Blueprint for MT Evolution
Reflections on "Elements of Style"

Jörg Schütz

IAI
Martin-Luther-Str. 14
D-66111 Saarbrücken
Germany
joerg@iai.uni-sb.de

Abstract

In this paper, organized in essay style, I first assess the situation of Machine Translation, which is characterized, on the one hand, by unsatisfied user expectations, and, on the other hand, by an ever increasing need for translation technology to fulfil the promises of the global knowledge society, which is promoted by almost all governments and industries worldwide. The assessment is followed by an outline of the design of a blueprint that describes possible steps of an MT evolution regarding short term, mid term and long term developments. Although some user communities might aim at an MT revolution, the evolutionary implementation of the different aspects of the blueprint fit seamless with the foundation that we are faced with in the assessment part. With the blueprint the thesis of this MT evolution essay is established, and the stage is opened for the antithesis in which I develop the points for an MT revolution. Finally, in the synthesis part I develop a combined view which then completes the discussion and the establishment of a blueprint for MT evolution.

Introduction

At various occasions, I presented the architecture of Star Trek's Universal Translator (UT) which is a device that is characterized by the numerous and different knowledge resources employed in the analysis of an unknown language for the purpose of an instant (text or speech) translation. The main feature of the UT is the generation of a translation matrix based on large samples of language usage patterns, vocabularies, syntactic, semantic, and cultural information, and other related information elements. The UT design is based on the assumption that a language (human or alien) is a code that represents a certain message or information element which the UT has to decipher.

This doesn't make sense because we all know that languages are not codes. Communications without reference prints do not necessarily illuminate what they are about, and a language cannot be learned by assembling a huge library of communications and conversations, and then analyzing them by a computer. Imagine, for example, what we can infer from observing a conversation between two people in a language of which we do not have any command.

Nevertheless, the lesson we can learn form the UT design is the variety of knowledge sources and their different facets that are employed for the translation task. Most of these resources or combinations of these resources are not or only to a limited extent employed in current MT system design. Moreover, the design of the commercially available systems have not been further evolved since their inception, and research systems have not yet found their way to industrial deployment.

In this paper, presented in essay style, I assess the current situation and I develop a blueprint for the evolution of MT within a short term, a mid term and a long term time frame. This blueprint is entirely purpose oriented. The purpose orientation is essential because it takes into account the needs and demands of different user communities and their intrinsic translation purposes. In contrast to the evolutionary MT development, the revolutionary track of the total abandoning or substitution of MT by multilingual generation is introduced. The developed blueprint then is based on current and emerging market needs and trends of the emerging global society, and research aspects are taken into account only to a limited extent although they are back stage in any blueprint, and thus acting as an innovation engine.

Current MT Situation

MT Focus: Rendering Formats

Over the last decade, which is mainly characterized by the evolution of the internet in the incarnation of what is termed the World Wide Web or WWW and Web for short, MT has received more and more recognition. Besides the need to organize and manage the huge Web information pool for effective and efficient access and assimilation, there is an ever increasing need for the translation and the access of this information into and from different and multiple human languages. Translation is also needed to produce and present information in multiple languages. In its first days the today well-known concept of the URI (Uniform Resource Identifier) was originally called Universal Resource Identifier, and just has the URI is concerned with the structure and location of a certain information element, the Universal Translator shall be concerned with the content of this information element. The separation between structure and content which is very well maintained by markup languages such as TeX, SGML and the youngest child of this family, XML, which today receives a lot of marketing hype, has been watered by the desktop publishing concept of WYSIWYG with a very strong focus on rendering formats only. This is why some of the markup evangelists have coined the acronym WYSIAYG (= What You See Is All You Get) to account for this content-ignoring focus. This conceptual shift had also a tremendous influence on the translation technology (MT as well as Translation Memory) developments: instead of focusing or concentrating on content related aspects of the translation

task, and incorporating new insights from research into the systems, the main development line was dedicated to implementing filter software for the various desktop publishing systems. This has certainl also widened the chasm between commercial systems and research systems.

MT Focus: Web Accessibility

Yet another orientation of MT, which again is geared by the Web, is the accessibility of the systems by Web browsers. This development mainly concerns the technical field on how to integrate an existing MT system architecture into the infrastructure of Web servers, and the possibility to communicate with their protocols and of course their formats. In this case, the format - HTML - is quite simple and based on an SGML DTD. These developments receive also great marketing hype, and most of the translation technology vendors have already started with the design of an MT ASP (Application Service Provider) model to gain additional revenues within this market segment. Although, the supported format is not proprietary as in the other cases that I discussed in the previous section, the separation between structure and content is very limited because the employed HTML tags are semantic-poor regarding their relationship to the actual content. The technical developments for the integration with Web infrastructure allow for more insights to the integration possibility of translation technology into different IT infrastructures which has not been followed closely by MT system vendors in the past. This is also documented by the small number of officially published APIs.

MT Focus: Resource Interchange

The fully-fledged deployment of an MT system in an industrial environment is very much dependent on how easy the feeding of a new or additional vocabulary can be accomplished. Most MT systems provide a facility to manually import new words and terms based on a proprietary format of the underlying system. It would be more appropriate to have a commonly agreed-upon exchange format to permit the interchange of resources between different MT systems. The interchange of resources also contributes to an enhanced throughput and robustness of a system, and thus has a certain influence on the overall quality of the generated translation output.

MT-Focus: Summary

In summary, over the last decade the focus of any commercial MT system development has been primarily on:

- supporting many proprietary DTP formats and HTML
- limiting system integration to Web deployment and ASP models (no real APIs)
- importing user-specific vocabulary is mainly based on proprietary formats
- disregarding an incorporation of research results such as more content-orientation within SGML/XML environments, ontology-based exchange formats, and others.

In the following sections, I will take especially these aspects into account for my design of an MT future blueprint.

MT Future Blueprint - Evolution

User-centric and purpose-oriented MT

The technical information available in different modern business information systems, including huge amounts of legacy data which also resides in translation memory (TM) systems, is more and more seen as the knowledge capital of an industrial organization, which has to be turned into a digital information and knowledge repository to fully leverage its added-value in different business processes. During the last decade, the development focus was on more appropriate information technology to facilitate the building of such repositories. Since these developments where mostly restricted to the structural level of the information, the envisaged cost reductions did not roll up as expected. Now, this fact gives raise to the emerging field of content management, and more recently termed knowledge management, which couples the structural extension of information with its associated semantic extension. This reorientation takes also place within the Web community with the advent of the so-called Semantic Web or Sweb for short. In the following sections, I discuss four technology cases that serve as main future MT enablers:

1. Controlled languages
2. SGML/XML as input format
3. Ontology-based vocabularies for MT lexicon import in an XML exchange format
4. Combining CLs and ontologies and enable the webizing of deployed systems and resources

Case 1: Controlled Language enabled MT

The question of how to cover the semantic dimension of a given information entity within a certain application scenario fostered the investigation into what is known as Controlled Language (CL) application. A CL application is concerned with the delimiting of the semantics of a human language expression in a well-defined context, such as service and repair. The concept certainly has its roots in the field of Fachsprachen (domain-specific languages), and today a CL is often seen as the long sought killer application of Language Technology (LT) and in particular of MT technology. Until today, CLs have received broad attention only in the air and space industry, where mostly English is the language of communication in different areas: maintenance literature, air control communication and joint international research and development. The CL is defined with a set of rules the language of the technical information has to comply with and the vocabulary that has to be used. Recently, organizations in other industrial sectors have started to evaluate the use of CLs; in particular in combination with the advent of new business information systems of the third generation, which focus on so-called business intelligence combined with Internet technology, and the demand to fully deploy MT as a cost reducing technology. Here, CLs play the role of an enabling technology that allows for better information production and information extraction in terms of comprehension and accuracy. Until now, there is only a limited technology breakthrough in this area. However, this is changing since today's industries are also confronted with a much wider customer spectrum regarding languages and cultures for which they

have to offer and to maintain technical and non-technical information in ever-shorter time frames to be successful in business-to-business, business-to-customer and business-to- consumer operations. Now, CLs and less strict more flexible approaches to language control are seen as the means to empower the full spectrum of an enabling technology for specifying and implementing the processes of information acquisition, production, translation, dissemination and assimilation more efficiently and cost effectively, particularly when coupled with smart automation technology of the information industry sector and the language industry sector.

To cover both incarnations of CLs, we created the term Controlled Language Authoring (CLA). In our definition, the CLA concept also covers or subsumes any quality assurance mechanism employed in the entire life cycle of technical information, and thus it includes the checking of the target language of a translation process (human and machine). This makes it fit perfectly with the actual needs and demands for MT deployment of different industrial users because the CL rules might also reflect MT-specific aspects.

Today, the rather small CL technology market is mainly focussing on the syntactic layer of language which, nevertheless, is no easy implementation task since the identified syntactic criteria have to be efficiently translated into computationally tractable components. Available CL products, however, also lack modularity and interoperability which makes the integration into existing production workflows a very hard, sometimes impossible task.

A new market for CL technology is emerging with the increasing interest in CL enabled tools and utilities. These tools and utilities are essentially based on the different needs and demands of the manufacturing industry for the production and processing of technical information. The automotive industry and the software industry in particular are looking for the deployment of semantics-based CL technology and translation technology to handle their information processing requirements because the employment of translation memory technology has reached a threshold where no direct further improvements seem possible. Besides the improvement of the source language quality and the translatability of technical information, their interest is also to maintain a direct link to design within virtual design, construction and modeling environments.

Different LT vendors are now investing into this emerging new market segment, which as an enabling technology is situated in two already existing market segments: content technology and translation technology. For these market segments a growth rate of approximately 38-45% is envisaged over the next three years until 2004. Within the overall LT market, the content technology segment and the translation technology segment hold a 56% share (31% content and 25% translation), the rest is concerned with various speech applications. The average growth of the content technology market is forecasted with 43%, and the translation technology market with 41%. Therefore, CLs are more and more in the focus of industry because they hold great promise as a method to support the complete information cycle including acquisition, creation and maintenance of technical information in several dimensions:

- terminological standardization
- accuracy of technical descriptions
- readability and comprehensibility of the information elements
- reusability and maintainability of the information elements
- cost-effective, controllable and benchmarkable translation (human and machine)
- reduced lead times and faster time-to-market of the human language product in multiple human languages.

Besides cost and time savings, this method also allows the representation of both knowledge and structural relations between the elements of this knowledge, which then permits better control and interchange of such information across the boundaries of business partners, knowledge domains, computer notations and multiple human languages. Last but not least, the employment of a CL is an enabling factor for MT, and thus a CL is perfectly combinable with MT as follows:

- CLA is the technology to check CL compliance and to compute a translatability index (a measure that supports the decision about a possible translation process) that triggers MT deployment
- CL compliant information elements provide the basis for successful MT

Case 2: XML as MT input format

SGML and its simplified version XML provide a well-suited input format specification for MT because of the unambiguously definable semantics of the employed markup vocabulary (SGML/XML tags). In addition, this vocabulary can also trigger some of the linguistic processes of the MT engine. I will further detail this in a presentation of the CL Authoring technology, which we have developed and which is deployed at several companies ranging from the auto industry to the printing machines industry and the software industry.

Case 3: Vocabulary interchange and ontologies for MT

To optimally deploy an MT system, it must be possible to import a terminology collection from another system such as a term database. The technology that allows for this kind of interchange relies upon the availability of a set of common vocabularies organized within an XML-based framework. Currently, work on such vocabularies is ongoing in the European SALT project (http://www.loria.fr/projets/SALT) and in the international industry-led initiative OLIF (http://www.olif.net/), which both are founded on the data categories for lexical and terminological data descriptions of the international standard ISO 12620. The OLIF approach also permits the exchange of translation related information in the form of lexical transfer rules.

A further step for a uniform resource exchange between different language and translation technologies is the employment of ontological information. An ontology is an explicit specification of some subject field. In the context of most projects, it is a formal and declarative representation which includes the vocabulary (or names) for referring to the concepts in the chosen subject area. It also includes logical statements that describe what the concepts are, and how they are related to each other. As such an ontology provides the vocabulary for representing

and communicating knowledge about a specific subject field in a structured way. Additionally, it specifies a set of relationships that hold among the concepts in that vocabulary. Ontologies have received broad attention within the artificial intelligence community, in particular in the field of knowledge representation. Today, industry is more and more focussing on ontologies for facilitating corporate knowledge management, which is a term that receives a lot of marketing hype. For consulting organizations, knowledge management is essentially about new business management techniques which are designed to address the fact that people and the expertise they possess are the primary assets in an increasingly knowledge-based economy. Although some vendors equate knowledge management with information management, knowledge is fundamentally different from information: the difference is that between knowing a thing versus simply having information about a thing. Ruggles (1997) claims that *"knowledge management covers three main knowledge activities: generation, codification, and transfer*". This means that an ontology can be regarded as a system for codification which is the necessary prerequisite for the development of tools and utilities that assist in the generation and the transfer of knowledge.

Case4: Combining CLs and ontologies

An ontology is a highly structured repository of knowledge that makes explicit the attributes and properties of individual concepts as well as the different relationships that exist between concepts. It defines the ways in which concepts are related, their relative significance, and their dependencies. The most significant relationship between concepts in the ontology is that of hyponymy/hyperonymy which determines if a concept belongs to the class defined by another concept. These pieces of knowledge are not simply listed but they are systematically linked. The concepts are organized in a hierarchy according to the main organization principle of specialization of more general concepts down to more specific concepts. The lowest conceptual level provides anchoring points where concrete terms are represented as instances of these concepts. They can be thought of as leaves which form the ending points of branches of an acyclic directed graph. The classification principle of specialization allows for the inheritance of characteristics of concepts defined on upper levels down to the lower levels. These instances profile the vocabulary of the controlled language deployed in a certain application scenario.

This vocabulary and the associated applications have to be used to bootstrap the Web to yet another deployment dimension by making use of these already existing different technologies. The process facilitating the step from a closed world application to an open world application, i.e. an open Web application, is known as "webizing". It is mainly concerned with a transformation of the closed world entity descriptions into a notation that incorporates URIs, namespaces, and so forth, and which can be accomplished by RDF descriptions and other similar formalisms such as Topic Maps, particularly the XML incarnation XTM.

The ontology view shall be contrasted with term databanks and thesauri which also offer resources for MT deployment. Conventional term databanks and specialized dictionaries do also have a sort of knowledge component which includes the definitions of terms, contextual examples, grammatical information, and others. However, term databanks do generally not provide structured knowledge about a given subject field. Additionally, the traditional media (databases, paper material) do not serve as an adequate means to support the high complexity of the knowledge structures. The usual way of representing knowledge stored in term databases is by linear alphabetical ordering of terms with links to synonyms, grammatical information, context examples, and others. A thesaurus is a collection of technical terms of a (specific technical) domain[1]. The relations between the terms are made explicit based, for example, on semantic characteristics.

A thesaurus is particularly appropriate for fields which have a hierarchical structure, such as names of objects, subjects, places, materials and disciplines. Technical terms are represented uniquely, which means that a concept is always represented by one (and not more) terms. This is achieved by defining so-called descriptors which are the allowed terms for a concept. These descriptors are linked with unapproved terms for the same concept. Non-allowed terms are mostly synonyms, for example, *adaptor lead* versus *adaptor cabl*e, or orthographic variants of a term, for example, *micro-switch* versus *microswitc*h. For indexers and searchers, a thesaurus is an information storage and retrieval tool: a listing of words and phrases authorized for use in an indexing system, together with relationships, variants and synonyms, and aids to navigation through the thesaurus.

A webization of these resources is accomplished in a similar way as described above, and would be particularly fostered if available in formats such as those developed within SALT and OLIF.

Evolution Blueprint

The cases introduced in the previous sections form the basis of an MT evolution scenario for the next two to five years with the following steps:

1. Development of several MT related style rules together with domain-specific style rules to account for different incarnations of CLs
2. Agreed-upon XML-based resource exchange formats for MT lexicons and terminologies (standardization)
3. Development of ontologies for the different styles and the MT resources to further support 2. and 1.
4. Agreed-upon XML-based description formats for information processing and workflows which are coupled with 2, and the webizing of the descriptions and associated applications.

The latter shall allow for the efficient and effective transfer of data between applications and the configuration of certain processes. The idea is to have a configurable network that includes resources and processing components, and to access them with standard

[1] An example for a non-technical, general language thesaurus is Roget's Thesaurus (Roget 1852) which attempts to structure the words of the English language according to semantic properties. It contains 1073 categories which are classified into fifteen broad classes.

formats and protocols. This then is also the basis for the communication between different processes. The benefits of this approach are that they allow to create messages and processes/programs that can be used by a "recipient" who has little or no prior knowledge about how they are/were created. This is different, subtle but important, from the way we do things today. In addition, a software as well as a lingware component "learns" automatically new procedures as they are required. This allows designers to decide which components to use based on technical aspects instead of market share/pressure. Webizing this approach then makes the deployment scenario an open world application scenario.

MT Future Blueprint: Revolution

Multilingual Generation

Yet another approach, which certainly can also be founded on the described technology cases, is to abandon MT entirely, and to deploy multilingual generation based on a formal language representation of a given information element (see above the link to design, construction and modeling). In principle, this is similar to the theoretical foundation of the international UNL project. UNL is a TLA for Universal Networking Languages which aims at breaking the language barriers on the Web.

The UNL representation model is anchored on the concept of Universal Words (UWs), a set of relations that can hold between UWs in several combinations, and additional attributes that further define or constrain the UWs and the relations. UWs are accessible through a central knowledge base. This UNL KB is dynamic in the sense that it evolves as information is added. The semantic links that build structures out of concepts are signaled in human language texts by different grammatical means such as word order, agreement, suffixes, and so forth, for different languages. The links can also be interrelated in complex ways to represent very complex relations between concepts or groups of concepts, for example, coordinated structures. Therefore, representations across sentence boundaries are also possible but they are not yet on the UNL development agenda.

Ideally the UWs should be organized in an ontology to account for different reading distinctions and granularities on the conceptual level that could be expressed by certain terms employed in the human language generation process. However, the UNL project does not maintain such an ontology yet, although it would foster the real breakthrough of the approach.

The opposite process is the translation from a human language into the UNL presentation, and as such UNL resembles very much some of the concepts of knowledge-based MT, which unfortunately has not found its way to any commercial product.

Obviously, the UNL formalism is webizable which allows a similar open world deployment scenario has described in my MT evolution section.

Revolution Blueprint

Within the next five to ten years, we will certainly have more developments in this direction because the concept of an ontology is more and more in the focus of industrial consideration. This particularly combined with the emerging developments of the XML community for the description of resources, processes, and exchange formats, as well as with the genesis of the Internet 2 which is a further evolutionray step of the ongoing Sweb initiatives.

Synthesis "Elements of Style"

Both blueprint paradigms, which I have briefly introduced, will coexist, and since both are founded on similar prerequisites they will benefit from each other. It is my opinion that mainly XML related aspects for data formats and process control, CL style rules for quality and performance, and ontologies for vocabulary maintenance and exchange will be among the gearing forces of future MT system development, until we have found the human language gene, which then opens of course other challenging opportunities and new fields in bioinformatics and biotechnology.

In summary, this blueprint demonstrates that the UT of Star Trek is possible with its originally defined knowledge resources but with the differing assumption that a human language is determined by different elements of style that allow for an efficient MT deployment and effectively enable MT as a thorough technology.

In addition, aspects of webizing an MT application offers yet another challenging approach to full-fledged MT deployment.

References

Gillies, James and Robert Caillian (2000). How the Web was Born. Oxford University Press, New York and London.

Gresh, Lois and Robert Weinberg, 1999. The Computers of Star Trek. Basic Books, New York.

Krauss, Laurence M., 1995. The Physics of Star Trek. Basic Books, New York.

Ruggles, R. (1997). Knowledge Management Tools. Butterworth-Heinemann, Boston, 1997.

Schütz, Jörg (1996a). Network-based Machine Translation Services. In: Proceedings of EAMT Workshop 1996, Vienna, Austria, pp. 145-157.

Schütz, Jörg (1996b). Combining Language Technology and Web Technology to Streamline an Automotive Hotline Support Service. In: Proceedings of AMTA96, Montreal, Canada.

Schütz, Jörg (1997). Utilizing Evaluation in Networked Machine Translation. In: Proceedings of TMI 97, Santa Fe, New Mexico, USA, pp. 208-215.

Schütz, Jörg (2000). *One Web, One Language: The Universal Networking Languag*e. In: Intelligence in Industry, Issue 2/2000, Unicom, Uxbridge, Middlesex, UK.

Schütz, Jörg (2001). Ontologies in Terminology Work - Enabling Controlled Authoring. In: Proceeding of TAMA 2001, Antwerp, Belgium.

Star Trek The New Generation Technical Manual, 1991. Pocket Books, New York.