Programme
STOA Workshop

State of the art of Machine Translation - current challenges and future opportunities

December 3, 2013. 09:30 - 12:30, Conference Room ASP A1G2, European Parliament, Brussels

The objectives of the workshop:

- present the different applications of MT today and what are the trends on a longer term,
- identify what are the different technologies and methodologies available for MT in use today, what are their respective advantages and disadvantages,
- identify what are the different obstacles in the development of MT, both from a technological point of view, but also from an organisational one
- identify if there are specific obstacles in obtaining an identical level of quality in all EU languages
- explore what are the technological and organisational options to consider by the key stakeholders in this field to further improve the quality MT with an identical level of quality for all EU languages.
- explore what are the possible policy options in support of the organisational and technological options considered above.

Programme

09.30 - 09.35 Welcome and introduction by Algirdas Saudargas, MEP

09.35 - 09.50 **MT as the New Lingua Franca**

Jaap van der Meer, TAUS (Translation Automation User Society)

09.50 - 10.05 **Europe's Languages in the Digital Age: Multilingual Technologies for overcoming Language Barriers and Preventing Digital Language Extinction**

Georg Rehm, DFKI, META-NET

10.05 - 10.20 **Attacking Quality Limitations: New Approaches to Translation Technologies**

Hans Uszkoreit, DFKI, META-NET

10.20 - 10.35 **The way forward with MT@EC in 2014 and beyond**

Daniel KLUVANEC, European Commission, Directorate General for Translation

10.35 – 10:45 Break

10.45 - 11.00 **MT approach at the European Parliament**

Pedro GARCIA-DIEGUEZ, European Parliament, Directorate General for Translation

11.00 - 11.15 **Current and future applications of Machine Translation from a business perspective**

Heidi Depraetere, CrossLang

11.15 - 11.30 **Speech Machine Translation – Current status and long term prospects**

Alexander Waibel, Karlsruhe Institute of Technnology (KIT)

11.30 - 11.45 **The future of MT - A paradigm change: From word-based statistical Machine Translation towards high-quality, hybrid systems**

Jan Hajic, Charles University in Prague

11.45 - 12.25 Round table discussion on policy options to promote technology for language and multilingualism in the digital age.

12.25 - 12.30 Closure and conclusions by Algirdas Saudargas, MEP

Registration before 26th November on the STOA website: http://www.europarl.europa.eu/stoa.
For registration, the following data are needed: name, birthdate, nationality and ID number.

Further information: Peter Ide-Kostic, STOA Administrator: peter.idekostic@ep.europa.eu

Biographical information

MT as the New Lingua Franca - Jaap van der Meer.

Since Google and Microsoft have started making (machine) translation freely available on the internet, the target audience for translation is shifting to billions of citizens, consumers, patients and tax payers. Translation is becoming a utility, always available, real-time and embedded in every app, on every screen and sign board, even built in to new devices like smart eye-ware. The translation industry is challenged to live up to these new expectations. The opportunities are tremendous. But how do we manage these changes effectively? The translation industry potentially plays a central role in the evolution of global economies in the coming years. Translators hold the key to connecting countries and cultures; connecting businesses with foreign businesses, with new customers; governments with citizens; consumers with consumers worldwide. Jaap van der Meer, director of TAUS (Translation Automation User Society) will present four essential action lines for industry-government shared services to prepare the future of MT as the new Lingua Franca.

Jaap van der Meer was the founder and CEO of some of the largest global translation and localization service companies in the 1980s and 1990s. In 2005 he founded the Translation Automation User Society (TAUS). TAUS is an innovation think tank and platform for industry-shared services for the global translation and localization sector. Many of the largest IT companies, government translation bodies and their suppliers of translation and localization services and technologies are members of TAUS. TAUS offers among others a platform for translation quality evaluation and benchmarking and a platform for pooling and sharing of translation memory data. Jaap van der Meer has written many articles over the years about the translation industry.

Europe's Languages in the Digital Age: Multilingual Technologies for overcoming Language Barriers and Preventing Digital Language Extinction - Georg Rehm.

Together with more than 200 contributing experts from all over Europe, the META-NET Network of Excellence (http://www.meta-net.eu) conducted a large and comprehensive study on 30 European languages and the level of support they receive through Language Technologies. This study was published in the 30 volumes of the META-NET Language White Paper series “Europe's Languages in the Digital Age” which also discusses the most urgent risks and chances of the languages in question. The series covers all official EU languages and several other

languages spoken in geographical Europe. The very alarming conclusion of the study is that more than 20 European languages are in danger of digital extinction. Published on 26 September 2012, the European Day of Languages, the corresponding press release had a tremendous impact. It generated more than 600 mentions in the international press and interest on a global scale, demonstrating that the European citizens are very interested and passionate about their languages. The presentation will first introduce the main goals of META-NET and then present the key findings of our study “Europe's Languages in the Digital Age” which has recently been updated and extended by taking into account several additional languages and also by enlarging the group of contributors with representatives of three different language-related stakeholders.

Georg Rehm works in the Language Technology Lab at the German Research Center for Artificial Intelligence (DFKI), in Berlin. He is the Network Manager of META-NET, an EU/EC-funded Network of Excellence consisting of 60 research centres from 34 countries, dedicated to building the technological foundations of a multilingual European information society (see http://www.meta-net.eu). He holds an M.A. in Computational Linguistics and Artificial Intelligence, Linguistics and Computer Science. After completing his PhD in Computational Linguistics, Georg Rehm joined the University of Tübingen, leading projects on the sustainability of language data sets and language technologies. In 2009, he worked at an award-winning internet startup in Berlin where he was responsible for product development and language technology. Georg Rehm joined DFKI in early 2010. He is (co-)author of more than 70 research publications and co-edited, together with Hans Uszkoreit, the META-NET White Paper series "Europe's Languages in the Digital Age" as well as the "META-NET Strategic Research Agenda for Multilingual 2020". Furthermore, Georg Rehm is the Manager of the German/Austrian Office of the World Wide Web Consortium (W3C).

Attacking Quality Limitations: New Approaches to Translation Technologies - Hans Uszkoreit.

High-quality translation is in greater demand today than ever before. Europe is the region where the lack of fast, affordable quality translation hurts the most. Despite considerable progress in machine translation (MT), which has enabled many new applications for automatic translation, the quality barriers for outbound translations (i.e., translations to be published or distributed outside of an organisation) have not yet been overcome. As a result, the volume of translation today falls far short of what is needed for optimal business operations and legal requirements. European industry, administration, and society all urgently need progress in translation technologies to fulfill existing translation needs, to extend multilingual communication to additional languages and services (e.g., to conquer untapped markets), and to reduce costs associated with commitments to linguistic diversity. Based on the work of META-NET with its "Strategic Research Agenda for Multilingual Europe 2020", a strategy paper prepared together with ca. 200 experts from research and

Industry, which recommends the priority research theme "Translingual Cloud" and the first results of the EU/EC-funded project QTLaunchPad, which is dedicated to overcoming quality barriers in machine and human translation and also in language technologies, the presentation will introduce QT21, our initiative towards High-Quality Machine Translation (http://qt21.eu).

Hans Uszkoreit is teaching as a professor of Computational Linguistics at Saarland University. Moreover, he serves as a Scientific Director at the German Research Center for Artificial Intelligence (DFKI) where he heads the DFKI Language Technology Lab. Hans Uszkoreit studied Linguistics and Computer Science at the Technical University of Berlin and the University of Texas at Austin. While he was studying in Austin, he also worked as a research associate in a large machine translation project at the Linguistics Research Center. After he received his Ph.D. in Linguistics from the University of Texas, he worked as a computer scientist at the Artificial Intelligence Center and was affiliated with the Center for the Study of Language and Information at Stanford University. Hans Uszkoreit has been involved in many EU/EC-funded Language Technology projects, especially in the field of Machine Translation, most notably EuroMatrix, EuroMatrix+ and QTLaunchPad. He is the Coordinator of META-NET, an EU/EC-funded Network of Excellence consisting of 60 research centres from 34 countries; dedicated to building the technological foundations of a multilingual European information society (see http://www.meta-net.eu).

The way forward with MT@EC in 2014 and beyond - Daniel Kluvanec

MT@EC (Machine Translation at the European Commission) as a state of the art statistical machine translation system becomes one of important building blocks for the e-governmental interoperability among EU Member States in many distinct ways. In order to understand what this technology can do and what cannot yet, several key concepts have to be mentioned. Besides technicalities like probabilistic pattern recognition within big corpora or cognitive incapacity of any present computational system, the opportunities for natural language processing are determined in particular by linguistic characteristics (morphology, syntax, semantics, pragmatics, etc.) and their differences among analytic, inflected and agglutinating languages or language families. These key concepts allow for an interdisciplinary synthesis about future needs still to be addressed by researchers.

Daniel Kluvanec is a Business Manager and Adviser for Machine Translation at the European Commission. As a generalist he holds two master's degrees, one in engineering (microprocessors based computational systems) as well as an interdisciplinary one in social sciences (didactics, psychology, quantitative sociology) and is a world-wide 3rd and 2nd prize winner in International Physics Olympiad. After several years spent in Slovak Technical University as a researcher, he started his first company in 1992 and served to fundraising organisations based mostly in Switzerland (like Unicef, WWF, Helvetas) or to royalties collecting agencies in the field of audio visual works (like Agicoa, GWFF, Suissimage) while being a CEO he

still kept his hands-on experience with scientific publishing, internetworked cloud solutions, large scale databases, sociology statistics, big data processing, data pattern recognition or phonetic search. With his extensive interdisciplinary experience in computing and linguistics, he was after all appointed in 2007 an EU civil servant at the Directorate General for Translation of the European Commission.

MT approach at the European Parliament - Pedro Garcia-Dieguez.

The MT approach of the European Parliament is complementary to the MT@EC service. The EC offers generic MT systems, and the EP will complement that with the development of domain-specific SMT (Statistical MT) systems optimised for restricted sublanguages and constrained by the EP's particular translation requirements. Domain-specific systems provide better results than generic systems for EP domains because of the EP context specificity and the proximity of sub-languages as well.

The EP-specific approach is usually satisfactory for pairs of structurally dissimilar, distant and under-resourced languages, such as the Finno-Ugric or Slavic families, Estonian and German. A hybrid approach consisting on probabilities guided by rules for certain constructions, which are a source of frequent errors, is also one of our research areas. Another line of research is to improve Translation Memory (TM) results with MT. In any case, MT will be used as a complement to TMs, and TM results will be offered in priority when previous human translations exist.

Pedro Garcia-Dieguez is a European civil servant, responsible for Machine Translation and the CAT tools at the European Parliament, previously project manager of workflow tools at the European Commission and head of service for Translation workflow tools at the Court of Justice. Degree in Computer science at the University of Deusto (Bilbao).

Current and future applications of Machine Translation from a business perspective - Heidi Depraetere.

This presentation will focus on how machine translation (MT) is used in a business context. It will address some of the technical and deployment challenges from a company perspective. It will also take a particular look at a different challenge: how do localisation professionals cope with MT? What sorts of (mis)perceptions do business users encounter when dealing with MT? Finally, this presentation will also explore some future use applications of MT in response to new business requirements driven by “everywhere, anytime, always on”.

Heidi Depraetere has over 20 years' experience in the localisation and language technology industries. She is a founder and director of CrossLang, a privately owned consulting and systems integration company dedicated to translation automation technology.

Heidi is familiar with the evolution of the industry from desktop CAT tools to continuous translation, from commercial software deployment to open-source solutions. User-centric Machine Translation evaluation is an area of active interest.

She has created and worked with international teams throughout Europe and the United States. Heidi holds a degree in linguistics from the University of Leuven and is based in Gent, Belgium

Speech Machine Translation – Current status and long term prospects
Dr. Alexander Waibel.

Dr. Waibel was one of the founders of C-STAR, the international consortium for speech translation research and served as its chairman from 1998-2000. His team has developed the JANUS speech translation system, the first American and European Speech Translation system, and more recently the first real-time simultaneous speech translation system for lectures. His lab has also developed a number of multimodal systems including perceptual Meeting Rooms, Meeting recognizers, Meeting Browser and multimodal dialog systems for humanoid robots. He directed the CHIL program (FP-6 Integrated Project on multimodality) in Europe and the NSF-ITR project STR-DUST (the first domain independent speech translation project) in the US. He is project coordinator of the IP EU-BRIDGE, funded by the EC and started on Feb. 01, 2012. In the areas of speech, speech translation, and multimodal interfaces Dr. Waibel holds several patents and has founded and co-founded several successful commercial ventures.

Dr. Alexander Waibel is a Professor of Computer Science at Carnegie Mellon University, Pittsburgh and at the Karlsruhe Institute of Technology (KIT) Germany. He directs InterACT, the International Center for Advanced Communication Technologies at both Universities with research emphasis in speech recognition, language processing, speech translation, multimodal and perceptual user interfaces. At Carnegie Mellon, he also serves as Associate Director of the Language Technologies Institute and holds joint appointments in the Human Computer Interaction Institute and the Computer Science Department.

Dr. Waibel received the B.S. in Electrical Engineering from the Massachusetts Institute of Technology in 1979, and his M.S. and Ph.D. degrees in Computer Science from Carnegie Mellon University in 1980 and 1986. His work on the Time Delay Neural Networks was awarded the IEEE best paper award in 1990. His contributions to multilingual and speech translation systems was awarded the "Alcatel SEL Research Prize for Technical Communication" in 1994, the "Allen Newell Award for Research Excellence" from CMU in 2002, and the Speech Communication Best Paper Award in 2002.

The future of MT - A paradigm change: From word-based statistical Machine Translation towards high-quality, hybrid systems - Jan Hajic

Machine translation has grown over the past 25 years from an IBM T. J. Watson Research lab in Yorktown through the creation of open source toolkits, such as Giza, Joshua and Moses, to a commercially successful field the results of which almost everyone takes for granted. However, apart from a few pairs of "big" languages (such as English and French) the quality of automatic translation systems, as currently represented mainly by the phrase-based statistical systems, is not at a level suitable for direct information dissemination. For some language pairs it is still below the threshold needed even for efficient postediting. While current systems still have room for improvement, it is long in the talks of the stakeholders involved that basic research into fundamentally different approaches might be appropriate – while not forgetting the proven advantages the statistical methodology has contributed to the field of machine translation in the recent past.

Jan Hajic is a full professor of Computational Linguistics, vice-director of the Institute of Formal and Applied Linguistics at the School of Computer Science at the Charles University in Prague, and the director of the Czech national language resource center, LINDAT/CLARIN. His interests cover morphology of inflective languages, machine translation, deep language understanding, and the application of statistical methods in natural language processing in general. He also has an extensive experience in building annotated language resources. His work experience includes both industrial research (IBM Research Yorktown Heights, 1991-1993) and academia (Charles University, Prague and Johns Hopkins University, Baltimore, MD, USA). He has published more than 100 papers both internationally and nationally. He has been the PI or Co-PI of several national and international grants and projects, and has served in journal editorial boards as well as various advisory and scientific boards, most notably the Research Board of the Technology Agency of the Czech Republic.