[From: Wolfgang Menzel & Cristina Vertan (eds.) *Natural Language Processing between Linguistic Inquiry and System Engineering*. Iaşi: Editura Universităţii "Alexandru Oan Cuza", 2003.]

Machine Translation - where are we going?

Bente Maegaard

Center for Sprogteknologi, Copenhagen, Denmark

bente@cst.dk

1 Introduction

Machine Translation is more than 50 years old, and during its last 15 years it has reached a maturity which was probably not expected by many watchers or prospective users 20 years ago. Researchers and developers did believe in machine translation 20 years ago, and funding agencies believed in it too, otherwise the systems that we are currently using would not have existed.

The 50 years of machine translation history have seen the development of several major approaches and, more recently, of a new enabling paradigm of statistical processing. Still, today, there is no dominant approach. Despite the commercial success of many MT systems, tools, and other products, the main problem remains unsolved, and the various ways of combining approaches and paradigms are only beginning to be explored.

2 Definition of the term MT

The term Machine Translation (MT) is normally taken in its restricted and precise meaning of fully automatic translation. However, here we consider the whole range of tools that may support translation and multilingual document production in general. This is especially important when considering the integration of other language processing techniques and resources with MT.

We therefore define Machine Translation to include any computer-based process that transforms (or helps a user to transform) written text from one human language into another. We define Fully Automated Machine Translation (FAMT) to be MT performed without the intervention of a human being during the process. In practice however, in this article, the term MT is used in both the broad sense defined above, and in the FAMT sense, but only where no confusion is possible.

Traditionally, two different set-ups for the translation task have been identified, namely assimilation and dissemination. Assimilation refers to the set-up for the

translation task in which an individual or organisation want to convert foreign material written in various languages into their own language. Dissemination refers to the set-up in which an individual or organisation want to disseminate their own material, written in one language, to the world in a variety of languages, A third set-up for the translation task, namely communication, has now become relevant. Communication refers to the setting in which two or more individuals are in more or less immediate interaction, typically via email or otherwise online, with an MT system mediating between them. Each set-up for the translation task has very different features, is best supported by different underlying technology, and is to be evaluated according to somewhat different criteria.

3 MT status

Thanks to ongoing commercial growth and the influence of new research, the situation today is different today from ten years ago. There has been a trend towards embedding MT as part of linguistic services, which may be as diverse as email across countries, foreign-language web searches, traditional document translation, and portable speech translators with very limited lexicons (for travelers, etc).

In organisations such as the European Commission, large integrated environments have been built around MT systems; cf. the European Commission Translation Service's Euramis (Theologitis, 1997). Here the translator has access to all types of tools: translation memory, MT, term bases, parallel texts, document repositories etc. The new environment has also given rise to a new workflow, including new tasks and new ways of distributing the tasks over staff types, e.g. the preparation of a text for translation, term look-up etc. can be done by less educated and therefore less expensive staff than translators.

MT services are offered via the Internet, often free for shorter texts; In addition, MT is increasingly being bundled with other web services, cf. the web site of AltaVista, which is linked to Systran.

4 Status of MT performance

General purpose vs. Domain-specific: Most (commercial) systems are meant to be general purpose. Although the performance is actually not always very good, the systems are used anyway. However, if the systems were better, MT would be used a whole lot more - given the explosion of information in the world, the demand for translation is booming, and the only possible answer to this demand is MT (in all its forms).

Domain-specific systems deliver better performance, as they can be tailor-made to specific text types. TAUM-METEO, for example, contains a lexicon of only 220

words, and produces translations of weather reports at 98% accuracy; PaTrans (Maegaard and Hansen, 1995) translates patents within chemistry and other technical domains at high quality. However, domain specific systems exhibit two drawbacks: they are only cost-effective in large-volume domains, and maintaining many domain-specific systems may not be manageable. This is the reason most commercial systems are general purpose as mentioned above.

5 MT methodology

One of the most pressing questions of MT results from the recent introduction of a new paradigm into Computational Linguistics. It had always been thought that MT, which combines the complexities of two languages (at least), requires highly sophisticated theories of linguistics in order to produce reasonable quality output.

However, DARPA evaluations in the early 1990es showed that MT systems using statistical techniques to gather their rules of cross-language correspondence were feasible competitors to traditional, purely hand-built ones. On the other hand, the statistics-only approach was clearly not the optimal path; in developments since 1994, it has included steadily more knowledge derived from linguistics. This left the burning question: which aspects of MT systems are best approached by statistical methods, and which by traditional, linguistic ones?

Since 1994, a new generation of research MT systems is investigating various hybridisations of statistical and symbolic techniques (Knight et al., 1995; Brown and Frederking, 1995; Dorr, 1997; Nirenburg et al., 1992; Kay et al., 1994). While it is clear by now that some modules are best approached under one paradigm or the other, it is a relatively safe bet that others are genuinely hermaphroditic, and that their best design and deployment will be determined by the eventual use of the system in the world. Given the large variety of phenomena inherent in language, it is highly unlikely that there exists a single method to handle all the phenomena— both in the data/rule collection stage and in the data/rule application (translation) stage - optimally. Thus one can expect all future non-toy MT systems to be hybrids. Methods of statistics and probability combination will predominate where robustness and wide coverage are at issue, while generalisations of linguistic phenomena, symbol manipulation, and structure creation and transformation will predominate where fine nuances (i.e., translation quality) are important.

If we look at how the techniques correspond to the tasks of translation as mentioned above, we may summarise as follows:

- assimilation tasks: lower quality, broad domains — statistical techniques predominate

- dissemination tasks: higher quality, limited domains - symbolic techniques predominate
- communication tasks: medium quality, medium domain - mixed techniques predominate

6 Multi-Engine MT

In recent years, several different methods of performing MT-transfer, example-based, simple dictionary lookup, etc. - have all shown their worth in the appropriate circumstances. A promising recent development has been the attempt to integrate various approaches into a single multi-engine MT system. The idea is very simple: pass the sentence(s) to be translated through several MT engines in parallel, and at the end combine their output, selecting the best fragment(s) and recomposing them into the target sentence(s).

7 Looking forward

7.1. Applications

One important trend, of which the first instances can be seen already, is the availability of MT for casual, one-off, use via the Internet. Such services can either be standalone MT or bundled with some other application, such as web access (as is the case with web site of Altavista and Systran), multilingual information retrieval in general, text summarisation, and so on.

A second trend can also be recognised: the availability of low-quality portable speech-to-speech MT systems. An experimental system constructed at Carnegie Mellon University in the USA was built for use in Bosnia. Verbmobil handles meeting scheduling in spoken German, French, and English. It is expected that these domains will increase in size and complexity as speech recognition becomes more robust.

It is difficult to set limits for the possibilities for MT applications in the future: 1) a newsreader device that translates articles form thousands of publications worldwide, delivering them as MP3 audio files, 2) Travel Sunglasses offer real-time translation of road signs, marquees, and menus into wearer's native language, 3) a Lexical Disambiguation System (LDS) is embedded into smartcards equipped with membrane microphones so travellers can converse with store clerks in dozens of languages, etc. Examples taken from (Demos et al. 2000) where they even appear with a proposed year: newsreader in 2002, sunglasses in 2008, LDS in 2012!

7.2 Methods

As analysis and generation theory and practice becomes more standardised and established, the focus of research will increasingly turn to methods of constructing low-quality yet adequate MT systems (semi-)automatically. Methods of automatically building multilingual lexicons and word lists involve bitext alignment and word correspondence discovery; see e.g. (Wu, 1995).

It is clear from the discussion above that future developments will include highly integrated approaches to translation (integration of translation memory and FAMT, hybrid statistical-linguistic translation, multi-engine translation systems, and the like). We are likely to witness the development of statistical techniques to address problems that defy easy formalisation and obvious rule-based behaviour, such as sound transliteration, word equivalence across languages (Wu, 1995), word sense disambiguation, etc.

8 Problems to solve

Semantics is needed! Without some level of semantic representation, MT systems will never be able to achieve high quality, because they will never be able to differentiate between cases that are lexically and syntactically ambiguous. Ongoing work on semantics in lexicons and ontologies will benefit MT (as it will other applications such as Summarisation and Information Extraction).

Secondly, an increasingly pressing bottleneck is the fact that essentially all MT systems operate at the single-sentence level. Except for a few experimental attempts, no systems have explored translating beyond the sentence boundary, using discourse representations. Typically, their only cross-sentence record is a list of referents for pronouns and other anaphora. Yet many phenomena in language span sentence boundaries.

Third, the treatment of so-called less-used languages requires additional attention. Not all languages are equally well covered by MT. The so-called major languages are reasonably well covered at present, and will certainly be well covered in the future, but users of less spoken languages need MT and other tools just as much or even more than users of English, Spanish, French and Japanese. For some languages the market is not sufficiently large, which means that users of those language will lack the tools which are otherwise available. This lack of tools will have an obvious economic effect, but also a cultural effect by excluding some languages from participating in an otherwise flourishing multilinguality.

For such languages, where market forces alone cannot solve the problem, governments, organisations and funding agencies have to take an active role in the development of tools, since this is the only way to protect and reinforce most languages in the world. The active role consists in at least creating the necessary amount of basic language resources, such as text repositories, corpora, speech databases etc.

9 Closing remarks

The future of MT is rosy. Thanks largely to the Internet and the growth of international commerce, casual (one-off) and repeated MT is growing at a very fast pace. Correspondingly, MT products are coming to market as well.

In parallel with this growth, it is imperative to ensure that research in MT continues. Without research the difficult problems of MT will not be solved; MT companies do not have enough financial leeway or in many cases the technical expertise required to make theoretical breakthroughs. Since market forces alone cannot solve the problem, governments and funding agencies have to take an active role in the protection and reinforcement of MT.

Acknowledgements

The present article draws upon the chapter on Machine Translation, edited by me, but with numerous other contributors, in Hovy et al. (ed.): *Multilingual Information Management: Current Levels and Future Abilities,* Istituti Editoriale e Poligrafici Internazionali, Pisa, 2001.

References

[1] Arnold, D.J. et al. *An Introduction to Machine Translation.* Oxford: Blackwell. 1994

[2] Boitet, C. *Factors for success (and failure) in Machine Translation - some lessons of the first 50 years of R&D.* In: MT Summit V Proceedings, Luxembourg 1995.

[3] Brown, P.P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossin. *A Statistical Approach to Machine Translation.* Computational Linguistics 16(2) (79—85)., 1990

[4] Brown, R., and R. Frederking. *Applying Statistical English language Modeling to Symbolic Machine Translation.* Proceedings of the Conference on Theoretical and Methodological Issues in MT (TMI-95), (221—239). 1995

[5] Carter, D., R. Becket, M. Rayner, R. Eklund, C. MacDermid, M. Wiren, S. Kirchmeier-Andersen, and C. Philp. *Translation Methodology in the Spoken language Translator: An Evaluation.* Proceedings of the Spoken Language Translation Meeting, (73—81). ACL/ELSNET, Madrid. 1997

[6] Church, K.W. and E.H. Hovy. *Good Applications for Crummy Machine Translation.* Journal of Machine Translation 8 (239—258). 1993

[7] Copeland, C, J. Durand, S. Krauwer, B. Maegaard (ed.): *The EUROTRA linguistic Specifications.* Studies in Machine Translation and Natural Language Processing, Vol. 1, Commission of the European Communities, Luxembourg 1991

[8] Copeland, C., J. Durand, S. Krauwer, B. Maegaard (ed.): *The EUROTRA Formal Specifications.* Studies in Machine Translation and Natural Language Processing, Vol. 2, Commission of the European Communities, Luxembourg 1991

[9] Demos, K., M. Frauenfelder: *Machine Translation's Past and Future.* Wired Digital Inc, 2000.

[10] Dorr, B.J. *Machine Translation Divergences: A Formal Description and Proposed Solution.* Computational Linguistics 20(4) (597—634). 1994

[11] Frederking, R., S. Nirenburg, D. Farwell, S. Helmreich, E. Hovy, K. Knight, S. Beale, C. Domanshnev, D. Attardo, D Grannes, R. Brown. *Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation System.* Proceedings of the First AMTA Conference, Columbia, MD (73—80). 1994

[12] Hutchins, J. *Compendium of Machine Translation Software.* Available from the International Association of Machine Translation (IAMT). 1999

[13] Kay, M., J.M. Gawron, and P. Norvig. *Verbmobil: A Translation System for face-to-face Dialog.* CSLI Lecture Notes No. 33, Stanford University. 1994

[14] Kay, M. *The proper place of men and machines in translation.* Machine Translation 23.1997

[15] Knight, K., I. Chander, M. Haines, V. Hatzivassiloglou, E.H. Hovy, M. Iida, S.K. Luk, R.A. Whitney, and K. Yamada. *Filling Knowledge Gaps in a Broad-Coverage MT System.* Proceedings of the 14th IJCAI Conference. Montreal, Canada. 1995

[16] Maegaard, B. and V. Hansen. *PaTrans, Machine Translation of Patent Texts, From Research to Practical Application.* Proceedings of the Second Language Engineering Convention, (1—8). London: Convention Digest. 1995

[17] Nagao, M. *A Framework of a Machine Translation between Japanese and English by Analogy principle,* (173—180). In Elithorn and Banerji (eds.), Artificial and Human Intelligence, North Holland. 1984

[18] Niemann, H., E. Noeth, A. Kiessling, R. Kompe and A. Batliner. *Prosodic Processing and its Use in Verbmobil.* Proceedings of ICASSP-97, (75—78). Munich, Germany. 1997

[19] Nirenburg, S., J.C. Carbonell, M. Tomita, and K. Goodman. *Machine Translation: A Knowledge-Based Approach.* San Mateo: Morgan Kaufmann. 1992

[20] Nirenburg, S., Project Boas: *"A Linguist in the Box" as a Multi-Purpose Language Resource.* Proceedings of the First International Conference on Language Resources and Evaluation (LREC), (739—745). Granada, Spain. 1998

[21] Theologitis, D. *Integrating Advanced Translation Technology.* In the 1997 LISA Tools Workshop Guidebook, (1/1—1/35). Geneva. 1997

[22] Tsujii, Y. *Multi-Language Translation System using Interlingua for Asian Languages.* Proceedings of International Conference organized by IPSJ for its 30th Anniversary. 1990

[23] Vossen, P., et al. *EuroWordNet.* Computers and the Humanities, 1999.

[24] White, J. and T. O'Connell. *ARPA Workshops on Machine Translation.* Series of 4 workshops on comparative evaluation. PRC Inc., McLean, VA. 1992-94.

[25] Wu, D. *Grammarless Extraction of Phrasal Translation Examples from Parallel Texts.* Proceedings of the Conference on Theoretical and Methodological Issues in MT (TMI-95), (354—372). 1995

[26] Yamron, J., J. Cant, A. Demedts, T. Dietzel, Y. Ito. *The Automatic Component of the LINGSTAT Machine-Aided Translation System.* In Proceedings of the ARPA Conference on Human Language Technology, Princeton, NJ (158—164). 1994