Toward a Hybrid Integrated Translation Environment

Michael Carl[1], Andy Way[2], and Reinhard Schäler[3]

[1] Laboratoire de Recherche Appliquée en Linguistique Informatique
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal, Montréal, Quebec, Canada
carl@iro.umontreal.ca
[2] School of Computer Applications,
Dublin City University, Dublin 9, Ireland
away@computing.dcu.ie
[3] Localisation Research Centre (LRC)
Department of Computer Science and Information Systems (CSIS)
University of Limerick, Limerick, Ireland
Reinhard.Schaler@ul.ie

Abstract. In this paper we present a model for the future use of Machine Translation (MT) and Computer Assisted Translation. In order to accommodate the future needs in middle value translations, we discuss a number of MT techniques and architectures. We anticipate a hybrid environment that integrates data- and rule-driven approaches where translations will be routed through the available translation options and consumers will receive accurate information on the quality, pricing and time implications of their translation choice.

1 A Model for the Use of MT

In this paper, we present a model for the future use of Machine Translation (MT) and Computer Assisted Translation (CAT) (cf. [22]). The model (see Figure 1) is based on the assumption that information can be categorized into three types.

At the bottom of the pyramid comes non-mission-critical information, the so-called gisting market. An example might be an article written in Japanese about Picasso on a website in Japan, of which an English speaker with no Japanese but interested in the Spanish painter wants a rough and ready translation. This is the ideal application scenario to ensure wide use of general purpose MT.

In the middle of the pyramid come large amounts of material that have to be translated accurately, where gisting is not acceptable. Examples of this type of information are technical manuals and other documentation. Most of these translations are domain-specific requiring specialized dictionaries with well defined meanings and/or specialized grammars. MT is currently being used at this level, although not widely.

At the top of the pyramid come small amounts of mission-critical or creative material to be read or referenced where accuracy and presentation are

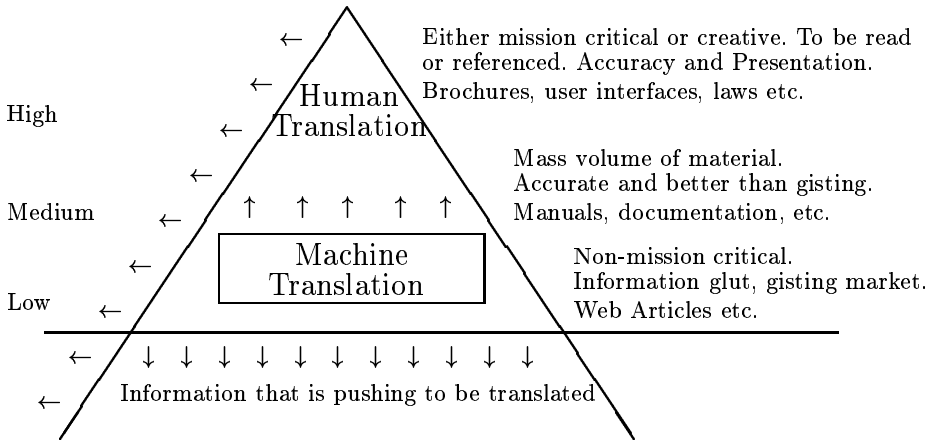

Fig. 1. A Model for the Future Use of Translation Technology

paramount. Examples of this include user interfaces, laws and creative literature.

The model presumes (1) that the shape of the pyramid is expanding in two directions and (2) that improvements in translation technology will open up new markets for developers of MT systems. The expansion of the pyramid will be driven by two factors: a growing demand for translated material because of the globalisation of the economy (horizontal expansion) and the increasing availability and accessibility of information in a variety of languages to end-users on the web (vertical expansion). At the same time, MT will push its way up the pyramid and be used for higher quality translation.

2 An Integrated Translation Environment

Translation service vendors will offer various translation facilities online, from high quality human translation to low-end, cheaper MT. In between we envisage a range of mixed options, including human-edited MT using specialized and fine-tuned lexical and semantic databases, a combination of TM and MT, and alignment and maintenance of previously translated material.

We anticipate a hybrid MT platform which integrates a number of applications, techniques and resources together. These include applications such as multilingual alignment, terminology management, induction of grammars and translation templates, and consistency checkers. These platforms will also integrate example-based, statistics-based and rule-based approaches to MT, together with a variety of other linguistic resources and corpora.

Some researchers have pondered the suitability of texts for MT. The work that we are aware of regarding translatability and MT focuses only on what texts should be sent to rule-based MT systems. One possible translatability indicator for the use of MT in general is the identification of (sets of) phenomena which are likely to cause problems for MT systems ([11], with respect to the LOGOS MT system). Based on their work with the PaTrans system, Underwood and Jongejan provide a definition of translatability:

> "the notion of translatability is based on so-called 'translatability indicators' where the occurrence of such an indicator in the text is considered to have a negative effect on the quality of machine translation. The fewer translatability indicators, the better suited the text is to translation using MT" [26, p.363].

In an integrated translation environment, these definitions have to be widened considerably. Future translatability indicators will have to be more fine-grained and MT systems will have to be adaptable to and learn from such indicators. Translatability indicators will have to detail why a text is not (yet) suited for automatic translation so that a tool may be triggered to render the text suitable for automatic translation. That is, a hybrid integrated translation environment has to provide a means of separating translatable from non-translatable parts of a text in a more sophisticated manner than TMs currently do. For each part one has to estimate the expected quality of the translation, the effort and cost of upgrading resources and/or improving the source text in order to improve translation quality. The integrated system has to be aware of gaps in the source text which it cannot tackle and provide intelligent inference mechanisms to generate solutions for bridging these gaps.

Translations will be routed through the available translation options according to criteria such as the type of text, the value of the information to be translated, the quality requirements of the consumers, and the resources available to them. Before they select one of the translation options, consumers will receive accurate information on the quality, pricing and time implications of their choice.

3 Enhancing Medium Value Translation Quality

Despite major efforts to build new translation engines and to increase the quality of automatic translations, a major breakthrough cannot be expected in the years to come through refined technologies alone. Rather, in order to enhance the quality of MT systems and to make it suitable for medium value translations (cf. Figure 1), MT systems need to be adjusted to the domain at hand. Controlled languages and controlled translations have a crucial role to play here.

Controlled languages have been developed since the early 70's as a compromise between the intractability of natural languages and the formal simplicity of artificial languages. Controlled languages define a writing standard for domain-specific documents. Linguistic expressions in texts are restricted to a subset of natural languages. They are characterized by simplified grammars and

style rules, a simplified and controlled vocabulary with well defined meanings, and a thesaurus of frequently occurring terms. Controlled languages are used to enhance the clarity, usability, and translatability of documents. According to Lehrndorfer and Schachtl [16, p.8], "the concept of controlled language is a mental offspring of machine translation". A number of companies (e.g. Boeing, British Airways, Caterpillar) use controlled language in their writing environment. Nor is this trend restricted to English: Siemens use controlled German (Dokumentationsdeutsch [16]), Aérospatiale use controlled French (GIFAS Rationalized French [4]), while Scania use controlled Swedish (ScaniaSwedish [2]).

We now examine how well rule-based and data-driven MT systems may be adapted to controlled languages.

3.1 Controlled Language and Rule-Based MT

Controlled languages have been developed for restricted domains, such as technical documentation for repair, maintenance and service documents in large companies (e.g. Siemens, Scania, GM). Caterpillar's Technical English, for instance, defines monolingual constraints on the lexicon, and constraints on the complexity of sentences. However, when using this controlled language for translation in the KANT rule-based MT (RBMT) system, it was found that: "[terms] that don't appear to be ambiguous during superficial review turned out to have several context-specific translations in different target languages" [14].

Van der Eijk *et al.* [27, p.64] state that "an approach based on fine-tuning a general system for unrestricted texts to derive specific applications would be unnecessarily complex and expensive to develop". Later work in METAL applications refers to there being "limits to fine-tuning big grammars to handle semi-grammatical or otherwise badly written sentences. The degree of complexity added to an already complex NLP grammar tends to lead to a deterioration of overall translation quality and (where relevant) speed" [1, p.595]. Furthermore, attempts at redesigning the Météo system, probably the biggest success story in the history of MT, to make it suitable for another domain (aviation) proved unsuccessful.

Controlled translation, therefore, involves more than just the translation of a controlled language. Passing a source language text through a controlled language tool is not sufficient for achieving high quality translation. Large general purpose MT systems cannot easily be converted to produce controlled translations. In a conventional transfer-based MT system, for instance, controlling the translation process involves controlling three processing steps: (i) the segmentation and parsing of the source text; (ii) the transfer of the source segments into the target language ; and (iii) the recombination and ordering of the target language segments according to the target language syntax. As the resources of each of these steps require independent knowledge resources, adjusting a conventional RBMT system to a new controlled language is non-trivial as domain-specific knowledge resources have to be acquired, adjusted and homogenized.

3.2 Controlled Language and Data-driven MT

It is widely acknowledged that data-driven MT systems can overcome the 'knowledge acquisition bottleneck' given that available translations can be exploited. In contrast to traditional approaches, data-driven MT systems induce the knowledge required for transfer from a reference text. To date, data-driven MT technologies have yet to tackle controlled languages: they have not supported the acquisition of controlled translation knowledge, nor have they provided an appropriate environment for controlled translation.

This is extremely surprising: the quality of data-driven translation systems depends on the quality of the reference translations from which the translation knowledge was learned. The more a reference text is consistent, the better the expected quality of the translations produced by the system, while translation knowledge extracted from noisy corpora has an adverse impact on overall translation quality. The only research we are aware of here attempts to detect mistranslations [21] or omissions in translations [9], [17]. However, in the context of data-driven MT, such methods have not been used so far to eliminate noisy or mistranslated parts of the reference text, nor to enhance the quality and consistency of the extracted translation units.

Controlled Language and TM Conventional TM systems are not suitable for controlled translation. TMs are essentially bilingual databases where translation segments are stored without a built-in possibility to automatically check the consistent use of terms and their translations or the complexity of the sentences.

Within the TETRIS-IAI project [13], controlled language was fed into a TM. It was found that controlling the source language without controlling the reference material does not increase the hit-rate of the TM and thus does not increase the chance of high quality translations — from a company's point of view, the bottom line is that the translation cost is not lowered. Methods for preparing and modifying reference texts to achieve better consistency on a terminological and syntactic level have therefore been proposed [24] and could also be a feasible way forward for TMs.

The translation process in a TM system may be distorted by two factors: the way entries are retrieved from the TM (fuzzy matching) and the contents of the TM (the chance that it contains noisy and inconsistent fragments). Where these factors co-occur, we can expect translation quality to deteriorate further.

Controlled Language and Statistics-Based MT Similarly, purely statistics-based MT (SBMT) is not an appropriate candidate for controlled translation. Owing to the size of the reference texts, one cannot usually expect consistent reference translations in SBMT. In many cases, texts from different domains are merged together to compute word translation probabilities for a language pair in various contexts. However, how words and phrases are used in different domains can differ greatly. [15] shows that the performance of a statistical MT system trained on one of the largest available bilingual texts — the Canadian Hansards — deteriorates when translating texts from a very specific domain.

Controlled Language and Example-Based MT In our view, the main potential of example-based MT (EBMT) lies in the possibility of easily generating special purpose MT systems. As in other data-driven MT systems, EBMT [8] extracts lexical and transfer knowledge from aligned texts. A number of systems (e.g. [18, 25, 29]) combine linguistic resources and dictionaries to support this acquisition process. Automatic and/or semi-automatic control mechanisms could be implemented at this stage to extract high quality and/or domain-specific translation equivalences [3, 29].

Controlling the translation in EBMT implies the careful selection of translation examples which are similar to the input. Given that only target fragments of the retrieved examples are recombined to build the translation, controlling EBMT is reduced to controlling the retrieval of appropriate analogous examples.

The more the domain of the text to be translated is restricted, the more these restrictions are well defined and the more high quality reference translations are available, so the analogy between the new text and the retrieved examples will become more obvious. In contrast to SBMT and TM, the potential of EBMT improves as the likelihood of producing high quality translations increases as more examples are added to the system database.

[6] shows that coverage can be increased by a factor of 10 or thereabouts if templates are used, but it would be fanciful to think that this would scale up to domain-independent translation. Even if EBMT systems were augmented with large amounts of syntactic information (e.g. [18, 30]), they would in all probability stop short of becoming solutions to the problems of translating general language. Nevertheless, it is our contention that EBMT systems may be able to generate controlled, domain-specific translations given a certain amount of built in linguistic knowledge and some preparation of the aligned corpus.

4 Integrating Different MT Paradigms

The various MT paradigms have different advantages and shortcomings. TMs are fed with domain-specific reference translations and are widely used as tools for CAT. TMs, however, run short of providing sufficient control mechanisms for more sophisticated translations. In contrast, RBMT systems are mostly designed for translating general purpose texts. As a consequence, they are difficult to adjust to specialized texts and consequently suffer from limited portability. Probabilistic approaches to MT are trained on huge bilingual corpora, yet portability of these systems remains low. As a compromise between the different approaches, EBMT systems have emerged as primarily data-driven systems but which may also make use of sophisticated rule-based processing devices at various stages of the translation process.

Given the different advantages and disadvantages of each MT paradigm, hybrid and multi-engine MT systems have been designed as an attempt to integrate the advantages of different systems without accumulating their shortcomings.

4.1 Multi-Engine Machine Translation Systems

In order to classify these systems, a distinction can be made as to whether entire translation engines are triggered in parallel or sequentially. In a parallel multi-engine scenario, each system is fed with the source text and generates an independent translation. The translations are then collected from their output and (manually or automatically) recombined.

Parallel Multi-Engine Translation Systems There are a number of projects which incorporate different MT components in parallel in a multi-engine system. Verbmobil [28] integrates the complementary strengths of various MT approaches in one framework, i.e. deep analysis, shallow dialogue act-based approach and simple TM technology. [19] shows that the performance of the integrated system outperforms each individual system. PanGloss [10] uses EBMT in conjunction with KBMT and a transfer-based engine.

While there is an element of redundancy in such approaches given that more than one engine may produce the correct translation (cf. [30]), one might also treat the various outputs as comparative evidence in favour of the best overall translation. Somers [23] observes: "what is most interesting is the extent to which the different approaches mutually confirm each other's proposed translations".

Sequential Multi-Engine Translation Systems In this approach, two or more MT components are triggered on different sections of the same source text. The output of the different systems is then concatenated without the need for further processing. This dynamic interaction is monitored by one system — usually the most reliable amongst the available systems. The reasoning behind this approach is that if one knows the properties of the translation components involved, reliable translations can be produced by using fewer resources than in a parallel multi-engine approach.

Integration of a TM with a rule-based component is a common strategy in commercial translation. A dynamic sequential interaction between a translation memory (TRADOS) and an MT system (LOGOS) is described in [12]. In the case where only poorly matching reference translations are available in TRADOS, the input sentence is passed to LOGOS for regular translation. The user is then notified which of the systems has processed the translation, since LOGOS is less likely to produce reliable results.

A similar scenario is described in [7]. Where a TM is linked with an EBMT system, the quality of translations is likely to be higher for EBMT translation than for TM translation, where the match score of the TM falls below 80%.

4.2 Hybrid MT Systems

While in multi-engine MT systems each module has its own resources and data structures, in a hybrid MT system the same data structures are shared among different components. Some components may, therefore, modify or adjust certain

processing resources of others in order to enhance the translation candidates with respect to coverage or translation quality.

Hybrid Statistics-Based and Rule-Based MT Systems Coupling (statistical) data and RBMT leads to a hybrid integration. In some such hybrid systems, statistical data is added to the lexical resources of the RBMT system in order to adjudge different translation candidates as more or less felicitous for a given thematic context. In particular, it has been shown that statistically enriched RBMT systems can handle collocational phenomena.

[20] describes an application of statistical data during the rule-based transfer phase. Statistical data are derived by manually scoring translation variants produced by the system. Since training is based on texts belonging to one specific subject field, typical mistakes made by the system can be corrected. The probability of a transfer candidate is calculated by means of the transfer probability and the probability of the resulting target structure. As such a multiplication of probabilities requires large amounts of data in order to be effective, these approaches are applicable only to very restricted subject fields where only a few examples may suffice in order to produce reliable data. In such cases, translation quality is traded for improved coverage.

Hybrid Example-Based and Rule-Based MT Systems In a hybrid stratificational integration of example-based and rule-based techniques, some processing steps are carried out by the rule-based component while examples are used in other stages.

[18] combines rule-based analysis and generation components with example-based transfer. [5] generates translation templates for new sentences on the fly from a set of alignments. The differing sections in the source template and the input sentence are identified and translated by a rule-based noun-phrase translation system.

However, even a very large data-driven MT system is unlikely to be able to translate a completely new sentence correctly, let alone an entire new text. However, such systems are able to 'learn' in that new examples can be added to the system database, so that subsequent encounters with previously unknown strings will be translated successfully. In RBMT systems there is no such analogous process. That is, they do not store translation results for later reuse, so that all post-editing effort is wasted: RBMT systems handle the same input in exactly the same way in perpetuity.

A hybrid system, in contrast, will be able to learn and adapt easily to new types of text. Furthermore, such systems are based on sophisticated language models as a property of the rule-based component. Consequently, one can envisage that even if none of the individual engines can translate a given sentence correctly, the overall system may be able to do so if the engines are allowed to interact. Even if the individual components improve, the integrated system should always outperform the individual systems with respect to either the quality of the translation, the performance, or the tunability of the system.

5 Conclusion

On various occasions in recent decades, MT companies have claimed that the linguistic technology developed by them has made human translation redundant. These claims have so far not had a significant impact on the reality of translation as a profession and as a business. The one technology that has had a considerable impact on translation has been TM — it has changed the way translators work, as can be seen when examining the impact it had in the localization industry, one of the largest employers of technical translators. Ironically, TM technology works without any of the sophisticated linguistic technologies developed over decades by MT developers.

Only recently, and driven by increased activities in the area of EBMT, has the interest shown by the linguistic tools industry in research results been reciprocated by the research community. One possible reason for this development is that although EBMT as a paradigm has been described in research papers as far back as the 1980's, and although it has managed to capture the interest and enthusiasm of many researchers, it has, so far, failed to reach the level of maturity where it could be transformed from a research topic into a technology used to build a new generation of MT engines — and new approaches, technologies and applications are badly needed in MT.

If data-driven MT is to find a niche amongst the different MT paradigms, we believe it has to offer the potential to easily adapt to new domains in a more controlled manner than TMs currently do. The adaptation process differs from TM technology with respect to how and what kind of translation knowledge is stored, how it is retrieved and how it is recomposed to build a new translation. This requires sophisticated processing based on linguistic resources and/or advanced numerical processing.

In this paper we developed a model for the future use of translation technology. We anticipate an integrated hybrid translation environment which unifies a number of MT technologies, linguistic and processing resources and the actual human translator. This setting would be a valuable aid to translators, capable of generating descriptive, controlled or general translations according to the needs of users and the effort they are willing to invest.

References

1. G. Adriens and D. Schreurs. From cogram to alcogram: toward a controlled English grammar checker. *Coling*, pp.595–601, Nantes, France, 1992.
2. I. Almqvist and A. Sagvall Hein. Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. *CLAW 96*, pp.159–164, Leuven, Belgium, 1996.
3. T. Andriamanankasina, K. Araki, and K. Tochinai. EBMT of POS-tagged sentences with inductive learning. In *[8]*.
4. K. Barthe. GIFAS rationalised French: Designing one controlled language to match another. *CLAW 98*, pp.87–102, Pittsburgh, PA, 1998.
5. F. Bond and S. Shirai. A hybrid and example-based method for machine translation. In *[8]*.

6. R. D. Brown. Example-Based Machine Translation at Carnegie Mellon University. *The ELRA Newsletter*, 5(1):10–12, 2000.
7. M. Carl and S. Hansen. Linking Translation Memories with Example-Based Machine Translation. *MT-Summit VII*, Singapore, 1999.
8. M. Carl and A. Way, editors. *Recent Advances in Example-Based Machine Translation.* Kluwer Academic Publishers, Boston/Dordrecht/London. forthcoming.
9. S. Chen. Building probabilistic models for natural language. PhD thesis, Harvard University, Cambridge, MA, 1996.
10. R. Frederking and S. Nirenburg. Three heads are better than one. *Proceedings of ANLP-94*, pp.95–100, Stuttgart, Germany, 1994.
11. C. Gdaniec. The LOGOS translatability index. *Proceedings of the First Conference for Machine Translation in the Americas*, pp.97–105, 1994.
12. M. Heyn. Integrating machine translation into translation memory systems. In *EAMT Workshop Proceedings*, pp.111–123, ISSCO, Geneva, 1996.
13. IAI, Saarbrücken, Germany. *Technologie-Transfer intelligenter Sprachtechnologie*, 1999. `http://www.iai.uni-sb.de/tetris/tetris_home.htm`.
14. C. Kamprath, E. Adolphson, T. Mitamura, and E. Nyberg. Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. *CLAW '98*, 1998.
15. P. Langlais. Terminology to the rescue of statistical machine translation: an experiment. *To appear*, 2002.
16. A. Lehrndorfer and S. Schachtl. Controlled Siemens Documentary German and TopTrans. *TC-FORUM*, 1998.
17. D. I. Melamed. *Empirical Methods for Exploiting Parallel Texts.* MIT Press, Cambridge, MA, 2001.
18. A. Menezes and S. D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *[8]*.
19. R. Nübel. End-to-End Evaluation in Verbmobil I. *MT-Summit*, San Diego, 1997.
20. M. Rayner and P. Bouillon. Hybrid transfer in an English-French spoken language translator. *Proceedings of the IA'95*, Montpellier, France, 1995.
21. G. Russell. Errors of Omission in Translation. *TMI 99*, pp.128–138, 1999.
22. R. Schäler. New media localisation - a linglink report for the European Commission DGXIII. Technical report, Luxembourg.
23. H. Somers. Review Article: Example-based Machine Translation. *Machine Translation*, 14(2):113–157, 1999.
24. H. Somers. The Current State of Machine Translation. *MT-Summit VI*, pp.115–124, San Diego, CA, 1993.
25. E. Sumita. An example-based machine translation system using DP-matching between word sequences. In *[8]*.
26. N. Underwood and B. Jongejan. Translatability checker: A tool to help decide whether to use MT. In *Proceedings of MT-Summit VIII*, pp.363–368, 1994.
27. P. Van der Eijk, M. d. Koning, and G. v. d. Steen. Controlled language correction and translation. In *CLAW 96*, pp.64–73, Leuven, Belgium, 1996.
28. W. Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer, Heidelberg, 2000.
29. H. Watanabe. Finding translation patterns from dependency structures. In *[8]*.
30. A. Way. LFG-DOT: A hybrid archirecture for robust MT. PhD thesis, University of Essex, Colchester, UK, 2001.