[*Translating and the computer 10*. Proceedings of a conference … 10-11 November 1988, ed. Pamela Mayorcas (London: Aslib, 1990)]

Text typology and machine translation: an overview

Douglas Arnold

Language and Linguistics, University of Essex, UK

The purpose of these introductory remarks is to indicate the main issues relating to text typology that arise in the context of machine translation, from both the theoretical and practical points of view.

There is no 'typology' or classification of texts to which MT practitioners or theoreticians can appeal, but it is clear that 'text type' is one of the critical parameters which can lead to success or failure in MT – for particular systems on particular occasions or in general, or perhaps ultimately even for the field as a whole.

It is clear why this should be the case, as Juan Sager emphasised in a paper given at this conference some seven years ago (Sager 1982). Firstly, because appreciating the kind of text one is dealing with (having regard to subject field, style, register, text format conventions and purpose for which the translation is intended) is the first job of a human translator. Secondly, because not every type of text is equally suitable for every kind of MT. There are a variety of reasons for this.

The intended purpose of a translation is important: is it to help *readers,* who are ignorant of the source language, or *writers*, who are ignorant of the target language? If it is for readers, what kind of information do they want to extract from the text? If it is for writers, what is the purpose of the translation? Is it for publication? Will it have any legal standing?

One must be clear about the level of demand for translation for a particular kind of text, and the shortage or otherwise of human translators with respect to that demand. (Setting up and maintaining an MT system involves an enormous overhead, which will be hard to justify unless there is a sufficient volume of material to translate.)

Some kinds of text are inherently difficult for MT. They may contain syntactic and morphological constructions which, if they cannot be eliminated, are problematic for all kinds of natural language processing (for example, N-N compounding, co-ordination, ellipsis, ‘unbounded dependencies’, the need for extensive resolution of pronominal reference[1]). Texts may be more or less well-behaved (e.g. free of mistakes such as typing and spelling mistakes); more or less restricted, or ‘open ended’ (e.g. open to extended or metaphorical language uses). Thus, poetry is difficult to translate, and weather reports are not.

Of course, asking these questions in relation to MT, as though that itself were a single thing, oversimplifies matters. One should be asking the question about different *kinds* of MT. But the general point remains – the practical success or otherwise of MT depends on matching up:

- the capabilities of current and projected MT systems
- the needs and purposes for translation
- the inherent properties of texts to be handled by MT.

When this equation is correct, as with METEO, the result is practical success.

I do not know of any work that seeks to provide a principled classification of texts in terms of ‘demand for’ versus ‘availability of translation’, or of the purposes for which translations are intended, and this is not the place to discuss the capacities and limitations of MT systems (even where these are known). Instead, I will direct the remaining space to the theoretical and practical issues that arise from restrictions on the inherent properties of texts.

The inherent properties which can be restricted seem to be the following:

1. SEMANTIC DOMAIN (domain of discourse/subject field). The texts to be translated can be restricted to ones dealing with, for example, weather reports, stock market or medical reports, magazine horoscopes, recipes, knitting patterns, word processor documentation, engineering or aviation manuals.
2. OVERALL DISCOURSE TYPE (this alone is sometimes called ‘text type’). Texts may be restricted to those that have a particular internal format or structure, for example, business letters, newspaper stories, technical abstracts, patent applications, legal or governmental proclamations.
3. DISCOURSE STRUCTURE. Texts may be such that they exclude, for example, pronominal references outside the sentence, or paragraph; that headings may always be noun phrases; that ‘descriptive’ and ‘imperative’ sections of text may be clearly separated.

4. SYNTAX AND MORPHOLOGY. The range of syntactic constructions may be limited such that they exclude all but declarative and imperative sentences; restrict N-N compounds to those which are listed as single items; limit the kinds of co-ordination that are allowed.
5. LEXIS. The vocabulary used may be limited in terms of the number of distinct words that can be used, or in terms of the range of uses or readings of each word e.g., the *Concise Oxford English Dictionary* assigns the verb ‘press’ 10 distinct senses, and the noun another seven, but in a word processor manual it will almost certainly be possible to limit the usage to the verbal meaning ‘to exert pressure on’ as in ‘Press (the) return (key)’.

If one can find texts which observe these restrictions, and if, in addition, one can see that the (restricted) language in which these texts are written is a true language in the sense of being systematic, productive (creatively usable), used by some community, adequate for its intended purpose, etc. then one has a *sublanguage,* and the chances of a successful application of (some form of) MT are particularly good[2].

To get some idea of how good, one only has to consider lexicon. A typical large general dictionary contains approximately 400,000 words; Lehrberger (1982 p83) suggested that the aviation manuals studied in the TAUM-AVIATION project contained around 40,000 and that for agriculture market reports the number may be as low as a few hundred words (the METEO dictionary contains less than 1,000 words [Kittredge 1982 p124]). A translation system for the general language covered by a normal dictionary would be an enormous undertaking, but one whose lexicon contains only a few hundred words looks distinctly feasible.

However, there are still some interesting practical and theoretical problems[3]:

1. What is the relation between a sublanguage and a general language? It is a matter of definition that there are words and constructions in the general language that are not part of the sublanguage. However, (despite the name) it does not follow that a sublanguage forms a subset of the general language. In particular, it is often the case that there are constructions in the sublanguage that are at best marginal in the general language. For example, it is common in instruction manuals to find locutions of the form: *confirm high tension circuit complete* (that is, confirm [that the] high tension circuit *is* complete). This construction is found in the general language (cf ‘I believe the *high tension circuit complete*’), but not with the verb *confirm* (Lehrberger 1982 p90). The practical point here is that, to be useful, a sublanguage must not just exclude some constructions, it must exclude some *difficult* constructions, and any marginal constructions it includes must not themselves be too difficult.

2. What is the relation between the grammar of the sublanguage and that of the general language? Here the answer is clear: there is no necessity for the grammar of a sublanguage to bear any interesting resemblance to that of the general language. TAUM-METEO recognises five main ‘sentence’ types (Lehrberger 1982 p100), *none* of which remotely resembles the rules conventionally assumed for English. Given that the system designer may not have a ‘native-speaker-like’ grasp of the sublanguage, writing a sublanguage grammar poses obvious practical and methodological problems and the theoretical problem of automatically inferring grammars from collections of texts becomes interesting (Hirschman 1986, Slocum 1986).

3. What kinds of relation are there between different sublanguages within one general language? (For example, is there any interesting relation between the language of weather reports and the language of stock market reports?) The corresponding practical question relates to the (non-) portability of sublanguage systems.

4. Given that sublanguage grammars may differ from each other, and from the grammar of the general language, does this mean there are different roles for, for example, morphology, syntax, semantics, and for the relationship of analysis, generation, (and if appropriate, transfer)? That is, are different architectures appropriate for sublanguage translation systems?

The practical importance of points 1 to 4 can be summarised as follows:

- one cannot guarantee that any system, or part of a system, which has been developed for one domain will be suitable for a different sublanguage (or, if it is suitable, that it will be ‘good’);

There are important practical problems in:

- discovering the restrictions that a sublanguage observes
- exploiting them.

5. What kinds of relation are there between the sublanguage of one general language, and those of another: is the English language of instruction manuals closely related to that of French instruction manuals (similarly for Japanese, Thai, Swahili, etc.)? Interestingly, there is considerable evidence that they are similar (see Kittredge 1982a, Teller, Kosaka and Grishman 1988). But if they are similar, why are they similar? Is it a result of contact between the communities who use the sublanguage, or is it somehow a result of pressure from the semantic domain?

6. In general, neither sublanguages, nor texts written in them are completely ‘closed’. Sublanguages (like all languages) are to some extent ‘permeable’, and typically allow ‘escapes’ into the more general language. For example, a text, or section of text that is basically imperative may contain stretches of description; a description of a drugs trial which is

basically restricted to a technical language and vocabulary, may at some point, describe everyday occurrences that affected a subject, and the vocabulary here will be quite unpredictable. Important questions are then: How ‘permeable’ is any given sublanguage to more general usage? Is there any way of (automatically) recognising whether a particular stretch of text is ‘general’ or ‘sublingual’?

7. Finally, the most obvious practical problem that arises for anyone who wants to exploit the apparent suitability of sublanguages for MT is the problem of discovering (or successfully defining) one. It is easy enough to enumerate semantic domains but there is no guarantee that there will be associated restrictions on discourse, syntax, morphology and lexis[4]. Practically speaking, given a proposal for a sublanguage (say a particular semantic domain), how can one investigate how tractable it is likely to be? (see Kittredge 1986).

CONCLUSION

It is clear from what I have said that, for practical MT, restrictions on text-type, including the limiting case of sublanguages, represent a major line of advance. Moreover, because I think practical experience is a necessary pre-requisite for theoretical advance, the same goes for MT research. But I would like to enter a *caveat.*

There is a real danger in restricting the inherent properties of the texts that one deals with: the danger that one will not be able to generalise from that type of text to the more general language (or even to any other restricted domain); the danger that the restrictions hide essential aspects of the ‘problem of translation’. In particular, the interest and importance of sublanguage-based MT should not be allowed to obscure the value of what one might call ‘theoretically-based’ or ‘phenomena-based’ work in MT: work which begins with an idea about translation, or studies the problems that arise in translating a particular construction, and which pursues that idea or construction in the full glory and awfulness of dealing with relatively unrestricted ‘general’ language.

NOTES

1. The following are examples of some of these:
   (a) N-N (noun-noun) compounds:
   Replacement exhaust service centre personnel manager.
   (b) Ellipsis:
   Company A *gave the men* a bonus, Company B . . . a pay rise.
   (c) Unbounded dependency – there is a dependency between two items, for example, a verb and its object, but they can be separated by an unbounded distance, in the sense that there can be an unlimited number of intervening sentences:

They do not employ people *who* (s the police suspect . . .)
They do not employ people *who* (s they know [s the police suspect. . . .])
They do not employ people *who* (s they know [s other companies think (s the police suspect. . .)]).

2. The restrictions in question may be imposed (for example, by a style sheet), or arise spontaneously (and so be 'discovered') in a family of texts (for example, as a result of a group of writers dealing with similar subject matter with similar aims and intentions). This does not matter so long as what result can be considered to be a 'language' in this sense.

   To be pedantic, a further requirement is that to constitute a sublanguage, a set of texts must be 'maximal' in the sense of being the largest set that satisfies the relevant restrictions. The point is that a sublanguage is not an arbitrary collection of texts, and texts must not be arbitrarily excluded. This means that new texts belonging to a sublanguage can always be produced, so a sublanguage is not generally a finite collection of texts.

   It is sometimes said that MT systems deal with sublanguages by *definition.* This is because the properties of the texts that can be automatically translated are controlled by definition: an MT system defines (at least) two languages (a language it will accept, and a language it will generate) and the relation between them. These languages are restricted and well-defined, in general. This is true, if by 'sublanguage' we mean (roughly) some controlled subset of a natural language. However, as will be clear, I think the term is usefully applied in a rather more precise and narrow way.
3. These problems arise to some extent with any restrictions on text type, whether or not they are sufficient to yield a sublanguage. See the articles in Kittredge and Lehrberger (1982), and Grishman and Kittredge (1986) for detailed discussion of these and other problems; Kittredge (1986) provides an overview.
4. Notice that the existence of a sublanguage of (say) English does not guarantee that there will be an equivalent in any other language.

REFERENCES

Hirschman, L. (1986) 'Discovering sublanguage structures' in Grishman and Kittredge (eds.) 211-234.

Grishman, R. and Kittredge, R.I. (eds) *Analyzing language in restricted domains: sublanguage description and processing* New Jersey: Hillsdale, Lawrence Erlbaum Associates, 1986.

Kittredge, R.I. (1982) 'Variation and homogeneity of sublanguages' in Kittredge and Lehrberger (eds) 107-137.

Kittredge, R.I. (1986) 'The significance of sublanguage for automatic translation' in Nirenburg, S. (ed.) *Machine translation systems* Cambridge: Cambridge University Press, 1987, 59-67.

Kittredge, R.I. and Lehrberger, J. 1982 (eds) *Sublanguage,* Berlin: Walter de Gruyter.

Lehrberger, J. 1982 ‘Automatic translation and the concept of sublanguage’ in Kittredge and Lehrberger (eds), 81-107.

Sager, J.C. 1982 ‘Types of translation and text forms in the environment of machine translation (MT)’ in V. Lawson (ed). *Practical experience of machine translation* Dordrecht: North Holland Publishing Co, 11-19.

Slocum, J. 1986 ‘How one might identify and adapt to a sublanguage: an initial exploration’ in Grishman and Kittredge (eds) 195-210.

Teller, V. Kosaka, M. and Grishman, R. 1988 ‘A comparative study of Japanese and English sublanguage patterns’ in S. Nirenburg (ed.) *Proceedings of the second conference on theoretical and methodological issues in MT* Carnegie Mellon University, Pittsburg, Pa., 1987

AUTHOR

Douglas Arnold, University of Essex, Department of Language and Linguistics, Wivenhoe Park, Colchester, Essex.