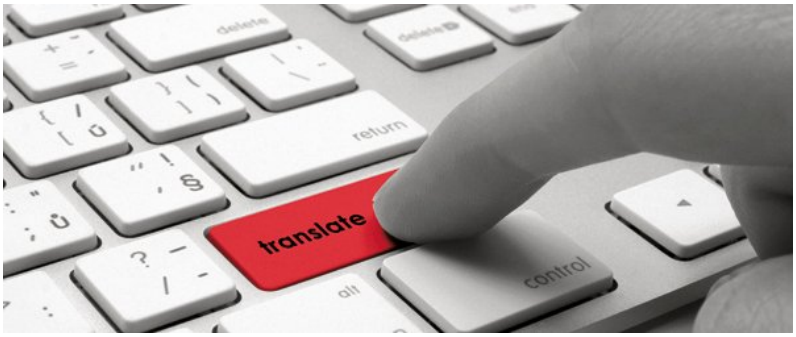

Translating and the Computer 32

Programme

08.30 Registration
09.15 Introduction by Chair: Olaf-Michael Stefanov, Austria

09.15 **Challenges at World Trade Organization: Evaluation and implementation of a Statistical Machine Translation System**
Olivier Pasteur, World Trade Organization, Switzerland
Mr Pasteur will address the ongoing challenges of integrating statistical machine translation into the translation workflow of an organization such as his, in which significant requirements for technical correctness must be balanced against those for political sensitivity, while fending with such matters as talented but yet not computer-oriented professional translators, budget constraints and complex workflows, tight deadlines, high quality requirements as well as those dealing with translators' reluctance to accept the fundamental changes in their work that the integration of machine translation entails.
10.00 discussion

10.05 **Creating Your Own Translation Memory Repository**
Ronan Martin, *SAS Institute A/S*, Denmark
At SAS we had been accumulating translation memories for some years but felt that we weren't getting maximum benefit from them. There were two areas that could potentially yield benefit for our translation community: the possibility of searching for existing translations in an entire TM repository, and locating existing TMs that would be suitable for attaching to new projects. Obviously you normally attach the TM from a previous version to a current localization project, but there were complex crossovers between our products and we found we were paying several times over for translating strings that existed somewhere in the TMs - though no-one was sure where. As SAS had some available programming resources we decided that it was feasible to build the functionality we desired into our existing workflow system. We wrote a script which searches the project repository for the txt TM files across all languages, and aggregates them into a set of language-specific silos. Creation of the language-specific TM silos is done using the SAS programming language, which is a statistical and data processing language. The primary purpose was to provide a TM repository search facility, but one of the major spin offs of the work has been the development of TM Discovery. This entails taking the current set of source strings for a new localization projet, and doing a concordance search against the whole repository for each language. This product has led to reduced translation costs. For completely new products SAS now has a reliable statistical method of predicting which TMs will yield the greatest number of pre-translations. This paper will present these developments at SAS.
10.45 discussion
10.50 Coffee

11.20 **Does Google know better? Translators and machine translation**
Dr Ignacio Garcia, University of Western Sydney, Australia
New translation memory tools and new versions of established ones offer the translator the option to post-edit machine generated text for segments for which no match is found in the memories. Translating no matches from the machine translation baseline is actually the default option of the Google Translator Toolkit, which in its Settings window further advises that "most users should not modify this". We imagine that at least for some language pairs and some tasks this should work well. Would it, however, for any language pair (for which Google Translate is available) and for any type of text? This is what Google assumes. This paper presents the results of research based on work with the English-Chinese language pair.
11.55 discussion

12.00 **Translation and technology: is it a jungle?**
Juliet Macan, Ic.Doc srl, Italy
Does technology make globalisation more complicated? In a sector where demand constantly exceeds supply, and time frames continue to shrink at a terrifying speed, it might help to take a closer look at the globalisation actors: the decision-makers at the various stages, those involved in the creation of original texts, through translation, review and

publishing without forgetting the tool vendors. A examination of the traditional processes might reveal why progress is so slow. A clear overall view of all the crucial aspects involved in the globalisation process is essential when deciding priorities and setting up an efficient workflow. These decisions are generally taken on the basis of overriding time and cost requirements. There is a real need to address the problem up-front, improving the quality of the source material for machine or computer assisted translation and placing terminology centre-stage. In this extremely complex and variegated situation, perhaps it is time to establish some kind of authority to analyse both requirements and solutions in an objective manner and offer all the actors involved disinterested advice about the technology which can facilitate their work. This paper will present ideas for the way forward.

12.35 discussion

12.40 Short presentations on new products:

ABBY LingvoPro

LTC

TRANSLI

13.30 Lunch

14.30 Afternoon session chaired by Professor Ruslan Mitkov, University of Wolverhampton, UK

14.30 **XTRF-TM – Managing teamwork**

Andrew Nedoma, Lido-Lang Technical Translations, Poland

This presentation shows the evolution of the translation management system that was presented at ASLIB 27 in 2005 under the name TROFFI and has now developed to a technically advanced management system now known as XTRF-TM. XTRF-TM involves a translation management system, a workflow management system, a CAT tool, and an e-mail software, linking together to create a real, online teamwork environment that provides a service to clients which conforms to the ISO 9001 and EN15038 requirements. The main tool in the daily work of many players co-operating in translation services, XTRF-TM manages all the processes and helps to establish a teamwork environment for its users. The aim of this presentation is to show real scenarios, involving several players, illustrating how people can work together through this system, how they can interact and how their work is reflected in other modules that are viewable by other players.

15.05 discussion

15.10 **Language Technology for Automatic Control in eLearning Tools for Translators and Second Language Students**

Paul Schmidt, Saarbrücken University, Germany

This paper introduces the results of a number of projects whose objective was to create elearning sites for students of translation and second language learners as autonomous learning systems. The paper presents a set of innovative tools that allow for automatic correction. The application of these tools has been widely tested with hundreds of students and dozens of teachers from different cultures at several universities. The areas in which the tools were tested were 'translation' and 'foreign language teaching'. The tools, however, can be used for any kind of textual input. The major innovation that will be presented in the paper is that though sophisticated language technology is to be used for creating automatic correction tools language teachers who do not have any computational linguistic skills may automatically generate their own automatic correction tools.

15.45 discussion

15.50 Tea

16.20 **Managing Translingual BPM Repository Content**

Jörg Schütz and Alexandra Weissgerber, bioloom group, Germany

In software engineering, internationalization is something to be done initially by system designers and developers based on existing standards and best practices to prepare the pathway for localizers and language professionals such as content creators, terminologists, translators and post-editors who join in at different stages with their trans-lingual experience and knowledge. In this presentation, the authors elaborate on this scenario with our use case being a BPM (Business Process Modelling/Management) environment, and in particular the content of the business process repository. The overall solution is based on combining linguistic intelligence with semantic technologies and machine learning technique to build a business-specific knowledge matrix that extends the existing business repositories with a machine understandable semantic layer, and that supports different reasoning capabilities to gain implicit knowledge from the integrated data resources. The presenters will also report on their findings regarding the employment of existing standards and best practices, and the conclusions drawn for their future R&D strategies and next practices.

16.55 discussion

17.00 End of Day One

DRINKS RECEPTION

PROGRAMME: DAY TWO

09.00 Introduction by Chair: Daniel Grasmick, Lucy Software and Services, Germany

09.00 Keynote TBC

09.45 discussion

09.50 **Next Generation Translation and Localization: Users Are Taking Charge**
Sharon O'Brien, Dublin City University and Reinhard Schäler, University of Limerick

The user is taking charge and examples for this are many. The best known example in the commercial sector is Facebook. The site grew from 34 million international users in early 2008 to well over 400 million by mid 2010. Facebook now "speaks" more than 100 languages – all supplied by its users. The *MTC India Youth Icons* used to be business magnates, cricket players or actors. In 2007, the award did not go to a person but to a website: Orkut, the social networking site translated by its users into Hindi, Marathi, Bengali, Tamil and Telugu. Industry experts such as Greg Oxton in *The Power and Value of On-line Communities* at the AGIS'09 event have been calling on digital publishers to give up "their illusion of control". The proposed paper will examine the case of The Rosetta Foundation as an example of a not-for-profit volunteer translation facilitator. The paper focuses on the *motivating factors* for volunteer translators. A survey was distributed to the several hundred volunteers who signed up as translators in the first few months of The Rosetta Foundation's launch. The paper will present the results of this survey and will explore how The Rosetta Foundation, and other not-for-profit crowdsourcing translation organisations, might better motivate volunteers to contribute their skills and expertise.

10.25 discussion

10.30 **Managing Social Translation: Online Tools for Translators' Communities**
Anas Tawileh, Meedan.net

The digital lifestyle of more than a billion citizens of the world is generating massive amounts of information stored in digital format and made available on the web. According to a recent study by consultancy firm International Data Corporation (IDC), the world's digital output currently stands at 8,000,000 petabytes and may surpass 1.2 zettabytes (a zettabyte is equal one million petabytes). This creates substantial demand for translators and translation services, but also poses challenges to the traditional approaches to translation management and conduct, and the way translators have organized in the past. While much has been said about developments in machine translation technologies, the reality remains these translations tend to be of lower quality that human translations. Frequently, content consumers are complaining that machine translation has not yet reached sufficient maturity levels to produce meaningful content. For this purpose, social translation builds on initial translation drafts generated by machine translation engines, and empowers translators' communities with the ability to vet the translation and make any necessary corrections and amendments. Another issue that relates directly to the translation quality is the capability of translators in the network to produce high quality output. Social translation tools and platforms address this by designing a portable translation reputation management system for translators, coupled with a rating mechanism for translated content. This paper will explore these ideas and approaches in greater depth, and will present examples for the social translation tools and platforms currently available on the web.

11.05 discussion

11.10 Coffee

11.30 **A Computational Framework for a Cognitive Model of Human Translation**
Michael Carl, Copenhagen Business School, Denmark

Human translation process research analyses the translation behaviour of translators such as, for instance, memory and search strategies to solve translation problems, types of units that translators focus on, etc., and determine the temporal (and/or contextual) structure of those activities or describe inter- and intra personal variation.
Cognitive models have been developed that explain translator's behaviour in terms of controlled and uncontrolled workspaces and with micro- and macro translation strategies. However, up to date no attempts have been made to ground and quantify such translation models in empirical user activity data. In order to close this gap, the paper elaborates a computational framework for a cognitive model of human translation. We investigate the structure of the translators keystroke and gaze data, discuss possibilities for their classification and visualisation and elaborate how the translation model can be grounded and tested in the empirical data. The insights gained from such a computational translation model does not only enlarge our knowledge about human translation processes, but has also the potential to enhance the design of interactive MT systems and help interpret user activity data in human-MT system interaction.

12.05 discussion

12.10 ***Mixed up* with Machine Translation: Multi-word Units Disambiguation Challenge**
Anabela Barreiro, Universitá degli Studi di Salerno, Italy

The Internet has helped machine translation (MT) to become increasingly popular within the general public. Considerable progress has also been made in qualitative terms because of the availability and use of large parallel corpora, the development of knowledge bases, the adoption of statistical models, and the integration with various computer assisted translation tools, particularly with translation memories. However, despite recent significant progress, lexical problems still represent a critical area in MT. Among lexical problems, multi-word units (MWU), are particularly difficult to be processed by MT systems. The aim of this paper is to provide a systematic qualitative evaluation of the shortcomings of existing MT systems with reference to the processing of MWUs with different degrees of variability, in order to point out benefits, strengths and weaknesses of distinct approaches. The paper will discuss the usage of combined Lexicon Grammar lexical resources and OpenLogos lexical resources together with

syntactico-semantic rules as a possible solution to overcome machine translation limitations with regards to the automated processing and translation of MWUs. What could the ideal MT evaluation tool look like to correctly evaluate the performance of MT engines with regards to MWUs?

12.45 discussion

12.50 **Identifying Fixed Expressions: A Comparison of SDL MultiTerm Extract and Déjà Vu's Lexicon**
María Fernández-Parra & Pius ten Hacken, Swansea University, UK

This paper will evaluate and compare the usefulness of two approaches that can be used for the identification of fixed expressions and are implemented in different CAT tools, SDL MultiTerm Extract and Déjà Vu's Lexicon. Both MultiTerm Extract and the Lexicon are designed to extract potentially useful strings from a text, but they differ in the method of extraction and the scope of use for the extracted strings so that it will be interesting to compare their relative merits in this particular task. The Lexicon works rather like a concordancer, whereas MultiTerm Extract is a statistical term extraction tool. Concordancing results in non-selective, total recall, whereas statistical term extraction aims to present a smaller selection while maximizing recall. In both cases, facilities for supporting manual confirmation of relevant expressions from the automatically produced selection should be taken into account in the evaluation. On the basis of the existing functionalities, the paper will aim to suggest how these functionalities can be optimized for the identification of fixed expressions.

13.25 discussion

13.30 Lunch

14.30 Afternoon session chaired by Chris Pyne, SAP, AG Germany

14.30 Panel Discussion: The Right to Access to Content in your Language needs to be extended beyond the G20
Moderator: Chris Pyne, SAP AG, Germany
Contributors include Reinhard Schäler, University of Limerick, Ireland

15.30 Tea

15.50 **Highlighting Matched and Mismatched Segments in Translation Memory Output through Sub-Tree Alignment**
Ventsislav Zhechev, EuroMatrix+, CNGL, School of Computing, Dublin City University

In this paper, we present a novel system that can automatically detect and highlight the segments that need to be modified in a TM-suggested translation. We base it on state-of-the-art sub-tree alignment technology that can produce aligned phrase-based-tree pairs from unannotated data. We chose this particular tool because can fully automatically capture translational equivalencies, without the need for any training data. The only auxiliary requirement it has is for a probabilistic dictionary for the languages that are being aligned. If such is not readily available, it can be automatically inferred from the data stored in the TM using an off-the-shelf open-source word-alignment tool. Our system can also be used to pre-translate the identified mismatched segments in the suggested translation using a (Statistical) Machine Translation system. Researchers are currently working on a prototype demonstrator that will showcase the described system. The ultimate goal, however, is to integrate this technology into an open-source or commercial TM tool to make it available to professional translators. In the future, we plan to perform user studies evaluating the impact our system has on post-editing speeds in real-life situations.

16.20 discussion

16.25 **A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions**
Nasredine Semmar, CEA, France

Translation lexicons play a vital role in machine translation and cross-language information retrieval. The high cost of bilingual lexicons development and maintenance is a major barrier for adding new languages pairs for these applications. Word alignment approaches are generally used to construct bilingual lexicons. In this paper, we present a hybrid approach to align single words, compound words and idiomatic expressions from bilingual parallel corpora. This approach combines linguistic and statistical information in order to improve word alignment results. The results obtained showed that the single-word aligner generates a translation lexicon corresponding to the corpus of 1 103 sentences with 90 % of precision and 81% of recall, and the multi-word aligner achieves a significant enrichment of the bilingual lexicon with compound words and idiomatic expressions.

16.55 discussion

17.00 End of the Conference

aslib
MANAGING INFORMATION