[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

Introduction

John M^{c}Naught*

July, 1990

1 Background

The workshop that gave rise to the following papers was organised under the auspices of the UK *Speech and Language Technology Club* (SALT Club) at UMIST, Manchester, in July 1990. The SALT Club, sponsored by the Information Technology Division of the Department of Trade and Industry, was formed to help bring together researchers in the Speech and Natural Language communities, to discuss common interests, learn about each other's techniques, explore the intersection between speech and natural language and provide a forum to enable the joint community to discuss and formulate strategies for research on the national scale.

Although there were several SALT Club workshops held previously to discuss issues of strategy, this was the first one to allow a topic to be discussed from the research point of view. Some 80 participants took part, including an encouraging number from the speech community. Machine translation is typically a text-centered activity, however, as was seen in this workshop, there are major efforts under way, particularly in Japan, to tackle machine translation of speech.

These proceedings have taken some time to appear: this was due to some papers being delivered late. We felt it was however important to wait for these papers.

2 Workshop contents

The workshop consisted of a mix of plenary papers, group discussion sessions, parallel short paper sessions and commented demonstrations, rounding up with a panel discussion. Background plenary papers were presented on the *State of the Art in Machine Translation* (H. Somers, UMIST), *Sublanguage and Machine Translation* (S. Ananiadou, UMIST) and the *Lexicon in Machine Translation* (L. Humphreys, Essex). Somers's paper takes a synchronic view, assessing the success of second generation MT systems and systems derived from this paradigm (poor), and arguing for a broadening of the scope of MT research and for a rejection of second generation tenets, belief in which he sees as being responsible for the lack of advance in many projects. He discusses various promising alternative approaches, adopted by relatively recent or little known projects.

In her paper, Ananiadou takes this discussion a step further, concentrating on the notion of *sublanguage,* which is seen as being a key notion for today's MT systems. She discusses the role of sublanguage in MT, and demonstrates how linguistic knowledge of the specialised languages of science and technology can be modularly constructed and exploited to constrain analysis and help us to obtain higher-quality translation than can be obtained by attempting to customise general language descriptions to serve specialised ends. She emphasises that parallel sublanguages of different natural languages are 'closer together' (i.e. it is easier to translate between them) than their respective

*Centre for Computational Linguistics, UMIST, Manchester M60 1QD, UK. E-mail: jock@uk.ac.umist.ccl

general languages. As sublanguage documents make up the bulk of texts for translation, it is crucial to take advantage of their special language characteristics. In particular, adoption of a sublanguage approach is seen to be crucial for the success of MT systems, as such an approach is oriented towards the characterisation and processing of 'real text'. She is dearly critical of approaches based on general language: it is extremely hard to formally define general language in such a way that resulting descriptions can be easily and successfully used for MT purposes.

Humphrey's paper discusses the several components of a computational lexicon for MT, detailing the types of lexical knowledge required in order to support MT. He emphasises how decisions regarding monolingual issues may affect multilingual issues. The examples and cases presented demonstrate that, perhaps contrary to received opinion, great care must be taken over the design of lexica and that lexica present a highly complex sub-field of MT endeavour.

A second set of papers dealt with MT projects and systems from an industrial viewpoint. The workshop was fortunate in having representatives of British Telecom (Stentiford, BTRL: *Work at British Telecom in Automatic Interpretation and Translation*), IBM UK (Sharman, IBM Scientific Centre, Winchester: *The Application of Information-Theoretic Models to Machine Translation*) and, from overseas, ATR (Morimoto, ATR Interpreting Telephony Research Laboratories, Japan: *Automatic Interpreting Telephony Research at ATR*) and Siemens Nixdorf (Thurmair, SNI, FRG: *The METAL Project*). It was especially interesting to learn about developments from our overseas guests. Stentiford's paper describes the variety of BT services in which MT could play a role, and which are a focus for BT sponsored MT research. He discusses the particular requirements for MT systems design when the user is monolingual, as opposed to bilingual (the usual case, when e.g. post-editing is called for). He reports on a number of MT and MT-related projects in which BTRL has an interest, including: systems to handle business correspondence, through sophisticated mixing of proformas and free text; fundamental research into typical linguistic patterns employed by French and British police forces in charge of security of the Channel Tunnel, with a view to building training and translation aids; and phrase-book translation aids, with text-to-speech output.

Sharman's paper expands on points raised by Somers, developing arguments for a different approach to MT than the traditional second generation approach. He examines the relevance of applying results from Information Theory, found useful for speech recognition, to the MT problem. In applying the notion of *encoding* and *decoding* texts, he illustrates how one may view translation as a task of discovering which source language sentence or sentences could *most probably* have been used to yield the encoded target language sentence. This leads to the utilisation of probabilistic language models in conjunction with parallel text corpora. Work on such models both in the UK and in the USA has shown that a MT system following such a model can perform at least as well as a traditional second generation MT system. Such results moreover challenge those who advocate heavy use of rich linguistic knowledge in MT to demonstrate the advantage and indeed cost-effectiveness of attempting to work with linguistic knowledge.

Thurmair's paper indirectly provides a response to this challenge. He describes SNI's METAL MT system, which is now fully operational and on the market. This is one of the best-known MT systems. Thurmair reminds us of the importance of ensuring that any MT system is integrated with a working environment. This is especially true for commercial systems, otherwise they will fail to improve the overall performance of the translation task. To this end, METAL has been the target of substantial work on interfaces, of interchange formats and the like to ensure smooth importing and exporting of texts, the preservation of text structure and appearance, etc. Thurmair then describes the METAL architecture in some detail, and presents the METAL view of linguistics: pragmatic and judicious use of linguistic theories. The problems of scaling up to large grammars and lexica are examined, and the METAL solution presented. In particular, the need for robust support tools is felt to enable linguists to organise linguistic knowledge and navigate through it. This paper then gives a detailed description of a major MT system, showing how good linguistic engineering can provide a system that performs with high quality, partly due to the user being able to exert flexible control over the system's behaviour to suit his ends.

Morimoto's paper gives an overview of the large-scale ATR Interpreting Telephony Research project, which aims at developing an automatic telephone interpreting system over 7 years. The ATR laboratories are experimenting with several approaches to the MT of speech, and with combinations of approaches, in order to develop the fundamental research needed to build a prototype system. The requirements of the project are extremely demanding: no pre- or post-editing; no bilingual human aid; resolution of ellipsis and anaphora through use of pragmatics; understanding of speaker's intentions; application of rich linguistic information to resolve ambiguities. Morimoto discusses three developments at ATR in detail. Firstly, he describes a new grammar-driven continuous speech recognition mechanism, combining a HMM phone model and a generalised LR parsing algorithm. Secondly, he reports on work on a method of translation that is intermediate between the well-known transfer and interlingua approaches, involving propositional and illocutionary analysis - translation is carried out in two stages, firstly by transferring the propositional meaning to a target language meaning, and subsequently by merging the target propositional meaning with the illocutionary meaning in generation. Thus, propositional meaning is seen to be language dependent, and illocutionary meaning is seen to be language independent. Thirdly, he describes SL-TRANS, which is capable of recognising Japanese speech, translating it into English and yielding synthesised English speech as output. Of interest to SALT Club members is the emphasis placed at ATR on the integration of speech and NLP research, and the various combinations of speech and NL techniques that have been tried.

A final plenary paper was given by J. Pugh (UMIST): *The EUROTRA Transition Programme.* Pugh briefly describes the background to and objectives of the EUROTRA project (sponsored by the CEC and, in the UK, by the DTI), which has been the major force in European MT for over a decade. Although she concentrates on informing us about current plans, involving collaboration between industry and academia and further applied MT research, she also emphasises the role of EUROTRA-related research within the CEC's Framework Programmes. Furthermore, she demonstrates how EUROTRA has been responsible for generating a large body of trained MT researchers. It is stressed that the original EUROTRA programme was not intended to deliver a fully operational system, contrary to popular belief, and that current plans are intended to effect the transition from the research prototype phase to the industrial development phase.

Four parallel paper sessions were held, which then fed into the later corresponding group discussion sessions. The overall themes of these sessions were:

1. Formalisms for MT (chair: J. Durand, Salford)
2. Discourse & linguistics for MT (chair: D. Arnold, Essex)
3. Speech/NL interaction and MT (chair: K. Morton, Essex)
4. Lexicon and MT (chair: J. M^{c}Naught, UMIST)

1. Formalisms for MT

Three papers were given in this session, dealing with *Interactive Translation Using Quasi Logical Forms* (H. Alshawi & D. Carter (SRI International, Cambridge), *Co-description and Transfer* (L. Sadler, Essex), and *Two-level morphology in a unification-based formalism* (V. Pirrelli, Salford).

Alshawi & Carter describe work then in progress at SRI International on using the Core Language Engine in a MT context. The CLE is capable of deriving Quasi Logical Form representations which can then support reasoning and moreover allow representation of contextually determined aspects of interpretation. They illustrate the capabilities of the CLE through reference to collaborative work on an English-Swedish (typed) conversation interpreter. Alshawi & Carter argue that QLF is appropriate for transfer as it is far enough removed from surface form to capture cross-linguistic mappings flexibly, and that the mapping to QLF, because it relies on linguistic-based processing, is efficient and feasible as no reasoning is involved to reach QLF. In other words, the CLE approach involves

linguistic processing to reach QLF, and then may take advantage of knowledge-based processing thereafter to resolve contextual interpretation difficulties. In the MT system under discussion, QLF is used as a basis to drive interactive disambiguation sessions involving the user. Several issues arising from the use of QLF are discussed, with exemplification.

Sadler, in her paper, examines a particular approach to transfer, based on the co-description mechanism of Lexical-Functional Grammar, which has aroused interest in the MT community as an alternative to the classical transfer-model approach. After a brief description of the LFG approach, she then highlights two problematic sets of cases for this approach. Co-description involves specifying and resolving sets of constraints between source and target language structures, by means of translation functions. These functions are established between several linguistic levels across languages, hence allow translation relations to be stated at several points, not, as in the classical approach, only at the level of transfer itself. Problems however arise in at least two sets of cases: head-switching and cases where monolingual translation units are not co-extensive with units for translation. Examples of both sets of cases are given, with detailed discussion, and it then appears that the LFG approach is infelicitous in a number of respects.

Pirrelli presents the basics of Koskenniemi's Two-level Model for morphology, which relies on finite state transducers to capture correspondences between lexical and surface strings, and is capable of handling morphographemic alternations in an elegant manner. He then looks at unification-based derivatives of the original model, which reserve the Two-level Model mainly for morphographemics, while employing a unification-based morphosyntactic rule component to describe word structure. He concludes that such derivatives are computationally inefficient, being prone to excessive backtracking. The bulk of his paper is devoted to discussion and exemplification of an approach couched in the EUROTRA formalism, which is claimed to offer advantages in terms of more intelligent search, better context dependent feature assignment and more modular processing which allows features to be left unspecified until appropriate information is available.

2. Discourse & linguistics for MT

This session contained three papers: *Large-scale discourse structures and MT* (K. Sparck Jones, Cambridge), *Running a robust MT dialogue system* (W. Black, UMIST) and *High quality translation for monolinguals* (D. Jones, UMIST). The papers, and the ensuing group session discussion, are described in the rapportage by D. Arnold (Essex), in these proceedings.

3. Speech/NL interaction and MT

This session contained three papers, *Speech* (M. Tatham, Essex), *Functional Architectures for Spoken MT* (J. Connolly, Loughborough) and *Machine Translation with syntactic neural networks* (S. Lucas, Southampton). The papers, and the ensuing group session discussion, are described in the rapportage by K. Morton (Essex) in these proceedings.

4. Lexicon and MT

This session contained 4 papers, *The Eurotra UK Monolingual Dictionary* (K. Maxwell & B. Nkwenti-Azeh, UMIST), *Dictionaries for machine and machine-aided translation* (F. Knowles, Aston), *The lexicon and MT: a position paper* (J. Clear, OUP) and *Mel'čuk's ECD and MT* (A. Way, Essex).

Maxwell & Nkwenti-Azeh report on the English lexicon designed and implemented by the British EUROTRA Group. They relate this lexicon to other types of lexicon, then enter into substantial detail about design considerations, rationale of the dictionary, the features and values used and the interaction between the lexicon and the grammar modules within EUROTRA. They emphasise that only a subset of the lexicon information is used for EUROTRA purposes: in fact, this lexicon is designed to be re-usable by other projects and for other purposes.

Knowles addresses the problem of creation, configuration, enhancing, calibration and deployment of dictionaries for MT and MAT. He discusses the many types of lexical knowledge needed for MT, and emphasises the role of corpus based lexicography in elaborating MT dictionaries. He notes that

such dictionaries not only require different types of information to human oriented dictionaries, but also need access to such information as extensive lists of trademarks, proper names, etc. His paper stresses the complex nature of lexicography for MT purposes.

Clear examines the particular role of publishers' machine readable dictionaries and machine readable corpora in aiding MT researchers to build their lexica, and asks whether these resources can respond to MT researchers' needs. He examines how reliable publishers' dictionaries are, how dictionary data can be represented for NLP system purposes, what size and composition requirements are to be considered regarding the use of text corpora for deriving lexical data, and whether manual compilation of dictionaries provide useful benefits for MT.

Way presents an overview of Mel'čuk's ECD (Explanatory Combinatorial Dictionary). This dictionary model is part of Mel'čuk's Meaning-Text Model. The ECD is intended to be fully comprehensive and consistent, avoiding circularity and being formal and explicit. A key component of the BCD is the Lexical Function. The notion of LF is explained and exemplified. Way then goes on to briefly describe implementations or adaptations of the ECD in various current MT projects. It is concluded that the ECD is a useful model for MT, although there are doubts over its degree of complexity and the time needed to elaborate entries. However, it is noted that this may be a price that has to be paid for fairly complete formalisation of lexical knowledge. However, it is further noted that the effort put into creating an ECD will be rewarded as the dictionary will be re-usable for other purposes.

Much discussion took place during presentation of the papers, and in the ensuing group discussion session, especially as a large amount of detail had been presented. It was encouraging to note the high level of interaction among lexicographers from publishing houses and those concerned with MT: each group learned substantially from the other and it is to be hoped that each is now aware of the design possibilities and requirements of MT lexica.

The demonstration sessions were divided in three sessions:

1. EUROTRA
2. Essex work in Machine Translation
3. UMIST work in Machine Translation

The EUROTRA demonstration was a joint effort between the two universities comprising the British EUROTRA Group (UMIST and Essex). After a preliminary demonstration of a EUROTRA English ↔ Dutch system (N. Underwood, J. Lindop (UMIST) & I. Crookston (Essex)), two short papers were given as follows:

- *Strategies for MT Synthesis in English in the E-Framework* (I. Crookston, Essex)
- *The Work of the Eurotra Linguistic Specifications Group* (J. Durand, Salford and P. Bennett, UMIST).

A background paper on EUROTRA was provided by A. Syea (UMIST).

In the session on Essex Work in MT, demonstrations were given of two MT systems (MiMo and CAT2), together with a short talk by L. Balkan on *Complex transfer,* with reference to EUROTRA.

In the UMIST session, J. Philips gave a paper on the *Automated production of domain-specific MT systems,* which examined the use of cluster analysis in helping to arrive at general collocational restrictions for lexical items, thus aiding in ambiguity resolution.

It is stressed that many other aspects of MT research at UMIST and Essex were presented throughout this workshop.

3 Conclusion

This workshop was, we believe, highly successful in bringing recent MT research results to the knowledge of the SALT community. It was clear that MT in the UK is being pursued at many levels and according to many models. This is all the more remarkable when one considers the quite small number of centres involved in MT in the UK. It is equally clear that MT is a very active field, and that encouraging results have been forthcoming. The EUROTRA project has had the single most significant impact on the field in Britain, and has been largely responsible for the existence of two key centres, Essex and UMIST.

Industrial interest and activity in MT appears to be growing, although somewhat slowly. Those industrial representatives who gave talks at this workshop demonstrated that much valuable research, both pure and applied, has been garnered in a few centres. It is evident that the UK has the expertise to achieve much in MT, however it would appear that further consolidation and expansion of links, particularly between academia and industry, is required if MT is to develop further at an industrial level.

4 Acknowledgements

Many people contributed to the smooth running of this workshop, to whom we extend our grateful thanks. We are particularly indebted to Helen Stanley (DTI/ITD, SALT Club secretary) for running the registration desk team and coping with participants' queries, to DTI/ITD who underwrote the workshop and to the Department of Language and Linguistics, UMIST, who provided computational and logistic support. Martin Earl (UMIST) deserves special mention for arranging installation of demonstration equipment. Lastly, we were pleasantly surprised by the volume and quality of papers given, and wish to record our gratitude to all speakers, chairpersons and rapporteurs, and to say an especial word of thanks to our two guest speakers from SNI and ATR.