George Talbot and Donncha Ó Cróinín, Dublin City University

Developing EIRETERM: The Termbank of the Eurotra Machine Translation Project

ABSTRACT: The paper describes the development of a termbank for the Eurotra research programme in Machine Translation (MT) and the issues involved in reuse of the termbank, after the project life of the MT research, in relation to human users and, more especially, as a resource for Natural Language Processing (NLP) applications.

1. Terminology and Machine Translation

Terminology provides both an opportunity and a challenge to Machine Translation (MT) research. On the one hand, given that there are clear and distinct concepts in various areas of scientific and technological enterprise which are not language-specific, but which are named by terms in the languages of many technologically advanced communities, it is theoretically possible to achieve the simple transfer of a particular concept from language A to language B. Thus, terms could be seen as components in a system of naming roughly analogous to the use of proper names. Ideally, within a particular sublanguage a term refers to one and only one concept. These concepts can then be given labels or numeric codes and the terms from each of the various languages can be coded along with the concept numbers in a kind of electronic dictionary. The opportunities are well known at this stage, in fact this approach formed the theoretical basis of the first patented mechanical translation system.[1] The challenge consists not in the theory but in the practical implementation of the theory in an MT system.

1.1 Terminology and the Eurotra MT Project

The literature on Eurotra is already voluminous and there is probably no need to add to it here with a description of the project. For present purposes it is enough to say that within the project, terminology was conceived of as an area of language in which simple transfer would apply. To that end, a corpus of texts in English was chosen and analyzed by terminologists and subject experts in order to extract terms and give them precise definitions. As the project developed it was decided that the sublanguage to be chosen for implementation of the software would be the sublanguage of telecommunications. So terminologists worked with a telecommunications engineer on a representative corpus, in order to identify the relevant conceptual content of the sublanguage. The terms ex-

tracted, defined and classified in this manner were sent to Dublin, where they were processed by Eurotra Ireland staff. These terms (some of which were identical to terms stored in Eurodicautom) and definitions were then sent on diskette to the other language groups, who worked with telecommunications experts in their language communities to produce equivalents in the other Eurotra languages.[2] These language-specific details (term, grammatical category, gender etc.) were then sent back to Dublin where they were eventually stored in the database which came to be known as EIRETERM. A record in EIRETERM consists in (i) a ten-digit ID number, (ii) the term (iii) the definition, (iv) documentary source of term, (v) source of definition (vi) a three-level classification system for rudimentary sublanguage analysis, (vii) grammatical code, (viii) gender code, (ix) equivalents for the term in the other languages and (x) sources for these equivalents. EIRETERM is therefore unique even among concept-based termbanks of non-trivial dimensions in that it was not designed to pragmatically record information primarily of use to humans: it was designed, primarily, as a machine-machine interface.

1.2 Physical and Logical Structure of EIRETERM

The termbank was created in 1990 using a commercial relational database management system, running on a Vax/Ultrix platform.[3] This package had been used in the development of the database portions of the experimental Eurotra MT software and it was primarily this compatibility which dictated the choice. PC-based packages were not considered to be suitable, particularly as it was also envisaged that the database would, at some stage, be made available on a Europe-wide academic network.

There are two main tables within the termbank through which data is entered and retrieved. One, EQUIVS, contains the data referred to in Section 1.1 above. The other, SYNON, which is accessed through a combination of term identification number and language code, stores synonym details for each term as well as information on permitted abbreviations of the term.

It would seem reasonable to assume that a database designed for efficient automated use would be easily used by humans also. This is not necessarily true. Automated insertion and extraction of data is easy in EIRETERM. However, the display of such data to a human user requires high-quality menu systems combined with an efficient and easy-to-use query language system. Many databases packages have been criticised for their poor quality of information retrieval, and the software used to create EIRETERM is no exception. The drawbacks caused by its limited Structured Query Language could probably have been overcome by the use of B-tree indexing, which allows queries based on ranges of values or on partial matches. And the menu system could have been expanded and improved upon, using menus to trigger programs written in the C programming language. Such enhancements were not considered necessary at the time since, as we have mentioned above, the emphasis was on creating a resource whose primary use was in automated situations.

So, while the termbank may be used by humans, its value as an on-line translation aid for human translators, for example, is practically nil. We have, however, used the Structured Query Language successfully for report production and for downloading data to files for use in other systems. The production of printed glossaries for translators is a simple procedure also.

In its basic form, then, the current version of EIRETERM fulfils its purpose as a module for use in a large-scale MT system.

It seems, though, that the current type of database management systems may not be sufficient to serve the language industry of the future. While the actual data stored on EIRETERM are valuable, the possibilities offered by new developments in object-oriented and semantic databases must be explored in order to extract full value from such a resource.

2. The Future of EIRETERM

We conceive of the future of EIRETERM in two different ways and (probably) in two different hardware/software environments. For the next stage of research and development in MT systems within the European Community, the emphasis will almost certainly be placed on the use of multi-tasking computers. The original version of EIRETERM has been converted from Ultrix onto a Sun SPARC2 system, using the new release of the commercial software package in which it had been written. We took the opportunity during the changeover to include greater scope for storage of lexical information, particularly in the definitions of terms. This should encourage further investigation into the possible integration of EIRETERM in whatever large-scale MT systems may be developed within the European research community over the next few years.

Parallel to any future work on EIRETERM as a Unix-based termbank for use within a large-scale MT system or systems, work on a PC-based termbank has passed feasibility and preliminary design stages.

2.1 EIRETERM-PC

Initially, the termbank is being developed within the National Centre for Language Technology (formerly Eurotra Ireland) on a PC network. The package chosen for prototype purposes is CDS-ISIS (for Computerized Documentation System/Integrated Set of Information Systems), which is an information storage and retrieval system distributed by Unesco. The flexibility of the package means that the finished product can be extremely complex in its structure, for example in its search facilities, yet very easy to use. On the development side, this enables us to design a large and powerful termbank, while on the user's side the package will remain constantly manipulable and user-friendly. The use of multimedia is also under investigation.

Further development, which will include the marketing of the termbank to the language industry, will require the collaboration of industrial partners.

2.2 Reusability for Natural Language Processing

The PC development is quite distinct from computational lexicography and can be regarded as a spin-off from the Eurotra project. There are clearly other ways in which EIRETERM, as a terminological resource, can be reused for the purposes of Natural Language Processing (NLP) and to provide a valuable basis for post-Eurotra work, but they involve more long term projects. We have observed that EIRETERM is unique

among termbanks of a non-trivial size, not because it is concept-based as opposed to term-based, but because it was designed as a machine-machine interface. That said, there are obstacles which must be overcome before it could be claimed that much useful information for NLP could be extracted (semi-) automatically from EIRETERM.

2.2.1 The Definitions

A limited amount of syntactic and categorial information is available, but the semantic information is contained in definitions which were originally written by subject experts for the purpose of indicating to their counterparts in other countries, the concept referred to by the English term. So, we have a somewhat anomalous situation: although the termbank was designed as a machine-machine interface, the definitions were written for humans, because the Eurotra Translation Software was not intended to make use of the definitions. Extracting semantic information even semi-automatically from these definitions is fraught with difficulties because the definitions were not formulated according to a set of in-house directions, and they were not cast in a restricted control language as, for example, are the definitions in the Longman Dictionary of Contemporary English (LDOCE).[4]

A project is currently underway, the aim of which is to restructure the definitions (in consultation with the expert who originally provided them) in terms of a control vocabulary so that semantic relations (e.g., ISA, PURPOSE, SYNONYM) can be triggered. The semantic relations will be used to construct a terminological network and/or hierarchy. These restructured definitions will be encoded in SGML to facilitate NLP application and/or data sharing.

EIRETERM, however, does already have a useful sublanguage analysis component (as mentioned above) which takes the form of a three-level classification. In the following table, which is an extract from the classification system, TCOMM is the only class_1, and within that particular class_1 there are twelve class_2s (GENTER to COSTS) and seven class_3s (PRODCT to CTRL). The different class_2s are mutually exclusive as are the class_3s, but any class_3 may be combined with any class_2, so the relationships are potentially both very flexible and very complex:

Telecommunications (TCOMM)
- General Telecommunications Term (GENTER)
- Transmission (TRANM)
- Propagation (PROPAG)
- Antennae (ANT)
- Radio Communication Systems (RADCOM)
- Space Systems (SPACEC)
- Broadcasting (BROADC)
- Ground Networks (GRONET)
- Measurement Techniques (MEAEQ)
- Components (COMPON)
- Quality (QUAL)
- Economics of Telecom Systems (COSTS)
  - Input/output (PRODCT)
  - Equipment/System (EQUIP)

Transmission Method (METHOD)
Natural Process (PROCESS)
Property of product/system (PROPTY)
Change in quality (PERF)
Monitoring (CTRL)

The following is an (abbreviated) example of the kind of entry available at present in EIRETERM:

ID: 2000000219 TERM: high capacity satellite
DEF: A satellite which is capable of dealing with a large number of channels and which would have a greater than average number of transponders.
Class_1: tcomm, Class_2: spacec, Class_3: equip.

Our hunch is that certain semantic relations may prove to run parallel to combinations of class_2 and class_3 within the subworld of telecommunications, and that hunch will be put to the test in the course of the project, the aim of which is to construct a knowledge-based module which will be implemented and tested in the new ET-6 (post-Eurotra) MT formalism.

Another issue which must be addressed is the matter of collocational information. It is well known that certain sublanguages incline towards morphological usage which is significantly different from the morphology of general language (e.g., the suffix *-itis* in medical terminology). Collocations within sublanguages may, to some extent, be determined morpho-syntactically, but there are instances (at least in relation to telecommunications, but we do not imagine that it is uniquely in relation to telecommunications) in which particular collocational strings are categorized as terms by the expert on the pragmatic basis of world knowledge. As far as possible, the pragmatics of expert knowledge must be incorporated into a Knowledge Representation module via corpus analysis and analysis of co-occurrence restrictions within the definitions.

2.2.2 *Standardization*

Another central problem involved in constructing a non-trivial reusable terminological resource, and one touched on above, is the problem of standardization. There are several aspects to the problem. Firstly there is the question of a representational formalism. The restructured definitions will be encoded in SGML so as to allow their conversion to a typed feature logic formalism – a necessary pre-requisite for implementation in the new generation of MT software. Secondly there is the question of classification. The three-level classification proposed by the telecommunications expert may be useful for sublanguage, but, may hamper the possibility of data-exchange in the long run. There are, of course, ISO recommendations on the classification of terminology, but the expert's three-level system is much more detailed in terms of eliminating ambiguity and (perhaps) giving expression to collocational restrictions. The long term status of the three-level classification is an imponderable at this stage. It may be the case that it would have little effect on data exchange, and it may even be the case that the information supplied by the restructured definitions concerning semantic relationships will make it redundant. For the moment it has to be seen as a potential risk.

3. Conclusion

In the foregoing text the notion of 'reusability' has been used uncritically. Following several authorities in the literature we shall, in conclusion, distinguish between two readings of 'reusable lexical resource' and discuss their relevance to the future of EIRETERM. Heid,[5] in relation to 'reusable lexical resource', has distinguished between (i) "an existing lexical resource in machine readable or machine tractable form which can be used outside its initial application or from which knowledge can be extracted which is fed to applications other than the initial one" and (ii) "a resource to be built in future, such that it can serve different purposes or different applications, or that it can be accessed from different theories". Clearly EIRETERM is a resource of the first type. However, the fact that it is not in use on a daily basis as a reference facility for human translators, unlike most non-trivial termbanks, means that it can be more readily adapted than other existing resources to embrace the objectives of the second type. With that in mind, work is in progress on refining the Achilles' heel of the termbank, the structure of the definitions, with the intention of turning EIRETERM into a significant reusable resource.

Notes

1 Cf. MAEGAARD, B. & PERSCHKE, S. (1991), p. 7.

2 By the "Eurotra languages" we mean the following: Danish, German, Greek, English, Spanish, French, Italian, Dutch and Portuguese.

3 We are grateful to Patrick Fogarty at Dublin City University (and formerly of Eurotra Ireland) for some of the information contained in this section.

4 Cf. BOGURAEV & BRISCOE (1989), p. 200ff.

5 Cf. HEID, U. (1991).

Bibliography

BOGURAEV, B. and BRISCOE, T. (1989): Computational Lexicography for Natural Language Processing. Longman, London and New York.

HEID, U. (1991): "Towards reusable lexical resources for natural language processing". In: Eurotra Internal Papers, presented at Eleventh International Conference 'Expert Systems and their Applications', Avignon, May 1991.

MAEGAARD, B. and PERSCHKE, S. (1991): "An Introduction to the Eurotra Programme". In: Studies in Machine Translation and Natural Language Processing. Ed. by Copeland et al. Official Publications of the European Communities, Luxembourg.

KEYWORDS: Eurotra, termbanks, terminology, reusability, Machine Translation, Natural Language Processing