Network-based Machine Translation Services

Dr. Jörg Schütz

Abstract

The emerging network technologies such as Intranets and the Internet, particularly the World Wide Web (WWW), will create a new model of computing which I will call *Network Computing (NC);* others, e.g. Forrester Research Inc. call it Internet Computing with the obvious stress on the Internet. Recent analyses of this emerging market conclude that Network Computing:

- Extends and improves client/server technology.
- Will be used by companies to build new customer connections.
- Will change the structure of the industry and the role of information technology as well as the role of language technology.

The focus of my presentation will be multilingual language technology and includes the following aspects:

- What is Network Computing?
- How Network Computing will emerge.
- How Network Computing will change the industry.
- What this means to vendors.
- Impact on users.

I will limit these aspects primarily to network-based translation services, which comprise translation on demand, multilingual communication, multilingual information search and retrieval, as well as customisable translation engines.

For details on Dr. Jörg Schütz and the IAI, please refer to Session 5 cover page.

0. Introduction

In order to thrive in the next millennium industries will have to overcome the confusing revolutions introduced by business and technological areas. On the one hand, the business revolution is characterized by increased competitiveness, the pressure to reduce costs, increased productivity and an increase of responsiveness; on the other hand, the technological revolution concerns the step-wise deployment of distributed systems, PCs, client/server technology, graphical user interfaces (GUI) and open systems. Currently, significant reengineering efforts are underway to taming the complexity of both business processes and computer information systems. The information systems that are needed for tomorrow's global markets must be far more robust, intelligent and user-centered than the data processing systems of today. With the increasing global competition of trade and industry, multilinguality is an additional asset of future information systems, particularly in combination with the ubiquitous information networks (or information highways) that form the digital foundation of the information society (personally I prefer to talk about the knowledge society, cf. below). Just as quantum physics and the relativity theory have changed the thinking of how to describe nature (particularly the recent works of Hawking and Penrose summarized in [Hawking & Penrose, 1996]), the Internet and its multimedial extension, the World Wide Web (WWW or Web), have changed the way of thinking about human communication and the world's globalisation in general (and of course we have the same problems of what this is all about).

Today, we are still faced with the Babel threat, especially when we are surfing on the Internet. Therefore the myth of Star Trek's universal translator is alive more than ever, especially because past experiences, in particular the various evaluation and validation projects, have shown that current machine translation (MT) approaches are not flexible enough to solve the diversity of the multilinguality problems of today's global business and communication situations. The time is ripe for the introduction of a new MT paradigm that is able to combine existing and emerging information technologies and language technologies in an integrated way, thus MT can be seen as a kind of specialized information system (as described above) that is able to contribute to overcoming the existing language and cultural barriers of the world-wide network community. For this, however, a better understanding of the concept of network-based MT is needed. This shall be the first dimension of our investigation.

The combination of Web, information and language technologies will also create a new model of computing, which will revolutionize the existing hardware and software markets. I will call this new model of computing *Networked Computing (NC)*[1]. Recent analyses of these emerging markets conclude that Networked Computing:

- Extends and improves current client/server technology.
- Will be used by companies to build new customer connections.
- Will change the structure of the industry and the role of information technology as well as the role of language technology.

[1] Although I have used *NC* as the abbreviation for *Networked Computing*, there is not an intended relationship with the Network Computer (also abbreviated NC) which has been specified by Sun Microsystems, Oracle and others as a low cost Internet PC.

In this model MT will obviously have its place, but its exact locus has yet to be defined. This constitutes the second dimension of our investigation.

When talking about a new MT paradigm founded on multilingual network-based communication capabilities we are not aiming at monolithic local machine translation systems, as they have been on the market for several decades, but at systems that deploy the potentials of the international networks in terms of language resources and software capabilities. Examples of the former type of systems are Logos, Metal and Systran (the real dinosaurs of MT in terms of the employed language and information technology), the huge amount of available PC-based translation systems such as Globalink´s Professional Translator, IBM's Personal Translator and Langenscheidt´s T1, and the new class of translation support systems known as example-based translation (EBT) or as translation memories (TM) and offered for example by Trados, IBM and others.

However, currently the network scenario with embedded machine translation functionality (MT plug-ins) is the vision which might turn MT (hitherto I will use this term to subsume fully automatic MT as well as machine aided translation and translation memories) to challenging new application domains such as interlingua-based email transmission, the translation of Web pages on demand, interactive machine translation and speech translation in multiparty teleconferencing situations.

Today, existing MT applications are often faced with serious problems when they are employed in a network environment. This can be exemplified by the following description which I received recently by email:

> *In short, the most efficient way to use* [... the TM system ...] *is by sharing a translation memory (TM) between several translators on a network. That way translators can benefit from repetition across manuals and they can also control consistency in both style and terminology. However, if your network software is either not state of the art or has too much traffic on it, then the response time for the Workbench is seriously reduced and any increase in throughput figures, thanks to leveraging from the TM, is canceled out. You're faced with even more trouble if your network crashes regularly (something which happens quite a lot in large companies!). One solution is to have a separate network server just for translation memories, but that's quite costly.*

In this example, the bottleneck of the application can be manifold ranging from the MT system proper, especially its network capabilities (e.g. concurrent requests), to the surrounding technical infrastructure, i.e. the local area network (LAN) in which it is operating. The MT infrastructure aspects then define the third dimension of our investigation.

To summarize, the specifications of a new MT paradigm which is capable of being deployed as a network-based MT service (local and wide area) have to comply with three dimensions:

1. Concept of network-based MT.
2. Locus of the MT engine and the MT resources (local vs. distributed).
3. Technical infrastructure of MT (hardware and software aspects including costs).

The first dimension has to take into account the application domain of an envisaged translation task, i.e. the purpose of the MT employment; this is also a serious aspect for the evaluation of MT systems. The second dimension is crucial for planning the future of an application and relies heavily on the existing and planned workflow cycles, e.g. the extensibility of the service to other tasks and domains or the treatment of multilingual documents within a company where the recent trends are characterized by systems that allow content-based access to documents (document databases). The third dimension interplays with the second dimension because it is concerned with the overall technical infrastructure in which the MT system is embedded.

The remainder of this presentation is organized as follows: in the first section we discuss Networked Computing in general, we will contrast it with current client/server technology and we will develop a calendar for its evolution within the next three years; in the second section we will extend the Networked Computing scenario with the concept of intelligent software agents which finally results in the definition of an intelligent translation assistant. The third section reports on how we can benefit from these developments yet to come into operation in existing machine translation applications such as outlined in the above example. The paper closes with some prospects of ongoing research and development activities.

1. Networked Computing and Machine Translation

1.1 Networked Computing: Listening vs. Conversation

Networked Computing (NC) can be defined as:

> *"Remote servers and clients cooperate over Intranets and/or the Internet to fulfill a certain task."*

Intranets are corporate networks that are based on the Internet protocol and Web technology; this means the look and feel of an Intranet application is similar to a Web application embedded in a browser such as Netscape, Mosaic, etc. In a NC application the user on the client end will connect to a Web site, but instead of asking for a file (the usual way today) the user will request a session and will receive client code. Once this code is loaded, the client computer and the (network) service server will cooperate, exchange data and communicate. While standard Web applications are based on listening, NC applications are based on conversation.

On the one hand, NC will extend the classical client/server applications which are local, prearranged and limited to a set group of users, because it will have global reach and a potentially massive scale, and on the other hand, it will also enhance Web applications, which handle static documents, by allowing clients and servers to carry on rich discussions based on Web technology, they will use LANs, the Internet, Web browsers and Web-enabled servers.

NC has the advantage that it is:

- **global,** this is facilitated by the standardized communication protocols;
- **frictionless** because maintenance and customization will be drastically reduced by the possibility to download and install client code from remote servers on demand;

- **modular** because NC will rely on software objects to minimize bandwidth and allow for flexible application design; and
- **scaleable,** this again is facilitated by the Internet technology.

The time seems to be ripe for NC because of the existing global standards such as the Internet protocols, browsers, HTML, etc., and the tens of millions of people that are on the Internet. However, NC certainly will be introduced step-by-step:

1996: experimentation; no applications with real customers; building of the basic technology and competences (Java, JavaScript, Tcl/Tk, Virtual Reality, etc.).

1997: deployment of NC applications with focus on Intranets (corporate users).

1998-1999: deployment of NC applications with focus on the Internet.

Obviously, NC will change industry:

- New systems will emerge that are not entirely under the control of current hardware and software leaders, e.g. systems based on VRML (Virtual Reality Modeling Language).
- New and different pricing models will appear, e.g. freeware.
- No software development cycles in the traditional sense will exist.
- New distribution channels will exist, e.g. downloading from the Internet, interactive system upgrades, etc.
- The industrial market will be driven by split-offs, venture-funded start-ups and new private companies, i.e. by a creative divergence.

For software vendors the impact of NC depends on their move towards NC applications which demands new user interfaces, especially multimodal and multimedia based user interfaces. Software will be customized according to predefined application scenarios, updates of software vendors will be done overnight. And last but not least new database players will emerge, in particular those who are concerned with object-oriented databases and those who optimize their products for Java and the emerging Scripting Languages such as Java-Script, Visual Basic Script, Lotus Script, etc.

Users will have the opportunity to run traditional client/server applications in parallel to NC applications. This certainly will demand new skills, both on vendor and on customer sites. Like in every application where time plays a critical role (cf. the example in the introduction section) speed will be a crucial factor. Therefore there is a need for organizational transition, which then also permits new mechanisms for client support.

1.2 NC environments for MT

Today, MT systems with acceptable translation results are still very costly and mostly available for high-end workstation platforms. The existing PC-based systems do not accomplish comparable translation results and they are only available for a limited number of language pairs; we particularly start getting into trouble when our application focuses on Asian and Arabic languages. However, similar problems exist

when we look for systems dealing with French as source language (one of the universal languages). This situation is clearly demonstrated on the Internet: multilinguality does simply not exist; the English language is the cyber-lingua of today's information highways.

Obviously, the idea for an MT service available as an Intranet/Internet browser plug-in is fascinating and is becoming more and more feasible with the emerging NC technology. Some MT vendors have already started to offer MT services via the Internet, e.g. the off-line translation of HTML documents (Systran and MTSU) or form/email-based translation services (e.g. CAT2). However, online translation is only provided by the Web site of Rivendell International Communications on a word-to-word basis (this Web site is a very useful resource of language related data; for this Rivendell maintains world-wide links to different resources and services). NC will not only provide online translation services but will also be a basis for other interactive services. Today, the Web is only mildly interactive with just hyperlinks to take the user from location to location.

This situation has changed a bit with the introduction of forms, for example, users can specify search queries or parameters and texts for off-line services. A form consists of two parts: the form itself, which is rendered in the browser, and a Common Gateway Interface (CGI) script or program located on the server. This script processes the user's input mainly to validate the correctness of the user data (e.g. online subscription). This form mechanism is also employed by most of the above mentioned Internet translation services. However, the server executes only one program (CGI script). Since the interaction is immediately direct over the network, this approach causes a higher load of the network. Incorrect and incomplete input can only be perceived on the server side, and the possibility to implement appropriate user interfaces is very limited. This situation led to the idea to execute some tasks on the client side similar to client/server programming. For this the Web browser must be able to run a program which has been implemented by the service provider and which is accessible via the offered HTML page.

Thus, the next step to interactivity is the use of so-called *browser plug-ins*. The employment of plug-ins also reduces the network traffic, which today still is a major problem. However, with the introduction of high-speed networks based on the Asynchronous Transfer Mode (ATM) and mobile communication capabilities via satellites, this will no longer be a real bottleneck, especially because the costs are going down in this technical area. Plug-ins can be installed by running a setup program supplied with the plug-in. Plug-ins reside on the local hard drive and are detected by the browser when it starts up. When the browser encounters data handled by a plug-in (either embedded in an HTML page or in a separate file), it loads the appropriate plug-in and gives access to all or a part of a window. The plug-in remains active until the associated page or file is closed. Currently the programming language Java is mostly used for this application, because it is platform independent. A Java program is compiled into Java bytecode, which can be embedded in a HTML page. A Java enabled browser on the client side loads this bytecode over the network and executes the code. This sort of program is called *Applet*. Other programming languages or scripting languages used for plug-in development are for example SafeTcl/SafeTk, Phantom and TeleScript.

Such a plug-in program can also use additional resources needed by the application and located somewhere on the Intranet/Internet. Thus, plug-ins are the ideal mechanism for handling the trade-off between local and distributed resources and the overall network infrastructure. Today's plug-ins can be seen as the first operational examples of NC.

2. Intelligent Software Agents

2.1 Current and future situation of network environments

As outlined in the previous section, the interactivity of current network-based applications and services is limited, i.e. mildly interactive. Current Intranet/Internet applications provide static Web documents, and the user can use, on the one hand, presentation services provided by Web browsers, and on the other hand, search services as maintained by Lycos, Yahoo, Inktomi, etc. This situation is presented in Figure 1.

With the evolution of NC a new type of software is emerging which will operate on behalf of a user or another program. Thus, it is called intelligent software agent. The Intranet/Internet scenario of tomorrow then might look like it is shown in Figure 2. In this scenario the intelligent software agent acts as the personal navigator of a user. Therefore, we talk about either a personal net assistant or a softbot (software robot).

This softbot cooperates with other software agents for the fulfilment of a specific task by the technical means of NC. The cooperating agents constitute the intelligent network infrastructure, which is the knowledge layer above the information entities, e.g. multimedial Web documents, of an Intranet and the Internet.

A softbot of a user is able to execute different tasks such as

- Information filtering according to user defined parameters, e.g. technical watch in terms of selective dissemination of information and competitive intelligence of large companies.
- Information condensing, e.g. for the storage in digital libraries and in multimedial databases.
- Information brokering, e.g. for online services and emergency applications.
- Telematics assistance, e.g. in telecooperation situations with working cooperations across time zones including the synchronous exchange of working materials and the asynchronous development.
- Translation services, e.g. machine translation and online speech translation for teleconferencing.
- Data analysis, e.g. automatic validation and evaluation of service results.
- Knowledge acquisition, e.g. distance service operations, and distance learning and training.

In order to do this the softbot must have information about the available resources on the Intranet/Internet, either on the basis of local knowledge or through the communication and co-operation with other software agents located on the intelligent infrastructure layer, and information about his master (user), for example, predefined and learned usage patterns, a user model or a specific consumer/customer profile.

Such an intelligent infrastructure service scenario is shown in Figure 3. In this scenario the personal softbot gets an order from the user for whom it is navigating, and according to the user's order and its knowledge about the user, the softbot negotiates with the software agent of a selected supplier who offers the appropriate information, goods or services the user is looking for.

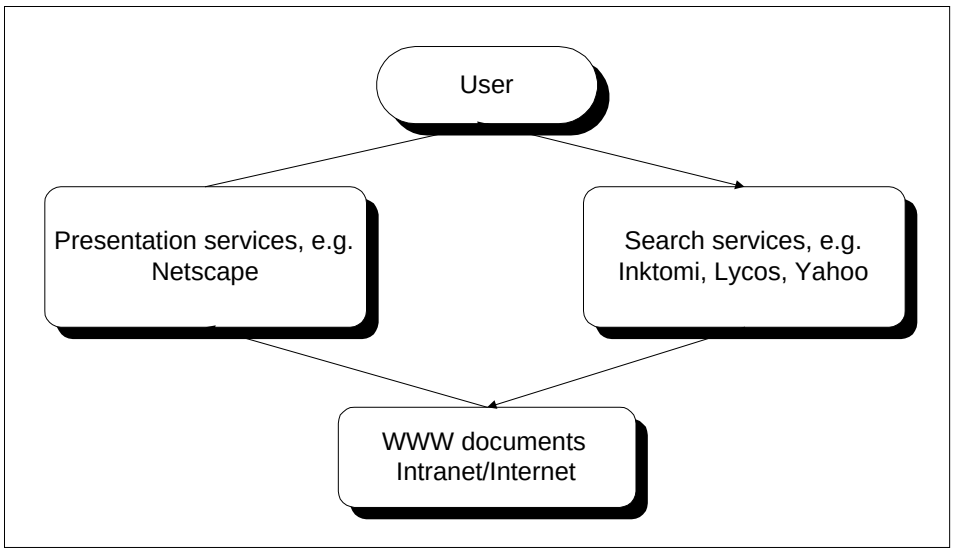

Figure 1: Intranet/Internet situation today

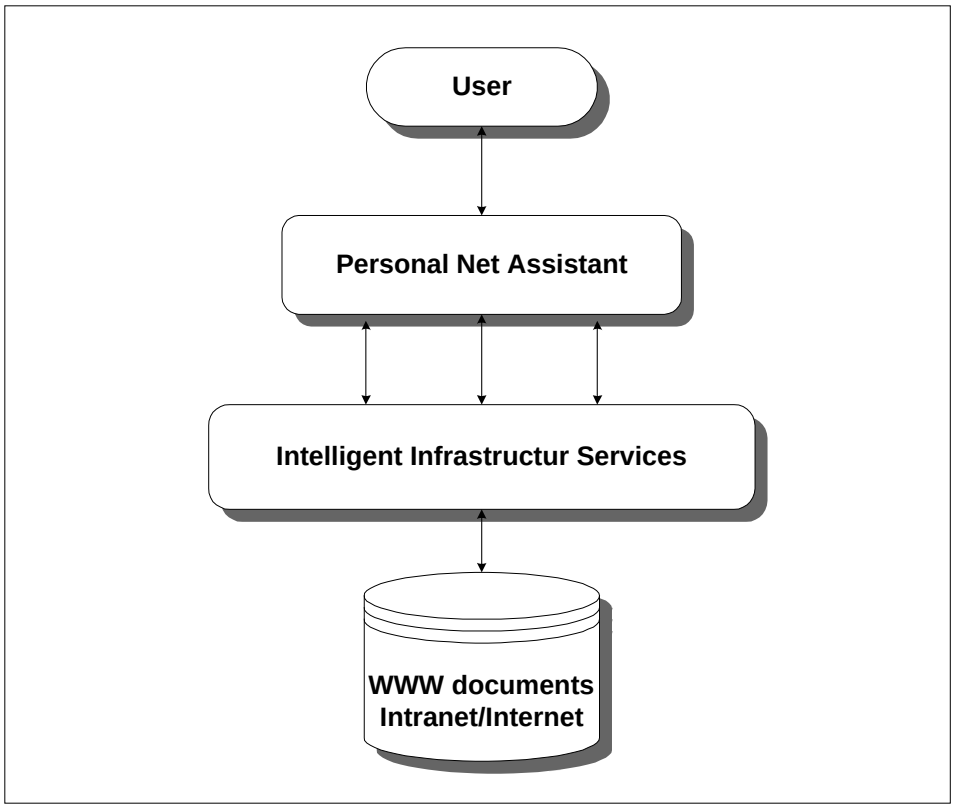

Figure 2: Intranet/Internet situation tomorrow

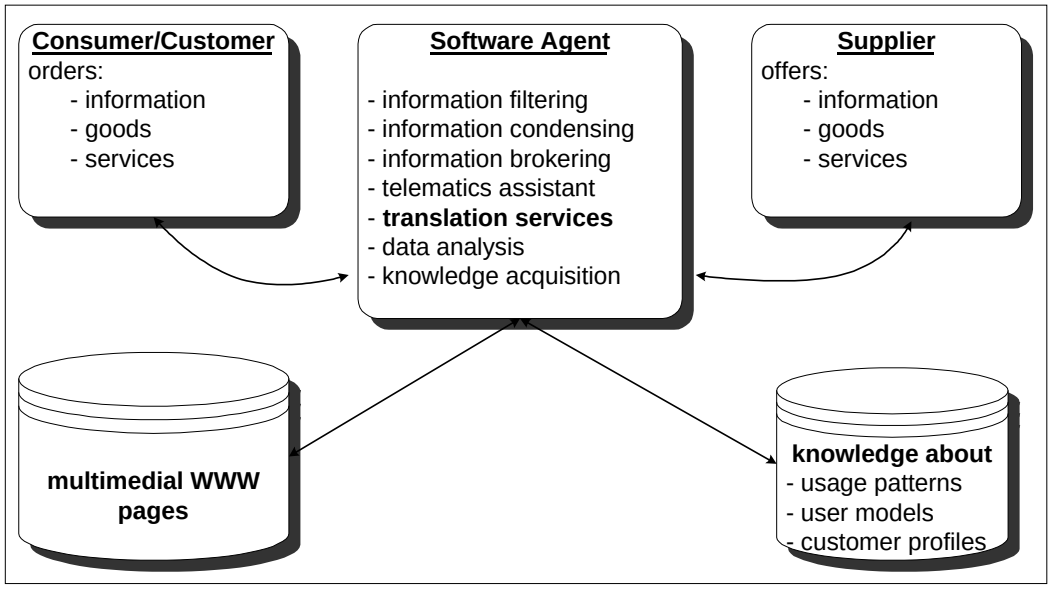

Figure 3: Intelligent infrastructure services

2.2 Intelligent translation assistants

Up to now we have described the technical basis of current and future network-based services, and we have demonstrated how the traditional MT system approaches could be an integrated service within this scenario. The feasibility of the integration in today's network infrastructure is also proven by the recent efforts of some of the international MT system vendors.

However, our envisaged concept of network-based MT aims at the design and specification of an intelligent translation agent, the translation broker, that on the one hand is able to identify and select translation services on the network infrastructure (Intranet as well as the Internet) for a specific application domain, and on the other hand provides a validation and evaluation of the translation results, which will be used in future decision making processes (MT service selection). In addition the MT service agent can inform the personal MT brokers about new features and enhancements of their services according to the principles of NC, i.e. communication and cooperation. This global scenario is shown in Figure 4. In the next section we will discuss this scenario and its realization on the basis of our three investigation dimensions.

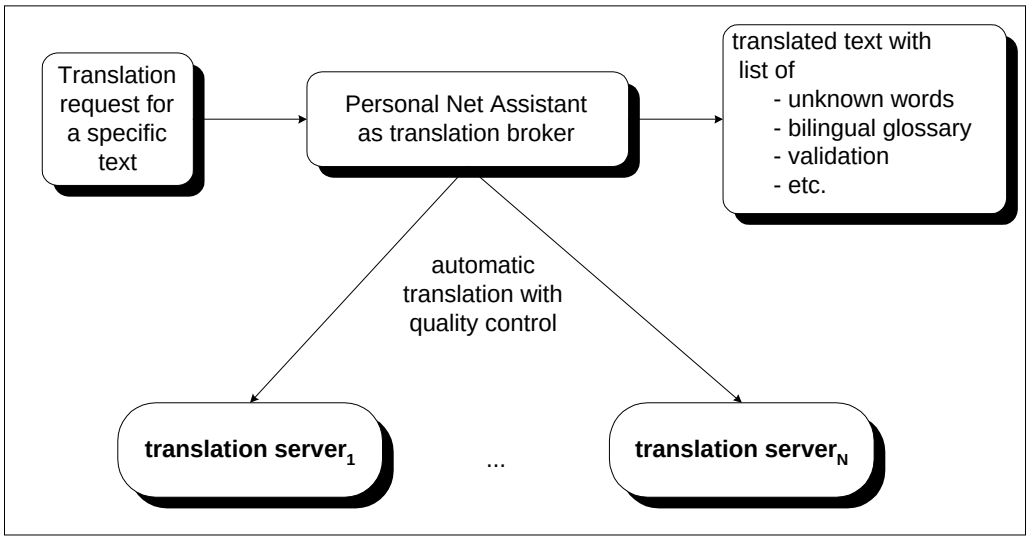

Figure 4: Intelligent translation assistant

3. A New Machine Translation Paradigm

To achieve the goal of an intelligent translation broker we will now develop the theoretical and functional foundation of a new MT paradigm. The need for a translation broker that operates on Intranets and the Internet is obvious, and such a softbot will certainly contribute to overcoming the language barriers of the international trade and commerce, as well as world-wide personal communications.

3.1 Technical infrastructure

The third dimension of our investigation, i.e. the technical infrastructure, can be achieved by the deployment of NC and high-speed network connections (ATM, satellites, etc.). Since the costs in this technical area are dropping drastically, the new investments of a company will be subsumed by their general investments in the field of Web technology. Today nearly every company is investigating or actually investing in the employment of Intranet technology, and most of these companies are already present on the Internet with Web pages or limited services. However, this implies detailed planning, and must be integrated in the general reengineering efforts of the industrial companies. Since the investments in Intranet technology already show an increase in corporate communication and information exchange productivity, as recently reported in market surveys and market researches, especially for the hardware and software market, as well as the pharmaceutical market (cf. [Bowen & Wong, 1996]), the extension to other Intranet/Internet services is prepared.

3.2 Locus of MT engines and MT resources

A remaining open question within this technical scenario is the openness of current MT system vendors with respect to Intranet/Internet based translation services, which go beyond their present efforts, and is of course a matter of competition. This leads us directly to the second dimension of our investigation, i.e. the locus of MT engines and MT resources.

MT users would certainly benefit if different systems and approaches would cooperate to achieve better translation throughput in terms of quality and speed

instead of competing with different approaches. Such a cooperation of MT system vendors could start with a standardization effort for an exchange format of lexical and terminological resources together with national and international projects in this area. This well-defined standard will have also an impact on other areas of network-based services such as multilingual search engines and multilingual text retrieval in general.

A second area of standardization could be the field of distributed environments, for example, the modularization of language resources and technical resources. The idea is to start with general lexical information and to enhance this source of information by additional terminological knowledge or general world knowledge based on shared ontologies (cf. the DARPA knowledge sharing initiative; [Gruber, 1993]) according to the specific translation task of a personal translation broker.

Additional standardization areas could include (automatic) evaluation procedures and processes, and the customization of a brokering task in a way similar to the customization of an intelligent software agent (cf. above). To some extent, these efforts can be compared with the ongoing initiatives in the field of distributed language engineering; there the focus is particularly in the area of modular grammar engineering. A cooperation of MT vendors and these initiatives would be of mutual benefit for each party.

Last but not least, the standardization and the sharing of the technical MT infrastructure and the linguistic MT infrastructure contributes to reducing the costs of both MT system development and MT system maintenance. This certainly will also allow for better translation results and thus contributes to the further development and evolution of the information society.

3.3 Networked Machine Translation

The overall concept of network-based MT, our first dimension, can now be defined as:

> *"Remote translation servers and translation brokers (clients) cooperate over Intranets and the Internet to fulfill a translation task."*

This definition implies the relationship between network-based MT and Networked Computing on the basis of intelligent software agents (cf. Chapter 2). Actually, in this scenario a translation broker will request a translation session from a MT server, which will provide a basic client code for a specified translation task. Once this client code is loaded, the translation broker and the MT server will cooperate (e.g. specific requirements for the translation result), exchange information (e.g. local lexical resources) and communicate (e.g. for the identification of additional resources).

To summarize, the technology for network-based MT is emerging with the recent efforts and developments in Networked Computing and language technology. I will call this new MT paradigm *Networked Machine Translation (NMT).* The success of this new MT paradigm depends on the steps the MT industry will take in this direction. Traditional MT approaches, including translation memory approaches, are integratable in this scenario. However, the actual effort for the design and the implementation of appropriate interfaces and APIs (Application Programming Interface) will differ from system to system, and will depend on their already existing network capabilities.

It should be noted that search engine developers have already initiated standardization efforts in the field of language resources together with language technology companies. Some of these efforts are spin-offs of former R&D projects funded by the European Commission.

In this context the on-going European project OTELO may also contribute to the further development and the realization of the presented new MT paradigm. The OTELO consortium is coordinated by Lotus Development Ltd. in Ireland and consists of several language technology related European companies, including MT vendors. According to the project description ([SAP, 1996]) the aims and goals of the OTELO consortium comprise:

- Defining standardized common lexical resource and text-handling formats.
- Providing an array of tools which will increase the quality and efficiency of machine translation.
- Developing a network infrastructure that makes NLP tools widely accessible.
- Integrating access to the OTELO network into the groupware framework.

I am looking forward to the first results of this project which certainly can contribute to the further evolution of Networked MT.

4. Conclusions and Prospects

In this presentation we have developed the initial prerequisites of a new machine translation paradigm based on the recent and emerging developments in the fields of information technology, Web technology and language technology. This new MT paradigm aims at combining the different technologies to achieve a better translation throughput in terms of translation quality and translation processing speed by means of cooperating and communicating network-based intelligent software agents, which make use of local and distributed language resources that are based on agreed standards. The language resources of this new kind of MT software is intended to be sharable with other Intranet/Internet based language technology applications and services, such as multilingual search engines, text retrieval, information brokering and telematics services.

Security and privacy issues have not been addressed in this presentation, although this topic is of crucial importance for both service providers and service customers. This is mainly due to space limits; an in-depth discussion of this topic would certainly beyond the scope of this presentation. However, the available Intranet/Internet security measures, such as Firewalls and Phil Zimmerman's PGP (Pretty Good Privacy) freeware, would be sufficient for the investigated new MT application.

With the further evolution of the information society and the need to overcome the language barriers of global trade and commerce, the time is ripe for further R&D work towards the actual realization of this new MT paradigm.

In closing this presentation, I would like to come back to the network problem cited in the introduction chapter. Is there any direct answer to their problem?

The described application is a strict LAN (Local Area Network) application and the network failures seem to be based on the used communication protocols but not on

the used hardware (old Ethernet connections are still very reliable). Therefore I would suggest the analysis of the network infrastructure on the protocol basis including the network capabilities of the employed TM software and its use of language resources. This analysis should also include the investigation of the portability of the TM to an Intranet application with the support of the TM vendor.

5. References and Suggested Further Readings

[Andreessen et al., 1996] Marc Andreessen & The Netscape Product Team, 1996. *The Netscape Intranet Vision and Product Roadmap.* Netscape Communications Corporation, Mountain View, CA.

[Bowen & Wong, 1996] Barry D. Bowen and Carolyn W.C. Wong, 1996. *Spinning the internal Web.* Sun World Online 4/96.

[Dean et al., 1996] Drew Dean, Edward W. Felton and Dan S. Wallach, 1996. *Java Security: From HotJava to Netscape and Beyond.* In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA.

[Fingar et al., 1996] Peter Fingar, Dennis Read & Jim Stikeleather, 1996. *Next Generation Computing: Distributed Objects for Business.* SIGS Books & Multimedia, SIGS Publications Inc., New York.

[Hawking & Penrose, 1996] Stephen W. Hawking and Roger Penrose, 1996. *The Nature of Space and Time*. Princeton University Press, Princeton.

[Gruber, 1993] Thomas R. Gruber, 1993. *A Translation Approach to Portable Ontology Specifications.* Knowledge Systems Laboratory Technical Report KSL92-71, Stanford University, Stanford, CA.

[SAP, 1996] SAP, 1996. *OTELO - EU Project LE-2703*. SAP AG, Walldorf, Germany.

[Stoll, 1995**]** Clifford Stoll, 1995. Silicon Snake Oil - Second Thoughts on the Information Highway. Macmillian, London.