (e.g. Carnegie Group's CLE). These could be adapted to check not only the well-formedness of formulas in an interlingua but the distribution and usage of the primitive terms. Again, a range of techniques have been developed at research centres to sense-tag texts against some given division of the lexical senses of words; so that each word in a text is tagged with one and only one sense tag that resolves its lexical ambiguity (e.g. Bruce et al. 1993). This technique could probably be extended to interlinguas, if their formulas were viewed as texts, so as to control the non-ambiguity of the interlingual forms. As we noted above, Guo has already performed this task for the prose definitions of LDOCE, and that task is not different in principle from what we are discussing here.

The motivation for all this, remember, is so that interlingual expressions can be controlled so that they are understood by native speakers of the language from which the interlingual was drawn and by others, where the latter group are far more important for accessibility of interlingual MT as a technique.

None of this is an argument against interlinguas, but a suggestion for treating them seriously, making them more tractable, in the way MRD-based research has made lexicons more serious and consistent than the old, purely a priori, ones.

Another possible way of dealing with the difficulty we diagnosed is Hovy and Nirenburg's (1992) argument that an interlingua could be extended by the union of primitives from the classifying ontologies for the relevant languages under definition. This would abolish at a stroke the difficulty of an interlingua as a whole being based upon a single natural language, but would not help any users understand the parts not in their language. The gain would be in equity: all users would now

be in the same position of not believing they understood all the symbols in the interlingua, but the basic problem would not be resolved.

It is vital to remember here that none of the above makes any sense if you are able to cling firmly to the belief that interlinguas are not using natural language symbols at all, but only manipulating words as "labels for concepts". If you believe that, then all the above is, for you, unnecessary and irrelevant, and some of my close colleagues are in that position. I appeal to them, however, to look again and see that the position is sheer self deception: and we have no access at all to concepts other than through their language names which are, irreducibly, in some language. Because of the convenience that computers, say, are objects to which we can all point, we may persuade ourselves that we all have the concept of computer and the name doesn't matter. This, consolation, however does not last once one notices that of the words used to define other words (e.g. the 2000 words of the LDOCE defining vocabulary - the very words that appear in interlinguas, of course) virtually none are the least like "computer": state, person, type, argument, form are not open to simple ostensive definition and their translations are matters of much dispute and complexity. I rest my case.

6 Evaluation as hegemony

I want now to move from one undiscussible subject to another, but at shorter length. We neglect at our peril the international aspects of evaluation systems and the way in which they become, or are perceived to be "hegemonic": in the sense of attempts to assert control over the R&D of another culture. There is strong resistance in the European Commission to any general regime for the evaluation of MT based on open competitions between entrants of the kind that has developed research so rapidly, at least in its initial stages, in the ARPA community in the US. There is a belief in the Commission that such competitions are wasteful and divisive, and that belief has clearly helped to keep some substandard research in Europe alive and well for many years.

Protracted negotiations on sharing linguistic resources (lexicons and corpora) between the US and the EU have not progressed well largely because of this issue of evaluation, largely because the US side wanted to tie exchange of resources to the idea of common evaluation. The US side stressed the value of competitive evaluations between groups that accepted the same regime (usually imposed by the funding agency).

The EU side stressed co-operative R&D and downplayed evaluation, pointing out the incestuous effects of groups that compete and co-operate too intensely . Evidence of the latter are the unexpected successes of EU groups that entered ARPA MUC and Speech competitions (Sussex, Siemens, Philips, LIMSI): one could say they opened up a gene pool that had become too incestuous.

The Commission side saw the US position as hegemonic in the sense defined here: the US saw the European position as wanting to be shielded from open competition and ungrateful in that it expected to get US resources (chiefly speech data) for no return. I retail this history not to show a right and wrong side--it is not so simple--but to note that international co-operation is a complex cultural matter, in MT as anywhere else, and we should be aware of the complex links between evaluation and resources as well as the more technical issues to do with the representations and interfaces we noted above.

7 Resource sharing in the future

Nonetheless, resources will be essential to the future of MT and resources for MT, almost by definition, come from diverse languages and so states and cultures. Ways round these difficulties must be found, and in a range of areas:

Resources:	corpora, lexicons, dictionaries
Standards:	(mark-up (e.g. SGML)), tag sets, for lexicon interchange
Software modules:	alignment, taggers etc.

In all of these areas there is progress: the EU has actively encouraged the spread of the first type, and the inhibitions tend to come far more from the commercial concerns of publishers that from governments. Resource and software distribution centres have sprung up (e.g. CLR and LDC in the US, Saarbrücken in the EU). Software modules like taggers from the US and segmenters like Kyoto university's JUMAN have become widely available through individual acts of corporate and individual good citizenship. The EDR in Japan and Cambridge University Press (with its new lexicon) in the EU have announced plans to make lexical data far more available than was normally the case.

The EU has a crucial role to play in future resource provision for MT, not only because, with its twelve major languages, its need for MT is so great but because it has funded such substantial resource projects (and tool projects to use resource) already: NERC, ELRA, MULTEXT, GENELEX, AQUILEX, PAROLE, EAGLES, the names are legion.

These are still early days, even though so much has been spent, in that it is still hard to actually get hold of genuinely reusable resources and tools: interface and format problems still bedevil real reuse. the EU is also haunted by the spectre of English: it is more than one of the twelve languages: it is the superlanguage, that provokes both utilisation and fear of take-over, and all tied in with the mixed attitudes to US culture that we noticed in connection with evaluation. This complex attitude has worked against the EU funding of specifically English resources, on the grounds that they are available from the US and that the UK has already put such great efforts into its learner's dictionaries (LDOCE, OALD, COBUILD, the new Cambridge Dictionary etc.) and its national corpora (The Bank of English, the British National Corpus etc.). Were it not for these last, English could easily be in the extraordinary position of being the only EU language, all of whose resources were from or controlled by sources outside the EU.

All this effort and activity has tended to downplay the ultimate need to build resources in major languages (e.g. Russian, Chinese, Arabic) that are neither ones own nor, at the moment, seem inclined to build their own electronic resources. Russia has such resources but they seem to have deteriorated in the short term with the economy itself. The issue of who builds such resources is also relevant, of course, and in the real world, tied up with perceived threats, commercial and military.

In spite of all this, we can be sure the resource issue will not now go away from MT, and that commercial and government interests will ensure that greater resources are built and maintained. What we, as researchers, need to work for is maximum availability and the way that such resources can serve international communication, politically, of course, but, crucially, within interlingual aspects of the R&D process itself.

References

Y. Bar-Hillel (1960) The present status of automatic translation of languages. In Alt (ed.) Advances in Computers (Vol. 1). Academic Press: New York.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer & P. Roossin (1988), A Statistical Approach to Language Translation, In Proceedings of the 12th International Conference on Computational Linguistics, Budapest.

B. Bruce, L. Guthrie & Y. A. Wilks (1993) Automatic Lexical Extraction: theories and applications, In F. Beckmann & G. Heyer, Theorie und Praxis des Lexikons: A Festschrift for Helmut Schnelle. Berlin: De Gruyter.

C. M. Guo (Editor), (1992) Machine Tractable Dictionaries: Design and Construction, Ablex: Norwood, NJ.

R. Grishman & J. Sterling (1989) Preference Semantics for Message Understanding, In Proceedings DARPA Speech and Natural Language Workshop. New York University.

E. Hovy and S. Nirenburg (1992) Approximating an interlingua in a principled way. In Proc. DARPA Speech and Language Workshop, Harriman, NY.

R. Johnson, M. King and L. desTombes 1985 EUROTRA: a multi-lingual system under development. Compuational Linguistics, 11.

F. Karlsson (1990) Constraint Grammar as a Framework for Parsing Running Text. In Proc. COLING90, Helsinki.

C. K. Ogden (1942) The General Basic English Dictionary W. W. Norton: New York.

P. Procter (1978) Longman Dictionary of Contemporary English (LDOCE) Longman Group Limited: Harlow, Essex, UK.

R. C. Schank (1973) Identification of Conceptualisations Underlying Natural Language. In Roger C. Schank and Kenneth M. Colby (Editors), Computer Models of Thought and Language, San Francisco: W. H. Freeman.

T. Utsuro, H. Ikeda, M. Yamanae, Y. Matsumoto, and M. Nagao (1994) Bilingual text matching using Bilingual Dictionary and Statistics, In Proc. COLING94, Kyoto.

Y. A. Wilks (1978) Making Preferences More Active, Artificial Intelligence (11)

Y. Wilks (1994) Developments in machine translation research in the US . In The ASLIB Proceedings (Vol. 46) (The Association for Information Management).

Y. Wilks, B. Slator, and L. Guthrie (1995) Electronic Words: dictionaries, computers and meanings. Cambridge, MA: MIT Press.